PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure de Paris

On Structured Lattices in Cryptology

Soutenue par
Henry Bambury
Le 18 novembre 2025

École doctorale n°386
Sciences Mathématiques de Paris Centre

Spécialité
Informatique

Composition du jury :

Phong Q. Nguyen ⋆ Inria ⋆ École normale supérieure	*Directeur de thèse*
Pierre-Alain Fouque ⋆ Université Rennes 1	*Rapporteur*
Antoine Joux ⋆ CISPA – Helmholtz Center for Information Security, Saarbrücken	*Rapporteur*
Léo Ducas ⋆ Centrum Wiskunde & Informatica, Amsterdam ⋆ Leiden University	*Examinateur*
Alexander May ⋆ Ruhr-Universität Bochum	*Examinateur*
Alice Pellet–Mary ⋆ CNRS ⋆ Université de Bordeaux	*Examinatrice*
Hugues Randriambololona ⋆ ANSSI ⋆ Télécom Paris	*Examinateur*
Damien Vergnaud ⋆ Sorbonne Université	*Examinateur*

1794
ENS | PSL

Résumé

La menace que représente l'arrivée prochaine d'un ordinateur quantique assez puissant pour briser la cryptographie du début du XXIe siècle a contraint les chercheurs à développer de nouveaux algorithmes pour le chiffrement et la signature numérique. La sécurité de ces protocoles repose sur la difficulté présumée de problèmes mathématiques autres que la factorisation ou du logarithme discret, dont le principal est le problème du plus court vecteur dans les réseaux euclidiens.

Pour des raisons d'efficacité, les réseaux utilisés dans les premières constructions élevées au rang de norme sont dotés d'une structure particulière, souvent issue de la géométrie des corps de nombres. Il est donc crucial de comprendre les attaques qui exploitent cette structure afin de choisir des paramètres assurant réellement leur sécurité.

Le premier grand axe de cette thèse est l'étude de différentes familles de réseaux utilisés en cryptologie, sous divers aspects :

- Sur le plan algorithmique, nous étudions des attaques prouvées et heuristiques sur le problème du plus court vecteur dans les réseaux NTRU et les réseaux hypercubiques.
- Sur le plan mathématique, nous étudions le comportement moyen des vecteurs les plus courts dans les réseaux issus d'idéaux de corps de nombres.

Le second grand axe est l'amélioration de potentielles normes post-quantiques pour les signatures numériques :

- Sur le plan offensif, nous élaborons une attaque de récupération de clé par réduction de réseaux sur le schéma de signature DEFI, prétendu post-quantique par ses auteurs.
- Sur le plan défensif, nous proposons un nouveau type de distribution pour cacher l'information secrète dans les schémas de signature de type Fiat-Shamir avec Rejets à base de réseaux. Cela nous permet d'élaborer sous les mêmes hypothèses de sécurité un schéma de signature plus compact que la norme Dilithium, mais qui contrairement à la norme Haetae n'utilise pas de distributions gaussiennes, difficiles à implémenter de façon sécurisée.

Cryptographie Post-Quantique ⋆ Réseaux Euclidiens ⋆ Cryptanalyse ⋆ Signatures ⋆ Théorie des Nombres

Abstract

Early 21st-century cryptography is threatened by the fact that a robust quantum computer could be built in the near future. This threat has forced researchers into designing new algorithms for encryption and digital signatures, whose security relies on the presumed computational hardness of solving mathematical problems different from factoring or discrete logarithm, such as the shortest vector problem in Euclidean lattices.

Due to efficiency concerns, the first standardised schemes rely on lattices with extra structure, often derived from the geometry of well-chosen number fields. Understanding the attacks that exploit this particular structure is crucial in order to select parameters that truly ensure security.

The first main topic of this thesis is the study of different families of lattices used in cryptology:

- On the algorithmic level, we study provable and heuristic attacks on the shortest vector problem in NTRU and hypercubic lattices.
- On a more mathematical level, we study the average behaviour of the shortest vectors in ideal lattices: lattices that inherit structure from ideals of number fields.

The second main topic is the improvement of potential post-quantum standards for digital signatures:

- From an attacker's perspective, we develop a key recovery attack using lattice reduction on the DEFI signature scheme, which its authors claimed to be post-quantum.
- From a defender's perspective, we propose a new type of distribution to hide the secret information in Fiat-Shamir with Aborts lattice-based signature schemes. This allows us to construct, under the same security assumptions, a more compact signature scheme than the Dilithium standard, but which, unlike the Haetae standard, does not use Gaussian distributions, known to be difficult to implement securely.

Post-Quantum Cryptography ⋆ Lattices ⋆ Cryptanalysis ⋆ Signatures ⋆ Number Theory

À mes grands-parents

Remerciements

Je voudrais d'abord remercier mon directeur de thèse Phong, pour sa passion, sa bienveillance et son encadrement d'une qualité bien au-delà de mes attentes pendant ces quelques années passées ensemble. Je garderai un très bon souvenir de tous les moments passés à discuter de réseaux, ou d'autres choses. Je te remercie pour tout ce que tu m'as transmis.

Ensuite, je souhaite remercier Pierre-Alain Fouque et Antoine Joux, qui me font l'honneur de rapporter ce manuscrit de thèse. Depuis mes études de master, j'ai toujours apprécié lire vos publications, dont la pertinence et la diversité m'ont toujours impressionnés. Merci pour votre temps.

Merci aussi à tous les membres du jury Léo, Alexander, Alice, Hugues et Damien. Par vos travaux ou à travers des discussions, vous m'avez tous beaucoup appris, et je serai ravi de travailler avec vous si l'occasion se présente un jour.

Merci à Christophe de m'avoir introduit à la cryptographie à base de réseaux, et à Fabien de m'avoir initié aux belles mathématiques afférentes à la cryptographie post-quantique : c'est pour toi que j'ai choisi d'inclure un chapitre d'isogénies dans une thèse sur les réseaux. Merci à tous les doctorants et chercheurs avec lesquels j'ai pu discuter en conférence et lors de séminaires, et en particulier à mes co-auteurs non-déjà cités précédemment Hugo, Antoine, Francesco, Benjamin, Thomas, Éric. Special thanks to Seungki for teaching me so much on the mathematics of random lattices. Merci aussi à Jérôme et Alice pour avoir fait partie de mon comité de suivi de thèse.

Merci à Tuong-Huy, Antonin et les autres membres de l'équipe à DGA-MI d'avoir gardé un œil sur le déroulement de ma thèse depuis Rennes. C'est un plaisir de vous rejoindre bientôt.

Merci à tous les membres (présents ou passés) de l'équipe CASCADE[1] pour votre présence au quotidien, et pour avoir fait du laboratoire un lieu vivant, humain et scientifiquement très stimulant. Sami, Roderick, Léonard, Amine, Hugo, Valentine, Nicolas, Alexandre, Sacha, Céline, Cédric, Wissam, Vincent, Lénaïck, Paul, Laurent, Antoine, Hung, Guirec, Jianwei, Hanyu, Jules, Florette, Brice, David, Huyen, Ky, Paola, Tu, David, Raj, Pierre, Michael, Mahshid, Éric, Robert, Erkan, Florian, Huy : je garde de bons souvenirs de moments passés avec chacun d'entre vous. Bon vent à ceux qui ont rejoint l'équipe en ce début d'année.

Merci à Lise-Marie, Diana, Sandra, Chantal, Mohammed, Éric, Mélanie pour leur assistance au niveau administratif. Merci également au personnel essentiel mais moins visible de l'ENS, en particulier aux agents de nettoyage du bâtiment Rataud et aux cuisiniers de la cantine.

Merci à mes parents, qui m'ont toujours fait confiance et soutenu dans mes choix, à mes sœurs pour leur bonne humeur, et à Lucie pour m'avoir supportée pendant ces trois années. Merci enfin à tous les proches et amis (dont je ne me risquerai pas à lister les noms de peur d'en oublier) qui ont contribué de manière indirecte à cette thèse, et bon courage à tous ceux d'entre eux qui soutiennent bientôt la leur.

[1] et les infiltrés des équipes QAT et SECURITY

Contents

Part I

Overview

Chapter

1 Résumé substantiel en Français

This chapter is a direct translation of the next chapter.

Chapter content

Dans le monde d'aujourd'hui, les données privées sont plus importantes et plus menacées que jamais. Nous confions à la technologie nos communications personnelles, nos transactions financières, nos dossiers médicaux et d'identité. Il est essentiel de protéger toutes ces informations sensibles. Pour ce faire, il est possible de recourir à la cryptographie[1], la science du secret.

La cryptographie dans l'histoire Au cours des millénaires, l'utilisation de codes par les individus et les sociétés a influencé le cours de l'histoire à plusieurs reprises. Cela s'est illustré au cours des trois événements historiques suivants, rendus célèbres grâce au cinéma :

- **La conspiration de Babington (1586) :** Mary Stuart, reine d'Écosse, a été exécutée par la reine Elizabeth I d'Angleterre après que ses cryptanalystes eurent décrypté des lettres qui prouvaient l'implication de Mary dans un complot d'assassinat visant à s'emparer du trône.
- **Le télégramme Zimmerman (1917) :** Pendant la Première Guerre mondiale, les Allemands ont envoyé un télégramme chiffré à leur ambassade au Mexique, demandant une alliance contre les États-Unis. L'indignation suscitée par la publication du télégramme déchiffré par les services secrets britanniques a largement accéléré l'entrée en guerre des États-Unis.
- **La machine Enigma et la Seconde Guerre mondiale (1939-1945) :** Les efforts conjoints des cryptanalystes polonais, français et britanniques ont permis de briser le chiffrement Enigma utilisé pour les communications sécurisées de l'armée allemande, donnant ainsi aux Alliés un avantage décisif dans la guerre. En même temps, la cryptanalyse des machines de Lorenz voit naître Colossus, le premier ordinateur électronique numérique au monde.

Les systèmes de chiffrement historiques au temps de la Rome antique reposaient sur des schémas de substitution simples, c'est-à-dire des méthodes qui remplacent systématiquement des lettres ou des groupes de lettres par d'autres dans l'alphabet, comme le chiffre de César, où

[1] Littéralement dérivé du grec ancien "écriture cachée". Le terme cryptologie est plus général et se réfère simultanément à la cryptographie et à la cryptanalyse, la partie offensive de la cryptologie.

chaque lettre est remplacée par la suivante. Les chiffrements par substitution étant trop faciles à casser, ils ont évolué vers des chiffrements polyalphabétiques plus sophistiqués employant des chiffrements par substitution simultanément sur plusieurs alphabets. Parmi les exemples notables, on peut citer le chiffre de Vigenère, créé au XVIe siècle et cassé seulement au XIXe siècle, et la célèbre machine Enigma. Avant le XXe siècle, la cryptanalyse nécessitait de l'intuition, de la patience, des prouesses linguistiques et un savoir-faire acquis par l'expérience. Cette situation a radicalement changé depuis le début des années 1980. La cryptographie d'aujourd'hui suit le principe de Kerkoffs : la sécurité d'un système de chiffrement ne doit pas reposer sur le secret de celui-ci.

Cryptographie prouvée La cryptographie moderne est passée d'un art à une science rigoureuse, où la sécurité peut être prouvée mathématiquement. Les protocoles cryptographiques pratiques dont la sécurité est garantie inconditionnellement semblent hors de portée. En contrepartie, pour certains schémas, nous sommes capables de prouver formellement que briser leur sécurité est au moins aussi difficile que de résoudre un problème mathématique bien étudié et calculatoirement difficile, comme la factorisation. À l'ère de l'information, la cryptographie est devenue la pierre angulaire de la cybersécurité, protégeant à la fois les infrastructures civiles et militaires.

La cryptographie fait régulièrement l'objet de nombreux conflits politiques. En France, jusqu'en 1999, une cryptographie trop securisée était illégale, même pour un usage domestique. Souvent, les régimes autoritaires souhaitent maîtriser la cryptographie afin de faciliter la lutte contre le crime organisé ou leurs opposants politiques. Ces objectifs vont directement à l'encontre de la demande croissante du public pour une meilleure protection de la vie privée.

Cryptographie symétrique L'envoi d'un message par internet est comparable au fait de crier en direction de quelqu'un dans une grande salle remplie d'auditeurs. Vous pouvez vous rapprocher de votre destinataire et chuchoter, mais vous ne saurez jamais vraiment si quelqu'un vous a écouté. La cryptographie à clé secrète permet à deux personnes, Alice et Bob, de communiquer en toute sécurité dans une telle pièce de la manière suivante. Ils se rendent d'abord en cachette dans une pièce privée et conviennent d'un mot de code secret partagé, appelé *clé secrète*. Une fois qu'ils le connaissent tous les deux, Alice brouille son message à l'aide d'une procédure publique qui dépend de la clé secrète, et crie le texte brouillé, le *chiffré* à travers la pièce. Bob, grâce à sa connaissance de la clé secrète, est le seul à pouvoir inverser la procédure et récupérer le message d'Alice. Les autres personnes présentes dans la pièce n'entendront que du charabia provenant d'Alice.

Cryptographie asymétrique La partie délicate de la cryptographie à clé secrète vient du fait que les deux entités communicantes doivent se rencontrer en personne pour échanger la clé secrète. Comment peuvent-elles partager ces informations en toute sécurité si elles ne peuvent pas se rencontrer en privé, par exemple lorsqu'elles communiquent sur l'internet ou lorsqu'elles tentent de communiquer avec un trop grand nombre de personnes en même temps ? Une solution a été trouvée dans l'article de Diffie et Hellman publié en 1976[2] "New Directions in Cryptography" [DH76]. Bob génère deux clés : une *clé publique* et une *clé privée*. Si cela est fait de manière à garantir que tout message chiffré à l'aide de la clé publique ne peut être déchiffré qu'à l'aide de la clé privée, alors Alice (ou toute autre personne) peut utiliser la clé publique de Bob pour lui envoyer un message secret en toute sécurité. Notez qu'Alice et Bob n'ont pas eu besoin de se rencontrer au préalable. La cryptographie asymétrique est généralement plus coûteuse que son homologue symétrique, mais elle offre une solution au problème de l'échange de clé. Dans la pratique, on utilise d'abord un système asymétrique pour établir une clé secrète partagée, puis la communication se fait par chiffrement symétrique grâce à ce secret partagé.

[2]Bien que des documents déclassifiés du GCHQ montrent que celle-ci avait déjà été découverte plusieurs années auparavant

RSA Dans le contexte de la cryptographie à clé publique, pour s'assurer que la génération des clés garantit que tout message chiffré via la clé publique ne peut être déchiffré que via la clé privée, nous nous appuyons sur des hypothèses calculatoires. Nous illustrons cela en décrivant un schéma de chiffrement très simple mais incroyablement polyvalent : le cryptosystème RSA.

- **Génération des clés :** Deux nombres premiers p, q sont générés de façon sécurisée. Les produits $N = pq$ et $\varphi(N) = (p-1)(q-1)$ sont ensuite calculés. Un entier e est choisi, premier avec $\varphi(N)$, puis d, l'inverse de e modulo $\varphi(N)$ est calculé. La clé publique est (N, e), et la clé privée (N, d).

- **Chiffrement :** Un message est encodé par un entier positif $m < N$. Le chiffré c est obtenu par le calcul de $m^e \pmod N$. Cette étape ne demande que la connaissance de N et e, qui est assurée par la connaissance de la clé publique.

- **Déchiffrement :** Le chiffré c est déchiffré par le calcul de $c^d \pmod N$. Ce calcul ne demande que la connaissance de N et d, qui sont donnés par la clé privée.

On peut vérifier que le déchiffrement fonctionne à l'aide du petit théorème de Fermat, car $ed \equiv 1 \pmod{\varphi(N)}$ implique que $m^{ed} \equiv m \pmod N$. Pour casser RSA, un adversaire doit réussir à inverser la fonction $f : m \mapsto m^e \pmod N$, un problème qui est aujourd'hui hors de portée des ordinateurs contemporains, si les facteurs de N ne sont pas connus. Il n'est pas difficile de voir qu'il est équivalent de connaître p et q ou bien de connaître d, ce qui permet d'inverser f. D'une certaine manière, la sécurité du chiffrement RSA est directement liée à la difficulté de l'une des questions les plus fondamentales des mathématiques : la factorisation des entiers. Jusqu'à récemment, cela suffisait, et RSA était utilisé avec la cryptographie à base de courbes elliptiques (ECC) pour sécuriser la quasi-totalité de l'internet.

Calcul quantique En 1994, Shor a publié un algorithme capable, à l'aide d'un ordinateur quantique, de résoudre efficacement les problèmes de factorisation des entiers et de calcul du logarithme discret [Sho94], compromettant RSA, ainsi que ECC. Les ordinateurs quantiques sont construits sur la base de briques élémentaires qui fonctionnent selon les lois de la physique quantique. Il en existe déjà de petits prototypes, mais ceux-ci sont notoirement difficiles à passer à l'échelle. Avec un ordinateur quantique suffisamment grand et suffisamment stable, la meilleure attaque contre RSA s'exécuterait en temps polynomial (en le nombre de bits de N), ce qui représenterait une amélioration dévastatrice par rapport aux attaques classiques dont le coût est sous-exponentiel. La majeure partie de la cryptographie contemporaine est en jeu, car un ordinateur quantique suffisamment grand permettrait à un attaquant de casser n'importe quel système de cryptographie à clé publique. Il convient de noter que la cryptographie symétrique est moins touchée par les attaques quantiques, car celles-ci ne semblent bénéficier que d'une accélération quadratique.

Cryptographie post-quantique Comme cela a toujours été le cas historiquement, lorsqu'un schéma est cassé, les cryptographes en inventent un nouveau, et le cycle se répète jusqu'à ce que tout le monde ait suffisamment confiance en l'hypothèse sous-jacente. La factorisation n'étant plus une option viable, les cryptographes ont proposé de nouvelles hypothèses et construit des schémas reposant sur ces nouvelles hypothèses. L'étude des protocoles cryptographiques qui sont supposés résister aux attaquants quantiques est appelée *Cryptographie post-quantique* (PQC). Les principales familles d'objets mathématiques étudiées par la cryptographie post-quantique sont les réseaux euclidiens, les codes correcteurs d'erreurs, les systèmes multivariés d'équations quadratiques ainsi que les isogénies entre courbes elliptiques. Il convient de noter que tous les algorithmes en PQC sont classiques, c'est-à-dire qu'il n'est pas nécessaire de supposer que le défenseur a accès à du matériel quantique, mais seulement l'attaquant. La PQC ne

doit pas être confondue avec la cryptographie quantique, qui obtient des garanties de sécurité inconditionnelles grâce à la non-clonabilité des états quantiques, et ne fera pas l'objet de cette thèse. Le déploiement de la cryptographie quantique serait beaucoup moins pratique et fiable que celui de la PQC.

Normalisation de la PQC S'il y a quelques années, l'anéantissement de la cryptographie actuelle par les ordinateurs quantiques pouvait encore passer pour une théorie du complot, elle est aujourd'hui devenue une menace sérieuse que toute organisation raisonnablement soucieuse de ses risques sera disposée à atténuer. La date exacte de ce "Q-day" (date à laquelle la cryptographie à base de RSA ou ECC devient vulnérable aux ordinateurs quantiques) n'est pas encore connue, mais certaines sources bien fondées comme [fSic25] s'attendent à ce que le "Q-day" se produise d'ici une vingtaine d'années. De nombreux acteurs ont déjà commencé à enregistrer le trafic réseau chiffré d'aujourd'hui, afin de le stocker en vue de le déchiffrer plus tard, lorsque les outils seront à leur disposition. Cette stratégie est connue sous le nom de "Harvest Now, Decrypt Later" (HNDL), et oblige les organisations possédant des secrets à moyen ou long terme à commencer leur transition dès maintenant.

La transition globale vers la PQC est une tâche monumentale, qui est facilitée par les efforts de normalisation rigoureux visant à assurer un déploiement interopérable à grande échelle. Le NIST, l'Institut national des normes et de la technologie des États-Unis, est à l'avant-garde de ces efforts de normalisation. Un premier appel à algorithmes PQC a été lancé en 2016 et, après plusieurs séries d'améliorations et de cryptanalyse, quatre algorithmes ont été sélectionnés en juillet 2022 : les schémas à fondés sur les réseaux euclidiens Kyber (pour l'échange de clé), Dilithium et Falcon (pour les signatures numériques), ainsi que le schéma de signature SPHINCS+ à base de fonctions de hachage. En mars 2025, le NIST a sélectionné le schéma HQC, à base de codes correcteurs, comme protocole supplémentaire pour l'échange de clé. Des appels similaires ont été lancés par d'autres pays, notamment la Corée du Sud et la Chine. Le processus de normalisation du NIST s'est déroulé dans un contexte très international, avec des chercheurs et des entreprises du monde entier qui ont fait équipe pour prouver ou infirmer la sécurité des algorithmes candidats. Cela a conduit à des surprises tardives, le schéma Rainbow à base de systèmes multivariés et le schéma SIKE à base d'isogénies subissant des attaques dévastatrices au dernier tour de la compétition. Le NIST et la NSA recommandent conjointement de finaliser la transition avant 2030, et prévoient d'interdire RSA et ECC dès 2035.

La plupart des pays semblent faire confiance aux algorithmes sélectionnés par le NIST, bien que de nombreux pays, y compris des entités européennes telles que l'ANSSI (France) et le BSI (Allemagne), recommandent également des schémas plus conservateurs et plaident en faveur d'une hybridation pré- et post-quantique, afin de maintenir la sécurité dans un monde où les nouvelles normes post-quantiques seraient cassés avant la création d'ordinateurs quantiques cryptographiquement pertinents.

Tout le monde semble s'accorder sur le fait que la cryptographie à base de réseaux soit l'alternative la plus prometteuse à la cryptographie pré-quantique. Ces réseaux seront le sujet principal de cette thèse.

1.1 Les Problèmes de Réseaux Illustrés

Nous commençons par illustrer les réseaux et leurs problèmes algorithmiques de façon ludique. Cette section s'adresse délibérément aux non-spécialistes, un lecteur expert préférera peut-être passer outre.

Réseaux euclidiens Imaginez-vous un échiquier infini. Vous êtes un magicien debout sur une tour, qui est une unique case enchantée et que nous appelons l'origine $(0, 0)$. Vous ne pouvez pas

vous éloigner de votre tour, mais vous aimeriez quand même explorer l'échiquier et enchanter autant de cases que possible. Pour vous aider à atteindre votre objectif, vous contrôlez une

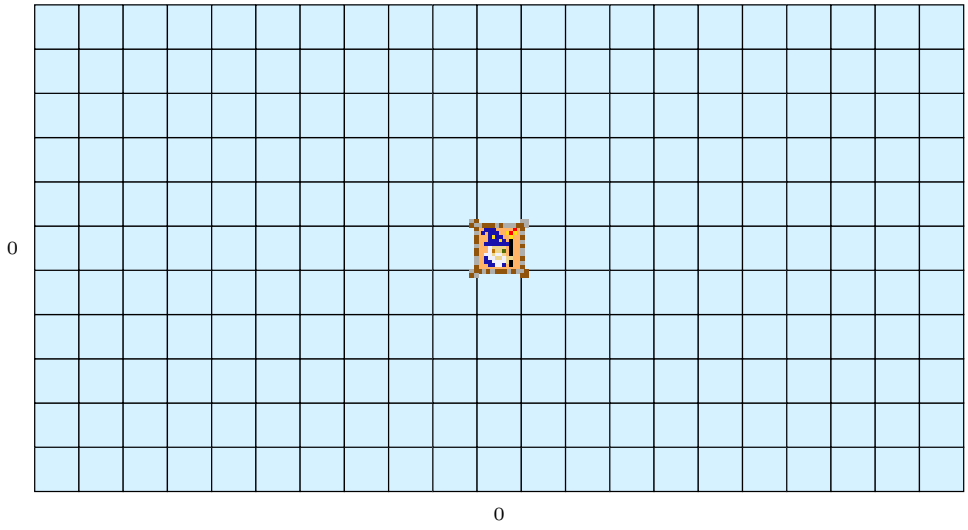


Figure 1.1: Les cases enchantées sont représentées en orange.

équipe de gnomes magiques que vous pouvez téléporter sur n'importe quelle case déjà enchantée par vous ou par l'un de vos gnomes. Chaque gnome de votre équipe se déplace, depuis sa case, dans une direction fixe, en avant ou en arrière, d'un nombre fixé de cases, ce qui lui permet d'enchanter de nouveaux territoires. Par exemple, le gnome rouge se déplace de façon similaire au cavalier du jeu d'échecs : il se déplace toujours de 1 pas vers la droite et de 2 pas vers le haut (ou la même chose dans la direction opposée : 1 pas vers la gauche et 2 pas vers le bas). Avec un gnome rouge dans votre équipe, vous pourrez explorer une infinité de cases, mais elles seront toutes alignées sur une même droite. Vous aurez besoin d'un gnome aux propriétés différentes

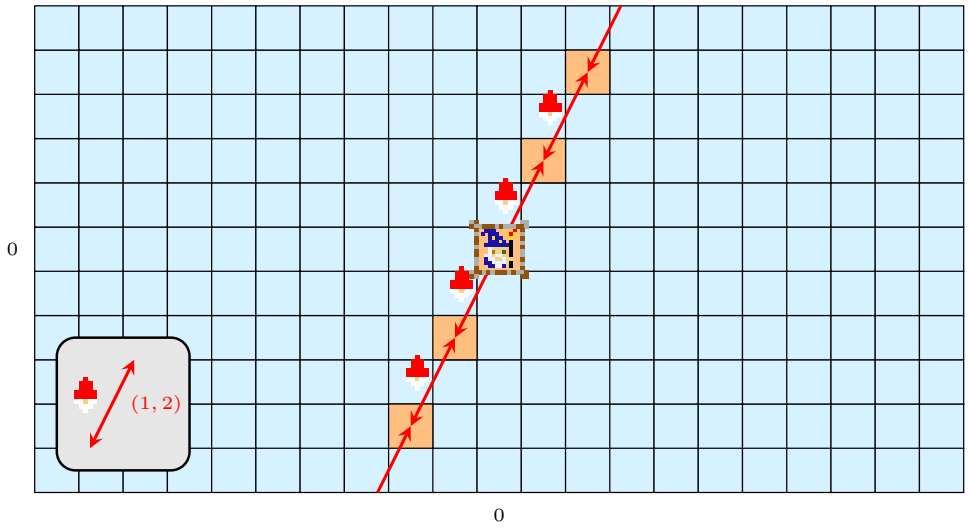


Figure 1.2: Cases atteignables par un gnome seul.

pour explorer plus de cases. Par exemple, le gnome bleu qui se déplace toujours de 2 pas vers la droite et de 1 pas vers le haut. A partir de maintenant, nous désignons les déplacements de chaque gnome par un vecteur, ainsi le gnome bleu se déplace selon le vecteur $(2, 1)$. Votre territoire est défini comme étant l'ensemble des cases que votre équipe de gnomes peut atteindre

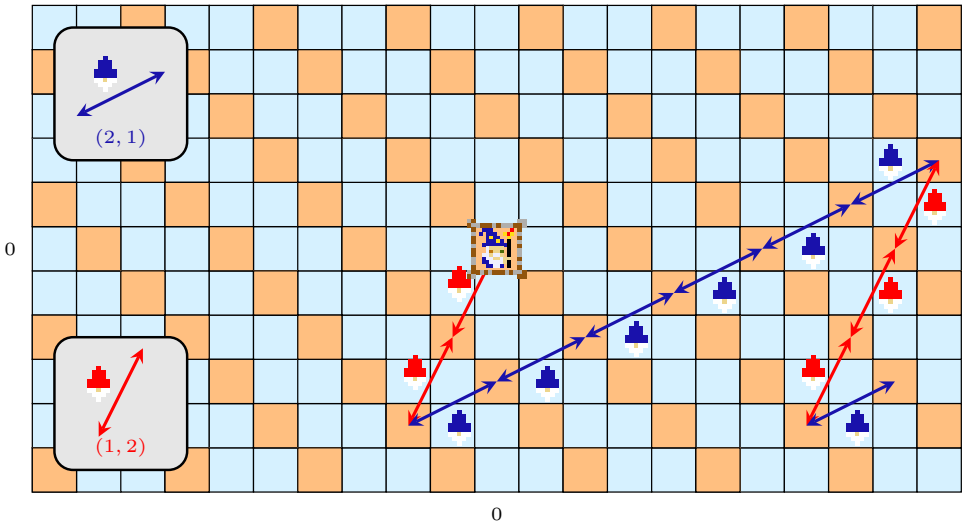


Figure 1.3: En utilisant les gnomes rouge et bleu, toutes les cases oranges sont atteignables.

(représentées en orange dans les illustrations de cette section). C'est exactement ce que nous appellerons un réseau : un ensemble discret et périodique de points dans l'espace. Notez qu'un réseau contient toujours la case du magicien, c'est-à-dire l'origine.

Bases et réduction Le nombre minimal de gnomes nécessaires pour enchanter un réseau donné est appelé son rang, et une équipe de taille minimale qui enchante exactement ce même réseau est appelée une base du réseau. Cette dernière définition suggère que certains gnomes pourraient être redondants. En effet, il n'est pas difficile de voir que le gnome rouge foncé avec un mouvement $(2, 4)$ est inutile par rapport au gnome rouge, puisque le gnome rouge peut recréer le mouvement du gnome rouge foncé en se déplaçant simplement deux fois. Dans d'autres cas, il n'est pas toujours évident de savoir si un gnome est utile ou non. Prenons par exemple le gnome violet avec un mouvement de $(11, 7)$. Est-il inutile ? C'est plus difficile à dire, car il n'est pas directement plus faible que les gnomes rouge et bleu. Cependant, une analyse attentive montre qu'il n'aide pas à enchanter de cases en dehors du réseau obtenu via la base constituée des gnomes rouge et bleu. Nous considérons maintenant le réseau obtenu en utilisant une équipe de gnomes violets et verts, où le gnome vert a un mouvement $(13, 8)$. C'est un bon exercice que de montrer que le territoire généré par le violet et le vert est en fait le même que celui généré précédemment par le rouge et le bleu. Il s'agit d'une observation fondamentale : un réseau peut avoir plusieurs bases ! Dans ce qui suit, nous supposerons que le réseau est de rang 2.

Volume d'un réseau Therefore, we say that the volume[3] of the corresponding lattice is 3. A great property of the volume is that it can be computed directly from any basis. If the lattice has rank 2, then its bases consist in two vectors (gnomes). The two vectors define a parallelogram, whose area happens to always equal exactly the volume. It is therefore not necessary to fully map the enchanted territory in order to deduce the proportion of enchanted squares.

Une quantité intéressante liée au réseau est son volume. Il s'agit d'un invariant fondamental du réseau qui mesure la proportion de cases de l'échiquier qui ont été enchantées. Visuellement, nous voyons que le territoire enchanté en Figure 1.3 occupe exactement un tiers de l'ensemble des cases. Par conséquent, nous disons que le volume[4] du réseau correspondant est 3. Une propriété

[3]The correct mathematical terminology here would be *co-volume*, which should feel right as we have inverted 1/3 under the hood.

[4]La terminologie mathématiquement correcte ici serait *covolume*, ce qui devrait sembler correct puisque nous

importante du volume est qu'il peut être calculé directement à partir de n'importe quelle base. Si le réseau est de rang 2, alors ses bases sont constituées de deux vecteurs (gnomes). Ces deux vecteurs définissent un parallélogramme dont l'aire est toujours exactement égale au volume. Il n'est donc pas nécessaire de cartographier entièrement le territoire enchanté pour en déduire la proportion de cases enchantables.

Problèmes algorithmiques sur les réseaux Bien qu'un réseau de l'échiquier infini soit un objet infini, une base de seulement 2 gnomes suffit pour stocker toute l'information qu'il contient. Cela signifie que 4 entiers suffisent pour représenter le réseau : les vecteurs de mouvement de chaque gnome. Imaginez maintenant que la case du réseau la plus proche de l'origine (sans être elle-même l'origine) contienne un coffre au trésor. Avec votre équipe de deux gnomes, comment pouvez-vous, en tant que magicien, utiliser les gnomes à votre disposition pour efficacement atteindre le coffre au trésor ? Cette question est en fait généralement appelée le problème du vecteur le plus court (SVP) : la principale hypothèse calculatoire sous-jacente à la cryptographie post-quantique.

Pourquoi ce problème est-il intéressant ? Deux magiciens différents ayant accès à des bases différentes du même réseau auront une expérience très différente en essayant de récupérer le coffre au trésor. Utilisons le même réseau que dans la section précédente. Le coffre au trésor se trouve aux coordonnées $(1, -1)$, mais les magiciens ne le savent pas. Notez que $(-1, 1)$ est tout aussi proche de l'origine, mais nous n'en tiendrons pas compte, car le proplème du plus court vecteur demande seulement de récupérer l'un des plus courts vecteurs non nuls du réseau.

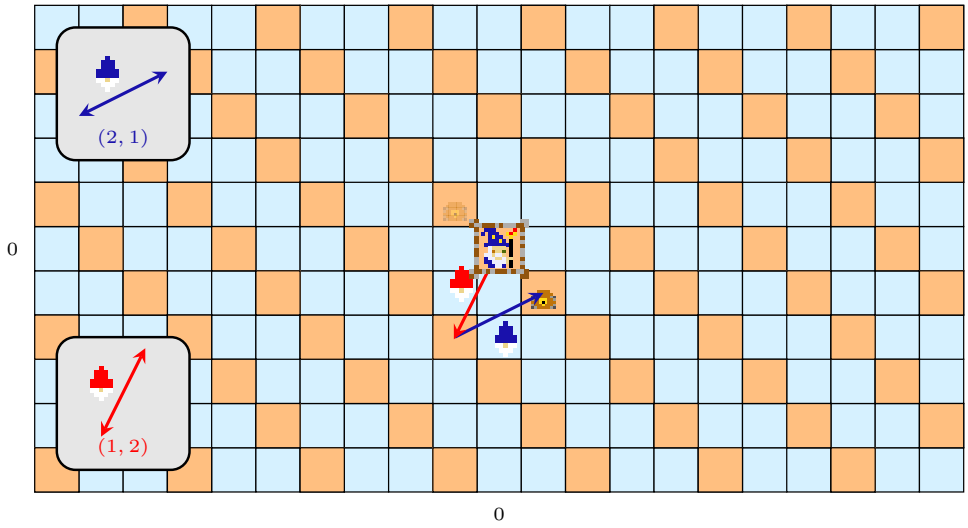


Figure 1.4: Un chemin simple vers le coffre au trésor.

- Un magicien ayant dans son équipe les gnomes rouge et bleu n'aura aucun mal à trouver le trésor. En effet, ses gnomes lui permettent d'explorer presque immédiatement la case $(1, -1)$ en utilisant d'abord le gnome rouge, puis le gnome bleu, comme on peut le voir en Figure 1.4. On dit que l'équipe rouge-bleue est une bonne base du réseau.

- Un magicien ne disposant que des gnomes violet et vert aura beaucoup plus de mal à trouver un chemin vers le trésor. En effet, les deux gnomes ont des mouvements longs et maladroits, presque parallèles, et le chemin le plus court vers le trésor nécessite une séquence compliquée de longs mouvements vers l'avant et vers l'arrière. Le chemin le plus

avons implicitement inversé 1/3.

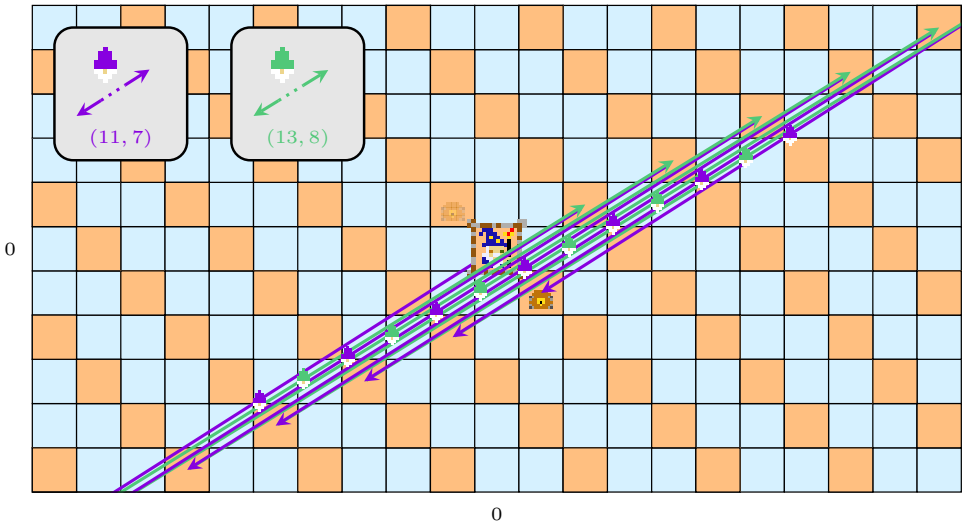


Figure 1.5: Un chemin plus compliqué.

court est caché dans l'une de ces séquences maladroites. En fait, on peut vérifier via la Figure 1.5 qu'un chemin vers le trésor utilisant les gnomes violet et vert nécessite au moins 13 déplacements ! En particulier, avec cette équipe en comparaison avec la précédente, la stratégie consistant à suivre une suite aléatoire de déplacements a moins de chances de passer par la case contenant le trésor, puisqu'il existe un grand nombre de chemins possibles à 13 déplacements. Nous dirons que la base violet-vert est une mauvaise base du réseau.

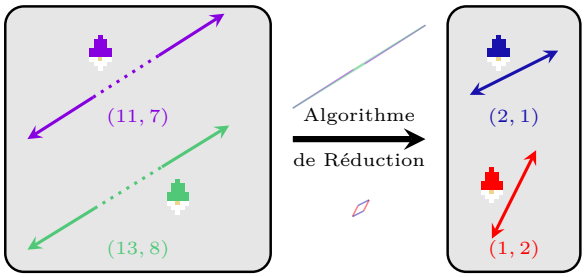


Figure 1.6: La réduction améliore la qualité d'une base.

Réduction de réseau Une mesure raisonnable pour juger de la qualité d'une base est la taille relative et l'orthogonalité des vecteurs de la base. En effet, les bases qui génèrent un parallélogramme très long et plat sont de piètre qualité. Il existe plusieurs façons de quantifier cette notion de *qualité*. Dans notre exemple en deux dimensions, on peut choisir de la mesurer en considérant la longueur normalisée de la grande diagonale du parallélogramme associé à notre base : une diagonale plus courte est associée à une meilleure qualité. La réduction de réseau désigne le processus suivant : au lieu de chercher directement le trésor à l'aide de son équipe de gnomes, le magicien peut mesurer la qualité de son équipe et, si cette qualité n'est pas satisfaisante, il lui applique un algorithme qui convertit la mauvaise équipe en une équipe de meilleure qualité. Il peut ensuite chercher le trésor avec une équipe améliorée, ce qui rend la recherche plus facile. Comprendre la réduction de réseaux – ce processus qui transforme une

mauvaise équipe de gnomes en une meilleure équipe – est le point clé de la cryptographie à base de réseaux, puisque si un adversaire était capable de trouver le trésor caché au niveau du plus court vecteur, il invaliderait la sécurité du protocole.

1.2 Réseaux et Algorithmes

En termes plus mathématiques, un réseau L est un sous-groupe additif discret de $\mathbb{R}^m$. La dimension de l'espace vectoriel réel engendré par les vecteurs de L est son rang n. Une base est une famille de vecteurs linéairement indépendants $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ dont les combinaisons $\mathbb{Z}$-linéaires génèrent L. Le (co)-volume d'un réseau L est défini par $\mathrm{vol}(L) = \sqrt{\det(\mathbf{B}\mathbf{B}^T)}$. Il est bien défini, puisque n'importe quelles bases $\mathbf{B}_1$ et $\mathbf{B}_2$ de L sont liées par une matrice de changement de base $\mathbf{U} \in \mathcal{M}_n(\mathbb{Z})$, avec $\det(\mathbf{U}) = \pm 1$, par $\mathbf{U}\mathbf{B}_1 = \mathbf{B}_2$, et il s'ensuit que

$$\det(\mathbf{B}_2\mathbf{B}_2^T) = \det(\mathbf{U}\mathbf{B}_1\mathbf{B}_1^T\mathbf{U}^T) = \det(\mathbf{U})\det(\mathbf{B}_1\mathbf{B}_1^T)\det(\mathbf{U}^T) = \det(\mathbf{B}_1\mathbf{B}_1^T),$$

ce qui confirme que le volume est indépendant de la base. Le volume peut aussi être vu comme le volume (en valeur absolue) du parallélépipède défini par les vecteurs de base $(\mathbf{b}_1, \ldots, \mathbf{b}_n)$. Une analyse dimensionnelle nous indique qu'il est naturel de comparer les longueurs des vecteurs du réseau avec la quantité $\mathrm{vol}(L)^{1/n}$. Le théorème de Minkowski affirme en effet que la longueur $\lambda_1(L)$ du plus court vecteur non nul du réseau doit satisfaire l'inégalité suivante

$$\lambda_1(L) \leq \sqrt{n} \cdot \mathrm{vol}(L)^{1/n}.$$

De fait, pour un réseau pris aléatoirement, on s'attend avec très forte probabilité à ce que l'égalité soit vérifiée à une constante près :

$$\lambda_1(L) \simeq \sqrt{\frac{n}{2\pi e}} \cdot \mathrm{vol}(L)^{1/n}.$$

Dans le problème du plus court vecteur (SVP), nous devons trouver un vecteur $\mathbf{v}$ du réseau L vérifiant $\|\mathbf{v}\| = \lambda_1(L)$, étant donnée une famille génératrice du réseau. Ce problème peut être relâché : dans le cas du problème γ-HSVP, un facteur d'approximation supplémentaire $\gamma > 0$ est imposé, et le nouvel objectif est de trouver un vecteur $\mathbf{v} \in L$ tel que $\|\mathbf{v}\| \leq \gamma \cdot \mathrm{vol}(L)^{1/n}$.

En dimension 1, le problème SVP est trivial étant donné une base. Avec une famille génératrice $(a, b) \in \mathbb{Z}^2$, le réseau correspondant est $a\mathbb{Z} + b\mathbb{Z} = \mathrm{pgcd}(a, b)\mathbb{Z}$, et ainsi le premier minimum λ_1 du réseau peut être obtenu par une application directe de l'algorithme d'Euclide.

Dans l'exemple imagé de la section précédente, le réseau considéré était de rang 2 ; dans cette situation, SVP peut être résolu efficacement grâce à l'algorithme de Lagrange-Gauss, cet algorithme pouvant être vu comme une généralisation de l'algorithme d'Euclide.

Pour les applications en cryptographie, les réseaux qui interviennent auront bien plus de dimensions, typiquement entre 500 et 1000, ce qui rend le problème bien plus dur, même pour un ordinateur quantique. Les meilleurs algorithmes pour la version exacte de SVP ont une complexité exponentielle en $2^{n+o(n)}$. Pourtant, le fameux algorithme LLL (Lenstra-Lenstra-Lovász, [LLL82]) est polynomial en le rang n et la taille en bits des vecteurs de la base $\mathbf{B}$ donnée en entrée. Il renvoie une nouvelle base réduite dont le premier vecteur $\mathbf{b}_1$ satisfait

$$\|\mathbf{b}_1\| \leq 2^{\frac{n-1}{4}} \cdot \mathrm{vol}(L)^{1/n}. \tag{1.1}$$

En d'autres termes, LLL résoud $2^{(n-1)/4}$-HSVP en temps polynomial. LLL fonctionne en résolvant SVP sur un ensemble choisi itérativement de sous-réseaux projetés de rang 2. Ce concept peut être généralisé : en supposant l'existence d'un oracle en boîte noire capable de résoudre SVP pour des réseaux de rang β, l'algorithme BKZ-β (Blockwise Korkine Zolotarev) appelle itérativement l'oracle sur des réseaux de dimension β bien choisis, et insère progressivement des

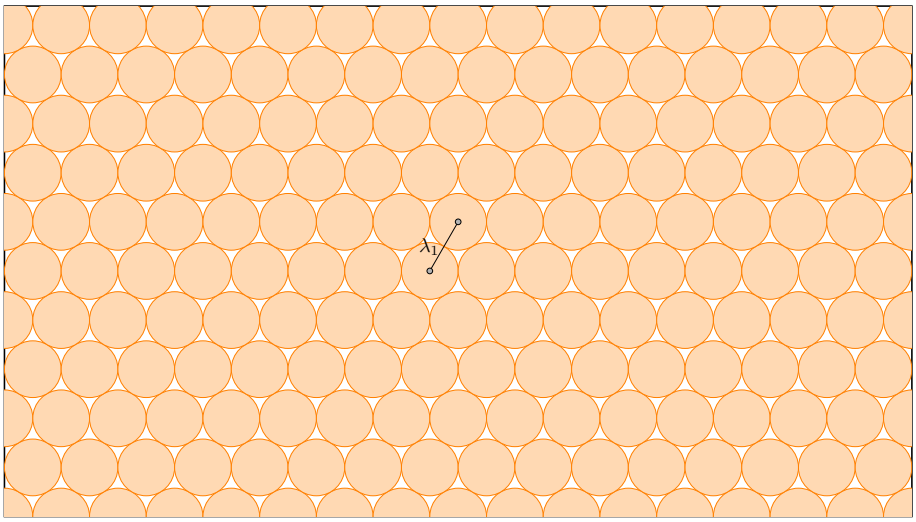


Figure 1.7: Un empilement de sphères selon le réseau hexagonal.

vecteurs de plus en plus courts dans la base initiale, jusqu'à ce que cette base de n vecteurs soit suffisamment réduite. Les algorithmes de type BKZ permettent un arbitrage entre temps d'exécution et facteur d'approximation : BKZ-β résoud SVP pour un facteur d'approximation $2^{O(n/\beta)}$ en temps $2^{O(\beta)}$.

Il est intéressant de remarquer que le paramètre clé de la cryptographie à base de réseaux, à savoir le premier minimum λ_1, est directement relié au problème connu de l'empilement de sphères. Ce problème pose la question suivante : quelle fraction de l'espace peut être recouverte par des boules ouvertes disjointes de même rayon sachant que leurs centres doivent être des points d'un réseau ? Pour un réseau donné L, le plus grand rayon possible pour des boules centrées en L est précisément $\lambda_1(L)/2$, comme on peut l'observer en Figure 2.7. Ces problèmes d'empilements sont aussi étudié par les physiciens et les crystallographes.

Une application de la réduction L'algorithme LLL a eu des applications révolutionnaires dans plusieurs branches des mathématiques. Il a été développé dans le but de factoriser efficacement des polynômes à coefficients rationnels, un problème qui pendant longtemps était supposé difficile par analogie avec la factorisation des entiers. Dans cette section, nous donnons un exemple d'une autre application classique de la réduction de réseaux, dans l'espoir de donner au lecteur un ressenti de la puissance de cet algorithme.

Question 1.2.1. *Peut-on trouver une combinaison linéaire entière courte des constantes π, e et 1 qui soit très proche de 0 ? Plus précisément, soit $\varepsilon > 0$. Peut-on trouver des entiers $A, B, C \in \mathbb{Z}$ tels que $|A\pi + Be + C| < \varepsilon$ et tels que la norme $\sqrt{A^2 + B^2 + C^2}$ soit petite ?*

La réponse est oui, à travers la construction d'un réseau L dans lequel une solution à notre problème est obtenue en calculant le plus court vecteur non nul de L. À cet effet nous posons une variable $\alpha \in \mathbb{Z}_{>0}$ qui sera fixée plus tard, et définissons

$$\mathbf{B}_\alpha = \begin{pmatrix} 1 & 0 & 0 & \lfloor \alpha\pi \rfloor \\ 0 & 1 & 0 & \lfloor \alpha e \rfloor \\ 0 & 0 & 1 & \lfloor \alpha \rfloor \end{pmatrix}.$$

Les vecteurs de $\mathbf{B}_\alpha$ génèrent un réseau L_α de $\mathbb{Z}^4$, de rang 3 et de volume

$$\mathrm{vol}(L_\alpha) = \sqrt{1 + \lfloor \alpha\pi \rfloor^2 + \lfloor \alpha e \rfloor^2 + \lfloor \alpha \rfloor^2} \leq \sqrt{1 + \alpha^2(\pi^2 + e^2 + 1)} \leq 2\alpha\sqrt{\pi^2 + e^2 + 1}.$$

En applicant LLL sur $\mathbf{B}_\alpha$, on obtient une nouvelle base dont le premier vecteur $\mathbf{v} = (A, B, C, D)$ est court et doit satisfaire $\|\mathbf{v}\| \leq \sqrt{2}\mathrm{vol}(L_\alpha)^{1/3}$ grâce à la borne de l'équation 1.1. Comme $\mathbf{v}$ est un vecteur du réseau, D vérifie $D = A\lfloor\alpha\pi\rfloor + B\lfloor\alpha e\rfloor + C\lfloor\alpha\rfloor$. On a donc

$$\begin{aligned}
|A\pi + Be + C| &= \frac{1}{\alpha}|A\{\alpha\pi\} + B\{\alpha e\} + C\{\alpha\} + D| \\
&\leq \frac{1}{\alpha}\|\mathbf{v}\|\sqrt{\{\alpha\pi\}^2 + \{\alpha e\}^2 + \{\alpha\}^2 + 1^2} \\
&\leq \frac{\sqrt{3}}{\alpha}\|\mathbf{v}\| \\
&\leq \frac{\sqrt{6}}{\alpha}(2\alpha\sqrt{\pi^2 + e^2 + 1})^{1/3} \\
&\leq \kappa \cdot \alpha^{-2/3},
\end{aligned}$$

pour une certaine constante $\kappa > 0$, où $\{x\} = x - \lfloor x\rfloor$ est la partie fractionnaire de x. Dans l'ordre, nous avons utilisé l'inégalité de Cauchy-Schwarz, le fait que $\{x\} \leq 1$, la borne sur $\|\mathbf{v}\|$ donnée par LLL, et la borne sur $\mathrm{vol}(L_\alpha)$ obtenue précédemment. Par conséquent il suffit de choisir $\alpha \geq (\kappa/\varepsilon)^{3/2}$ pour s'assurer que $|A\pi + Be + 1| < \varepsilon$. On peut même obtenir une borne sur $\sqrt{A^2 + B^2 + C^2}$ en remarquant que cette quantité est bornée supérieurement par $\|\mathbf{v}\|$. La borne ainsi obtenue est proportionelle à $\varepsilon^{-1/2}$.

Par exemple, en posant $\varepsilon = \frac{1}{2025}$, on trouve

$$19\pi + 6e = 76.000049.$$

Ce résultat pourrait être perçu comme une simple coïncidence, mais la capacité à trouver de manière efficace de telles quasi-relations est l'un des outils les plus puissants de la cryptanalyse moderne. Les paramètres des protocoles cryptographiques à base de réseaux sont calculés à partir du coût estimé des meilleures attaques existantes. Ainsi comprendre la réduction de réseaux est essentiel à la sécurité de ces protocoles. La première moitié de ce manuscrit (Chapitres 4,5,6) étudie la réduction de réseau appliquée à des réseaux utilisés par de tels protocoles.

1.3 Signatures Numériques à Base de Réseaux

La difficulté des problèmes de réseaux conditionne la sécurité du protocle d'échange de clé ML-KEM (Kyber), et des schémas de signature ML-DSA (Dilithium) et FN-DSA (Falcon), désormais tous les trois normalisés par le NIST.

Chacune de ces catégories d'algorithmes remplit des fonctions bien précises :

- L'objectif principal d'un protocole d'échange de clé (KEMs) est d'assurer la *confidentialité* de la donnée chiffrée.
- Les principales fonctions des signatures numériques sont l'*authentification*, l'*integrité* et la *non-répudiation* : celles-ci assurent respectivement l'identité de l'émetteur, le fait que le message n'ait pas été modifié, et le fait que l'émetteur ne puisse pas nier l'avoir envoyé.

Si la protection des signatures numériques face aux adversaires quantiques est essentielle, celle-ci n'est pas aussi urgente que la protection des KEM. En effet, dans le contexte du chiffrement, la menace HNDL met en danger les données à la fois présentes et futures. En revanche, pour les signatures, la vulnérabilité n'est pas rétroactive, une signature émise aujourd'hui à l'aide d'un ordinateur classique restera valable. Ce n'est que lorsqu'un ordinateur quantique sera opérationnel qu'il sera possible pour un adversaire de forger une signature. Cela nous laisse encore un peu de temps pour réfléchir à la conception des signatures, ce qui sera l'un des sujets principaux de ce manuscrit (Chapitres 6 et 7). Dans ce contexte, le NIST à émis un nouvel

appel à candidatures pour des schémas de signatures post-quantiques en 2023. Il faut noter cependant qu'un calendrier de transition devrait également prendre en compte à la fois le fait que les changements à l'échelle d'une organisation prendront inévitablement du temps, et que les clés de nombreux systèmes embarqués ne pourront pas facilement être remplacées, ce qui pourrait compromettre l'authentification à long terme.

Nous faisons ici une présentation rapide des deux principales techniques qui peuvent être utilisées pour concevoir des protocoles de signature fondés sur la difficulté des problèmes de réseaux.

Hash and Sign Hash-and-Sign est une technique de signature, dans laquelle le signataire a accès à une fonction de trappe à sens unique, c'est à dire une fonction f qui est facile à calculer, et dont l'inversion est computationellement impossible sans connaissance d'une information secrète bien particulière appelée la trappe. Le message μ est haché à l'aide d'une fonction de hachage cryptographique H, le signataire utilise alors sa fonction de trappe à sens unique pour calculer un σ tel que

$$f(\sigma) = H(\mu).$$

σ est alors la signature, et quiconque ayant accès à (μ, σ) est en capacité de vérifier la signature en évaluant l'équation. Voici deux exemples d'utilisation du procédé Hash-and-Sign.

- **Signatures RSA :** la fonction de trappe à sens unique pour RSA est l'exponentiation modulaire $x \mapsto x^e \pmod N$, qui est difficile à inverser sans la connaissance de d, l'inverse de e modulo $\varphi(N)$.

- **Signatures de type GPV :** la clé privée est une bonne base du réseau, qui agit comme une trappe dans le calcul d'une solution courte $\mathbf{s} \in \mathbb{Z}^n$ à l'équation $\mathbf{As} \equiv H(\mu) \pmod q$, où μ est le message, et q et $\mathbf{A}$ sont un entier et une matrice qui définissent un réseau. La vérification teste si $\mathbf{s}$ est effectivement court, et si l'équation est vérifiée. Dans le contexte de la cryptographie à base de réseaux, le problème de trouver un $\mathbf{s}$ court qui satisfasse l'équation est appelé le problème de la "petite solution entière" (SIS), et peut être reformulé comme le problème de trouver un vecteur proche d'un certain réseau défini par $\mathbf{A}$ et q, un problème analogue à SVP.

Fiat-Shamir avec Rejet Cette approche utilise la transformée de Fiat-Shamir générique pour convertir un schéma d'identification (ID) en un schéma de signature.

- **Schéma d'identification (ID) :** Le but d'un tel protocole est de permettre à un *Prouveur* de démontrer de manière interactive à un *Vérifieur* qu'il connaît une clé secrète `sk` (qui est liée à son identité, car il est le seul à la connaître), sans révéler aucune information sur `sk`. Les schémas d'identification sont interactifs : le Prouveur s'engage d'abord sur une valeur aléatoire r et envoie un témoin $w =$ `Commit(r)` au Vérifieur. Ceci garantit que le Prouveur ne peut pas tricher sur sa valeur de r sans pour autant la révéler. Le Vérifieur génère ensuite un défi aléatoire c et l'envoie au Prouveur. Finalement, le Prouveur envoie sa preuve $z =$ `Prove(sk,r,c)` au Vérifieur. Le Vérifieur utilise alors la clé de vérification publique `vk` associée à `sk` pour vérifier si `Verify(vk,z,c)`. Le schéma est *complet* si un Prouveur connaissant `sk` peut convaincre un vérifieur honnête ; ainsi, `Verify()` doit renvoyer vrai si z est généré conformément au protocole. Le schéma doit également être *correct* (en anglais, *sound*), c'est-à-dire qu'un Prouveur malhonnête ne connaissant pas `sk` ne devrait pas être en mesure de convaincre le Vérifieur, sauf avec une probabilité négligeable.

- **Transformée de Fiat-Shamir :** La transformée de Fiat-Shamir convertit le protocole interactif ci-dessus en une version non interactive. Pour ce faire, le Prouveur (renommé le Signataire) définit lui-même le défi c de manière déterministe, comme le haché de l'engagement concaténé avec le message. Si la fonction de hachage se comporte comme un oracle aléatoire, cette étape simule la génération d'un c aléatoire par le Vérifieur. Finalement, la preuve z sert de signature à vérifier. Il est à noter que c doit également être inclus dans la signature, sans quoi il serait impossible pour le Vérifieur de le deviner.

L'utilisation classique de Fiat-Shamir remonte aux schémas d'identification et de signature de Schnorr et repose sur la difficulté du logarithme discret : le Prouveur convainc le Vérifieur qu'il connaît un exposant secret s. La clé de vérification est g^s, où g est un générateur d'un groupe cyclique. Dans ce contexte, $z = r + cs$ masque parfaitement cs, car r est choisi de manière uniforme et aléatoire, et l'application d'une translation à la distribution uniforme résulte en la même distribution uniforme.

De manière assez contrariante dans le cas des réseaux euclidiens, l'information secrète est typiquement un vecteur court $\mathbf{s}$, et l'ajout d'un vecteur aléatoire $\mathbf{y}$ à l'information secrète $c\mathbf{s}$ ne la masque pas aussi bien que dans l'exemple précédent. En effet, $\mathbf{y}$ est échantillonné dans un ensemble infini (au lieu d'un ensemble fini pour r), des choix doivent donc être faits concernant sa distribution. Par exemple, $\mathbf{y}$ pourrait être échantillonné uniformément dans un hypercube de rayon R, en échantillonnant chaque coordonnée indépendamment dans $[-R, R] \cap \mathbb{Z}$. Cependant, il est aisé de voir que pour certaines valeurs de $\mathbf{y}$, $\mathbf{z} = \mathbf{y} + c\mathbf{s}$ divulgue de l'information : par exemple, si $\mathbf{z}$ a une coordonnée qui se situe en dehors de l'intervalle $[-R, R]$, le Vérifieur obtient de l'information sur la coordonnée correspondante de $c\mathbf{s}$. Pour corriger cette vulnérabilité, les schémas Fiat-Shamir à base de réseaux ont recours à l'échantillonnage avec rejet : si le protocole produit un élément $\mathbf{z}$ qui divulgue de l'information sur $c\mathbf{s}$, alors $\mathbf{z}$ est rejeté et le protocole recommence jusqu'à ce qu'un $\mathbf{z}$ satisfaisant soit trouvé, d'où le nom de Fiat-Shamir *avec Rejet*. La conception de tels schémas exige de choisir et d'analyser méticuleusement les distributions, ainsi que de sélectionner des paramètres de manière à minimiser le nombre attendu de rejets, tout en préservant la sécurité.

Discussion La signature Falcon utilise le paradigme Hash-and-Sign et présente les performances les plus impressionnantes parmi les nouvelles normes à base de réseaux. Cependant, le schéma est très difficile à implémenter de manière sécurisée, car la trappe nécessite l'utilisation de distributions gaussiennes, qui requièrent de l'arithmétique en virgule flottante de haute précision. À l'inverse, Dilithium utilise Fiat-Shamir avec Rejet et évite explicitement d'utiliser des distributions gaussiennes, ce qui le rend beaucoup plus facile à implémenter de manière sécurisée, bien qu'en contrepartie Dilithium ait des signatures de plus grande taille. Dilithium choisit une distribution uniforme dans un hypercube pour la distribution de $\mathbf{y}$. Nous consacrons une part importante de ce manuscrit à l'étude des signatures numériques, en présentant la cryptanalyse d'un schéma Hash-and-Sign dans le Chapitre 6, et la conception d'un schéma Fiat-Shamir avec Rejet avec un choix de distribution différent et motivé pour $\mathbf{y}$ dans le Chapitre 7.

1.4 Réseaux Particuliers en Cryptographie

Réseaux réels Les réseaux de rang plein dans $\mathbb{R}^n$ et de co-volume un peuvent être identifiés à l'espace

$$X_n = \mathrm{SL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{Z}),$$

de telle sorte qu'un point $[\mathbf{B}] \in X_n$ engendre le réseau $\mathbb{Z}^n\mathbf{B}$. Cette définition est cohérente, car deux représentants de la même classe sont égaux à une matrice unimodulaire $\mathbf{U} \in \mathrm{SL}_n(\mathbb{Z})$ près, et $\mathbf{B}$ et $\mathbf{BU}$ engendrent le même réseau. L'espace des réseaux X_n est un groupe topologique

localement compact, ce qui signifie qu'il possède une unique mesure de Haar μ_n (à constante près) : une mesure invariante par multiplication à droite[5] par les éléments de $\mathrm{SL}_n(\mathbb{Z})$. Siegel [Sie45] a prouvé que $\mu_n(X_n) < \infty$, on peut donc la normaliser pour en faire une mesure de probabilité. Cette mesure est sans doute la manière la plus naturelle de définir les réseaux réels aléatoires, pour lesquels certaines propriétés moyennes peuvent être établies. Les travaux classiques de Siegel, Rogers et d'autres étudient le nombre moyen de points d'un réseau dans une boule centrée à l'origine, une quantité liée à la norme des plus courts vecteurs du réseau.

Réseaux entiers Les cryptographes n'aiment pas représenter les nombres réels sur un ordinateur, c'est pourquoi ils ne considèrent généralement que des sous-réseaux de $\mathbb{Z}^m$, dont les coefficients sont tous entiers. Étant donné un groupe abélien fini G, on considère l'espace de réseaux suivant :

$$L(G) = \{L \subseteq \mathbb{Z}^m : \mathrm{rank}(L) = m, \mathbb{Z}^m/L \cong G\}.$$

$L(G)$ est fini, et l'on peut donc y échantillonner des réseaux selon la distribution uniforme. Dans son article de 1996 intitulé "Generating Hard Instances of Lattice Problems", Ajtai [Ajt96] a prouvé une réduction du cas le pire au cas moyen pour les problèmes sur les réseaux de $L((\mathbb{Z}/q\mathbb{Z})^n)$. Ce résultat a été fondamental pour la cryptographie à base de réseaux. Il montre que si un attaquant peut résoudre SVP pour un réseau aléatoire (cas moyen), alors il est également capable de résoudre SVP dans le cas le pire. Ceci est une preuve solide que SVP est aussi difficile dans tous les réseaux de cette classe. Ce résultat est rénéralisé par les auteurs de [GINX16] dans le cas où (G_n) est une suite de groupes abéliens telle que $|G_n|$ croît suffisamment vite vers l'infini. Dans ce cas, les auteurs de [EO06] démontrent que la distribution obtenue en échantillonnant un réseau uniforme dans $L(G_n)$ et en le normalisant à co-volume 1 converge vers la mesure de Haar μ_n. Ainsi, pour n assez grand, on s'attend à ce que les résultats en moyenne soient valables pour les deux distributions.

Réseaux NTRU L'année 1996 fût bonne pour la cryptographie à base de réseaux ; en effet, elle marque aussi l'invention du cryptosystème NTRU par Hoffstein, Pipher et Silverman. La sécurité de NTRU dépend également de la facilité à retrouver des vecteurs courts dans un réseau, mais cette fois le réseau Λ possède une structure très particulière : il est invariant par une certaine permutation cyclique de ses coordonnées. Par exemple, si le vecteur

$$\mathbf{v} = (1, 2, 3, 4, 5, 6) \in \Lambda$$

est un vecteur du réseau, alors les vecteurs suivants le sont aussi : $(3, 1, 2, 6, 4, 5)$ et $(2, 3, 1, 5, 6, 4)$. Ceci permet une représentation plus compacte du réseau via des polynômes : le vecteur $\mathbf{v} \in \mathbb{Z}^6$ ainsi que ses décalages sont représentés simultanément par un unique vecteur de polynômes $(f, g) \in (\mathbb{Z}[X]/(X^3-1))^2$. Dans notre exemple, on aurait $f = 1+2X+3X^2$ et $g = 4+5X+6X^2$. Le quotient indique essentiellement que $X^3 = 1$, ce qui signifie que les deux décalages de $\mathbf{v}$ sont exactement $(X \cdot f, X \cdot g)$ et $(X^2 \cdot f, X^2 \cdot g)$. Cette représentation polynomiale permet un stockage plus efficace et, pour des anneaux de polynômes bien choisis et un module additionnel q, elle autorise l'utilisation de la NTT (Number Theoretic Transform), une opération similaire à la transformée de Fourier rapide qui accélère la multiplication polynomiale, rendant les schémas avec une telle structure plusieurs ordres de grandeur plus efficaces.

Malheureusement, $X^3 - 1$ n'est pas irréductible dans $\mathbb{Q}[X]$, ce qui signifie que l'anneau quotient $\mathbb{Q}[X]/(X^3 - 1)$ ne définit pas un corps et peut en fait être décomposé en un produit de corps plus petits : les attaquants peuvent utiliser cette structure additionnelle. Pour cette raison, les schémas plus récents à base de réseaux utilisent des réseaux provenant de corps de nombres : un corps de nombres $\mathbb{Q}(\alpha)$ est le plus petit corps contenant 1 et α, où α est un nombre algébrique.

[5] Il se trouve que μ_n est également invariante à gauche.

Cela signifie que α est la racine d'un polynôme unitaire irréductible f à coefficients dans $\mathbb{Z}$, ce qui implique que $\mathbb{Q}(\alpha)$ peut aussi être représenté par l'anneau de polynômes $\mathbb{Q}[X]/f(X)$.

Réseaux idéaux De la même manière qu'un corps de nombres $K = \mathbb{Q}(\alpha)$ généralise les rationnels $\mathbb{Q}$, son anneau des entiers $\mathcal{O}_K$ généralise les entiers relatifs $\mathbb{Z}$. Les idéaux de $\mathcal{O}_K$ sont des ensembles $I \subseteq \mathcal{O}_K$ qui sont stables par multiplication par les éléments de $\mathcal{O}_K$: $\mathcal{O}_K \cdot I \subseteq I$. Cette définition s'applique déjà dans $\mathbb{Z}$, où les idéaux sont exactement les ensembles de la forme $a\mathbb{Z}$, pour $a \in \mathbb{Z}$. Un fait fondamental de la branche de la théorie des nombres appelée géométrie des nombres est que les idéaux et les réseaux sont essentiellement la même chose : les éléments d'un corps de nombres K de degré n sont envoyés canoniquement sur des points de $\mathbb{R}^n$, muni du produit scalaire usuel, et les idéaux sont envoyés sur des réseaux. Ce plongement – généralement appelé plongement canonique ou de Minkowski – préserve la structure algébrique des éléments du corps et nous permet de voir certains types de réseaux, appelées réseaux idéaux, comme des objets algébriques représentés par des idéaux du corps de nombres. Le problème de la recherche de vecteurs courts (où court se réfère ici à la norme euclidienne dans le réseau plongé) dans les idéaux est appelé Ideal-SVP. Le problème de l'apprentissage avec erreurs sur les anneaux (Ring-LWE) [SSTX09; LPR10] – une instanciation du problème LWE de Regev [Reg05] sur les corps de nombres – s'est avéré être prouvablement au moins aussi difficile que les pires instances d'Ideal-SVP.

L'ajout d'une structure supplémentaire à une primitive cryptographique annonce un compromis clair : cette structure algébrique permet des schémas plus efficaces au prix d'un affaiblissement de la sécurité. En effet, de nouvelles attaques pourraient exploiter cette structure. Dans le cas des réseaux idéaux, une belle série de travaux [CGS14; BS16; CDPR16; CDW17; BEFGK17; CDW21; PHS19; BR20; BLNR21] a découvert et étudié des attaques aussi bien classiques que quantiques contre (approximate)-Ideal-SVP. Cependant, aucune n'est praticable pour les dimensions et les facteurs d'approximation cryptographiquement pertinents, car ces attaques sont moins performantes que l'algorithme de réduction de réseau non structuré BKZ dans ce régime. La compréhension des attaques par les unités et S-unités était le sujet principal de mon mémoire de master [Bam22].

Réseaux modules Bien que des attaques réussies contre Ideal-SVP ne briseraient pas Ring-LWE mais seulement la réduction de sécurité, il est préférable que la sécurité soit fondée sur un problème plus difficile qu'Ideal-SVP : pour cela, les idéaux sont remplacés par des $\mathcal{O}_K$-modules de rang supérieur.

Tout comme les idéaux de $\mathcal{O}_K$ généralisent les entiers de $\mathbb{Z}$, les $\mathcal{O}_K$-modules de rang r généralisent les vecteurs de r entiers. Pour cette raison, les idéaux de $\mathcal{O}_K$ sont simplement des $\mathcal{O}_K$-modules de rang 1. L'objet géométrique obtenu en appliquant le plongement canonique à un module composante par composante est un réseau module. Notons que NTRU peut être vu comme un problème sur des réseaux modules de rang 2. Un module de rang r sur un corps de nombres de degré d se plonge dans un réseau de $\mathbb{R}^{rd}$. Toute l'information contenue dans un vecteur d'un réseau-module de rang r peut être représentée par r^2 éléments de $\mathcal{O}_K$. Dès que $r > 1$, c'est plus que ce qui est nécessaire pour représenter les réseaux idéaux, mais si r est constant, le gain reste énorme par rapport aux réseaux sans structure.

Du point de vue de la sécurité cependant, les attaques algébriques qui exploitent la structure des idéaux ne fonctionnent plus dès que $r \geq 2$. Actuellement, les meilleures attaques contre Module-SVP pour $r \geq 2$ sont les mêmes algorithmes que ceux utilisés pour attaquer le SVP général non structuré. Par conséquent, le problème Module-LWE est prouvablement [LS15] au moins aussi difficile que la résolution de Module-SVP, un problème que nous ne savons pas attaquer plus efficacement que SVP. C'est cet écart de difficulté perçu qui fait que toutes les nouvelles normes à base de réseaux (Kyber, Dilithium, Falcon, Haetae) reposent sur les réseaux modules.

Réseaux hypercubiques Pour clore cette section sur les réseaux particuliers, nous terminons avec le plus simple et le plus naturel de tous : $\mathbb{Z}^n$. Étant donné une (mauvaise) base $\mathbf{B}$ de $\mathbb{Z}^n$, trouver des vecteurs courts dans $\mathbb{Z}^n$ est un problème trivial : nous les connaissons déjà, ce sont les vecteurs de la base canonique

$$(0, \cdots, 0, \pm 1, 0, \cdots, 0).$$

Cependant, si l'on nous donne une mauvaise base $\mathbf{B}$ de Λ, où $\Lambda = \mathbf{O} \cdot \mathbb{Z}^n$ est le réseau obtenu après rotation de $\mathbb{Z}^n$ par une matrice orthonormale $\mathbf{O} \in \mathcal{O}_n(\mathbb{R})$, alors retrouver les vecteurs de la base canonique de Λ est (peut-être de manière surprenante) conjecturé comme étant un problème difficile : le Problème d'Isomorphisme de Réseaux sur $\mathbb{Z}$ ($\mathbb{Z}$LIP). On appelle de tels réseaux Λ des réseaux hypercubiques.

La sécurité du schéma de signature HAWK, un candidat du deuxième tour de l'appel additionnel du NIST pour les signatures post-quantiques, repose sur la difficulté d'une variante module de $\mathbb{Z}$LIP.

1.5 Aperçu Technique et Contributions Principales

Cette thèse contient des travaux liés au domaine de la cryptographie post-quantique. Elle se concentre sur deux sujets principaux et un troisième sujet secondaire. Les sujets principaux concernent la cryptographie à base de réseaux et peuvent être résumés par les questions suivantes :

Question 1.5.1. *Les réseaux utilisés pour concevoir des schémas cryptographiques efficaces possèdent une structure très particulière. Cette structure peut-elle être exploitée pour la cryptanalyse ?*

Question 1.5.2. *La conception de signatures numériques aux performances équilibrées est l'un des défis majeurs de l'effort de transition post-quantique. Est-il possible d'améliorer ces performances ?*

Le troisième sujet est différent des deux précédents et concerne l'étude de structures mathématiques qui apparaissent dans les graphes d'isogénies ordinaires. Ces graphes sont liés à la cryptographie à base d'isogénies, où la sécurité ne repose pas sur les réseaux, mais sur la difficulté de trouver des isogénies entre des courbes elliptiques.

Sujet Principal 1 : Sécurité de la cryptographie à base de réseaux particuliers

Pour comprendre la difficulté calculatoire d'un schéma cryptographique, il faut comprendre le coût précis des meilleures attaques. Ces attaques sont des algorithmes qui peuvent être étudiés de différentes manières :

- Complexité concrète vs asymptotique : en informatique, la complexité d'un algorithme est généralement mesurée de manière asymptotique, par rapport à un paramètre qui varie. Alors que l'analyse asymptotique des attaques donne un aperçu important de la puissance réelle d'une attaque lorsque le paramètre de sécurité tend vers l'infini, l'analyse concrète dénombre le nombre exact d'opérations requises par un attaquant pour briser un schéma avec des paramètres donnés.

- Analyse prouvable vs heuristique : l'analyse d'un algorithme peut conduire à des garanties mathématiquement prouvées sur le résultat ou le temps d'exécution. Dans ce cas, on dit que l'algorithme est prouvable. Cependant, lorsqu'une preuve entièrement rigoureuse est

hors de portée, il est toujours possible d'estimer la performance par des simulations et des conjectures. Si tel est le cas, on dit que l'analyse est heuristique[6].

Les complexités concrètes et asymptotiques sont toutes deux importantes. Nous nous concentrons principalement sur la complexité asymptotique car elle fournit une vue d'ensemble et permet de comprendre comment les différentes parties contribuent au coût final. Habituellement, les algorithmes heuristiques surpassent en performance les algorithmes prouvables, donc seuls les algorithmes heuristiques devraient être pris en compte pour la sécurité, mais les algorithmes prouvables fournissent des garanties. S'il existe un écart important entre les meilleurs algorithmes prouvables et heuristiques pour un problème, cela signifie généralement que le problème (ou l'algorithme) n'est pas encore bien compris.

Dans la section précédente, nous avons fait la connaissance de diverses familles de réseaux, tous dotés d'une structure particulière par rapport à ce que l'on attendrait d'un réseau réel aléatoire. Nous nous concentrons sur les algorithmes qui visent à retrouver un plus court vecteur non nul du réseau, en supposant l'accès à un oracle pour SVP en dimension β. Par analogie avec l'algorithme BKZ, nous appelons ce paramètre β la taille de bloc, et nous posons la question suivante.

Question 1.5.3. *Pour une suite de réseaux (Λ_n) donnée, paramétrée par un paramètre n (généralement $n = \mathrm{rank}(\Lambda_n)$), quelle est la plus petite taille de bloc $\beta(n)$ telle qu'il existe un algorithme ayant accès à $\mathrm{poly}(n)$ appels à un oracle SVP en dimension $\beta(n)$ qui résout SVP dans ce réseau ?*

Nous remarquons que les réseaux NTRU et les réseaux hypercubiques partagent les propriétés particulières suivantes :

1. Ils ont des vecteurs exceptionnellement courts.
2. Ils ont de nombreux plus courts vecteurs non nuls.
3. Ils sont essentiellement auto-duaux.

En gardant (1.) à l'esprit, nous donnons une réponse précise à la question 1.5.3 sous des hypothèses heuristiques relatives à l'attaque primale, et montrons que dans ce cadre, les termes du premier ordre dans le développement asymptotique de la taille de bloc ne sont pas affectés par (2.). Nous nous concentrons ensuite sur (3.), où nous comparons le réseau Λ_n à son dual

$$\Lambda_n^\vee = \{\mathbf{y} \in \mathbb{R}^n : \forall \mathbf{x} \in \Lambda_n, \langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{Z}\}.$$

NTRU Pour des réseaux aléatoires, on s'attend à ce que $\lambda_1(\Lambda_n)\lambda_1(\Lambda_n^\vee) \sim \frac{n}{2\pi e}$, pourtant pour les réseaux NTRU et hypercubiques, cette valeur est bien plus petite. Si ce produit est suffisamment petit, nous concevons une variante de l'algorithme de réduction de [GN08a] qui nous permet de prouver que la valeur prouvable de $\beta(n)$ est au plus $n/2$. Pour la plupart des instanciations classiques de NTRU, cela conduit à une attaque prouvable qui n'utilise qu'un nombre polynomial d'appels à un oracle en dimension moitié de la dimension du réseau complet, résolvant ainsi une conjecture de Gama, Howgrave-Graham et Nguyen de 2006 [GHN06].

Réseaux hypercubiques Les réseaux hypercubiques ont des propriétés supplémentaires, liées au fait que tous leurs vecteurs courts sont orthogonaux. Ducas a montré dans [Duc23] que $\beta(n)$ est prouvablement inférieure à $n/2 + O(1)$ pour les réseaux hypercubiques. Nous montrons que pour que son algorithme fonctionne, il est suffisant d'avoir accès à des oracles pour approx-SVP

[6] Nous parlerons également d'algorithmes heuristiques pour désigner les algorithmes qui ont une analyse heuristique.

avec un facteur $(\sqrt{2}-\varepsilon)$. Nous esquissons aussi une attaque qui montre de manière heuristique que la connaissance de vecteurs de norme $O(\log(n))$ du réseau hypercubique Λ_n suffit à retrouver les vecteurs de la base canonique en temps quasi-polynomial. Cette réduction est bien plus pratique que celle présentée dans [Jia+23].

Réseaux idéaux Enfin, nous cherchons à étudier le comportement moyen des réseaux idéaux, en particulier leur premier minimum, une quantité qui influence fortement le comportement des algorithmes pour SVP. Contrairement aux réseaux réels dont le premier minimum peut être arbitrairement court, cela ne peut pas être le cas pour les réseaux idéaux : en effet, si I est un idéal d'un corps de nombres K de degré n avec un discriminant Δ_K, alors pour tout $x \in I$ non nul, nous avons :

$$\|x\| \geq \sqrt{n} \cdot |N_{K/\mathbb{Q}}(x)|^{1/n} \geq \sqrt{n} \cdot N(I)^{1/n} = \sqrt{n} \cdot |\Delta_K|^{-\frac{1}{2n}} \cdot \mathrm{vol}(\sigma(I))^{1/n},$$

où $\sigma(I)$ est le réseau obtenu en plongeant I dans $\mathbb{R}^n$. Pour un x tel que $\|x\| = \lambda_1(\sigma(I))$, on obtient une borne inférieure absolue sur $\lambda_1(\sigma(I))$ qui ne dépend que du corps.

Question 1.5.4. *Comment se comporte la valeur moyenne du premier minimum d'un réseau idéal aléatoire ? Est-elle significativement différente de celle d'un réseau réel aléatoire ?*

L'inégalité ci-dessus suggère que les réseaux idéaux pourraient présenter des propriétés moyennes différentes de celles des réseaux aléatoires qui suivent la distribution imposée par la mesure de Haar. Des résultats plus avancés en théorie des nombres, qui utilisent des outils qui dépassent le cadre de cette thèse, semblent indiquer que lorsque le degré et/ou le discriminant du corps de nombres augmentent, des quantités telles que le nombre moyen de points d'un réseau idéal dans une boule centrée à l'origine convergent vers le résultat attendu dans le cas des réseaux réels aléatoires. Nous commençons par étudier le cas des corps quadratiques réels, où nous donnons un algorithme pour calculer les valeurs exactes des moments du premier minimum sur les réseaux idéaux aléatoires d'une classe donnée, en intégrant sur des orbites géodésiques sur la surface modulaire.

Nous nous tournons ensuite vers le problème du calcul de la valeur moyenne de sommes de la forme

$$\sum_{\mathbf{v}\in\Lambda} f(\mathbf{v})$$

sur l'espace des réseaux idéaux, pour des fonctions $f : \mathbb{R}^n \to \mathbb{R}$. C'est une tentative de trouver une version analogue de la formule de Siegel, qui affirme que pour une fonction f suffisamment régulière,

$$\int_\Lambda \sum_{\mathbf{v}\in\Lambda} f(\mathbf{v}) d\mu_n = f(0) + \int_{\mathbb{R}^n} f(x)dx.$$

En prenant pour f l'indicatrice d'une boule centrée, on obtient directement ce que l'on appelle communément l'*heuristique gaussienne* : l' espérance du nombre de points dans une boule de volume V est exactement $1 + V$. Il est intéressant de noter que de tels résultats sont tout aussi pertinents pour la recherche de bornes inférieures sur les empilements de sphères autour de réseaux, où des résultats moyens sur des réseaux avec des premiers minima plus contraints peuvent aider à prouver l'existence d'un réseau avec un premier minimum exceptionnellement grand [Ven13]. Nous esquissons notre tentative de preuve et obtenons une formule qui ne peut être calculée que dans de petites dimensions. Pendant que nous travaillions sur ce problème, Gargava et Viazovska [GV24] ont démontré une formule bien plus naturelle mais néanmoins complémentaire pour la même quantité. Ils concluent que pour les corps cyclotomiques de degré suffisamment grand, l'heuristique gaussienne devrait rester valable pour les réseaux idéaux aléatoires.

Sujet Principal 2 : Amélioration des signatures numériques

Le deuxième grand sujet de cette thèse est motivé par le fait que, si les normes pour le chiffrement devraient être déployés dès que possible, il reste encore un peu de temps pour améliorer la conception des schémas de signature, du moins jusqu'à ce que les ordinateurs quantiques deviennent assez puissants pour forger les signatures classiques.

Dans un certain sens, comme la plupart des signatures numériques à base de réseaux intègrent des réseaux particuliers comme choix de conception principal, on peut dire que le Sujet 2 inclut le Sujet 1. Cependant, ici, au lieu d'étudier la difficulté du problème sous-jacent, nous examinons les choix de conception au niveau du protocole.

DEFI Nous étudions d'abord une signature de type Hash-and-Sign à base de formes quadratiques isotropes et proposée par Feussner et Semaev : DEFI. Cette signature surpasse légèrement en performance presque tous les schémas de signature considérés par le NIST sur la plupart des métriques raisonnables, et c'est pourquoi il semblait important de l'analyser. Le schéma repose sur des variantes d'un problème intéressant que nous définissons formellement comme l'Équivalence de Formes Quadratiques (QFE), et sa version module. QFE peut être vu comme une généralisation directe de LIP à des formes quadratiques qui ne sont pas nécessairement définies positives. Dans le cas de DEFI, la clé privée est une matrice 4×4 $\mathbf{B}$ à coefficients dans $R = \mathbb{Z}[X]/(X^{64} + 1)$, et la clé de vérification est

$$\mathbf{C} = \mathbf{B}^T \mathbf{J} \mathbf{B},$$

où $\mathbf{J} = \mathrm{diag}(1, 1, -1, -1)$. Notons que pour que cela constitue une généralisation de module-LIP, l'opération de transposition devrait être remplacée par l'opération adjointe, plus naturelle. Il est intéressant de noter que DEFI est présenté comme un schéma multivarié. Bien que la nouvelle hypothèse utilisée par DEFI puisse sembler prometteuse, nous montrons que le schéma n'est pas sécurisé en raison de la manière dont la trappe est générée. Chaque signature fournit une équation linéaire sur R, dont les solutions sont liées à la clé secrète. Grâce à la réduction de réseaux bien choisis (qui se trouvent avoir une structure de module !), nous sommes en mesure de retrouver la clé secrète en pratique en quelques minutes, en utilisant moins de dix paires (message, signature). Malheureusement, cela signifie que DEFI n'est pas sécurisé.

Patronus Nous décidons alors de nous concentrer sur l'étude d'une classe de signatures dont la sécurité se réduit de façon prouvée à la difficulté de problèmes algorithmiques sur les réseaux : les signatures à base de réseaux dans le paradigme de Fiat-Shamir avec Rejet. Rappelons du début de l'introduction que de tels schémas nécessitent de choisir la distribution d'un paramètre de masquage $\mathbf{y} \in \mathbb{R}^n$. Ce paramètre est ajouté à l'information secrète $c\mathbf{s} \in \mathbb{R}^n$, et l'élément résultant $\mathbf{z} = \mathbf{y} + c\mathbf{s}$ est rejeté s'il tombe en dehors de la "zone de rejet", la zone où il divulguerait de l'information sur $c\mathbf{s}$. Pour des raisons techniques, la norme euclidienne d'un $\mathbf{z}$ non rejeté est directement liée à la taille de la signature. La norme sélectionnée par le NIST, Dilithium, utilise ce cadre et choisit d'échantillonner $\mathbf{y}$ uniformément dans un hypercube. La "zone de rejet" résultante s'avère être également un hypercube. Il a été montré par [DFPS22] que pour minimiser la norme attendue de $\mathbf{z}$, la meilleure distribution pour $\mathbf{y}$ est soit une distribution gaussienne, soit une distribution uniforme sur une boule euclidienne. Cela a motivé la conception du vainqueur de la compétition PQC sud-coréenne, Haetae. Cependant, si Haetae surpasse Dilithium en termes de taille de signature, il a l'inconvénient de recourir à l'échantillonage de distributions gaussiennes, une opération très coûteuse et potentiellement vulnérable.

Question 1.5.5. *Avec des garanties de sécurité équivalentes, est-il possible de réduire la taille des signatures dans* Dilithium *sans avoir besoin d'échantillonnage gaussien ou d'arithmétique en virgule fixe ?*

Dans notre travail, nous visons à répondre à la question 1.5.5 en généralisant le cadre de Fiat-Shamir avec Rejet pour les signatures sur réseaux à des polytopes plus généraux. Nous définissons un polytope particulier de rayon r et en dimension n en intersectant les boules ℓ_1 et ℓ_∞ comme suit :

$$\mathcal{H}_r^n := B_n^{(\infty)}(r) \cap B_n^{(1)}(r\sqrt{n}).$$

Le choix d'un hyperoctaèdre n'est pas anodin : c'est le dual de l'hypercube. En fait, à mesure que la dimension augmente, la masse dans $\mathcal{H}_r^n$ se concentre vers le centre de la boule ℓ_1, ce qui conduira à des $\mathbf{z}$ de norme euclidienne plus courte. C'est le dual du fait bien connu que les points aléatoires dans un hypercube de grande dimension se concentrent vers les coins, ce qui, d'une certaine manière, suggère que l'hypercube est un choix de distribution malheureux. Le calcul suivant aide à comprendre la différence : nous calculons les valeurs moyennes du carré de la norme euclidienne d'un point uniforme dans $B_n^{(\infty)}(r)$, $B_n^{(1)}(r\sqrt{n})$ et $B_n^{(2)}(r)$.

$$\int_{B_n^{(\infty)}(r)} \|\mathbf{z}\|^2 d\mathbf{z} = \frac{nr^2}{3}; \quad \int_{B_n^{(1)}(r\sqrt{n})} \|\mathbf{z}\|^2 d\mathbf{z} = \frac{2n^2r^2}{(n+1)(n+2)}; \quad \int_{B_n^{(2)}(r)} \|\mathbf{z}\|^2 d\mathbf{z} = \frac{nr^2}{n+2}.$$

En ignorant les subtilités techniques liées au fait que l'espérance d'une racine carrée n'est pas directement la racine carrée de l'espérance, l'intégrale de gauche nous dit qu'un point uniformément aléatoire dans $B_n^{(\infty)}(r)$ a une norme euclidienne d'environ $\sqrt{n/3}\cdot r$ en moyenne. L'intégrale de droite nous dit qu'un point uniformément aléatoire dans $B_n^{(1)}(r\sqrt{n})$, qui a la même boule euclidienne inscrite, a une norme euclidienne d'environ $\sqrt{2}\cdot r$ en moyenne, ce qui est à un facteur constant du même résultat pour la boule euclidienne, le cas optimal. Si l'échantillonnage de points entiers dans $\mathcal{H}_r^n$ est facile, cela signifie que remplacer l'hypercube par $\mathcal{H}_r^n$ dans Dilithium devrait conduire à une belle amélioration de la taille de la signature, sans nécessiter de changements majeurs dans la preuve de sécurité. Nous montrons qu'en effet, l'échantillonnage de points entiers uniformes dans $\mathcal{H}_r^n$ est possible et peut être rendu isochrone : en échange d'un échantillonneur légèrement plus compliqué, la signature devient un peu plus lente que pour Dilithium, mais ne nécessite pas beaucoup plus d'aléa.

Nous appelons notre nouveau schéma de signature Patronus[7]. Il atteint des tailles de signature plus courtes par rapport au schéma normalisé Dilithium. Assez logiquement, les signatures Patronus restent plus grandes que celles de Haetae, mais le processus d'échantillonnage est beaucoup plus rapide et, surtout, n'implique pas de gaussiennes.

Sujet 3 : Volcans d'isogénies

Notre dernier sujet est l'étude d'une structure qui ne se rapporte pas à la cryptographie à base de réseaux mais à celle à base d'isogénies. Nous considérons les courbes elliptiques dont l'anneau d'endomorphismes $\mathcal{O}$ est strictement plus grand que $\mathbb{Z}$; on dit qu'elles sont à multiplication complexe (CM). Alors qu'une courbe elliptique sur $\mathbb{C}$ peut être vue comme un réseau dans le plan complexe, les courbes elliptiques CM, vues sur $\mathbb{C}$, ont des symétries de rotation supplémentaires : ce sont en fait des réseaux idéaux pour des idéaux de $\mathcal{O}$, leur anneau d'endomorphismes.

Pour $p > 5$ et $\ell \neq p$ deux nombres premiers, nous définissons le graphe $\mathcal{G}_\ell(\mathbb{F}_p)$ dont les sommets $\mathcal{V}$ sont les j-invariants correspondant aux classes d'isomorphisme sur $\overline{\mathbb{F}}_p$ de courbes elliptiques ordinaires. Pour $j, j' \in \mathcal{V}$, $\mathcal{G}_\ell(\mathbb{F}_p)$ a autant d'arêtes entre j et j' que la multiplicité de (j, j') comme racine du ℓ-ième polynôme modulaire $\Phi_\ell(X, Y)$. Les arêtes peuvent aussi être vues comme des applications de degré ℓ entre courbes elliptiques : des ℓ-isogénies. Les polynômes modulaires sont des objets fondamentaux, utilisés notamment dans l'algorithme de Schoof-Elkies-Atkin pour le comptage de points sur les courbes elliptiques [FM02]. Le graphe d'isogénies ordinaires a été utilisé par Couveignes, Rostovtsev et Stolbunov [Cou06; RS06] pour

[7] Mot latin pour "protecteur".

proposer un protocole d'échange de clé, précurseur de schémas plus modernes à base d'isogénies comme CSIDH [Cas18].

La structure du graphe d'isogénies ordinaires a été comprise par Kohel dans sa thèse de doctorat (encore une fois, en 1996) [Koh96]. De tels graphes présentent des structures en forme de "volcans" :

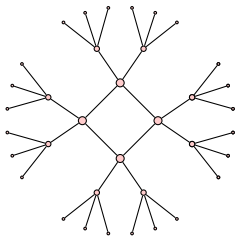


Figure 1.8: Un volcan avec un cratère de taille 4.

À chaque sommet de ce graphe, on peut attacher l'anneau d'endomorphismes des courbes correspondant à ce sommet. Les anneaux d'endomorphismes pour les courbes elliptiques ordinaires sont des ordres dans des corps quadratiques imaginaires, ce qui en fait des réseaux idéaux de rang 2. Descendre (resp. monter) dans le graphe via des ℓ-isogénies agit sur ces réseaux en les convertissant en sous-réseaux d'indice ℓ (resp. sur-réseaux d'indice ℓ^{-1}). Des aspects calculatoires intéressants des volcans d'isogénies ont été explorés par [IJ10; Sut13].

Dans notre travail, nous posons la question suivante, que nous voyons comme un problème inverse :

Question 1.5.6. *Quels graphes en forme de "volcan" peuvent effectivement apparaître en tant que composantes connexes d'un graphe d'isogénies ordinaire $\mathcal{G}_\ell(\mathbb{F}_p)$?*

En reliant la question 1.5.6 à des équations diophantiennes classiques et à des résultats de la théorie des corps de classes, nous montrons que tous les volcans raisonnables peuvent être construits, et nous donnons un algorithme naïf pour trouver les valeurs minimales de p et ℓ qui réalisent un volcan cible. Malheureusement, cet algorithme n'est pas particulièrement efficace car il nécessite des calculs de groupes de classes.

Structure de la thèse Cette thèse est organisée en quatre parties, chacune véritablement composée de deux chapitres.

- La Partie I commence par cette introduction au Chapitre 1, ainsi que sa traduction anglaise au Chapitre 2, puis se poursuit avec le Chapitre 3, où nous introduisons les notions préliminaires requises pour le corps de la thèse ;
- La Partie II se concentre sur les attaques contre les réseaux NTRU et hypercubiques, en mettant l'accent sur les attaques heuristiques dans le Chapitre 4 et les attaques prouvables dans le Chapitre 5 ;
- La Partie III est consacrée aux signatures numériques ; nous y présentons notre cryptanalyse de DEFI dans le Chapitre 6, et notre construction Fiat-Shamir avec Rejet à base de polytopes dans le Chapitre 7 ;
- La Partie IV aborde des questions mathématiquement plus complexes : le Chapitre 8 étudie les réseaux aléatoires et plus particulièrement les réseaux idéaux aléatoires, et enfin le Chapitre 9 présente notre solution au problème du volcan inverse dans les graphes d'isogénies ordinaires sur $\mathbb{F}_p$.

Les chapitres des Parties II, III et IV sont en grande partie indépendants et peuvent être lus dans n'importe quel ordre. Bonne lecture !

Chapter

2 Introduction

Chapter content

In today's world, private data is more important and at risk than ever. We trust technology with our personal communications, financial transactions, medical and identity records. Protecting all this sensitive information is essential. This can be achieved through cryptography[1], the science of secrets.

Historical cryptography Throughout millennia, the use of ciphers by individuals and societies has influenced the course of history multiple times. This is best verified when important encrypted messages are intercepted and decrypted by eavesdroppers, which happened in the following three movie-famous historic events:

- **The Babington plot (1586):** Mary, Queen of Scots was executed by Queen Elizabeth I of England when Elizabeth's cryptanalysts decrypted letters that provided evidence of Mary's involvement in an assassination plot to capture the throne.
- **The Zimmerman telegram (1917):** During World War I, the Germans sent an encrypted telegram to their embassy in Mexico, asking for an alliance against the United States. The outrage generated when British intelligence made the decryption of the telegram public contributed largely to the US's declaration of war on Germany later in 1917.
- **The Enigma machine and World War II (1939-1945):** Joint efforts by Polish, French and British cryptanalysts led to a practical break of the Enigma cipher used for secure communication by the German military, giving the Allies a decisive edge in the war. At the same time, cryptanalysis of the Lorenz ciphers led to construction of the Colossus computers, regarded as the world's first programmable, electronic, digital computers.

Historical ciphers dating back to ancient Rome relied on simple substitution ciphers - methods that systematically replace letters or groups of letters with others - such as Caesar's cipher where each letter in the alphabet is replaced by the following one in order. Substitution ciphers were too easily broken, and they evolved into more sophisticated polyalphabetic ciphers employing substitution ciphers simultaneously over multiple alphabets. Notable examples include the Vigenère cipher, that was created in the 16th century and broken only in the 19th century,

[1] Literally derived from Ancient Greek "hidden writing". The term cryptology is more general and refers simultaneously to cryptography and cryptanalysis, the science of breaking cryptography.

and the famous Enigma machine. Before the 20th century, breaking ciphers required intuition, patience, linguistic prowess, and craftsmanship acquired through experience. This has changed radically from the early 1980s onwards. All cryptography now follows Kerkoffs's principle: the security of a system should not rely on the fact that its specifications are secret.

Provable cryptography Modern cryptography has transitioned from an art into a rigorous science, where security can be proved mathematically. In fact practical cryptographic schemes that can achieve unconditional mathematical proofs of security are seemingly too much to hope for. Instead for some schemes, we can now formally prove that breaking security is at least as hard as solving a well-studied computationally intractable mathematical problem, such as factoring. In the Information Age, cryptography can now be trusted as the fundamental building block of all of cybersecurity, protecting both civil and military infrastructure.

There has recently been a lot of political conflict related to cryptography. Until 1999 in France, strong cryptography was illegal, even for domestic use. Law enforcement agencies and authoritarian regimes are interested in controlling cryptography, in order to facilitate eavesdropping on organised crime syndicates or political opponents. This clashes directly with the public's growing demand for better privacy.

Symmetric cryptography Sending a message over the internet is like shouting to someone across a large room full of eavesdroppers. You can get closer and whisper, but you never really know if someone was listening. *Secret-key* cryptography allows two individuals Alice and Bob to communicate securely in such a room in the following way. They first sneak off to a private room and agree on a shared secret codeword, called the *secret key*. Once both of them know it, Alice scrambles her message using a public procedure that depends on the secret key, and shouts the scrambled ciphertext across the room. Bob, using his knowledge of the secret key is the only one able to reverse the procedure and recover Alice's message. Anyone else in the room will only hear gibberish from Alice.

Asymmetric cryptography The tricky part in secret-key cryptography is that the two communicating parties must meet in person to exchange the secret key. How can they share this information securely if they cannot meet privately, for example when communicating over the internet, or when attempting to communicate with too many people at the same time? The solution was given in Diffie and Hellman's 1976 paper[2] "New Directions in Cryptography" [DH76]. Bob generates two keys: the *public key* and the *private key*. If this is done in a way that ensures that any message encrypted using the public key can only be decrypted using the private key, then Alice (or anyone else) can use Bob's public key to securely send a secret message. Note that Alice and Bob did not have to meet beforehand. Asymmetric cryptography is usually more expensive than its counterpart, but it provides a solution to the key-exchange problem. What happens in practice is that an asymmetric scheme is used first to establish a shared secret key, and then later communication happens through symmetric encryption with the shared secret.

RSA In the context of public-key cryptography, in order to make sure that key generation ensures that any message encrypted using the public key can only be decrypted using the private key, we rely on computational assumptions. We illustrate this by describing the blueprint for a very simple yet incredibly versatile scheme: the RSA cryptosystem.

- **Key generation:** Two primes p, q are securely generated. The products $N = pq$ and $\varphi(N) = (p-1)(q-1)$ are computed. An integer e is chosen coprime to $\varphi(N)$, and the inverse d of e modulo $\varphi(N)$ is computed. The public key is (N, e), and the private key consists of (N, d).

[2] Although declassified material from GCHQ show that it had been discovered several years earlier.

- **Encryption:** A message is encoded as a positive integer $m < N$. The ciphertext c is obtained by taking $m^e \pmod N$. This only requires knowledge of N and e, which is provided by the public key.
- **Decryption:** The ciphertext c is decrypted by computing $c^d \pmod N$. This only requires knowledge of N and d, which is provided by the private key.

Correctness of RSA follows from Fermat's little theorem as $ed \equiv 1 \pmod{\varphi(N)}$ implies that $m^{ed} \equiv m \pmod N$. Breaking RSA encryption requires an eavesdropper to invert the function $f : m \mapsto m^e \pmod N$, a problem that is computationally infeasible today without knowledge of the factors of N. It is not difficult to see that knowing p or q is enough to compute d and invert f. In some sense, the hardness of encryption relates directly to the hardness of one of the most fundamental questions in mathematics: factoring. Until recently, this was sufficient, and RSA was used along side Elliptic Curve Cryptography (ECC) to secure essentially the whole of the internet.

Quantum computing In 1994, Shor published an algorithm that uses a quantum computer to efficiently solve the problems of factoring integers and computing discrete logarithms [Sho94], compromising RSA, as well as ECC. Quantum computers are built from special hardware that harnesses physical properties of matter that cannot be explained by the rules of classical physics, but rather quantum physics. Small prototypes already exist, however they are notoriously difficult to scale. Assuming a sufficiently large and accurate quantum computer, the best attack on RSA would run in polynomial time (in the number of bits of N) instead of sub-exponential, a devastating improvement. Most of contemporary cryptography is at stake, as a sufficiently large quantum computer would allow an attacker to break any public-key cryptosystem. It is worth noting that symmetric cryptography is less affected by quantum attacks, because attacks only seem to benefit from a quadratic speed-up.

Post-quantum cryptography As has always been the case in history, when a scheme is broken, cryptographers invent a new one, and the cycle repeats until everyone is confident enough on the underlying assumption. As factoring is not a viable option any more, cryptographers have proposed new assumptions and built schemes from those new assumptions. The science of cryptographic schemes that are conjectured to resist against quantum attackers is called *Post-Quantum Cryptography* (PQC). The main families of mathematical objects studied by post-quantum cryptographers are lattices, error-correcting codes, multivariate systems of quadratic equations, isogenies of elliptic curves. Note that all PQC algorithms are classical, *i.e.* we do not need to assume that the defender has access to quantum hardware, only the attacker. PQC should not be confused with *quantum cryptography*, which obtains provable security guarantees through unclonability of quantum states, and will not be the subject of this thesis. Deployment of quantum cryptography would be orders of magnitude less practical and reliable than that of PQC.

Standards and PQC While quantum computers breaking cryptography might still have sounded like abstract conspiracy theory several years ago, it has now evolved into a serious threat that any reasonably risk averse organisation will be willing to mitigate. The exact date for Q-day - the date where RSA or ECC based cryptography becomes vulnerable to quantum computers - is yet unknown, yet many well-grounded sources such as [fSic25] expect Q-day to happen a couple decades from now. Many actors have started recording today's encrypted traffic, in order to store it for future decryption, when the tools become available to them. This strategy is known as "Harvest Now, Decrypt Later" (HNDL), and forces organisations with medium to long term secrets to start transitioning now.

Global transition to PQC is a monumental task, that is made smoother by the rigorous standardisation efforts that aim to ensure interoperable widespread deployment. At the forefront of the standardisation efforts is NIST: the United States' National Institute of Standards and Technology. A first call for PQC algorithms was issued in 2016, and after multiple rounds of improvements and cryptanalysis, four algorithms were selected in July 2022, the lattice-based schemes Kyber (for key-exchange), Dilithium and Falcon (for digital signatures), as well as the hash-based signature scheme SPHINCS+. In March 2025, NIST selected the code-based scheme HQC as an additional key-exchange protocol. Similar calls have been issued by other countries, most notably South Korea and China. The NIST standardisation process happened in a very international context, with researchers and companies from all over the world teaming up to break or prove security of candidate schemes. This lead to late surprises with multivariate-based scheme Rainbow and isogeny-based scheme SIKE falling to attacks in the final round of the process. The NIST and NSA jointly recommend completing the PQC transition before 2030, with all of RSA and ECC based cryptography being deprecated after 2030, and disallowed after 2035.

Most countries seem to trust the algorithms selected by NIST, although many countries including European entities such as ANSSI (France) and BSI (Germany) also recommend more conservative schemes, and advocate for pre- and post-quantum hybridisation, in order to maintain security in a world where the new post-quantum standards would be broken before the creation of cryptographically relevant quantum computers.

One thing that everyone seems to agree on is the choice of lattice-based cryptosystems as the most promising alternative to pre-quantum cryptography. Lattices for cryptography will be the main topic of this thesis.

2.1 Picturing Lattice Problems

We start by illustrating lattices and lattice problems in a gamified way. This section is deliberately intended towards non-specialists, an expert reader might prefer skipping ahead.

Lattices Imagine an infinite chessboard. You are a wizard standing in your tower, which is a single enchanted square that we call the origin $(0, 0)$. You cannot move away from your tower, but you would still like to explore the chessboard and enchant as many squares as possible.

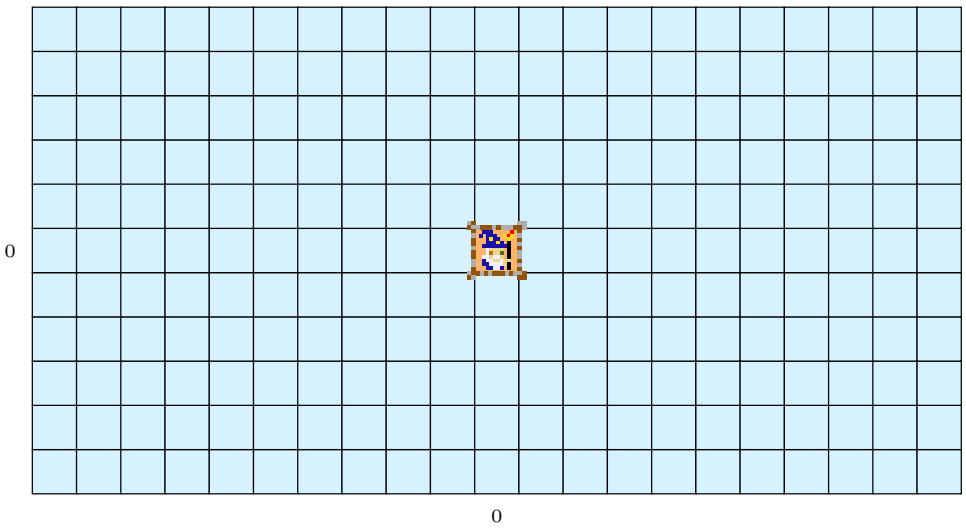


Figure 2.1: Enchanted squares are highlighted orange.

In order to help you reach your goals, you control a squad of magical gnomes that you can teleport to any square that has already been previously enchanted by you or one of your gnomes. Each gnome in your squad walks in a fixed direction, forwards or backwards according to its own personal move, allowing it to enchant new territory. For example the red gnome has a "knight-like" move: it always moves 1 step right and 2 steps upwards (or the same in the opposite direction: 1 step left and 2 steps downwards). With a red gnome in your squad, you can eventually explore infinitely many squares, but they will all lie on the same line.

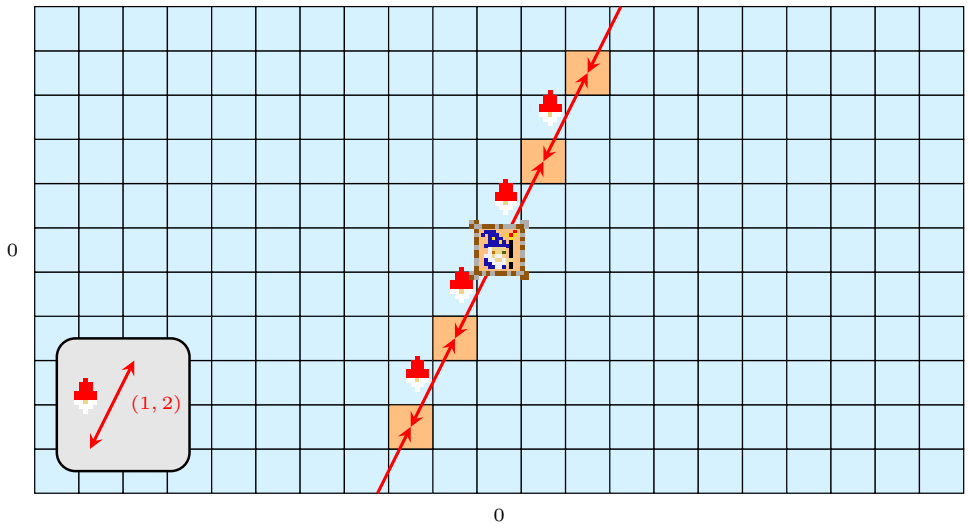


Figure 2.2: One-Gnome Squad.

You need an different gnome to explore more territory. For example the blue gnome that always moves 2 steps right and 1 step upwards. From now on we denote the moves using vectors, so the blue gnome has movement $(2, 1)$.

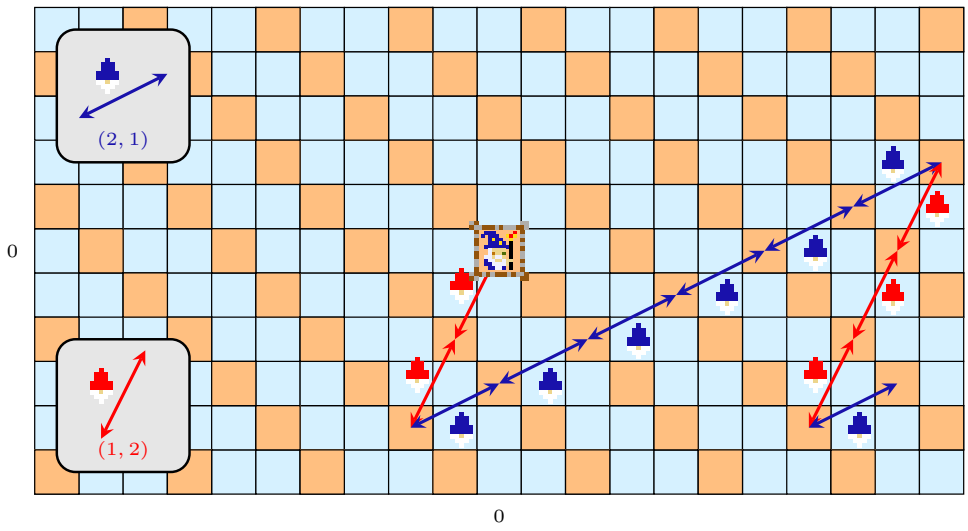


Figure 2.3: Using the red and blue gnomes, all orange tiles can be reached.

Your territory is defined as the set of squares that are reachable by your squad of gnomes (represented by orange tiles in the illustrations of this section). This is exactly what we will refer to as a lattice: a discrete and periodic set of points in space. Note that a lattice must always contain the wizard's tile, *i.e.* the origin.

Lattice bases and basis reduction The minimal number of gnomes that are needed to enchant a given lattice is called its rank, and a squad of minimal size that still enchants exactly the same lattice is called a basis of the lattice. This last definition suggests that some gnomes might be redundant. Indeed, it is not hard to see that the gnome with movement $(2, 4)$ is useless with respect to the red gnome, as the red gnome can recreate its movement by simply moving twice. In other cases it might not immediately be clear if a gnome is useful or not. For example consider the purple gnome with movement $(11, 7)$. Is it useless? This is harder to tell, as it is not directly weaker than either of the red or blue gnomes. However careful analysis shows that it does not help enchant territory outside of the lattice with basis the red and blue gnomes. We now consider the lattice obtained by using a squad of purple and green gnomes, where the green gnome has movement $(13, 8)$. It is a good exercise to show that the territory generated by purple and green is in fact the same as the one previously generated by red and blue. This is a fundamental observation: a lattice can have multiple bases! In what follows, we we assume that the lattice has rank 2.

Volume of a lattice An interesting quantity related to the lattice is its volume. It is a fundamental invariant of the lattice that measures the proportion of squares on the chessboard that have been enchanted. Visually, we see that the enchanted territory in Figure 2.3 covers a third of all total squares. Therefore, we say that the volume[3] of the corresponding lattice is 3. A great property of the volume is that it can be computed directly from any basis. If the lattice has rank 2, then its bases consist in two vectors (gnomes). The two vectors define a parallelogram, whose area happens to always equal exactly the volume. It is therefore not necessary to fully map the enchanted territory in order to deduce the proportion of enchanted squares.

Lattice problems Even though lattices of the infinite chessboard are infinite objects, a basis of only 2 gnomes is enough to store all the information they contain. This means that 4 integers are enough to represent the lattice: the movement vectors for each gnome. Imagine now that the square of the lattice that is closest to the origin (without itself being the origin) contains a treasure chest. With your squad of two gnomes, how can you, the wizard, use the gnomes at your disposal to reach the treasure chest as fast as possible? This question is in fact usually called the Shortest Vector Problem (SVP): the main hardness assumption underlying post-quantum cryptography.

Why is this problem interesting? Two different wizards with access to different bases of the same lattice will have a very different experience in recovering the treasure chest. Let's use the same lattice as in the previous section. The treasure chest will be at coordinates $(1, -1)$, but this is unknown to the wizards. Note that $(-1, 1)$ is also equally as close to the origin, but we will ignore it, as SVP only asks to recover one of the shortest non-zero vectors of the lattice.

- A wizard with the red and blue gnomes in its squad will have no problem finding the treasure. Indeed its gnomes allow it to almost immediately explore the square $(1, -1)$ by using the red gnome first and then the blue gnome, as in Figure 2.4. We say that the red-blue squad is a good basis of the lattice.

- A wizard with only the purple and green gnomes will find it much harder to find a path to the treasure. Indeed both gnomes have long and clumsy nearly-parallel moves, and the shortest path to the treasure requires a tricky sequence of long forward and backward moves. In fact one can check via Figure 2.5 that a path to the treasure using the purple and green gnomes requires at least 13 moves! In particular with this squad compared to

[3]The correct mathematical terminology here would be *co-volume*, which should feel right as we have inverted 1/3 under the hood.

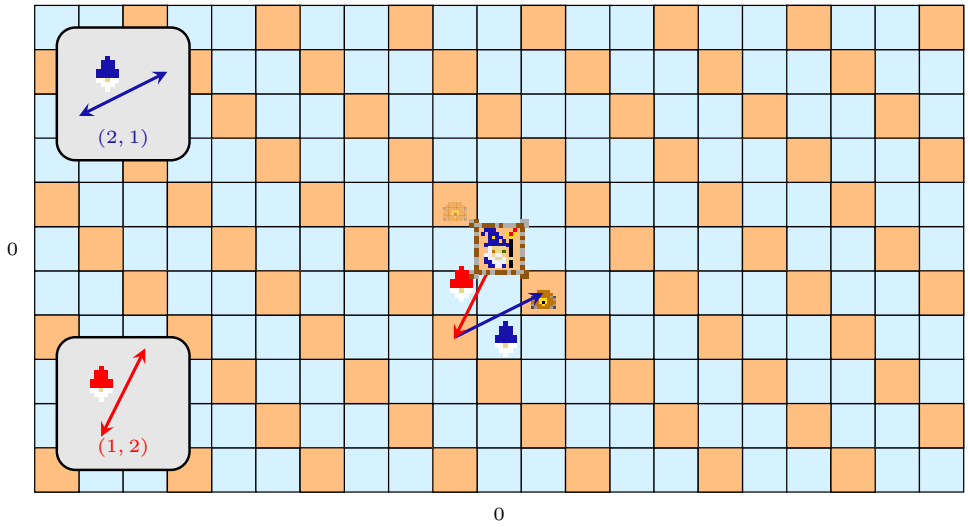


Figure 2.4: An easy path towards the treasure.

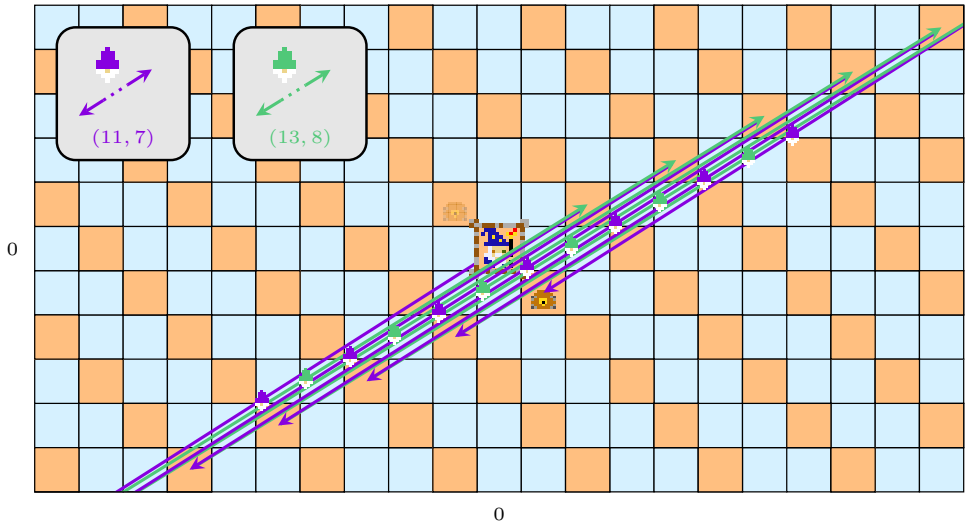


Figure 2.5: A more contrived path towards the treasure.

the previous one, the strategy consisting in applying a random sequence of moves is much less likely to be lucky and stumble onto the treasure chest, given the sheer amount of possible 13-move paths. We say that the purple-green basis is a bad basis of the lattice.

Lattice reduction A reasonable metric to judge how good a basis is is the relative size and orthogonality of the basis vectors. Indeed bases that generate very long and flat parallelograms are of poor quality. There are many ways to quantify this idea of *quality*. In our two dimensional example, we can choose to look at the normalised length of the long diagonal of our basis parallelogram: a shorter diagonal represents better basis quality. Lattice reduction denotes the following process: instead of directly searching for the treasure using his squad of gnomes, the wizard can measure the quality of his squad, and if the quality is not satisfying, then he can apply an algorithm that converts the bad squad into a squad with better quality. Then he can search the treasure with his improved squad of gnomes: the search phase is made more straightforward. Understanding lattice reduction - the process that converts bad gnome squads

into better gnome squads - is critical to lattice-based cryptography, as any adversary that is able to find the treasure chest hidden with the shortest vector of the lattice will be able to effectively break the scheme.

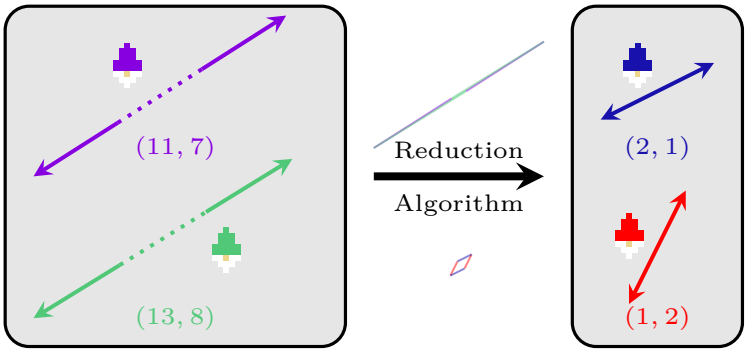


Figure 2.6: Reduction improves the quality of a basis, un-flattening the basis parallelogram.

2.2 Lattice Algorithms

In more mathematical terms, a lattice L is a discrete additive subgroup of $\mathbb{R}^m$. The dimension of the real vector space spanned by the vectors of L is its rank n. A basis is a family of linearly independent row[4] vectors $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ whose $\mathbb{Z}$-linear combinations generate L. The (co)-volume of a lattice is defined by $\mathrm{vol}(L) = \sqrt{\det(\mathbf{B}\mathbf{B}^T)}$. It is well-defined as any two bases $\mathbf{B}_1$ and $\mathbf{B}_2$ of L are related by a change of basis matrix $\mathbf{U} \in \mathcal{M}_n(\mathbb{Z})$, with $\det(\mathbf{U}) = \pm 1$, by $\mathbf{U}\mathbf{B}_1 = \mathbf{B}_2$, and it follows that

$$\det(\mathbf{B}_2\mathbf{B}_2^T) = \det(\mathbf{U}\mathbf{B}_1\mathbf{B}_1^T\mathbf{U}^T) = \det(\mathbf{U})\det(\mathbf{B}_1\mathbf{B}_1^T)\det(\mathbf{U}^T) = \det(\mathbf{B}_1\mathbf{B}_1^T),$$

confirming that the volume is independent of the basis. The volume is also the (positive) volume of the parallelepiped defined by the basis vectors $(\mathbf{b}_1, \ldots, \mathbf{b}_n)$. Dimensional analysis says that we should compare vector lengths with the quantity $\mathrm{vol}(L)^{1/n}$. Minkowski's theorem states that the length $\lambda_1(L)$ of the shortest non-zero vector of a lattice L must satisfy the inequality

$$\lambda_1(L) \leq \sqrt{n} \cdot \mathrm{vol}(L)^{1/n}.$$

In fact for random lattices we expect this inequality to be tight up to a constant, with very high probability:

$$\lambda_1(L) \simeq \sqrt{\frac{n}{2\pi e}} \cdot \mathrm{vol}(L)^{1/n}.$$

In the Shortest Vector Problem (SVP), we are given a generating set of the lattice L, and we are required to output a lattice vector $\mathbf{v} \in L$ such that $\|\mathbf{v}\| = \lambda_1(L)$. This problem can be relaxed: the approximate Hermite Shortest Vector Problem (γ-HSVP) provides an additional factor $\gamma > 0$ called the approximation factor, and asks to find a lattice vector $\mathbf{v} \in L$ such that $\|\mathbf{v}\| \leq \gamma \cdot \mathrm{vol}(L)^{1/n}$.

SVP is easy in dimension one: it is trivial if we are given a basis. If instead we are given two generating vectors $a, b \in \mathbb{Z}$, then the lattice is given by $a\mathbb{Z} + b\mathbb{Z} = \gcd(a, b)\mathbb{Z}$, and therefore the first minimum can be obtained simply by applying Euclid's algorithm.

[4] Both row and column vectors are used in the literature.

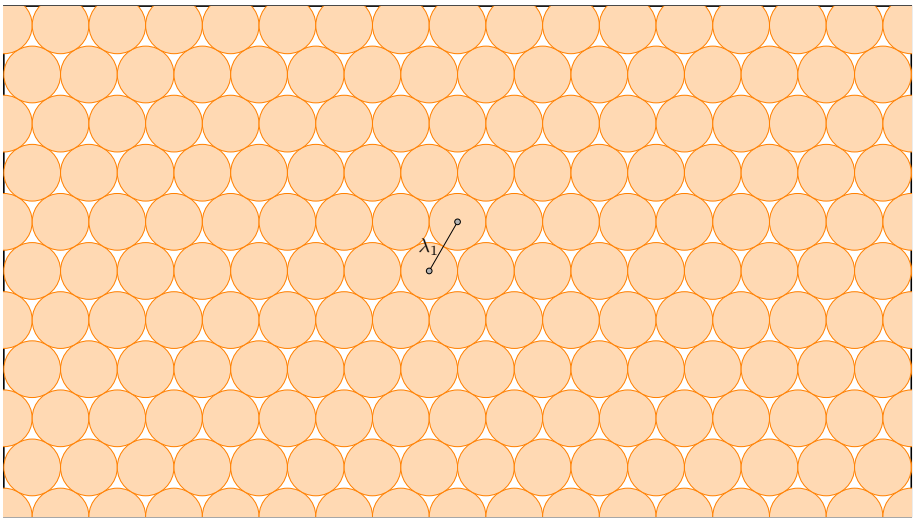


Figure 2.7: Two dimensional packing via the hexagonal lattice.

In our gamified example from the previous section, the lattice had rank 2, a situation in which SVP can be solved efficiently using the Lagrange-Gauss algorithm, an algorithm that can be viewed as a generalisation of Euclid's algorithm.

For cryptographic applications, lattices in play have many more dimensions, say 500 to 1000, making the problem plausibly hard, even for quantum computers. In this setting, the best algorithms for exact SVP run in exponential time $2^{n+o(n)}$. Yet, the celebrated LLL (Lenstra-Lenstra-Lovász, [LLL82]) algorithm runs in time polynomial in the rank n and bitsize of the input basis vectors $\mathbf{B}$, and returns a new reduced basis whose first vector $\mathbf{b}_1$ satisfies

$$\|\mathbf{b}_1\| \leq 2^{\frac{n-1}{4}} \cdot \mathrm{vol}(L)^{1/n}. \tag{2.1}$$

In other words, LLL solves $2^{(n-1)/4}$-HSVP in polynomial time. Conceptually, one can see LLL as an algorithm that solves SVP on a set of iteratively chosen smaller lattices of rank 2, progressively extracting a shorter and shorter basis. This idea can be generalised: assuming a black-box oracle that can solve SVP in lattices of rank β, the BKZ-β (Blockwise Korkine Zolotarev) algorithm iteratively calls the oracle on well-chosen β-dimensional lattices, progressively inserting shorter and shorter vectors in the basis, until this n-dimensional basis is reduced enough. The BKZ hierarchy of algorithms provides a trade-off between runtime and approximation factor, reaching an approximation factor of $2^{O(n/\beta)}$ in time $2^{O(\beta)}$.

Interestingly, the key player in lattice-based cryptography that is the first minimum λ_1 is directly related to the famous lattice sphere packing problem. It asks the following question: what fraction of space can be covered by equal balls that are not allowed to overlap except along their boundaries, given that all balls are centred around lattice points? For a given lattice L, the maximum tolerable radius of the balls is precisely $\lambda_1(L)/2$, as can be seen in Figure 2.7. Such problems are also of interest to physicists and crystallographers.

Lattice reduction exemplified The LLL algorithm led to revolutionary applications in many fields of mathematics, with its initial intent being efficient factoring for polynomials with rational coefficients, a problem long believed hard by analogy to the more classical problem of factoring integers. In the rest of this section we illustrate another classical application of lattice reduction, with the hope that it can convey to the reader how powerful this algorithm really is.

Question 2.2.1. *Can we find small integer linear combinations of the constants π, e and 1 that approach 0? More precisely, let $\varepsilon > 0$. Can we find integers $A, B, C \in \mathbb{Z}$ such that $|A\pi + Be + C| < \varepsilon$ and the norm $\sqrt{A^2 + B^2 + C^2}$ is small?*

The answer is yes, by building a lattice L in which the solution to our problem in given by computing the shortest vector of L. For this we choose a value of $\alpha \in \mathbb{Z}_{>0}$ to be fixed later, and define

$$\mathbf{B}_\alpha = \begin{pmatrix} 1 & 0 & 0 & \lfloor \alpha\pi \rfloor \\ 0 & 1 & 0 & \lfloor \alpha e \rfloor \\ 0 & 0 & 1 & \lfloor \alpha \rfloor \end{pmatrix}.$$

The vectors of $\mathbf{B}_\alpha$ generates a rank 3 lattice L_α of $\mathbb{Z}^4$, of volume

$$\mathrm{vol}(L_\alpha) = \sqrt{1 + \lfloor \alpha\pi \rfloor^2 + \lfloor \alpha e \rfloor^2 + \lfloor \alpha \rfloor^2} \leq \sqrt{1 + \alpha^2(\pi^2 + e^2 + 1)} \leq 2\alpha\sqrt{\pi^2 + e^2 + 1}.$$

Applying LLL on $\mathbf{B}_\alpha$ returns a new basis whose first vector $\mathbf{v} = (A, B, C, D)$ is short and must satisfy $\|\mathbf{v}\| \leq \sqrt{2}\mathrm{vol}(L_\alpha)^{1/3}$ because of the bound from Equation 2.1. We notice that because $\mathbf{v}$ is a lattice vector, D must be of the form $D = A\lfloor \alpha\pi \rfloor + B\lfloor \alpha e \rfloor + C\lfloor \alpha \rfloor$. We now get

$$\begin{aligned} |A\pi + Be + C| &= \frac{1}{\alpha}|A\{\alpha\pi\} + B\{\alpha e\} + C\{\alpha\} + D| \\ &\leq \frac{1}{\alpha}\|\mathbf{v}\|\sqrt{\{\alpha\pi\}^2 + \{\alpha e\}^2 + \{\alpha\}^2 + 1^2} \\ &\leq \frac{\sqrt{3}}{\alpha}\|\mathbf{v}\| \\ &\leq \frac{\sqrt{6}}{\alpha}(2\alpha\sqrt{\pi^2 + e^2 + 1})^{1/3} \\ &\leq \kappa \cdot \alpha^{-2/3}, \end{aligned}$$

for some constant $\kappa > 0$, where $\{x\} = x - \lfloor x \rfloor$ is the fractional part of x. We have used, in order, the Cauchy-Schwarz inequality, the fact that $\{x\} \leq 1$, the bound on $\|\mathbf{v}\|$ from LLL, and the bound on $\mathrm{vol}(L_\alpha)$ derived previously. Therefore it is sufficient to take $\alpha \geq (\kappa/\varepsilon)^{3/2}$ to ensure that $|A\pi + Be + 1| < \varepsilon$. We can even derive a bound on $\sqrt{A^2 + B^2 + C^2}$ by noticing that this quantity is bounded above by $\|\mathbf{v}\|$. We get a bound that is proportional to $\varepsilon^{-1/2}$.

For example, by setting $\varepsilon = \frac{1}{2025}$, we find that

$$19\pi + 6e = 76.000049.$$

This might look like an insignificant coincidence to a mathematical mind, but the ability to efficiently find such relations or almost-relations is one of the most powerful tools in modern cryptanalysis. Parameters of lattice schemes are derived from the estimated complexity of the best attacks, *i.e.* understanding lattice reduction is critical to security. The first half of this manuscript (Chapter 4,Chapter 5,Chapter 6) studies lattice reduction applied to lattices that are used in cryptographic schemes.

2.3 Digital Signatures from Lattices

The hardness of lattice problems conditions the security of the now NIST-standardised key exchange mechanism ML-KEM (Kyber), and of the signature schemes ML-DSA (Dilithium) and FN-DSA (Falcon). Both categories of algorithms have very specific purposes:

- The primary function of key-exchange mechanisms (KEMs) is to ensure *confidentiality* of encrypted data.
- The primary functions of digital signatures are *authentication*, *integrity* and *non-repudiation*: this ensures that the transmitted message was sent by a specific entity, that it has not been tampered with, and that the sender cannot deny having sent the message.

While protecting digital signatures against quantum adversaries is critical, it is not as urgent as protecting KEMs. Indeed in the context of encryption, HNDL threats apply, making both present and future data at risk. For signatures however, the vulnerability is not retroactive: a signature that is issued today with a classical computer will remain valid. Only when quantum computers become available will adversaries start forging signature. This leaves us with a bit more time to think about the design of signatures: one of the main topics of this manuscript (Chapter 6,Chapter 7). For this reason, NIST issued an additional call for post-quantum signature schemes in 2023, a process that has now advanced to its second round. Note however that a transition timeline should also incorporate both the facts that organisation-wide changes will take time, and that many embedded systems have hardcoded keys that cannot be changed easily, possibly compromising authentication on Q-day.

Two chapters of this manuscript study digital signatures. We give a quick presentation of the two main techniques that can be used to derive signature protocols from the hardness of lattice problems.

Hash and Sign Hash-and-Sign is a technique for signing messages in which the signer has access to a trapdoor one-way function, a function f that is easy to compute, and computationally hard to invert unless one has access to some secret information: the trapdoor. The message μ is hashed using a cryptographic hash function H, and the signer then uses his trapdoor one-way function to find a σ such that

$$f(\sigma) = H(\mu).$$

σ is the signature, and anyone with access to (μ, σ) can verify the signature by checking the equation. We give two examples of signatures that use Hash-and-Sign.

- **RSA signatures:** the RSA trapdoor one-way function is the modular exponentiation $x \mapsto x^e \pmod N$, this is hard to invert unless you know d, the modular inverse of e modulo $\varphi(N)$.
- **GPV-style signatures:** the private key is a good basis of a lattice, which acts as a trapdoor to find a small solution $\mathbf{s} \in \mathbb{Z}^n$ to the equation $\mathbf{As} \equiv H(\mu) \pmod q$, where μ is the message, and q and $\mathbf{A}$ are an integer and a matrix that define a lattice. Verification checks that $\mathbf{s}$ is indeed small and that the equation holds. In the context of lattice cryptography, the problem of finding a short $\mathbf{s}$ that satisfies the equation is called Short Integer Solutions (SIS), and can be rephrased as finding a vector that is close to a certain lattice defined by $\mathbf{A}$ and q, a problem similar to SVP.

Fiat Shamir with Aborts This approach uses the generic Fiat-Shamir transform to convert an identification (ID) scheme into a signature scheme.

- **ID scheme:** The aim of this protocol is for a *Prover* to interactively prove to a *Verifier* that he knows a secret key `sk` (which is linked to his identity as he is the only person to know it), without revealing any information on `sk`. ID schemes are interactive: the Prover first commits to some randomness r, and sends a witness $w =$ `Commit(`r`)` to the Verifier. This ensures that the Prover cannot cheat on his value of r, without revealing r. The Verifier then generates a random challenge c, and sends c to the Prover. Finally, the Prover sends his proof $z =$ `Prove(sk,`r`,`c`)` to the Verifier. The Verifier then uses the public verification key `vk` associated with `sk` to check if `Verify(vk,`z`,`c`)`. The scheme is *complete* if a Prover that knows `sk` can convince an honest verifier, so `Verify()` should return true if z is generated according to the protocol. The scheme should also be *sound*, *i.e.* a cheating Prover that does not know `sk` should not be able to convince the Verifier unless with negligeable probability.

- **Fiat-Shamir transform:** The Fiat-Shamir transform converts the interactive protocol above into a non-interactive version by having the Prover (renamed the Signer) defining the challenge c himself deterministically as a hash of the commitment concatenated with the message. If the hash behaves as a random oracle, then this step simulates the Verifier generating a random c. Ultimately, the proof element z acts as the signature to verify. Note that c should also be included in the signature, otherwise it would be impossible for the Verifier to guess it.

Textbook use of Fiat-Shamir goes back to Schnorr's ID and signature schemes, and relies on the hardness of discrete logarithm: the Prover convinces the Verifier that he knows a secret exponent s. The verification key is g^s, where g is a generator of a cyclic group. In this setting, $z = r + cs$ happens to perfectly hide cs, as r is chosen uniformly as random, and applying a translation to the uniform distribution results in the same uniform distribution.

Rather annoyingly in the case of lattices, we expect our secret information to be a short vector $\mathbf{s}$ in some lattice, and adding a random vector $\mathbf{y}$ to the secret information $c\mathbf{s}$ does not hide it as well as in the previous example. Indeed $\mathbf{y}$ is sampled from an infinite set (instead of a finite set for r), so choices should be made regarding the distribution of $\mathbf{y}$. For example $\mathbf{y}$ could be sampled uniformly in a hypercube of radius R, by sampling each coordinate in $[-R, R] \cap \mathbb{Z}$ independently. However it is not hard to see that for some values of $\mathbf{y}$, $\mathbf{z} = \mathbf{y} + c\mathbf{s}$ leaks information: for instance if $\mathbf{z}$ has a coordinate that lies outside of the interval $[-R, R]$, the Verifier learns information on the corresponding coordinate of $c\mathbf{s}$. To patch this vulnerability, lattice-based Fiat-Shamir schemes resort to rejection sampling: if the protocol produces an element $\mathbf{z}$ that leaks information on $c\mathbf{s}$, then $\mathbf{z}$ is discarded, and the protocol restarts until a satisfying $\mathbf{z}$ is found, hence the name Fiat-Shamir *with Aborts*. Designing such schemes requires carefully choosing and analysing distributions, and choosing parameters in such a way that minimises the expected number of aborts, while preserving security.

Discussion The Falcon signature uses Hash-and-Sign, and has the most impressive performances among the new lattice standards. However, the scheme is very hard to implement securely, because the trapdoor requires the use of Gaussians, which require high precision floating point arithmetic. On the other hand, Dilithium uses Fiat-Shamir with Aborts, and explicitly avoids using Gaussian distributions, making it much easier to implement securely, although in exchange Dilithium has larger keys and signature sizes. Dilithium chooses uniform hypercubes for the distribution of $\mathbf{y}$. We dedicate an important fraction of this manuscript to the study of digital signatures, demonstrating cryptanalysis of a Hash-and-Sign scheme in Chapter 6, and the design of a Fiat-Shamir with Aborts scheme with a different well-motivated choice of distribution for $\mathbf{y}$ in Chapter 7.

2.4 Special Lattices in Cryptography

Real lattices Full-rank lattices in $\mathbb{R}^n$ of unit co-volume can be identified with the space

$$X_n = \mathrm{SL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{Z}),$$

in such a way that a point $[\mathbf{B}] \in X_n$ generates the lattice $\mathbb{Z}^n\mathbf{B}$. This definition makes sense as two representatives of the same class must be equal up to a unimodular matrix $\mathbf{U} \in \mathrm{SL}_n(\mathbb{Z})$, and $\mathbf{B}$ and $\mathbf{BU}$ generate the same lattice. The space of lattices X_n is a locally compact topological group, which means that it possesses a unique (up to scaling) Haar measure μ_n: a measure that is invariant under right-multiplication[5] by elements of $\mathrm{SL}_n(\mathbb{Z})$. Siegel [Sie45] proved that $\mu_n(X_n) < \infty$, so we can assume it is rescaled to define a probability measure. This measure is

[5] μ_n happens to also be left-invariant.

arguably the most natural way to define random real lattices, for which some average properties can be derived. For example, classical works of Siegel, Rogers and others study the average number of lattice points in an origin-centred ball, which relates to the lengths of short lattice vectors.

Integer lattices Cryptographers do not like representing real numbers on a computer, so instead they usually only consider sublattices of $\mathbb{Z}^m$, whose entries are all integral. Given a finite abelian group G we consider the following space of lattices:

$$L(G) = \{L \subseteq \mathbb{Z}^m : \text{rank}(L) = m, \mathbb{Z}^m/L \cong G\}.$$

$L(G)$ is finite, and therefore we can sample lattices according to the uniform distribution on $L(G)$. In his 1996 paper "Generating Hard Instances of Lattice Problems" [Ajt96] Ajtai was able to prove a worst-case to average case reduction for lattice problems with lattices in $L((\mathbb{Z}/q\mathbb{Z})^n)$. This result was foundational for lattice-based cryptography. It shows that if an attacker can solve SVP for a random lattice (average-case), then he is also able to solve SVP in the worst case. This is strong evidence that SVP is hard in all lattices of the class. This was generalised by [GINX16] to (G_n) a sequence of abelian groups such that $|G_n|$ grows to infinity fast enough. Additionally, the authors of [EO06] prove that the distribution obtained by sampling a uniform lattice of $L(G_n)$ and rescaling it to unit covolume converges towards the Haar measure μ_n, so for large enough n we expect average results to hold for both distributions.

NTRU lattices It seems like 1996 was a great year for lattice cryptography, indeed that year also marks the invention of the NTRU cryptosystem by Hoffstein, Pipher and Silverman. The security of NTRU also depends on how easy it is to recover short vectors in a lattice, but this time the lattice Λ has a very special structure: it is invariant under a certain cyclic permutation of its coordinates: for example if the vector

$$\mathbf{v} = (1, 2, 3, 4, 5, 6) \in \Lambda$$

is a lattice vector, then so are $(3, 1, 2, 6, 4, 5)$ and $(2, 3, 1, 5, 6, 4)$. This allows for a more compact representation of the lattice through polynomials: the vector $\mathbf{v} \in \mathbb{Z}^6$ as well as its shifts are simultaneously represented by a single vector of polynomials $(f, g) \in (\mathbb{Z}[X]/(X^3 - 1))^2$. In our example we would have $f = 1 + 2X + 3X^2$ and $g = 4 + 5X + 6X^2$. The quotient essentially only says that $X^3 = 1$, which means that the two shifts of $\mathbf{v}$ are exactly $(X \cdot f, X \cdot g)$ and $(X^2 \cdot f, X^2 \cdot g)$. This polynomial representation leads to more efficient storage, and for well-chosen polynomial rings and an extra modulus q, it enables the use of the NTT (Number Theoretic Transform), a Fast Fourier Transform-like operation that speeds up polynomial multiplication, making schemes with such structure orders of magnitude more efficient.

Unfortunately, $X^3 - 1$ is not irreducible in $\mathbb{Q}[X]$, which means that the quotient ring $\mathbb{Q}[X]/(X^3 - 1)$ does not define a field, and can in fact be decomposed as a product of smaller fields: attackers can use this additional structure. For this reason, more recent lattice-based schemes use lattices that arise from number fields: a number field $\mathbb{Q}(\alpha)$ is the smallest field containing 1 and α, where α is algebraic. This means that α is the root of a monic irreducible polynomial f with coefficients in $\mathbb{Z}$, implying that $\mathbb{Q}(\alpha)$ can also be represented using the polynomial ring $\mathbb{Q}[X]/f(X)$.

Ideal lattices In the same way a number field $K = \mathbb{Q}(\alpha)$ generalises the rationals $\mathbb{Q}$, its ring of integers $\mathcal{O}_K$ generalises the rational integers $\mathbb{Z}$. Ideals of $\mathcal{O}_K$ are sets $I \subseteq \mathcal{O}_K$ that are closed under multiplication by elements of $\mathcal{O}_K$: $\mathcal{O}_K \cdot I \subseteq I$. This definition already applies in $\mathbb{Z}$, where the ideals are exactly the sets of the form $a\mathbb{Z}$, for $a \in \mathbb{Z}$. A fundamental fact in the part of number theory that is called geometry of numbers is that ideals and lattices are essentially the

same thing: elements of a degree n number field K are canonically mapped to points in $\mathbb{R}^n$, with the standard inner product, and ideals map to lattices. This mapping - usually referred to as the canonical or Minkowski embedding - preserves the algebraic structure of the field elements, and allows us to essentially view classes of lattices, called ideal lattices, as algebraic objects represented by ideals in the number field. The problem of finding short vectors (where here short refers to the Euclidean norm in the embedded lattice) in ideals is called Ideal-SVP. The Ring Learning with Errors problem (Ring-LWE) [SSTX09; LPR10] - an instantiation of Regev's LWE problem [Reg05] over number fields - was shown to be provably at least as hard as worst-case instances of Ideal-SVP.

There is a clear trade-off that needs to be made when extra structure is added to a cryptographic primitive: the extra algebraic structure allows for more efficient schemes at the cost of weakening security. Indeed, new attacks could exploit the extra structure. In the case of ideal lattices, a beautiful line of work [CGS14; BS16; CDPR16; CDW17; BEFGK17; CDW21; PHS19; BR20; BLNR21] discovered both classical and quantum attacks on (approximate)-Ideal-SVP. However, neither are practical for cryptographically relevant dimensions and approximation factors, as those attacks perform worse than the unstructured BKZ lattice reduction algorithm in this regime. Understanding unit and S-unit attacks was the main topic of my masters thesis [Bam22].

Module lattices Although successful attacks on Ideal-SVP would not break Ring-LWE but only the security reduction, it is preferable for security to be based on a harder problem than Ideal-SVP: for this ideals are replaced by $\mathcal{O}_K$-modules of higher rank.

Just as ideals of $\mathcal{O}_K$ generalise integers in $\mathbb{Z}$, $\mathcal{O}_K$-modules of rank r generalise vectors of r integers. For this reason, ideals of $\mathcal{O}_K$ are simply rank 1 $\mathcal{O}_K$-modules. The geometric object obtained by applying the canonical embedding to a module component-wise is a module lattice. Note that NTRU can be seen as a problem over rank 2 module lattices. A rank r module over a degree d number field embeds into a lattice of $\mathbb{R}^{rd}$. All the information contained in a rank r module lattice vector can be represented with r^2 elements of $\mathcal{O}_K$. As soon as $r > 1$, this is more than what is needed to represent ideal lattices, but if r is constant the gain remains huge compared to lattices without any structure.

Security-wise however, the algebraic attacks that exploit the structure of ideals break down as soon as $r \geq 2$. Currently, the best attacks against Module-SVP for $r \geq 2$ are the same algorithms as those used to attack general unstructured SVP. Consequently, the Module-LWE problem is provably [LS15] at least as hard as solving Module-SVP, a problem that we do not know how to attack better than SVP. This perceived hardness gap is what makes all new lattice-based standards (Kyber, Dilithium, Falcon, Haetae) rely on module lattices.

Hypercubic lattices To end this section on special lattices, we conclude with the simplest and most natural of all: $\mathbb{Z}^n$. Given a (bad) basis $\mathbf{B}$ of $\mathbb{Z}^n$, finding short vectors in $\mathbb{Z}^n$ is a trivial problem: we know them already, they are the unit vectors

$$(0, \cdots, 0, \pm 1, 0, \cdots, 0).$$

However if instead we are given a bad basis $\mathbf{B}$ of Λ, where $\Lambda = \mathbf{O} \cdot \mathbb{Z}^n$ is the lattice obtained after rotating $\mathbb{Z}^n$ by an orthonormal matrix $\mathbf{O} \in \mathcal{O}_n(\mathbb{R})$, then recovering the unit vectors of Λ is (maybe surprisingly) conjectured to be a hard problem: the $\mathbb{Z}$-Lattice Isomorphism Problem ($\mathbb{Z}$LIP). We call such lattices Λ hypercubic.

The security of the signature scheme HAWK, a second round candidate to NIST's additional call for post-quantum signatures relies on the hardness of a module variant of $\mathbb{Z}$LIP.

2.5 Technical Overview and Main Contributions

This thesis contains material related to the field of Post-Quantum Cryptography. It focuses on two main topics and a secondary third topic. The main topics relate to lattice-based cryptography, and can be summarised by the following questions:

Question 2.5.1. *Lattices used to design efficient cryptographic schemes have very special structure. How does this impact cryptanalysis?*

Question 2.5.2. *The design of well-rounded digital signatures is one of the major challenges of the post-quantum transition effort. Is it possible to improve their performances?*

The third topic is orthogonal to the previous two, and concerns the study of mathematical structures arising in ordinary isogeny graphs. Those graphs are related to isogeny-based cryptography, where security does not rely on lattices, but on the hardness of finding isogenies between elliptic curves.

Main Topic 1: Security of cryptography based on special lattices

In order to understand the computational hardness of a cryptographic scheme, one has to understand the precise cost of the best attacks. Those attacks are algorithms that can be studied in different ways:

- Concrete vs Asymptotic complexity: in computer science, the complexity of an algorithm is usually measured asymptotically, with respect to a varying parameter. While the asymptotic analysis of attacks gives important insight into how strong an attack really is as the security parameter goes to infinity, the concrete analysis counts the exact number of operations required by an attacker to break a scheme with given parameters.
- Provable vs Heuristic analysis: the analysis of an algorithm can lead to mathematically proven guarantees on the output or the runtime. In this case we say that the algorithm is provable. However, when a fully rigorous proof is out of reach, it is still possible to estimate the performance through experiments and conjectures. If this is the case, we say that the analysis is heuristic[6].

Both concrete and asymptotic complexities are important. We focus mostly on asymptotic complexity as it provides the bigger picture, and gives insight as to how different parts contribute to the final cost. Usually, heuristic algorithms outperform provable algorithms, so only heuristic algorithms should be taken into account for security, but provable algorithms provide guarantees. If there is a large gap between the best provable and heuristic algorithms for a problem, this usually means that the problem (or the algorithm) is not yet well understood.

In the previous section, we have been acquainted with a diverse cast of lattices, all with special structure compared to what one would expect of a random real lattice. We focus on algorithms that aim to recover a shortest non-zero vector of the lattice, assuming access to an oracle for SVP in dimension β. By analogy to the BKZ algorithm, we refer to this parameter β as the blocksize, and ask the following question.

Question 2.5.3. *For a given sequence of lattices (Λ_n) parametrised by a parameter n (usually $n = \mathrm{rank}(\Lambda_n)$), what is the smallest blocksize $\beta(n)$ such that there exists an algorithm with access to $\mathrm{poly}(n)$ calls to an SVP oracle in dimension $\beta(n)$ that solves SVP in that lattice?*

We notice that both NTRU and hypercubic lattices share the following special properties:

1. They have unusually short vectors.

[6] We will also speak of heuristic algorithms to denote algorithms that have a heuristic analysis.

2. They have many shortest non-zero vectors.

3. They are essentially self-dual.

With (1.) in mind, we give a precise answer to Question 2.5.3 under heuristic assumptions relating to the primal attack, and show that in this setting the first order terms of the asymptotic blocksize are not impacted by (2.). We then focus on (3.), where we compare the lattice Λ_n with its dual

$$\Lambda_n^\vee = \{\mathbf{y} \in \mathbb{R}^n : \forall \mathbf{x} \in \Lambda_n, \langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{Z}\}.$$

NTRU For random lattices we expect $\lambda_1(\Lambda_n)\lambda_1(\Lambda_n^\vee) \sim \frac{n}{2\pi e}$, yet for NTRU and hypercubic lattices this value is much smaller. If this product is small enough, we design a variant of the slide reduction algorithm that allows us to prove that the provable value of $\beta(n)$ must be at most $n/2$. For most classical instantiations of NTRU, this leads to a provable attack that only uses polynomially many calls to an oracle in dimension a half of the dimension of the full lattice, solving a conjecture of Gama, Howgrave-Graham and Nguyen from 2006 [GHN06].

Hypercubic Hypercubic lattices have additional properties, relating to the fact that all of its short vectors are orthogonal. Ducas showed in [Duc23] that $\beta(n)$ is provably less than $n/2+O(1)$ for hypercubic lattices. We show that for his algorithm to work, it is sufficient to have access to $(\sqrt{2}-\varepsilon)$ approximate SVP oracles. We also sketch an attack that heuristically shows that knowing $O(\log(n))$ sized vectors of the hypercubic lattice Λ_n is enough to recover unit vectors in quasi-polynomial time. This reduction is much more practical than the one presented in [Jia+23].

Ideal Finally, we aim to study average behaviour of ideal lattices, especially their first minimum, a quantity that heavily influences the behaviour of SVP algorithms. Contrarily to real lattices whose first minimum can be arbitrarily short, this cannot be the case for ideal lattices: indeed, if I is an ideal of the degree n number field K with discriminant Δ_K, then for any non-zero $x \in I$, we have:

$$\|x\| \geq \sqrt{n} \cdot |N_{K/\mathbb{Q}}(x)|^{1/n} \geq \sqrt{n} \cdot N(I)^{1/n} = \sqrt{n} \cdot |\Delta_K|^{-\frac{1}{2n}} \cdot \mathrm{vol}(\sigma(I))^{1/n},$$

where $\sigma(I)$ is the lattice obtained by embedding I into $\mathbb{R}^n$. For x such that $\|x\| = \lambda_1(\sigma(I))$, we get a hard lower bound on $\lambda_1(\sigma(I))$ that only depends on the field.

Question 2.5.4. *How does the average value of the first minimum of a random ideal lattice behave ? Is it significantly different to that of a random real lattice ?*

The inequality above suggests that ideal lattices might exhibit different average properties than Haar-random lattices. In fact, advanced results in number theory that use tools far outside the scope of this thesis seem to indicate that when the degree and/or discriminant of the number field increase, quantities such as the average number of ideal lattice points in an origin centred ball converge towards the result one would expect for real lattices. We start by focusing on quadratic fields, where we give an algorithm to compute the exact values of the moments of the first minimum over random ideal lattices in a given class, by integrating over geodesic orbits on the modular surface.

We then turn to the problem of computing the expected value of sums of the form

$$\sum_{\mathbf{v} \in \Lambda} f(\mathbf{v})$$

over the space of ideal lattices, for functions $f : \mathbb{R}^n \to \mathbb{R}$. This is an attempt at finding an analogous version of Siegel's mean value formula, that claims that for well-behaved f,

$$\int_\Lambda \sum_{\mathbf{v}\in\Lambda} f(\mathbf{v}) d\mu_n = f(0) + \int_{\mathbb{R}^n} f(x)dx.$$

Plugging in the indicator of a centred ball for f directly gives what is commonly known as the *Gaussian heuristic*: the expected number of points in a ball of volume V is exactly $1 + V$. Such results happen to also be relevant for lower bounds on lattice packings, where average results on lattices with more constrained first minima can help prove the existence of a lattice with unusually large first minimum [Ven13]. We sketch our attempt at a proof, and obtain a formula that can be computed only in very small dimensions. While we were working on this problem, Gargava and Viazovska [GV24] successfully derived a much more natural but complementary formula for the same quantity. They conclude that for cyclotomic fields of large enough degree, the Gaussian heuristic should still hold for random ideal lattices.

Main Topic 2: Improving digital signatures

The second main overarching topic of this thesis is motivated by the fact that while standards for encryption should be deployed as soon as possible, there is still a bit of time left to improve upon the design of signature schemes, at least until quantum computers become powerful enough to forge classical signatures.

In a certain sense, because most lattice-based digital signatures incorporate special lattices as their main design choice, the general goal of this section includes the Topic 1, however here instead of studying the hardness of the underlying lattice problem, we investigate design choices at the protocol level.

DEFI We first study a Hash-and-Sign signature based on isotropic quadratic forms and proposed by Feussner and Semaev: DEFI. This signature slightly outperforms essentially all signature schemes considered by NIST on most reasonable metrics, and for this reason it seemed important to analyse it. The scheme relies on variants of an interesting problem which we formally define as Quadratic Form Equivalence (QFE), and its module version. QFE can be seen as a direct generalisation of LIP to quadratic forms that are not necessarily positive definite. In the case of DEFI, the private key is a 4×4 matrix $\mathbf{B}$ with coefficients in $R = \mathbb{Z}[X]/(X^{64} + 1)$, and the verification key is

$$\mathbf{C} = \mathbf{B}^T\mathbf{J}\mathbf{B},$$

where $\mathbf{J} = \mathrm{diag}(1, 1, -1, -1)$. Note that for this to constitute a generalisation of module-LIP, the transpose operation should be replaced by the more natural adjoint operation. Interestingly, DEFI is presented as a multivariate scheme. Although the new assumption that DEFI uses might seem promising, we show that the scheme is insecure because of the way the trapdoor is generated. Each signature provides a linear equation over R, whose solutions are related to the secret key. Through lattice reduction used on well-chosen lattices (that happen to have module structure!), we are able to practically recover the secret key in a few minutes, using less than ten (message, signature) pairs. Unfortunately this means that DEFI is not secure.

Patronus We decide to now focus instead on the study of a class of signatures that have provable security reductions from hard lattice problems: lattice-based signatures in the Fiat-Shamir with Aborts paradigm. Recall from earlier in the introduction that such schemes require choosing the distribution of a masking parameter $\mathbf{y} \in \mathbb{R}^n$. This masking parameter is added to the secret information $c\mathbf{s} \in \mathbb{R}^n$, and the resulting element $\mathbf{z} = \mathbf{y} + c\mathbf{s}$ is rejected if it falls outside of the "rejection zone", the zone where it would leak some information on $c\mathbf{s}$. For technical

reasons, the Euclidean norm of a non-aborted $\mathbf{z}$ is directly related to the size of the signature. The NIST-selected standard Dilithium uses this framework and chooses to sample $\mathbf{y}$ uniformly from a hypercube. The resulting "rejection zone" turns out to also be a hypercube. It was shown by [DFPS22] that in order to minimise the expected norm of $\mathbf{z}$, the best distribution for $\mathbf{y}$ is either a Gaussian distribution, or a uniform distribution over a Euclidean ball. This motivated the design of the now-winner of the South Korean PQC competition Haetae. However if Haetae beats Dilithium on signature sizes, it has the drawback of requiring samples from Gaussian distributions, a very costly and potentially vulnerable operation.

Question 2.5.5. *With equivalent security guarantees, is it possible to improve signature sizes in* Dilithium *without the need for Gaussian sampling or fixed point arithmetic?*

In our work, we aim to answer Question 2.5.5 by generalising the Fiat-Shamir with Aborts framework for lattice-based signatures to more general polytopes. We define a special polytope with radius r and in dimension n by intersecting ℓ_1 and ℓ_∞ balls as follows:

$$\mathcal{H}_r^n := B_n^{(\infty)}(r) \cap B_n^{(1)}(r\sqrt{n}).$$

The choice of a cross-polytope is not innocuous: it is the dual of the hypercube. In fact as the dimension grows, the mass in $\mathcal{H}_r^n$ concentrates towards the centre of the ℓ_1 ball, and this will lead to $\mathbf{z}$ with shorter Euclidean norm. This is dual to the well-known fact that random points in a high-dimensional hypercube concentrate towards the corners, which in some sense says that the hypercube is an unfortunate choice of distribution. The following computation helps to understand the difference: we compute the expected squared Euclidean norms of a uniform point in $B_n^{(\infty)}(r)$, $B_n^{(1)}(r\sqrt{n})$ and $B_n^{(2)}(r)$.

$$\int_{B_n^{(\infty)}(r)} \|\mathbf{z}\|^2 d\mathbf{z} = \frac{nr^2}{3}; \quad \int_{B_n^{(1)}(r\sqrt{n})} \|\mathbf{z}\|^2 d\mathbf{z} = \frac{2n^2r^2}{(n+1)(n+2)}; \quad \int_{B_n^{(2)}(r)} \|\mathbf{z}\|^2 d\mathbf{z} = \frac{nr^2}{n+2}.$$

Ignoring technicalities related to the fact that the expected value of a square root is not directly the square root of the expected value, the integral on the left tells us that a uniform random point in $B_n^{(\infty)}(r)$ has Euclidean norm approximatively $\sqrt{n/3} \cdot r$ on average. The integral on the right tells us that a uniform random point in $B_n^{(1)}(r\sqrt{n})$, which has the same inscribed Euclidean ball, has Euclidean norm approximatively $\sqrt{2} \cdot r$ on average, which is a constant factor away from the same result for the Euclidean ball. If sampling integral points in $\mathcal{H}_r^n$ is easy, then this means that replacing the hypercube by $\mathcal{H}_r^n$ in Dilithium should lead to a nice improvement in signature size, without needing any major changes in the security proof. We show that indeed, sampling of uniform integral points in $\mathcal{H}_r^n$ is possible and can be made isochronous: in exchange of a slightly more complicated sampler, signing becomes a bit slower than for Dilithium, but doesn't require much more randomness.

We call our new signature scheme Patronus[7]. It achieves shorter signature sizes compared to the standardised scheme Dilithium. Not illogically, Patronus signatures remain larger than those of Haetae, but the sampling process is much faster, and most importantly does not involve Gaussians.

Topic 3: Isogeny volcanoes

Our last topic is the study of a structure that relates not to lattice-based but to isogeny-based cryptography. We consider elliptic curves with endomorphism ring $\mathcal{O}$ strictly greater than $\mathbb{Z}$, they are said to have complex multiplication (CM). While an elliptic curves over $\mathbb{C}$ can be seen

[7] Latin word for "protector".

as a lattice in the complex plane, CM elliptic curves when viewed over $\mathbb{C}$ have extra rotational symmetries: they are in fact ideal lattices for ideals of $\mathcal{O}$, their endomorphism ring.

For $p > 5$ and $\ell \neq p$ two prime numbers, we define the graph $\mathcal{G}_\ell(\mathbb{F}_p)$ whose vertices $\mathcal{V}$ are j-invariants corresponding to $\overline{\mathbb{F}}_p$-isomorphism classes of ordinary elliptic curves. For $j, j' \in \mathcal{V}$, $\mathcal{G}_\ell(\mathbb{F}_p)$ has as many edge between j and j' as the multiplicity of (j, j') as a root of the ℓ-th modular polynomial $\Phi_\ell(X, Y)$. Edges can also be seen as degree ℓ maps between elliptic curves: ℓ-isogenies. The modular polynomials are fundamental objects, used notably in the Schoof-Elkies–Atkin algorithm for point counting on elliptic curves [FM02]. The ordinary isogeny graph was used by Couveignes, Rostovtsev, and Stolbunov [Cou06; RS06] to propose a key exchange protocol, precursor to more modern isogeny-based schemes like CSIDH [Cas18].

The structure of ordinary isogeny graph was understood by Kohel in his PhD thesis (again, in 1996) [Koh96]. Such graphs exhibit "volcano"-like structures:

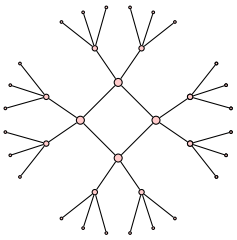


Figure 2.8: A volcano with size 4 crater.

To each vertex in this graph, one can attach the endomorphism ring of curves corresponding to the vertex. Endomorphism rings for ordinary elliptic curves are orders in imaginary quadratic fields, which in fact makes them ideal lattices of rank 2. Moving down (resp. up) in the graph through ℓ-isogenies acts on those lattices by converting them into index ℓ sublattices (resp. index ℓ^{-1} superlattices). Interesting computational aspects of isogeny volcanoes were explored by [IJ10; Sut13].

In our work we ask the following question, which we see as an inverse problem:

Question 2.5.6. *Which "volcano"-looking graphs actually appear as connected components of an ordinary isogeny graph $\mathcal{G}_\ell(\mathbb{F}_p)$?*

By relating Question 2.5.6 to classical Diophantine equations and results from class field theory, we show that all reasonable volcanoes can be constructed, and give a naïve algorithm to find minimal values of p and ℓ that realise a target volcano. Unfortunately, this algorithm is not particularly efficient as it requires class group computations.

Thesis structure This thesis is arranged into four parts, both truly consisting of two chapters.

- Part I starts with a French translation of this introduction as Chapter 1, the present chapter Chapter 2, and then continues with Chapter 3, where we introduce preliminary material required for the main body of the thesis;
- Part II focuses on attacks on NTRU and hypercubic lattices, with emphasis on heuristic attacks in Chapter 4 and provable attacks in Chapter 5;
- Part III relates to digital signatures, we present our cryptanalysis of DEFI in Chapter 6, and our polytope-based Fiat-Shamir with Aborts framework in Chapter 7;
- Part IV deals with more mathematically involved questions, Chapter 8 studies random and especially random ideal lattices, and finally Chapter 9 presents our solution to the inverse volcano problem in ordinary isogeny graphs over $\mathbb{F}_p$.

Chapters from Parts II, III and IV are mostly self-contained, and can be read in any order. Happy reading!

Chapter

3 Preliminaries

Chapter content

3.1 General Notations

Sets, vectors, matrices

$\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ denote the sets of integer, rational and real numbers respectively. To avoid confusion, we prefer the notation $\mathbb{Z}_{\geq 0}$ to $\mathbb{N}$ for the set of non-negative integers. For $x \in \mathbb{R}$, we use $\lfloor x \rfloor$ for its lower integer part, and $\{x\}$ for its fractional part. For $z \in \mathbb{C}$, $\Re(z)$ and $\Im(z)$ denote respectively the real and imaginary parts of z. Unless otherwise specified, n is an integer and denotes an ambient dimension. We use $\subseteq$ to denote inclusion of sets, and $\subset$ to denote strict inclusion of sets.

For a set $\mathcal{P} \subseteq \mathbb{R}^n$ we note $\mathrm{vol}(\mathcal{P}) \in \mathbb{R}_{\geq 0} \cup \{\infty\}$ its volume when appropriate, and $\mathrm{card}(\mathcal{P}) \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$ or $\#\mathcal{P}$ its cardinality. For a set $X \subseteq \mathbb{R}^n$, $X_{\mathbb{Z}} = X \cap \mathbb{Z}^n$ denotes its restriction to the integers and $\mathrm{conv}(X)$ denotes its convex hull, the smallest convex region of $\mathbb{R}^n$ containing X.

Vectors are written in bold lowercase $\mathbf{v}$, and matrices in bold uppercase $\mathbf{M}$. We use row convention for writing matrices. For a set of vectors $V \subseteq \mathbb{R}^n$, we write $\mathrm{span}(V)$ the real vector space generated by V. The Euclidean scalar product of $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^n$ is written $\langle \mathbf{a}, \mathbf{b} \rangle$. We write $\mathrm{span}(V)^{\perp}$ for the set of vectors $\mathbf{w} \in \mathbb{R}^n$ such that $\langle \mathbf{w}, \mathbf{v} \rangle = 0$ for all $\mathbf{v}$ in V. π_V denotes the orthogonal projection onto $\mathrm{span}(V)$.

For a ring R and a positive integer $n \in \mathbb{Z}_{>0}$, $M_n(R)$ denotes the set of $n \times n$ matrices with entries in R, and $\mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ denotes the diagonal matrix of $M_n(R)$ with coefficients (α_i). $\mathrm{GL}_n(R)$, $\mathrm{SL}_n(R)$ and $\mathrm{PSL}_n(R)$ denote the sets of matrices with determinant respectively non-zero, ± 1, and 1. We use $[n]$ as a notation for $\{1, \ldots, n\}$. If $z_1 \in R$ and $\mathbf{z} = (z_2, \ldots, z_n) \in R^{n-1}$, we will use $(z_1 || \mathbf{z})$ to denote the concatenated vector $(z_1, \ldots, z_n) \in R^n$.

Norms, balls and polytopes

The Euclidean norm of a vector $\mathbf{v} \in \mathbb{R}^n$ is denoted by $\|\mathbf{v}\|_2$, or more plainly $\|\mathbf{v}\|$. For $\mathbf{x} \in \mathbb{R}^n$ and $p \in \mathbb{R}_{>0} \cup \{\infty\}$, we let $\|\mathbf{x}\|_p$ denote the ℓ_p norm of $\mathbf{x}$. The p-ball (*resp.* p-sphere) of radius r centred at $\mathbf{c}$ in ambient dimension n is denoted by $\mathcal{B}_n^{(p)}(r, \mathbf{c})$ (*resp.* $\mathcal{S}_{n-1}^{(p)}(r, \mathbf{c})$) or $\mathcal{B}_n^{(p)}(r)$ (*resp.* $\mathcal{S}_{n-1}^{(p)}(r)$) for $\mathbf{c} = \mathbf{0}$. Omitting (p) will implicitly refer to the Euclidean case $p = 2$.

Table 3.1: Volume and cardinality of ℓ_1, ℓ_2 and ℓ_∞ balls.

	$B_n^{(1)}(r)$	$B_n^{(2)}(r)$	$B_n^{(\infty)}(r)$
$\mathrm{vol}(B)$	$\frac{(2r)^n}{n!}$	$\frac{\pi^{n/2}r^n}{\Gamma\left(1+\frac{n}{2}\right)}$	$(2r)^n$
$\mathrm{card}(B_\mathbb{Z})$	$\sum_{i=0}^{\min(r,n)} \binom{n}{i}\binom{r}{i}2^i$	$\approx \mathrm{vol}(B_n^{(2)}(r))$	$(2r+1)^n$

Table 3.1 provides a brief summary of known closed forms for the volume of common ℓ_p balls as well as their number of integral points. Precise estimates for $\mathrm{card}(\mathcal{B}_n(r)\cap\mathbb{Z}^n)$ are notoriously difficult to obtain in general. For large enough r we refer to the estimates of [Ste17].

Definition 3.1.1 (Polytope)**.** A set $\mathcal{P}\subset\mathbb{R}^n$ is a polytope if it is the convex hull of a finite set of vertices of $\mathbb{R}^n$. If n and v are integers, then the set $(\mathbf{x}_i)_{1\le i\le v}\in(\mathbb{R}^n)^v$ is the set of vertices of the polytope $\mathcal{P}\subset\mathbb{R}^n$ in dimension n if $\mathcal{P}$ is the convex hull of $(\mathbf{x}_i)_{1\le i\le v}$ and if there is no strict sub-family of $(\mathbf{x}_i)_{1\le i\le v}$ whose convex hull is equal to $\mathcal{P}$. If in addition, the linear span of the vertices is $\mathbb{R}^n$, then the polytope is referred to as full-rank. Lastly, a polytope $\mathcal{P}$ with vertices in $\mathbb{Z}^n$ is said *integral*.

Unless stated otherwise, by polytope we always mean full-rank polytope. A polytope can also be defined as an intersection of half-spaces of finite volume. The hyperplanes in this equivalent definition define the *facets* of the polytope. The vertices of a polytope are unique up to ordering, and we write $\mathcal{V}(\mathcal{P}):=\{(\mathbf{x}_i)_{1\le i\le v}\}$ for the set of vertices of $\mathcal{P}$.

Definition 3.1.2 (Translation and dilation of polytopes)**.** For a polytope (or any subset of $\mathbb{R}^n$) $\mathcal{P}\subseteq\mathbb{R}^n$, a centre $\mathbf{c}\in\mathbb{R}^n$, and a dilation factor $r\in\mathbb{R}$, we define $\mathcal{P}_{r,\mathbf{c}}:=\{r\mathbf{x}+\mathbf{c}\ :\ \mathbf{x}\in\mathcal{P}\}$. We will omit r if $r=1$ and $\mathbf{c}$ if $\mathbf{c}=0$. It follows that $\mathcal{V}(\mathcal{P}_{r,\mathbf{c}})=\{r\mathbf{x}+\mathbf{c}\ :\ \mathbf{x}\in\mathcal{V}(\mathcal{P})\}$.

Definition 3.1.3 (Symmetric and inscribed polytopes)**.** A full-rank polytope $\mathcal{P}$ is *symmetric* if $\mathcal{P}=\mathcal{P}_{-1}$ (or equivalently if $\mathcal{V}(\mathcal{P})=-\mathcal{V}(\mathcal{P})$). A full-rank polytope $\mathcal{P}$ is an *inscribed* polytope if for all $\mathbf{x},\mathbf{y}\in\mathcal{V}(P)$, $\|\mathbf{x}\|_2=\|\mathbf{y}\|_2$. The radius of an inscribed polytope or *circumradius* is the ℓ_2 norm of its vertices.

Definition 3.1.4 (Circumscribed polytope)**.** A polytope $\mathcal{P}$ is *circumscribed* if there exists a Euclidean ball that is tangent to all of the facets of $\mathcal{P}$. The radius of this ball is the *inradius* of $\mathcal{P}$.

Proposition 3.1.5 (Intersection of polytopes [Brø83, (Section 1)])**.** *The intersection of two (full-rank) polytopes is a (not always full-rank) polytope. If the intersection contains a non-trivial open ball, then it is also full-rank.*

Proposition 3.1.6 (Polytope vertices characterisation [Brø83, (Theorem 7.2)])**.** *Let $\mathcal{P}$ be a polytope. Then, $\mathbf{x}\in\mathcal{P}$ is not a vertex of $\mathcal{P}$ if and only if there exist vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{P}$ such that $\mathbf{x}\in(\mathbf{a},\mathbf{b})$, where*

$$(\mathbf{a},\mathbf{b})=[\mathbf{a},\mathbf{b}]-\{\mathbf{a},\mathbf{b}\}=\{t\mathbf{a}+(1-t)\mathbf{b}:t\in[0,1]\}-\{\mathbf{a},\mathbf{b}\}.$$

Asymptotic notations

We use the standard asymptotic notations $o(\cdot),O(\cdot),\Theta(\cdot)$ and $\omega(\cdot)$. As n goes to infinity, we use the notation $a_n\sim b_n$ as shorthand for $a_n=b_n+o(b_n)$. We use $<_{\mathrm{poly}}$ as follows: $a_n<_{\mathrm{poly}}b_n$ if there exists a monic polynomial P of constant degree such that for any large enough n, $a_n<b_n-\frac{1}{P(n)}$.

For a predicate P, we write $[\![P]\!]=1$ if P is true and 0 otherwise.

Modular arithmetic

For $q \in \mathbb{Z}_{>0}$ and $x \in \mathbb{Z}_q := \mathbb{Z}/q\mathbb{Z}$, we write $x \bmod^+ q$ the unique representative of the class of x in the interval $[0, q)$, and $x \bmod^\pm q$ the unique representative in $(-q/2, q/2]$.

We extend this definition to vectors entry-wise. For $x \in \mathbb{Z}_q$, we define $|x| := |x \bmod^\pm q|$. For any $q, n \in \mathbb{Z}_{>0}$, $p \in \mathbb{R}_{>0} \cup \{\infty\}$ and $\mathbf{x} \in \mathbb{Z}_q^n$, we define $\|\mathbf{x}\|_p$ as the ℓ_p norm of $|\mathbf{x}| \in \mathbb{R}^n$, where $|\cdot|$ is taken component-wise.

3.2 Probabilities

We denote the expectation of a random variable by $\mathbb{E}(\cdot)$, and probabilities by $\Pr(\cdot)$. For $\mathcal{D}$ a distribution, we write $\mathbf{z} \leftarrow\!\$\ \mathcal{D}$ to say $\mathbf{z}$ is sampled according to the distribution $\mathcal{D}$. In case $\mathcal{D}$ is not a distribution but a set, we use the convention that $\mathbf{z} \leftarrow\!\$\ \mathcal{D}$ means uniformly sampling $\mathbf{z}$ inside $\mathcal{D}$. If $\mathcal{D}$ is a probability distribution with values in a set X, we let $\mathsf{Supp}(\mathcal{D}) \subseteq X$ denote the support of $\mathcal{D}$.

Distribution of a projection

The squared norm of the projection of a unit vector of $\mathbb{R}^n$ onto a random k-dimensional subspace of $\mathbb{R}^n$ follows the *Beta distribution* $B(k/2, (n-k)/2)$. In particular, the expected squared norm of this projection is k/n. The cumulative distribution function of $B(a, b)$ is the *regularised incomplete beta function* $I_x(a, b)$. Asymptotic expansions of the regularised incomplete beta function rely on the *complementary error function* $\mathrm{erfc}(z) := \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2} dt$. When z goes to infinity, $\mathrm{erfc}(z) \sim \pi^{-1/2} z^{-1} e^{-z^2}$.

Rényi divergence

The Rényi divergence is a way of measuring closeness of two distributions. We will only need the order ∞ Rényi divergence, defined below.

Definition 3.2.1 (Rényi divergence)**.** Let $\mathcal{P}, \mathcal{Q}$ be two distributions such that $\mathsf{Supp}(\mathcal{P}) \subseteq \mathsf{Supp}(\mathcal{Q})$. The Rényi divergence of order ∞ is defined as follows:

$$R_\infty(\mathcal{P} \parallel \mathcal{Q}) = \max_{x \in \mathsf{Supp}(\mathcal{P})} \frac{\mathcal{P}(x)}{\mathcal{Q}(x)}.$$

Remark 3.2.2. The order ∞ Rényi divergence between two uniform distributions on measurable sets X_s and X_t, $X_t \subseteq X_s$, with non-zero volume (*resp.* on finite sets) is exactly the ratio of their volumes (*resp.* cardinalities).

Rejection sampling

Rejection sampling is a technique used to generate samples from a target distribution D_t based on samples from a source distribution D_s under the condition that the support of D_t is (almost) contained within the support of D_s.

Lemma 3.2.3 (Rejection sampling (from [Lyu12, (Lemma 4.7)]))**.** *Let D_s be a source distribution and D_t be a target distribution with $\mathsf{Supp}(D_t) \subseteq \mathsf{Supp}(D_s)$. If there exists $M > 1$ such that $\mathcal{R}_\infty[\, D_t \parallel D_s \,] \leq M$ then the output distributions of algorithms $\mathcal{A}$ and $\mathcal{F}$ are statistically indistinguishable.*

$\mathcal{A}$	$\mathcal{F}$
1 : $\mathbf{z} \hookleftarrow\$ D_s$	1 : $\mathbf{z} \hookleftarrow\$ D_t$
2 : *with probability* $\min\left(\frac{D_t(\mathbf{z})}{M \cdot D_s(\mathbf{z})}, 1\right)$:	2 : *with probability* $1/M$:
3 : **return** $\mathbf{z}$	3 : **return** $\mathbf{z}$
4 : **return** $\perp$	4 : **return** $\perp$

Notably, $\mathcal{A}$ outputs $\mathbf{z}$ with probability $\frac{1}{M}$.

Haar measure

A locally compact (Hausdorff) topological group $\mathcal{G}$ has a unique (up to scalars) non-zero left-invariant measure μ. For any measurable $\mathcal{H} \subseteq \mathcal{G}$ and $g \in \mathcal{G}$, it satisfies

$$\mu(g\mathcal{H}) = \mu(\mathcal{H}).$$

For example, the Haar measure on Euclidean space is the Lebesgue measure, and the Haar measure on the multiplicative group $\mathbb{R}^{\times}_{>0}$ is given by dx/x.

3.3 Lattices

A *lattice* L is a discrete subgroup of $\mathbb{R}^m$. Alternatively, we can define a lattice as the set $\mathcal{L}(\mathbf{b}_1, \ldots, \mathbf{b}_n) = \{\sum_{i=1}^n x_i \mathbf{b}_i \ : \ x_i \in \mathbb{Z}\}$ of all integer combinations of n linearly independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n \in \mathbb{R}^m$. This sequence of vectors is known as a *basis* of the lattice L. All the bases of L have the same number n of elements, called the dimension or rank of L, and the n-dimensional volume of the parallelepiped $\{\sum_{i=1}^n a_i \mathbf{b}_i \ : \ a_i \in [0,1)\}$ they generate. We call this volume the covolume, volume, or determinant, of L, and denote it by $\mathrm{vol}(L)$. If $\mathbf{B}$ is a basis of L, we have $\mathrm{vol}(L) = \sqrt{\det(\mathbf{B}\mathbf{B}^T)}$. The lattice L is said to be *full-rank* if $n = m$. We denote by $\lambda_1(L)$ the first minimum of L, defined as the norm of a shortest non-zero vector of L. $\lambda_1(L)$ is upper-bounded by Minkowski's theorem.

Theorem 3.3.1 (Minkowski)**.** *Let $L \subset \mathbb{R}^m$ be a lattice of rank n. Then*

$$\lambda_1(L) \leq \sqrt{n} \cdot \mathrm{vol}(L)^{\frac{1}{n}}.$$

Orthogonalisation

For a basis $B = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ of a lattice L, and an index $1 \leq i \leq n$, we denote by π_i the orthogonal projection on $\mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{i-1})^{\perp}$. The *Gram-Schmidt orthogonalisation* (GSO) of the basis B is defined as the orthogonal sequence of vectors $B^\star = (\mathbf{b}_1^\star, \ldots, \mathbf{b}_n^\star)$, where $\mathbf{b}_i^\star := \pi_i(\mathbf{b}_i)$. The projection of a lattice is not always a lattice, but for all $i \in \{1, \ldots, n\}$, $\pi_i(L)$ is a lattice of dimension $n + 1 - i$ generated by the basis $\pi_i(\mathbf{b}_i), \ldots, \pi_i(\mathbf{b}_n)$, and its volume is given by $\mathrm{vol}(\pi_i(L)) = \prod_{j=i}^n \|\mathbf{b}_j^\star\|$. When we speak of the (log) Gram-Schmidt profile, we refer to the plot of the quantities $(\ln \|\mathbf{b}_1^\star\|, \ldots, \ln \|\mathbf{b}_n^\star\|)$. It represents how well reduced the lattice basis is.

Duality

For any lattice L, its *dual lattice* $L^\vee$ is defined by

$$L^\vee := \{\mathbf{w} \in \mathrm{span}(L) : \langle \mathbf{w}, \mathbf{v} \rangle \in \mathbb{Z} \text{ for all } \mathbf{v} \in L\}.$$

If L has rank $n > 0$, then $L^\vee$ also, and $\mathrm{vol}(L) = \mathrm{vol}(L^\vee)^{-1}$. If $B = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ is a basis of L, then there is a unique *dual basis* $(\mathbf{d}_1, \ldots, \mathbf{d}_n)$ of $L^\vee$ such that $\langle \mathbf{b}_i, \mathbf{d}_j \rangle = \delta_{i,j}$ (Kronecker symbol) for all i, j. Duality is related to GSO as $\langle \mathbf{b}_i^\star / \|\mathbf{b}_i^\star\|^2, \mathbf{b}_i \rangle = 1$ implies that

$$\frac{\mathbf{b}_i^\star}{\|\mathbf{b}_i^\star\|^2} \in \mathcal{L}(\mathbf{b}_1, \ldots, \mathbf{b}_i)^\vee.$$

In particular, $\mathbf{d}_n = \mathbf{b}_n^\star / \|\mathbf{b}_n^\star\|^2$ and $\|\mathbf{d}_n\| = \|\mathbf{b}_n^\star\|^{-1}$. A lattice such that $\Lambda = \Lambda^\vee$ is called *unimodular* or *self-dual*.

Primitivity

A sublattice L' of L is called *primitive* if $L' = \mathrm{span}(L') \cap L$ or equivalently if its bases can be completed into a basis of L. This is also equivalent to L/L' being torsion free. We identify the quotient L/L' with the projection $\pi_{L'^\perp}(L)$, where by $L'^\perp$ we mean $\mathrm{span}(L')^\perp$. We will make heavy use of the following identity: if L' is a primitive sublattice of L,

$$L/L' = \pi_{L'^\perp}(L) = (L^\vee \cap L'^\perp)^\vee. \tag{3.1}$$

We refer [Mar03, Chapter 1] for a proof as well as a more complete presentation of the interconnections between duality and primitivity.

Gaussian mass

The Gaussian weight of a vector $\mathbf{x} \in \mathbb{R}^m$ with parameter $s > 0$ is

$$\rho_s(\mathbf{x}) = e^{-\pi \|x\|^2 / s^2}.$$

When the sum converges, this definition extends to a set X. For $s > 0$, the Gaussian mass of a set X is defined as

$$\rho_s(X) = \sum_{\mathbf{x} \in X} \rho_s(\mathbf{x}).$$

Definition 3.3.2 (Smoothing parameter)**.** Let $\varepsilon \in \mathbb{R}_{>0}$ be a real number and L be a lattice, then the *smoothing parameter* is defined as

$$\eta_\varepsilon(L) := \inf\{s > 0 : \rho_{1/s}(L^\vee \backslash \{\mathbf{0}\}) \leq \varepsilon\}.$$

Random lattices and Gaussian heuristic

The space $X_n = \mathrm{SL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{Z})$ of covolume 1 real lattices has a unique $\mathrm{SL}_n(\mathbb{Z})$-invariant Haar probability measure μ_n, that defines the mathematically correct way of thinking about a random lattice. The expected value for such a random lattice's first minimum is sometimes referred to as the Gaussian heuristic radius for lattices. For a rank n lattice L, we denote

$$\mathrm{GH}(L) := \mathrm{vol}(B_n(1))^{-1/n} \mathrm{vol}(L)^{1/n} = (1 + o_n(1)) \sqrt{\frac{n}{2\pi e}} \mathrm{vol}(L)^{1/n},$$

where $B_n(1)$ is the n-dimensional ℓ_2 ball of radius 1. The following theorem makes things more precise.

Theorem 3.3.3 ([LN24])**.** *Let L be a random lattice of X_n. Then with probability $1 - o(1)$ as n grows to infinity,*

$$1 - \frac{\ln \ln n}{n} \leq \frac{\lambda_1(L)}{\mathrm{GH}(L)} \leq 1 + \frac{\ln \ln n}{n}.$$

Lattice problems

Let $\gamma \geq 1$. The most famous lattice problem is the *approximate shortest vector problem* (γ-SVP or SVP if $\gamma = 1$), which asks to find a non-zero lattice vector of norm less than $\gamma\lambda_1(L)$. A γ-SVP-oracle (or SVP-oracle when $\gamma = 1$) is an algorithm that takes a lattice L as input, and outputs a non-zero vector of L of norm less than $\gamma\lambda_1(L)$. Currently, the fastest known algorithms for worst-case SVP have runtime $2^{n+o(n)}$ ([ADS15; AS18]).

Another lattice problem that has recently achieved significant cryptographic interest is the *Lattice Isomorphism Problem* (LIP), and in particular its specialisation to rotations of $\mathbb{Z}^n$ ($\mathbb{Z}$LIP): given the image of $\mathbb{Z}^n$ under a linear orthogonal map (or rotation) $O \in \mathcal{O}_n(\mathbb{R})$, $\mathbb{Z}$LIP asks to recover O. It is not hard to see $\mathbb{Z}$LIP reduces to recovering unit vectors of the rotation, making $\mathbb{Z}$LIP at least as easy as SVP. Indeed, [BGPS23] and [Duc23] propose $2^{n/2+o(n)}$ algorithms for $\mathbb{Z}$LIP.

Two lattices Λ_1, Λ_2 such that there exists $O \in \mathcal{O}_n(\mathbb{R})$ for which $\Lambda_1 = O \cdot \Lambda_2$ are called *isomorphic*. We call *hypercubic* any lattice of $\mathbb{R}^n$ that has a $\mathbb{Z}$-basis consisting of unit vectors which are pairwise orthogonal. Full rank hypercubic lattices of $\mathbb{R}^n$ are exactly lattices that are isomorphic to $\mathbb{Z}^n$. In addition, a hypercubic lattice Λ is unimodular: $\Lambda = \Lambda^\vee$.

Lattice reduction

The celebrated LLL algorithm [LLL82] solves 2^n-SVP in polynomial time. Blockwise algorithms such as BKZ [SE94; CN11] and its variants [GN08a; MW16; ALNS20] approximate SVP within better factors, using polynomially many calls to an exact (or near-exact) SVP oracle in rank less than an input parameter called the *blocksize*. Following [GN08a; MW16; ALNS20], we call γ-SVP-reduction any algorithm which outputs a basis whose first vector solves γ-SVP. Similarly, we call γ-DVSP-reduction (where D stands for dual) any algorithm which outputs a basis whose last Gram-Schmidt vector solves γ-SVP in the dual lattice. Given a γ-SVP-oracle, it is possible to γ-SVP-reduce or γ-DSVP-reduce in polynomial time (see [GN08b; MW16]).

Reduced bases

Lattice reduction algorithms aim to transform an input basis into a "high quality" basis. There are many ways to quantify the quality of bases produced by lattice reduction algorithms. One popular way is to consider the Gram-Schmidt norms $\|\mathbf{b}_1^\star\|, \ldots, \|\mathbf{b}_n^\star\|$. Intuitively speaking, a good basis is one in which this sequence does not decay too fast. In practice, it turns out that the Gram-Schmidt coefficients of bases produced by the main reduction algorithms (such as LLL or BKZ) have a certain "typical shape", assuming the input basis is sufficiently random. This property was thoroughly investigated in [GN08b; NS06]. This typical shape is often used to estimate the running time of various algorithms. In particular, many theoretical asymptotic analyses (as introduced by Schnorr [Sch03]) assume for simplicity that this shape is given by $\|\mathbf{b}_i^\star\|/\|\mathbf{b}_{i+1}^\star\| = q$ where q depends on the reduction algorithm; although less precise, this approximation called the *geometric series assumption (GSA)* is close to the shape observed in practice. It is heuristically[1] estimated [CN11; Che13; LN24] that the BKZ algorithm with blocksize β, given as input a basis of an n-rank lattice L outputs a basis whose first vector has norm approximately equal to $\delta_\beta^n \mathrm{vol}(L)^{1/n}$, where $\delta_\beta = \left(\frac{\beta}{2\pi e}(\pi\beta)^{1/\beta}\right)^{\frac{1}{2(\beta-1)}}$. Combining this with the GSA and the fact that $\mathrm{vol}(L) = \prod_{i=1}^n \|\mathbf{b}_i^\star\|$ gives estimates of the Gram-Schmidt norms: for $1 \leq i \leq n$,

$$\|\mathbf{b}_i^\star\| \approx \delta_\beta^{n-\frac{2n}{n-1}(i-1)} \mathrm{vol}(L)^{1/n}.$$

Such a heuristic model is widely used in security estimates of lattice-based schemes, see the survey of Albrecht and Ducas [AD21] for more details.

[1] By replacing Hermite's constant by a Gaussian heuristic estimate.

The primal attack

Parameters of lattice-based cryptosystems are chosen after careful study of known attacks. One of the most important attack that people consider today (with the rise of dual attacks [MAT22; CMST22], it is now unclear which attack is the most efficient, although all record computations still use the primal) is called the *primal attack*, which runs the BKZ blockwise reduction [SE94; CN11] with a sufficiently high blocksize. Building upon [GN08b; CN11], the authors of [ADPS16] proposed to heuristically estimate the blocksize required by this attack to recover a short vector $\mathbf{s}$ in a rank n lattice L, by comparing the expected norm of $\pi_{n-\beta+1}(\mathbf{s})$ to the expected value of $\|\mathbf{b}^\star_{n-\beta+1}\|$. Using the GSA, as soon as

$$\sqrt{\frac{\beta}{n}}\|\mathbf{s}\| < \delta_\beta^{2\beta-n-1}\mathrm{vol}(L)^{1/n} \tag{3.2}$$

holds, the projection $\pi_{n-\beta+1}(\mathbf{s})$ is either $\mathbf{0}$ and then $\mathbf{s}$ lives in the subspace generated by the first $n-\beta$ vectors of the reduced basis, or it has a high chance of being shorter than $\|\mathbf{b}^\star_{n-\beta+1}\|$, making it such that the SVP oracle on the last block of size β will recover it. Albrecht, Göpfert, Virdia and Wunderer [AGVW17] and Dachman-Soled, Ducas, Gong, Rossi [DDGR20] refine and experimentally confirm this framework in the case of LWE. It should be stressed that the analysis of the primal attack remains very much heuristic.

The original NTRU cryptosystem

The NTRU cryptosystem [HPS98], proposed by Hoffstein, Pipher and Silverman, works in the ring $\mathcal{R} = \mathbb{Z}[X]/(X^n-1)$. An element $f = \sum_{i=0}^{n-1} f_i x^i = [f_0, f_1, \ldots, f_{n-1}] \in \mathcal{R}$ is seen as a polynomial or a row vector. To select keys, one uses the set $\mathcal{L}(d_1, d_2)$ of polynomials $F \in \mathcal{R}$ such that d_1 coefficients are equal to 1, d_2 coefficients are equal to -1, and the rest are zero. There are two small coprime moduli $p < q$, such as $q = 128$ and $p = 3$.

Historically, the secret keys were $f \in \mathcal{L}(d_f, d_f - 1)$ and $g \in \mathcal{L}(d_g, d_g)$ for some integers d_f and d_g significantly smaller than n, but other NTRU instantiations [HHHW09; Hof+17; Che+20] use different parameters for $\mathcal{L}$, such as binary polynomials $\mathcal{L}(d, 0)$. To illustrate, we focus on the NTRU-HPS parameters of NTRU's NIST submission [Che+20], one of the seven finalists: f is a random polynomial in $\{0, \pm 1\}^n$, and $g \in \mathcal{L}(d_g, d_g)$ where $2d_g = q/8 - 2$. With high probability, f is invertible mod q. The public key $h \in \mathcal{R}$ is defined as $h = g/f \mod q$. Thus, in the ring $\mathcal{R}/q\mathcal{R}$ which we represent by $\mathbb{Z}_q^n$, we have $f * h = g$. In this thesis, there is no need to know exactly how NTRU encryption or signature work. The polynomial h defines the so-called NTRU lattice Λ_h, formed by all pairs of polynomials $(u, v) \in \mathcal{R}^2$ such that $v * h \equiv u \mod q$. Here, we follow the definition of [How07], but other papers may use a variant of Λ_h, using a permutation of the coordinates. Λ_h is generated by the rows of the following lower-triangular matrix, which is its Hermite normal form:

$$\begin{pmatrix} qI_n & 0 \\ H & I_n \end{pmatrix},$$

where H is the circulant matrix for the polynomial $h \equiv g/f = \sum_{i=0}^{n-1} h_i x^i$:

$$H = \begin{pmatrix} h_0 & h_1 & \ldots & h_{n-1} \\ h_{n-1} & h_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_1 \\ h_1 & \ldots & h_{n-1} & h_0 \end{pmatrix}.$$

The lattice Λ_h contains by definition the following set of n secret short vectors

$$\mathcal{S}_h = \{(x^i * g, x^i * f), 0 \le i \le n-1\}$$

formed by the secret vector (g, f) and its $n-1$ rotations.

Coefficient embedding

Let $R = \mathbb{Z}[X]/P(X)$ be a polynomial ring, where P is a degree n monic irreducible polynomial. Elements of R are represented by their *coefficient embedding*:

$$\text{coef} : \begin{cases} R & \to \mathbb{Z}^n \\ a = \sum_{i=0}^{n-1} a_i X^i & \mapsto (a_0, \dots, a_{n-1}) \end{cases}.$$

When speaking of the shortness of $a \in R$, we mean the shortness of the corresponding vector $\text{coef}(a) \in \mathbb{Z}^n$. This allows us to naturally extend ℓ_p norms to R. More generally, we extend both norms to direct products of R by considering concatenations of the coefficient vectors. If R is also taken modulo an integer q, then ℓ_p norms are extended in the same way by taking the representative given by $\bmod^{\pm}$.

NTRU variants

Variants of the original NTRU [HPS98] choose to use different polynomial rings $\mathcal{R} = \mathbb{Z}[X]/P(X)$, for a unitary degree n polynomial $P \in \mathbb{Z}[X]$. Without giving an exhaustive list, examples of cryptosystems that use such variants include NTRU Prime [BCLvV16], NTRU+ [KP23], as well as the Falcon signature scheme [Fou+19]. In these cases, the public key $h \in \mathcal{R}/q\mathcal{R}$ is also defined as $h = g/f \mod q$, where $(f, g) \in \mathcal{R}^2$ is the secret key. In the most general case, the NTRU lattice is obtained by embedding the rank 2 $\mathcal{R}$-module that we call the *NTRU module*

$$M_h := \{(u, v) \in \mathcal{R}^2 : hu \equiv v \mod q\mathcal{R}\}$$

into $\mathbb{C}^{2n}$ via an embedding map $\sigma : \mathcal{R} \to \mathbb{C}^n$. The secret key is usually of small norm after embedding, that is $\|(\sigma(g), \sigma(f))\|$ is small. Most commonly, as is the case in the aforementioned cryptosystems, σ is simply the coefficient embedding. It has the advantage of being simple as it is easy to implement, as its image is integral. Other embeddings can also be of cryptanalytic interest. Most notably, the *canonical embedding* is obtained by evaluating a polynomial of $\mathcal{R}$ at all complex roots of P. This embedding is more complicated to deal with on a computer, but is a ring homomorphism and therefore behaves well with multiplication, which is usually not the case with the coefficient embedding. In particular if P is irreducible, then $\mathcal{R}$ is the ring of integers of a number field and the *canonical embedding* coincides with the *Minkowski embedding*.

Sampling short elements in polynomial rings

There are many ways to sample short elements in $\mathcal{R}$. While some schemes might sample each coefficients independently and with the same distribution, others like NTRU or DEFI fix a number of non-zero coordinates λ, which they choose uniformly at random, and uniformly sample all non-zero coordinates in a small set of values, such as $\{\pm 1, \pm 2\}$ in DEFI, we note D_u the distribution from which u is sampled. Another possibility is to sample coefficients following a centred binomial. The probability distribution Binom denotes the centred binomial of parameter $(2, 0.5)$ on $\{-1, 0, 1\}$ such that $\Pr(\text{Binom} = 0) = 1/2$ and $\Pr(\text{Binom} = -1) = \Pr(\text{Binom} = 1) = 1/4$. By extension, Binom^n denotes the probability distribution of a dimension n vector whose coefficients independently follow Binom. Applying coef^{-1} to Binom^n allows to sample short elements in $\mathcal{R}$. If $q \in \mathbb{Z}_{>0}$ is an integer, we will sometimes work in $\mathcal{R}_q := \mathcal{R}/(q\mathcal{R})$.

3.4 Lattice Signatures

We give the following simplified definition of a signature scheme.

Definition 3.4.1 (Signature Scheme)**.** A signature scheme is a triple of probabilistic algorithms (KeyGen, Sign, Verify) satisfying:

- KeyGen takes a security parameter 1^λ as input and returns a public-private key pair $(\mathtt{pk}, \mathtt{sk})$.
- Sign takes as input the secret (or signing) key $\mathtt{sk}$ and a message μ, and returns a signature σ.
- Verify takes the public (or verification) key $\mathtt{pk}$, message μ and signature σ as inputs, and outputs 1 is σ is a valid signature for μ, and 0 otherwise.

Security properties and lattice hardness assumptions

To prove that a signature is secure, we prove that no adversary can forge signatures even when choosing messages. This notion is called Unforgeability under Chosen Message Attacks or alternatively $\mathsf{UF\text{-}CMA}$.

Definition 3.4.2 (Unforgeability under chosen message attack ($\mathsf{UF\text{-}CMA}$))**.** We define a signature scheme $\mathcal{S} = (\mathsf{KeyGen}, \mathsf{Sign}, \mathsf{Verify})$ that uses a quantum random oracle H and $\mathcal{O}_{\mathsf{sign}}(\mathsf{sk})$ an oracle which on input m computes $\sigma \leftarrow \mathsf{Sign}(\mathsf{sk}, m)$ and returns (m, σ).

We define the advantage $\mathsf{Adv}^{\mathsf{ufcma}}_{\mathcal{S},\mathsf{q_H},\mathsf{q_S}}(\mathcal{A})$ of a quantum adversary that uses at most $\mathsf{q_H}$ quantum queries to H and $\mathsf{q_S}$ signatures queries to $\mathcal{O}_{\mathsf{sign}}(\mathsf{sk})$ against the $\mathsf{UF\text{-}CMA}$ security game as:

$$\mathsf{Adv}^{\mathsf{ufcma}}_{\mathcal{S},\mathsf{q_H},\mathsf{q_S}}(\mathcal{A},\mathcal{S}) = \Pr\left[\begin{array}{ll} \mathsf{Verify}(\mathsf{vk}, m, \sigma) = 1 & (\mathsf{sk},\mathsf{vk}) \leftarrow \mathsf{KeyGen}() \\ \wedge (m,\sigma) \text{ not given by } \mathcal{O}_{\mathsf{sign}}(\mathsf{sk}) & : \ \wedge\ (m,\sigma) \leftarrow \mathcal{A}^{\mathcal{O}_{\mathsf{sign}}(\mathsf{vk}),|\mathsf{H}\rangle} \end{array}\right].$$

While proving $\mathsf{UF\text{-}CMA}$ is fairly standard for FSwA signatures in the Random Oracle Model (ROM), its proof counterpart in the Quantum Random Oracle Model (QROM) is not trivial to obtain. This gave birth to a nice line of work [KLS18; DFPS23; Bar+23] to prove $\mathsf{UF\text{-}CMA}$ with QROM specifically for signatures based in the FSwA paradigm. For this, they prove that an alternative version of $\mathsf{UF\text{-}CMA}$ for the specific case of No Message Attack ($\mathsf{UF\text{-}NMA}$) in the QROM implies $\mathsf{UF\text{-}CMA}$ in the QROM.

Definition 3.4.3 (Unforgeability under no message attack ($\mathsf{UF\text{-}NMA}$))**.** We define a signature scheme $\mathcal{S} = (\mathsf{KeyGen}, \mathsf{Sign}, \mathsf{Verify})$ that uses a quantum random oracle H.

We define the advantage $\mathsf{Adv}^{\mathsf{ufnma}}_{\mathcal{S},\mathsf{q_H}}(\mathcal{A})$ of a quantum adversary that uses at most $\mathsf{q_H}$ quantum queries to H against the $\mathsf{UF\text{-}NMA}$ security game as:

$$\mathsf{Adv}^{\mathsf{ufcma}}_{\mathcal{S},\mathsf{q_H},\mathsf{q_S}}(\mathcal{A},\mathcal{S}) = \Pr\left[\mathsf{Verify}(\mathsf{vk}, m, \sigma) = 1 \ : \ \begin{array}{l} (\mathsf{sk},\mathsf{vk}) \leftarrow \mathsf{KeyGen}() \\ \wedge\ (m,\sigma) \leftarrow \mathcal{A}^{|\mathsf{H}\rangle} \end{array}\right].$$

Both $\mathsf{UF\text{-}CMA}$ and $\mathsf{UF\text{-}NMA}$ are proven accordingly to the signature structure and more specifically on underlying hardness assumptions given the signature design.

We define the well-studied lattice hardness assumptions MLWE (Module-Learning with Errors) and MSIS (Module-Short Integer Solutions). For the sake of readability, we omit the explicit mention of the modulus q and the dimension $n = 256$ associated with $\mathcal{R}_q = \mathbb{Z}_q[X]/(X^n + 1)$ when defining the parameters of the problems. We refer to elements of $\mathcal{R}_q^k$ as elements of the $\mathcal{R}_q$-module of rank k and consider them with their embedding in $\mathbb{Z}_q^{k\times n}$. For $m, k \in \mathbb{Z}_{>0}$ and a distribution χ with $\mathsf{Supp}(\chi) \subseteq \mathcal{R}_q$, we define the distribution $\mathcal{D}^{\mathsf{mlwe}}_{m,k,\chi}$ on $\mathcal{R}_q^{m\times k} \times \mathcal{R}_q^m$ as follows:

$$(\mathbf{A}, \mathbf{b}) \hookleftarrow \mathcal{D}^{\mathsf{mlwe}}_{m,k,\chi} \ \Leftrightarrow\ \mathbf{A} \hookleftarrow \mathcal{R}_q^{m\times k}, \mathbf{b} = \mathbf{A}\mathbf{s} + \mathbf{e} \text{ for } (\mathbf{s}, \mathbf{e}) \hookleftarrow \chi^{k+m}.$$

Definition 3.4.4 (Decisional MLWE problem)**.** Given a set of parameters $m, k \in \mathbb{Z}_{>0}$ and a distribution χ with $\mathsf{Supp}(\chi) \subseteq \mathcal{R}_q$, the advantage $\mathsf{Adv}^{\mathsf{d\text{-}mlwe}}_{m,k,\chi}(\mathcal{A})$ of any probabilistic polynomial time algorithm $\mathcal{A}$ in solving the decisional $\mathsf{d\text{-}MLWE}$ problem over $\mathcal{R}_q$ is:

$$\left|\Pr[\mathcal{A}(\mathbf{A}, \mathbf{b}) = 1 : (\mathbf{A}, \mathbf{b}) \leftarrow \mathcal{D}^{\mathsf{mlwe}}_{m,k,\chi}] - \Pr[\mathcal{A}(\mathbf{A}, \mathbf{b}) = 1 : (\mathbf{A}, \mathbf{b}) \hookleftarrow \mathcal{R}_q^{m\times k} \times \mathcal{R}_q^m]\right|.$$

Definition 3.4.5 (Search MLWE problem)**.** Given a set of parameters $m, k \in \mathbb{Z}_{>0}$ and a distribution χ with $\mathsf{Supp}(\chi) \subseteq \mathcal{R}_q$, the advantage $\mathsf{Adv}^{\mathsf{s-mlwe}}_{m,k,\chi}(\mathcal{A})$ of any probabilistic polynomial time algorithm $\mathcal{A}$ in solving the search s-MLWE problem over $\mathcal{R}_q$ is:

$$\Pr[\mathcal{A}(\mathbf{A}, \mathbf{A}\cdot\mathbf{s}+\mathbf{e}) = \mathbf{s} : \mathbf{s} \hookleftarrow\$\ \chi^k, \mathbf{A} \hookleftarrow\$\ \mathcal{R}_q^{m\times k}, \mathbf{e} \hookleftarrow\$\ \chi^m].$$

Definition 3.4.6 (MSIS problem)**.** Given a set of parameters $l, k \in \mathbb{Z}_{>0}$ and $\beta > 0$, the advantage $\mathsf{Adv}^{\mathsf{msis}}_{l,k,\beta}(\mathcal{A})$ of any probabilistic polynomial algorithm $\mathcal{A}$ in solving the MSIS problem over $\mathcal{R}_q$ is:

$$\Pr\left[(\ \mathbf{A} \mid \mathbf{I}_k)\,\mathbf{y} = \mathbf{0}\ \ \wedge\ \ 0 < \|\mathbf{y}\|_\infty < \beta : \mathbf{A} \leftarrow \mathcal{R}_q^{k\times l}\ \ \wedge\ \ \mathbf{y} \in \mathcal{R}_q^{k+l} \leftarrow \mathcal{A}(\mathbf{A})\right].$$

3.5 Number Theory

We use some notions from algebraic number theory. The reader can refer to [ST16; Coh93; Neu99a] for a more complete introduction. For material relating to elliptic curves, we refer to [Hus04; Sil09].

Number fields

A number field K is a finite extension of the rationals $\mathbb{Q}$. The degree $\deg K$ of K is the dimension of the extension. The ring of algebraic integers contained in K is called the ring of integers of K and denoted $\mathcal{O}_K$. Given $\alpha \in K$, the algebraic trace and norm of α are the trace and determinant of the multiplication by α endomorphism $x \mapsto \alpha x$ of K, seen as a $\mathbb{Q}$-vector space. The trace is denoted $\mathrm{Tr}(\alpha)$ and the norm $\mathcal{N}(\alpha)$. For any $\mathbb{Z}$-basis $(\omega_1, \ldots, \omega_n)$ of $\mathcal{O}_K$, the value of $\det(\mathrm{Tr}(\omega_i\omega_j))_{i,j}$ is the same, and is called the discriminant of K, which we denote Δ_K. Loosely speaking, $|\Delta_K|$ measures the size of the number field. Units in $\mathcal{O}_K$ are the group $\mathcal{O}_K^\times$ of elements of $\mathcal{O}_K$ that have algebraic norm 1. An order in K is a subring $\mathcal{O} \subseteq \mathcal{O}_K$ that contains a $\mathbb{Q}$-basis of K. The discriminant $D(\mathcal{O})$ of an order can be defined through a basis of $\mathcal{O}$, in the same way as for $D(\mathcal{O}_K) := \Delta_K$. The ring of integers is also called the maximal order of K.

Ideals

Let $\mathcal{O}$ denote an order of K. An integral ideal $I \subseteq \mathcal{O}$ is an additive subgroup of $\mathcal{O}$, such that for any $r \in \mathcal{O}$ and $x \in I$, $rx \in I$. An ideal $\mathfrak{p}$ of $\mathcal{O}$ is prime if it is not $\mathcal{O}$ and if for any $a, b \in \mathcal{O}$ such that $ab \in \mathfrak{p}$, $a \in \mathfrak{p}$ or $b \in \mathfrak{p}$. The norm of an ideal is $\mathcal{N}(\mathfrak{a}) = \mathrm{card}(\mathcal{O}/\mathfrak{a})$. A fractional ideal of $\mathcal{O}$ is a set $J \subset K$ such that there exists a non-zero $r \in \mathcal{O}$ such that rJ is an integral ideal of $\mathcal{O}$. In this case we define its norm by $\mathcal{N}(J) = \frac{\mathcal{N}(rJ)}{|\mathcal{N}(r)|}$, which generalises the norm for integral ideals.

The set $\mathcal{I}_\mathcal{O}$ of invertible fractional ideals of $\mathcal{O}$ is an abelian group, with identity $\mathcal{O}$. A fractional ideal is principal if it can be generated by a single element. In this case we denote by (α) the ideal generated by the element $\alpha \in K$. If $u \in \mathcal{O}^\times$, the ideal (u) is just $\mathcal{O}$. Because $\mathcal{O}_K$ is what is called a Dedekind domain, any fractional ideal $\mathfrak{a} \in \mathcal{I}_{\mathcal{O}_K}$ can be factored uniquely up to reordering into

$$\mathfrak{a} = (\mathfrak{p}_1 \cdots \mathfrak{p}_r)(\mathfrak{q}_1 \cdots \mathfrak{q}_s)^{-1},$$

where $\mathfrak{p}_1, \ldots, \mathfrak{p}_r, \mathfrak{q}_1, \ldots, \mathfrak{q}_s$ are prime (integral) ideals of $\mathcal{O}_K$.

Class group

The class group $\mathrm{Cl}(\mathcal{O}) = \mathcal{I}_\mathcal{O}/\mathcal{P}_\mathcal{O}$ is the quotient of the group of invertible fractional ideals of $\mathcal{O}$ and the subgroup of principal fractional ideals of $\mathcal{O}$. It is a finite group, of cardinality the class number $h(\mathcal{O})$. Two (fractional) ideals $\mathfrak{a}$ and $\mathfrak{b}$ are equivalent if they have the same class $[\mathfrak{a}] = [\mathfrak{b}]$. If $\mathcal{O}$ is the ring of integers, we will use $\mathrm{Cl}(K)$ and h_K instead of $\mathrm{Cl}(\mathcal{O}_K)$ and $h(\mathcal{O}_K)$.

Quadratic fields

Number fields that have degree 2 are called quadratic fields, and are of the form $K = \mathbb{Q}(\sqrt{D})$ for a squarefree integer $D \in \mathbb{Z}_{\neq 0}$. If $D > 0$, then K is a real quadratic field, and otherwise K is an imaginary quadratic field. The ring of integers and discriminant of K depend on the congruence of D mod 4. If $D \equiv 2, 3 \pmod 4$, then $\mathcal{O}_K = \mathbb{Z}[\sqrt{D}]$ and $\Delta_K = 4D$, and otherwise if $D \equiv 1 \pmod 4$, then $\mathcal{O}_K = \mathbb{Z}\left[\frac{1+\sqrt{D}}{2}\right]$ and $\Delta_K = D$.

Cyclotomic fields

Cyclotomic fields are number fields of the form $K_n = \mathbb{Q}(\zeta_n)$, where $\zeta_n = \exp(2i\pi/n)$ is a primitive root of unity. The n-th cyclotomic field (defined by $K_n = \mathbb{Q}(\zeta_n)$) has degree $m = \varphi(n)$ and ring of integers $\mathcal{O}_{K_n} = \mathbb{Z}[\zeta_n]$. For $n > 2$, the discriminant of K_n is $\Delta_{K_n} = (-1)^{m/2} \frac{n^m}{\prod_{p|n} p^{m/(p-1)}}$.
Cyclotomic fields are one of the most studied families of number fields, and the most convenient for cryptography. K_n can be viewed as the polynomial quotient of $\mathbb{Q}(X)$ with the n-th cyclotomic polynomial. For a more complete introduction, see [Was97].

3.5.1 Ideal Lattices

Minkowski embedding

A number field K of degree $n = [K : \mathbb{Q}]$ has r_1 real embeddings and r_2 pairs of complex embeddings, where $n = r_1 + 2r_2$. They correspond to all field homomorphisms from K into $\mathbb{C}$. We let $K_\mathbb{R} = K \otimes \mathbb{R} \cong \mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$. The n embeddings together form the *canonical* or *Minkowski* embedding, defined by

$$\sigma : \begin{cases} K & \to K_\mathbb{R} \\ \alpha & \mapsto (\sigma_i(\alpha))_{i \in [n]} \end{cases},$$

where the σ_i are all real embedding. Note that we are making a slight abuse of notation resolved by the fact that pairs of complex embeddings can be seen as mapping to $\mathbb{R}^2$. The canonical embedding map σ is multiplicative in the sense that $\sigma(\alpha\beta)$ is the coordinate-wise product of $\sigma(\alpha)$ and $\sigma(\beta)$. It sends ideals of K to lattices of $\mathbb{R}^n$. In particular for a fractional ideal[2] $\mathfrak{a}$ of $\mathcal{O}_K$ we have

$$\mathrm{vol}(\sigma(\mathfrak{a})) = \mathcal{N}(\mathfrak{a})\sqrt{|\Delta_K|}.$$

While an element $\alpha \in K$ can be measured using its algebraic norm $\mathcal{N}(\alpha) = \prod_i \sigma_i(\alpha)$, it can also be measured through the Euclidean norm of its embedding $\|\sigma(\alpha)\|$. The inequality between the quadratic mean and geometric mean of the coordinates of $\sigma(\alpha)$ gives

$$\|\sigma(\alpha)\| \geq \sqrt{n} \cdot |\mathcal{N}(\alpha)|^{1/n},$$

which in turn leads to

$$\lambda_1(\sigma(\mathfrak{a})) \geq \sqrt{n} \cdot \mathcal{N}(\mathfrak{a})^{1/n}$$

if $\alpha \in \mathfrak{a}$. This condition constrains the size of the shortest vectors in lattices obtained by embedding ideals. Note that an element $\alpha \in K$ is sent through σ to an element on a hyperbola in $\mathbb{R}^n$, as the product of the coordinates of $\sigma(\alpha)$ is exactly $\mathcal{N}(\alpha)$. In particular none of the coordinates can be zero if $\alpha \neq 0$.

Remark 3.5.1. In general, the Minkowski and coefficient embedding do not agree, except in the case of power of 2 cyclotomic fields, which is the most frequently used in cryptography.

[2] We only consider proper ideals that lead to full-rank lattices.

The idèle class group

The complex embeddings defined above account for the infinite places of the number field K. Each embedding $\sigma_i : K \to \mathbb{C}$ defines an (Archimedean) absolute value $|\cdot|_{\sigma_i}$ defined by $|\alpha|_{\sigma_i} = |\sigma_i(\alpha)|$. The completion K_{σ_i} of K with respect to $|\cdot|_{\sigma_i}$ is either $\mathbb{R}$ if σ_i is a real embedding, or $\mathbb{C}$ otherwise. The infinite places are not the only valuations of K, and in order to account for all valuations (or all reasonable ways to measure elements of K), we must also consider finite places: if $\mathfrak{p}$ is a prime ideal of $\mathcal{O}_K$ dividing $p \in \mathbb{Z}_{>0}$, then we can define a (non-Archimedean) absolute value $|\cdot|_\mathfrak{p}$ defined by $|\alpha|_\mathfrak{p} = p^{-v_\mathfrak{p}(\alpha)}$ for $\alpha \in K$, where $v_\mathfrak{p}(\alpha)$ is the largest power of $\mathfrak{p}$ that divides the principal ideal (α). The completion of K with respect to $|\cdot|_\mathfrak{p}$ is the $\mathfrak{p}$-adic extension $K_\mathfrak{p}$. For a place ν, we write $\nu|\infty$ if it is an infinite place, and $\nu \nmid \infty$ otherwise. In both cases we refer to the associated completion as K_ν. Completions at finite places have their own rings of integers called valuation rings $\mathcal{O}_\nu := \{\alpha \in K_\nu : |\alpha|_\nu \leq 1\}$. Ostrowski's theorem says that any absolute value on K must be equivalent to one of those mentioned above. This enables us to introduce adèles and idèles, modern mathematical language for objects that are used to perform analysis on all completions K_ν simultaneously. If the reader is unfamiliar with this theory, we encourage them to follow along assuming $K = \mathbb{Q}$. Our language will differ slightly to the one in [dBDPW20], even though we will be considering the same objects. Our formalism will be closer to the one used in [DK22].

Definition 3.5.2 (Adèles)**.** The ring of adèles $\mathbb{A}_K$ of the number field K is defined via a restricted direct product of all K_ν

$$\mathbb{A}_K = \left\{ (x_\nu)_\nu \in \prod_\nu K_\nu : x_\nu \in \mathcal{O}_\nu \text{ for all but finitely many places } \nu \right\}.$$

As a subring of $\prod_\nu K_\nu$, the adèle ring is a locally compact topological ring.

K injects diagonally into $\mathbb{A}_K$: if $\alpha \in K$, then $(\alpha, \alpha, \ldots) \in \mathbb{A}_K$. The units $\mathbb{A}_K^\times$ are not a topological group since inversion would not be continuous at $x = (1, 1, \ldots)$, so we need to change the topology.

Definition 3.5.3 (Idèles)**.** The group of idèles $\mathbb{I}_K$ of K is the group of units of $\mathbb{A}_K$, equipped with the topology obtained by seeing elements $x \in \mathbb{I}_K$ as pairs $(x, x^{-1}) \in \mathbb{A}_K^2$.

For the topology described above, the idèles $\mathbb{I}_K$ are equipped with the following canonical Haar measure.

Definition 3.5.4 (Tamagawa measure)**.** The Tamagawa measure $d^\times x$ is defined via a product of local measures at each finite place, normalised as follows:

$$d^\times x = \prod_{\nu \nmid \infty} (1 - \mathcal{N}(\nu)^{-1})^{-1} \frac{dx_\nu}{|x_\nu|_\nu} \prod_{\nu|\infty} \frac{dx_\nu}{|x_\nu|_\nu},$$

where dx_ν is the Haar measure on K_ν normalised such that the measure of $\mathcal{O}_\nu$ is 1 at finite places, and dx_ν is exactly the Lebesgue measure at real infinite places, and twice the infinite measure at complex infinite places.

Remark 3.5.5. The normalisation by $(1 - \mathcal{N}(\mathfrak{p})^{-1})^{-1}$ at finite places $\nu = \mathfrak{p}$ is chosen such that the multiplicative Haar measure $|x_\mathfrak{p}|_\mathfrak{p}^{-1} dx_\mathfrak{p}$ on $K_\mathfrak{p}^\times$ normalises to 1 on the local units $\mathcal{O}_\mathfrak{p}^\times$. Indeed on $\mathcal{O}_\mathfrak{p}^\times$, we have $|x_\mathfrak{p}|_\mathfrak{p} = 1$ so the multiplicative measure is the same as the additive measure $dx_\mathfrak{p}$. Notice that $\mathcal{O}_\mathfrak{p}$ is the disjoint union of its units $\mathcal{O}_\mathfrak{p}^\times$ and maximal ideal $\mathfrak{p}$, we conclude from the normalisation of $dx_\mathfrak{p}$ that $\mathcal{O}_\mathfrak{p}^\times$ has multiplicative measure 1 minus that of $\mathfrak{p}$, which is $\mathcal{N}(\mathfrak{p})^{-1}$.

$K^\times$ injects into $\mathbb{I}_K$, forming the discrete subgroup of principal idèles in $\mathbb{I}_K$. This defines the idèle class group $\mathbb{I}_K/K^\times$.

By forgetting about the infinite places, idèles map surjectively to the set $I_{\mathcal{O}_K}$ of fractional ideals of $\mathcal{O}_K$: an idèle $(x_\nu)_\nu \in \mathbb{I}_K$ is mapped to the ideal $\prod_\mathfrak{p} \mathfrak{p}^{v_\mathfrak{p}(x_\mathfrak{p})}$. If we call ϕ this map, the kernel of ϕ is exactly the set of idèles classes whose associated ideal is principal. If $x \in \mathbb{I}_K$ is such a representative, there exists $\alpha \in K^\times$ such that $\phi(x) = (\alpha)$. Therefore up to multiplication by $\alpha^{-1} \in K^\times$, we deduce that there is a representative of the idèle class $[x]$ with units at all finite places. Therefore there is an exact sequence of the form

$$0 \to K_\mathbb{R}^\times/\mathcal{O}_K^\times \to \mathbb{I}_K/K^\times \to \mathrm{Cl}(K) \to 0, \tag{3.3}$$

where $K_\mathbb{R}^\times$ is meant component-wise. Dirichlet's unit theorem gives the structure of $\mathcal{O}_K^\times$.

Theorem 3.5.6 (Dirichlet's unit theorem)**.** *$\mathcal{O}_K^\times$ is finitely generated of rank $r_1 + r_2 - 1$, where r_1 is the number of real embeddings of K and r_2 the number of pairs of complex embeddings of K:*

$$\mathcal{O}_K^\times \cong \mu_K \times \mathbb{Z}^{r_1+r_2-1},$$

where μ_K is the set of roots of unity of K.

In order to define a Haar measure on the space of ideal lattices, we need to view ideal lattices as a locally compact topological group, however the idèle class group is not locally compact. Compactness can be achieved through imposing a norm condition.

Definition 3.5.7 (Idèle norm)**.** The idèle norm is the map $\|\cdot\|_K : \mathbb{I}_K \to \mathbb{R}_{>0}$ defined for $x = (x_\nu)_\nu \in \mathbb{I}_K$ by

$$\|x\|_K = \prod_\nu |x_\nu|_\nu.$$

The kernel of the idèle norm map defines the norm 1 idèles:

$$\mathbb{I}_K^1 = \{x \in \mathbb{I}_K : \|x\|_K = 1\}.$$

The product formula states that for any element $\alpha \in K^\times$ viewed as an idèle, $\|\alpha\|_K = 1$. This can easily be checked in the case of $K = \mathbb{Q}$. Because the product formula holds, $K^\times$ embeds in $\mathbb{I}_K^1$ as the subgroup of principal idèles.

Definition 3.5.8 (Arakelov class group)**.** The Arakelov class group $\tilde{\mathrm{Cl}}(K)$ of K is defined as the quotient of the norm 1 idèles by the principal idèles:

$$\tilde{\mathrm{Cl}}(K) = \mathbb{I}_K^1/K^\times.$$

The Arakelov class group fits in the following exact sequence, a variant of Equation 3.3.

$$0 \to K_\mathbb{R}^{(1)}/\mathcal{O}_K^\times \to \tilde{\mathrm{Cl}}(K) \to \mathrm{Cl}(K) \to 0, \tag{3.4}$$

where $K_\mathbb{R}^{(1)} = \{a \in K_\mathbb{R} : |\mathcal{N}(a)| = 1\}$, where $\mathcal{N}$ is naturally extended to $K_\mathbb{R}$. From Equation 3.4 and Dirichlet's unit theorem, we can read that $\tilde{\mathrm{Cl}}(K)$ behaves as h_K copies of the compact torus $K_\mathbb{R}^{(1)}/\mathcal{O}_K^\times$, and therefore is compact.

The Tamagawa measure on $\mathbb{I}_K = \mathbb{I}_K^1 \times \mathbb{R}_{>0}$ decomposes through the idèle norm map as $d^\times x = d\mu \times (d\beta/\beta)$ into Haar measures on the respective components. The measure $d\mu$ is the one we consider for $\mathbb{I}_K^1$, and by extension to $\mathbb{I}_K^1/K^\times$. For this measure $d\mu$, the measure of the norm one idèle class group evaluates to the residue at 1 of the Dedekind zeta function of the number field K:

$$\mu(\mathbb{I}_K^1/K^\times) = \mathrm{Res}_{s=1}\, \zeta_K(s). \tag{3.5}$$

The residue has a well-known expression as a function of number field constants, and can be computed through [BF14].

Random ideal lattices

The general definition of an ideal lattice relates to the Arakelov class group.

Definition 3.5.9 (Ideal Lattices)**.** We define the normalised space of ideal lattices to be the subsets of $K_{\mathbb{R}}$ of the form $\Lambda = x \cdot \mathfrak{a}$, where $x \in K_{\mathbb{R}}^{\times}$ and $\mathfrak{a}$ is a fractional ideal of K, identified up to isometries and rescaling by a real $c > 0$.

The normalised set of lattices can be identified with the Arakelov class group, essentially by taking x to be the infinite part of the idèle and $\mathfrak{a}$ to be its finite part. See [Sch08; dBoe22] for a more complete treatment. An idèle $g \in \mathbb{I}_K$ maps to an ideal lattice Λ_g with volume

$$\mathrm{vol}(\Lambda_g) = \|g\|_K \cdot \sqrt{|\Delta_K|}. \tag{3.6}$$

The Arakelov class group was used in [dBDPW20] to give a natural definition of a "random ideal lattice of fixed volume", and then prove random self-reducibility results for SVP on ideal lattices. Indeed by compactness, the uniform (Haar) measure on $\tilde{\mathrm{Cl}}(K)$ exists and defines the correct distribution for random ideal lattices. It is possible to approximate this distribution by first sampling a prime ideal uniformly among all prime ideals of bounded norm, and then perturbing the result with a small element of $K_{\mathbb{R}}^{(1)}$. After rescaling and for well-chosen parameters, this gives a distribution that is statistically close to the desired Haar measure under the extended Riemann hypothesis. We refer to [FPS22, Algorithm 2.1], and [dBDPW20, Theorem 3.3] for further details. Because of Equation 3.6, for the normalisation of ideal lattices to make sense with μ from Equation 3.5, we must consider ideal lattices normalised to have covolume $\sqrt{|\Delta_K|}$. In cases where we prefer the covolume to be normalised differently, this is a minor issue that can be dealt with by appropriately rescaling both the measures and lattices. This $\sqrt{|\Delta_K|}$ is the same one that appears in the formula for $\mathrm{vol}(\sigma(I))$.

3.5.2 Elliptic Curves and Isogeny Graphs

Let k be either a finite field or a number field (which we suppose to be embedded in $\mathbb{C}$). An elliptic curve E over k (shorthand E/k) is a smooth projective curve of genus one with a specified base point. Its set of points $E(k)$ forms an abelian group under a geometric law. The j-invariant of an elliptic curve E/k is a fundamental invariant $j(E) \in k$ defined from the equation for E. Two elliptic curves over the algebraic closure $\overline{k}$ are isomorphic if and only if they have the same j-invariant. For E_1, E_2 two elliptic curves, a morphism $\varphi : E_1 \to E_2$ that is a group homomorphism is called an isogeny. The degree of a separable isogeny φ is the size of its kernel. It admits a unique dual isogeny $\hat{\varphi} : E_2 \to E_1$ whose degree is the same and such that $\hat{\varphi} \circ \varphi = [\deg \varphi]$ is multiplication by $\deg \varphi$. By ℓ-isogeny between two elliptic curves over k we mean an isogeny of degree ℓ defined over $\overline{k}$. The set of isogenies over $\overline{k}$ from the elliptic curve E to itself is the *endomorphism* ring $\mathrm{End}_{\overline{k}}(E)$, and it always contains $\mathbb{Z}$ through multiplication maps. An elliptic curve E/k is said to have complex multiplication (CM) by a ring $\mathcal{O}$ if $\mathrm{End}_{\overline{k}}(E) \cong \mathcal{O}$ and $\mathcal{O}$ strictly contains $\mathbb{Z}$. The structure of these rings is well classified. An elliptic curve over a finite field is *ordinary* if $\mathcal{O}$ has rank 2 over $\mathbb{Z}$, and it is *supersingular* if $\mathcal{O}$ has rank 4 over $\mathbb{Z}$. These are the only two possible cases, see for instance [Sil09, Chapter III, Corollary 6.4 and Chapter V, Theorem 3.1.]. In the ordinary case, which will be the main focus of Chapter 9, the endomorphism ring $\mathcal{O}$ is an order in an imaginary quadratic field K. This means that $\mathcal{O}$ is of the form $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$ where $\mathcal{O}_K$ is the ring of integers of K and $f = |\mathcal{O}_K : \mathcal{O}|$ is an integer called the *conductor* of $\mathcal{O}$. Every imaginary quadratic order is uniquely determined by its discriminant $D(\mathcal{O}) := f^2\Delta_K$ where Δ_K is the discriminant of the ring of integers of K.

Proposition 3.5.10. *Let $\varphi : E_1 \to E_2$ be an ℓ-isogeny of ordinary elliptic curves over k with geometric endomorphism rings $\mathcal{O}_1$ and $\mathcal{O}_2$. If $\mathrm{char}(k) \neq \ell$ then either $\mathcal{O}_1 = \mathcal{O}_2$, or one is included in the other with index ℓ.*

Proof. See [Koh96, Proposition 21] or [Sut13, §2.7]. □

Definition 3.5.11. Using the notations of Proposition 3.5.10, we say that φ is *horizontal* if $\mathcal{O}_1 = \mathcal{O}_2$. Otherwise φ is *vertical*; *descending* if $[\mathcal{O}_1 : \mathcal{O}_2] = \ell$, *ascending* if $[\mathcal{O}_2 : \mathcal{O}_1] = \ell$.

Remark 3.5.12. The dual of a horizontal isogeny is horizontal, and the dual of a vertical isogeny is vertical, ascending if the initial isogeny was descending, and vice versa.

Definition 3.5.13. Let $\mathcal{O}$ be an imaginary quadratic order and let E/k be an elliptic curve with $\mathrm{End}_{\overline{k}}(E) \cong \mathcal{O}$. For every $\mathcal{O}$-ideal $\mathfrak{a}$ the $\mathfrak{a}$-*torsion subgroup* of E is defined as

$$E[\mathfrak{a}] := \bigcap_{\alpha \in \mathfrak{a}} \ker \alpha \subseteq E(\overline{k}).$$

The $\mathfrak{a}$-torsion subgroups correspond to isogenies $\varphi_{\mathfrak{a}} : E \to E/E[\mathfrak{a}]$ with kernel $E[\mathfrak{a}]$ and degree $\deg(\varphi_{\mathfrak{a}})$. Here, $E/E[\mathfrak{a}]$ denotes the quotient elliptic curve of E by the subgroup $E[\mathfrak{a}]$. If $\mathfrak{a}$ is an invertible ideal of $\mathrm{End}_{\overline{k}}(E)$, then $\mathrm{End}_{\overline{k}}(E/E[\mathfrak{a}]) \cong \mathrm{End}_{\overline{k}}(E)$. For the details we refer the reader to [Wat69, Proposition 3.9 and Theorem 4.5] when k is a finite field and to the proof of [Sil94, II, Proposition 1.2 (ii)] when k is a number field (note that this proof works also when the endomorphism ring of the considered elliptic curve is not a maximal order).

We thus obtain an action of the group of invertible $\mathcal{O}$-ideals on the set of $\overline{k}$-isomorphism classes of elliptic curves with complex multiplication by $\mathcal{O}$, given by

$$\mathfrak{a} * E := E/E[\mathfrak{a}].$$

This action factors via a faithful and transitive action of the class group $\mathrm{Cl}(\mathcal{O})$ of the order $\mathcal{O}$. If k is a number field, we can write $E(\mathbb{C}) \cong \mathbb{C}/\Lambda$ where $\Lambda \subseteq K := \mathrm{Frac}(\mathcal{O})$ is an invertible ideal, and then one has

$$(E/E[\mathfrak{a}])(\mathbb{C}) \cong \mathbb{C}/\mathfrak{a}^{-1}\Lambda.$$

We then see that the above action corresponds to the classical one over complex numbers described for instance in [Sil94, II, Proposition 1.2] for maximal orders. We summarise this discussion in the following theorem.

Theorem 3.5.14 (Main theorem of Complex Multiplication)**.** *Let K be an imaginary quadratic field, and let $\mathcal{O}$ be an order in K. Then, for every ideal class $[\mathfrak{a}] \in \mathrm{Cl}(\mathcal{O})$ and every elliptic curve $E/\overline{k}$ with $\mathrm{End}_{\overline{k}}(E) \cong \mathcal{O}$ the association*

$$\mathfrak{a} * E := E/E[\mathfrak{a}]$$

defines a simply transitive action of the class group on the set of $\overline{k}$-isomorphism classes of elliptic curves with complex multiplication by $\mathcal{O}$.

Proof. If k is a number field the statement is proved in [Lan87, Chapters 8 and 10]. If k is a finite field, the statement can be deduced from the previous case using Deuring's lifting theorem [Lan87, Chapter 13, Theorem 14]. □

In fact, Theorem 3.5.14 says that elliptic curves with CM are ideal lattices for potentially non-maximal orders, whereas general elliptic curves can be seen as lattices without this extra $\mathcal{O}$-module structure.

Definition 3.5.15. Let E and E' be elliptic curves over k. We say that two ℓ-isogenies $\varphi, \psi : E \to E'$ are equivalent if $\ker \varphi = \ker \psi$. This is the same as requiring that there exists an $\alpha \in \mathrm{Aut}_{\overline{k}}(E')$ such that $\alpha \circ \varphi = \psi$.

In Chapter 9, Definition 9.2.1, we will define the isogeny graph as a graph whose vertices are j-invariants, and directed edges are given by isogenies between curves with those j-invariants. We illustrate in the following example how there can exist multiple edges between two vertices of the graph.

Example 3.5.16. *This example is taken from [Sch95, page 238]. Consider the elliptic curve $E : y^2 = x^3 - 1$ defined over $\mathbb{F}_7$. This is an ordinary elliptic curve with $j(E) = 0$, whose CM order is generated over the integers by the automorphism $[\zeta_3] : (x, y) \mapsto (2x, y)$. The four cyclic subgroups of order 3 in $E(\overline{\mathbb{F}}_7)$ are given by*

$$C_0 = \langle (0, i) \rangle, \;\; C_1 = \langle (\sqrt[3]{4}, 2i) \rangle, \;\; C_2 = \langle (2\sqrt[3]{4}, 2i) \rangle, \;\; C_3 = \langle (4\sqrt[3]{4}, 2i) \rangle$$

where $i, \sqrt[3]{4} \in \overline{\mathbb{F}}_7$ are respectively a fixed square root of -1 and cube root of 4. Note that the automorphism $[\zeta_3]$ acts transitively on the subgroups C_i with $i = 1, 2, 3$. For this reason, $j(E/C_i) = 3$ is the same for all $i = 1, 2, 3$, leading to multiple edges between the same two vertices of the isogeny graph.

Lemma 3.5.17. *Let k be a finite field of characteristic p and let E/k be an elliptic curve with geometric endomorphism ring isomorphic to an order $\mathcal{O}$ in an imaginary quadratic field K. Denote by f the conductor of $\mathcal{O}$. Then p does not divide f.*

Proof. The proof is given in [Lan87, Chapter 13, Theorem 5 (ii)]. □

Corollary 3.5.18. *Let k be a finite field of characteristic p and let $j(E) \in k$ be the j-invariant an elliptic curve E/k with complex multiplication by an order $\mathcal{O}$ in the imaginary quadratic field K. If $h(\mathcal{O})$ denotes the class number of $\mathcal{O}$, then there are exactly $h(\mathcal{O})$ distinct elements of k that are the j-invariants of elliptic curves with complex multiplication by $\mathcal{O}$.*

Proof. See [BCP24, Corollary 2.8]. □

The j-invariants of the curves described in Corollary 3.5.18 are called singular moduli. For an order $\mathcal{O}$ with discriminant D, the Hilbert class polynomial $H_D(X)$ is defined as the monic polynomial whose roots are the $h(\mathcal{O})$ j-invariants of elliptic curves with CM by $\mathcal{O}$. It has integer coefficients and can be computed via [Coh93, Algorithm 7.6.1].

Part II

Attacks on NTRU and Hypercubic Lattices

Chapter

4 Heuristic Attacks on Near-Hypercubic Lattices

Abstract Lattice-based cryptography typically uses families of lattices with special properties to improve efficiency. NTRU and hypercubic lattices are two types of lattices that are used in some of the strongest candidate schemes for post-quantum standardisation. These lattices are special in at least two ways: their first minima are both unusually short compared with random lattices, and they both have a linear number of vectors that reach this first minimum, which again is very different to what is expected in the random case. In this chapter we revisit the heuristic analysis of what seems to be the best practical algorithm to recover short vectors in such lattices: the primal attack. While such results are already widely used to assess security of lattice-based schemes, our analysis differs slightly in that it aims to provide asymptotic results. We argue that the presence of linearly-many short vectors leads to no asymptotic gain. Finally, we present a new quasi-polynomial time algorithm that recovers unit vectors in a hypercubic lattice provided access to a basis of poly-logarithmic sized vectors. This new algorithm generalises to longer bases, but its efficiency decreases as the norm grows, in a way that is comparable to the pattern guessing attack of May and Silverman on the NTRU cryptosystem [MS01].

This chapter incorporates material that was published in the conference paper [BN24].

Chapter content

4.1 Introduction

Lattice-based cryptography has emerged as the main alternative to classical public key cryptography based on factoring and discrete logarithm: it can provide resistance to quantum computers and offer new functionalities such as fully-homomorphic encryption. However, for efficiency reasons, the lattices used in concrete cryptosystems are usually not random lattices: they have special properties, to improve keysize and/or speed up operations and/or enable extra operations. For instance, all the lattices used in NIST's new post-quantum standards are special: module

lattices for Kyber [Ava+19] and Dilithium [Duc+21], and NTRU lattices for Falcon [Fou+19]. Recently, even hypercubic lattices [Szy03], which are simply rotations of $\mathbb{Z}^n$, have been proposed in [DvW22; DPPvW22; BGPS23] as the basis of concrete cryptosystems, with HAWK [DPPvW22] being submitted to the new NIST call for post-quantum signatures.

Accordingly, it is crucial to understand if these special properties make the underlying lattice problems easier to solve, and if so, by how much. In the case of module lattices, this remains very much an open problem, except for the case of ideal lattices, for which better algorithms have been found [CDW17; PHS19; BR20].

In the primal attack [ADPS16], lattices are reduced by progressively reducing smaller lattices of dimension β. We call this parameter β the blocksize, and focus on asymptotic estimates of blocksizes required to reduce certain families of lattices. In the case of $\mathbb{Z}^n$, this was done by [DPPvW22], where the authors state without giving many details that a blocksize of $n/2+o(n)$ is heuristically sufficient to recover a shortest vector. We show more generally that for any n-dimensional lattice L such that $\lambda_1(L) = O(\mathrm{vol}(L)^{1/n})$, the primal attack heuristically recovers a shortest lattice vector using a blocksize $n/2 + \Theta(n/\ln n)$: this result applies to both $\mathbb{Z}^n$ and NTRU lattices. For these lattices, there are actually multiple shortest vectors, even a linear number: somewhat surprisingly, we show that the heuristic asymptotic blocksize required by the primal attack remains $n/2 + \Theta(n/\ln n)$, even though in practice, it is somewhat easier.

However, in the case of NTRU, there is a twist, due to the existence of q-vectors, which have a single non-zero coordinate (equal to q). If we ignore these q-vectors, then the heuristic asymptotic blocksize required by the primal attack is $n/2+\Theta(n/\ln n)$. But if we take into account the fact, for a given blocksize, these q-vectors yield better reduced bases, then the heuristic asymptotic blocksize required by the primal attack on NTRU is reduced to $4n/9 + \Theta(n/\ln n)$. This means that in the case of NTRU, unlike $\mathbb{Z}^n$, there is a noticeable difference between the best theoretical algorithm and the best heuristic algorithm (see Chapter 5 for details on the provable side). We note that Bernstein [Ber24] also proposed an asymptotic analysis of the heuristic primal attack on a q-ary lattice containing one very short vector: applied to NTRU, his analysis also gives a heuristic blocksize $4n/9$, but it does not take into account multiple short vectors.

Focusing exclusively on hypercubic lattices, we design a new algorithm that recovers a unit vector from the knowledge of approximations of the shortest vector. The analysis of this heuristic reduction from $\mathbb{Z}$SVP to γ-$\mathbb{Z}$SVP relies also on estimations for the primal attack. It runs in time polynomial in the dimension when γ is a constant, yet much more practical than the reduction of [Jia+23] (best paper at Asiacrypt 2023), and quasi-polynomial when $\gamma = \ln n$.

Technical overview

Our analysis of the primal attack relies on the same equation as the one introduced in [ADPS16], however it differs slightly from the literature [AGVW17; DDGR20]. In the primal attack, it is crucial to estimate the projection of a short vector onto random subspaces related to lattice reduction. Previous work [AGVW17; DDGR20] restricted to a short vector from LWE, whose coordinates are independent Gaussians. However, we argue that this model does not match $\mathbb{Z}^n$ nor NTRU. So instead of the χ^2 distribution, we rely on the Beta distribution related to classical sphere statistics. And we heuristically extend the analysis to the case of linearly many short vectors, by studying order statistics of the expected shortest projection of a secret vector onto a random subspace. Our asymptotic analysis shows that having linearly many short vectors does not substantially improve the primal attack. This means that it is reasonable to focus on a single short vector as a target, especially if it can decrease our lattice rank. In general specialising our search to a single target does not profit from the existence of many short vectors (as would be the case in Kannan embedding attacks on module-LWE). The following contribution is not in [BN24]. We show that in the specific case of $\mathbb{Z}^n$, if we can guess 0-coordinates of a unit vector, then by extracting an appropriate orthogonal sublattice, we restrain the search to a single target

in a heuristically easier-to-reduce lattice. This guessing strategy does not seem to be efficient in general, but when vectors of norm $\ln n$ are known, it allows for quasi-polynomial-time recovery of a unit vector.

Roadmap

In this chapter, we derive the asymptotic behaviour of the heuristic minimal blocksizes required to break lattice problems such as $\mathbb{Z}$LIP and NTRU (Section 4.2). In both cases, the quantity $\mathrm{vol}(L)^{1/n}/\lambda_1(L)$ is a constant, whereas we would expect it to be $\Theta(n^{-1/2})$ for a generic lattice. They nicely complement Section 5.2 from the following chapter, because they provide an opportunity to compare the best known provable and heuristic reduction algorithms for $\mathbb{Z}$LIP and NTRU. Those results using the primal attack approach could be considered folklore, but we think it profitable to write them down clearly. We then tweak the primal attack framework to incorporate the fact that special lattices like the hypercubic and NTRU lattice have not just one, but many shortest vectors (Section 4.3), and comment on the asymptotic and concrete impact on the blocksize. In Section 4.4, we present our new reduction from $\mathbb{Z}$SVP to γ-$\mathbb{Z}$SVP.

4.2 Primal Attack Asymptotics for a Single Short Vector

We recall that primal attack estimations use Equation 3.2. Under the heuristic assumption that this equation governs the success of the primal attack, we compute the asymptotic blocksizes required for the attack to succeed.

Proposition 4.2.1. *Let $c = \Theta(1)$ be a positive constant. If $\beta = \omega(1)$ satisfies the equation:*

$$\sqrt{\frac{\beta}{n}} = \delta_\beta^{2\beta-n-1}\sqrt{c},$$

then

$$\beta = \frac{n}{2} - \frac{\ln(2c)}{4}\frac{n}{\ln n} + o\left(\frac{n}{\ln n}\right).$$

Proof. All equivalents denote asymptotics as n goes to infinity. Let $0 < \beta < n$ be a solution to the equation for which $\beta = \omega(1)$. Because

$$\frac{\beta}{n} = \left(\frac{\beta}{2\pi e}(\pi\beta)^{\frac{1}{\beta}}\right)^{\frac{2\beta-n-1}{\beta-1}} c \sim \left(\frac{\beta}{2\pi e}\right)^{\frac{2\beta-n-1}{\beta-1}} c,$$

we obtain

$$\left(\frac{\beta}{n}\right)^{\beta-1} \sim \left(\frac{\beta}{2\pi e}\right)^{2\beta-n-1} c^{\beta-1}.$$

It is clear from $\beta = \omega(1)$ and the above expression that $\beta = n/2 + o(n)$. In what follows we write $\beta = (1/2-\varepsilon)n$, where $\varepsilon = o(1)$. We get

$$(1/2-\varepsilon)^{(1/2-\varepsilon)n-1} \sim \left(\frac{(1/2-\varepsilon)n}{2\pi e}\right)^{-2\varepsilon n-1} c^{(1/2-\varepsilon)n-1},$$

and by taking the log of the ratio, we must have

$$((1/2-\varepsilon)n-1)(\ln(1/2-\varepsilon) - \ln c) + (2\varepsilon n+1)\ln\left(\frac{(1/2-\varepsilon)n}{2\pi e}\right) \longrightarrow 0.$$

The dominating terms of the expression above are $2\varepsilon n\ln(n)$ and $-\frac{n}{2}\ln(2c)$ so they must cancel out, leaving us with $\varepsilon = \frac{\ln(2c)}{4\ln(n)} + o(\ln(n)^{-1})$.

□

Remark 4.2.2. We pay no concern to β having to be an integer. We choose to replace the inequality in Equation 3.2 by an equality as we are interested in the largest value of β such that the inequality still holds.

Corollary 4.2.3. *Let L be a rank n lattice and $\mathbf{s}$ a short vector of L for which $\|\mathbf{s}\|/\mathrm{vol}(L)^{1/n} =: c^{-1/2} = O(1)$. The primal attack framework heuristically predicts that applying BKZ with blocksize $\beta = n(1/2 - \ln(2c)/4\ln n + o(1/\ln n))$ recovers a vector of norm $\|\mathbf{s}\|$ or less with high probability. In particular, this condition holds for hypercubic and* NTRU *lattices.*

Proof. The main point follows directly from Proposition 4.2.1. In the case of hypercubic lattices, $\mathrm{vol}(L) = \|\mathbf{s}\| = 1$. For NTRU lattices, $\mathrm{vol}(L)^{1/n} = \sqrt{q} = \Theta(\sqrt{n})$, and $\|\mathbf{s}\| = \Theta(\sqrt{n})$, where $\mathbf{s} = (\mathbf{g}, \mathbf{f})$ is the secret key and q is the NTRU modulus.

□

The authors of the HAWK signature specifications [DPPvW22; Bos+23] use the primal attack to heuristically evaluate the security of their scheme. They obtain from Equation 3.2 that the optimal blocksize for secret key recovery is $n/2 + o(n)$. Corollary 4.2.3 helps with understanding the hidden contribution.

Remark 4.2.4. In the case of NTRU, the Gram-Schmidt norms after reduction behave differently to the GSA because of the presence of q-vectors. If the q-vectors are shorter than the predicted length of the first basis vectors achieved by BKZ, then projecting against those q-vectors will lead to an improved profile. This phenomenon is documented, e.g. in [DDGR20]. We study the asymptotic impact of this change in Proposition 4.2.5.

Proposition 4.2.5. *Let L be a rank n* NTRU *lattice, with modulus $q = \Theta(n)$, and secret vector $\mathbf{s}$ such that $\|\mathbf{s}\| = \Theta(\sqrt{n})$. Then the primal attack heuristic predicts that $\|\mathbf{s}\|$ can be recovered after running BKZ with blocksize $\beta = \frac{4}{9}n + o(n)$.*

Proof. Let B denote a BKZ-β reduced version of the Hermite Normal Form basis of Λ. We can apply GSA heuristics in two different ways:

- **Ignoring q-vectors:** In this model, the Gram-Schmidt norms $\|\mathbf{b}_1^\star\|, \ldots, \|\mathbf{b}_n^\star\|$ follow a geometric progression. Assuming usual heuristics on the quality of $\|\mathbf{b}_1\|$ after BKZ-β, we can assume that
$$\|\mathbf{b}_i^\star\| = \delta_\beta^{n - \frac{2n}{n-1}(i-1)} \mathrm{vol}(\Lambda)^{1/n},$$
for all $1 \leq i \leq n$, and where $\delta_\beta = \left(\frac{\beta}{2\pi e}(\beta\pi)^{1/\beta}\right)^{\frac{1}{2(\beta-1)}}$. In this case, the geometric factor is
$$\alpha_\beta := \frac{\|\mathbf{b}_{i+1}^\star\|}{\|\mathbf{b}_i^\star\|} = \delta_\beta^{-\frac{2n}{n-1}}.$$
In this model, the primal attack heuristic leads to a blocksize $\beta \approx \frac{n}{2}$, as studied previously in Corollary 4.2.3.

- **Using q-vectors:** In this second model, we choose to apply the usual GSA, but where we replace any Gram-Schmidt norm that is estimated to be larger than q by a q-vector. This changes the shape of the log-profile from a decreasing affine line to a horizontal line (representing the q-vectors), followed by an affine line with a slope equal to that of the previous GSA model. In practice, if we assume the horizontal line contains k q-vectors, then the volume loss in the first k coordinates results into an increase in the volume of the last $n - k$ coordinates, which might result in a substantial increase in the value of $\|\mathbf{b}_{n-\beta+1}^\star\|$. We denote by $\|\hat{\mathbf{b}}_i^\star\| = c \cdot \|\mathbf{b}_i^\star\|$ the value of the norm of the i-th Gram-Schmidt vector in this second model. According to the previous discussion, we expect $c > 1$.

We start by estimating the value of k in this model. The product of the Gram-Schmidt norms in the volume of Λ, hence

$$\mathrm{vol}(\Lambda) = q^k \prod_{i=k+1}^{n} \|\hat{\mathbf{b}}_i^\star\| = q^k \|\mathbf{b}_1^\star\|^{n-k} c^{n-k} \alpha_\beta^{k+\ldots+(n-1)}. \tag{4.1}$$

The second condition on k and c is that both lines should intersect, meaning that $q = \|\hat{\mathbf{b}}_k^\star\|$. This leads to

$$q = \|\hat{\mathbf{b}}_k^\star\| = c\|\mathbf{b}_k^\star\| = c\|\mathbf{b}_1^\star\|\alpha_\beta^{k-1}.$$

Therefore, $c\|\mathbf{b}_1^\star\| = q\alpha_\beta^{1-k}$. Recall that $\mathrm{vol}(\Lambda) = q^{n/2}$ and inject the previous expression into Equation 4.1:

$$\begin{aligned} q^{n/2} &= q^k (q\alpha_\beta^{1-k})^{n-k} \alpha_\beta^{k+\ldots+(n-1)} \\ &= q^n \alpha_\beta^{(1-k)(n-k)+\frac{n(n-1)}{2}-\frac{k(k-1)}{2}}. \end{aligned}$$

Reordering and taking the logarithm gives

$$\left((1-k)(n-k) + \frac{n(n-1)}{2} - \frac{k(k-1)}{2}\right) \ln \alpha_\beta + \frac{n}{2} \ln q = 0,$$

which simplifies down to

$$-\frac{2n}{n-1}\left(n - k - nk + k^2 + \frac{n^2}{2} - \frac{n}{2} - \frac{k^2}{2} + \frac{k}{2}\right) \ln \delta_\beta + \frac{n}{2} \ln q = 0,$$

and this can be rewritten as

$$k^2 + k(-2n-1) + n^2 + n - \frac{n-1}{2}\frac{\ln q}{\ln \delta_\beta} = 0.$$

Solving this equation in k gives

$$k = n + \frac{1}{2} - \frac{1}{2}\sqrt{1+4A}, \tag{4.2}$$

where $A = \frac{n-1}{2}\frac{\ln q}{\ln \delta_\beta}$. Using the expression for δ_β, $\beta = \gamma n$ and $q = \Theta(n)$, we obtain

$$\begin{aligned} 1 + 4A &= 1 + 4(n-1)(\beta-1)\frac{\ln q}{\ln\left(\frac{\beta}{2\pi e}(\pi\beta)^{1/\beta}\right)^{\frac{1}{2(\beta-1)}}} \\ &= 4\gamma n^2 - 4\gamma \ln\left(\frac{\gamma}{2\pi e}\right)\frac{n^2}{\ln n} + o\left(\frac{n^2}{\ln n}\right). \end{aligned}$$

Injecting this into Equation 4.2, we get

$$k = n(1-\sqrt{\gamma}) + \frac{1}{2}\sqrt{\gamma}\ln\left(\frac{\gamma}{2\pi e}\right)\frac{n}{\ln n} + o\left(\frac{n}{\ln n}\right).$$

Now that we know k, we can deduce the value of $\|\hat{\mathbf{b}}_{n-\beta+1}^\star\|$ and write the usual primal attack equation:

$$\begin{aligned}\|\mathbf{s}\|\sqrt{\frac{\beta}{n}} &= \|\hat{\mathbf{b}}^\star_{n-\beta+1}\| \\ &= c \cdot \|\mathbf{b}^\star_{n-\beta+1}\| \\ &= c \cdot \delta_\beta^{2\beta-n-1}\mathrm{vol}(\Lambda)^{1/n}.\end{aligned}$$

Recall that $c = \frac{q\alpha_\beta^{1-k}}{\|\mathbf{b}_1^\star\|}$, and $\|\mathbf{b}_1^\star\| = \delta_\beta^n \mathrm{vol}(\Lambda)^{1/n}$. Therefore the primal attack equation becomes

$$\|\mathbf{s}\|\sqrt{\gamma} = q\delta_\beta^{-n}\left(\delta_\beta^{-\frac{2n}{n-1}}\right)^{1-k}\delta_\beta^{2\beta-n-1}.$$

Asymptotically, the dominant terms on both sides are in n^{cst}, and therefore must be equal. This yields

$$\frac{1}{2} = 1 - \frac{1}{2\gamma} + \frac{1-\sqrt{\gamma}}{\gamma} + \frac{2\gamma-1}{2\gamma},$$

which can be simplified down to

$$\frac{1}{\sqrt{\gamma}} = \frac{3}{2},$$

and therefore $\gamma = \left(\frac{2}{3}\right)^2 = \frac{4}{9}$.

□

Remark 4.2.6. The second model leads to an asymptotic heuristic blocksize $\beta = \frac{4}{9}n < \frac{n}{2}$, beating our first analysis. In fact, this value corresponds exactly to the recent asymptotic analysis of Bernstein [Ber24, Th. 1.2.1], where in the spirit of [MS01], the author only keeps κ coordinates before applying the primal attack heuristic.

4.3 Primal Attack for Many Short Vectors

4.3.1 An Asymptotic Analysis

Hypercubic and NTRU lattices have multiple shortest (non-zero) vectors. The primal attack framework as described by Equation 3.2 does not take this into account, as it only relies on the expected value of the norm of the projection of a single vector. We only need one projection to be smaller than the expected Gram-Schmidt norm $\|\mathbf{b}^\star_{n-\beta+1}\|$ for the SVP oracle on the last BKZ block of size β to recover said projection. And because the squared norms of the projections onto random subspaces follow Beta distributions, we can estimate the expected value of the minimal projection and slightly lower the blocksize. See Figure 4.1 for an illustration. For smaller dimensions, we observe how considering more short vectors improves the double-intersection phenomenon described in [AGVW17].

The *Leaky-LWE estimator* of [DDGR20] has an option to account for the presence of multiple shortest vectors, however this option is not discussed in detail in [DDGR20]. Our new framework (although the same in spirit), addresses this issue differently, offering asymptotic insights as well as specifically isolating the impact of this condition on the blocksize.

In the literature on the primal attack, authors have never used any special property of the Beta distribution other than its mean. The authors of [DDGR20] use a probabilistic model in which the squared norms of the projections are approximated using a χ^2 distribution. Even

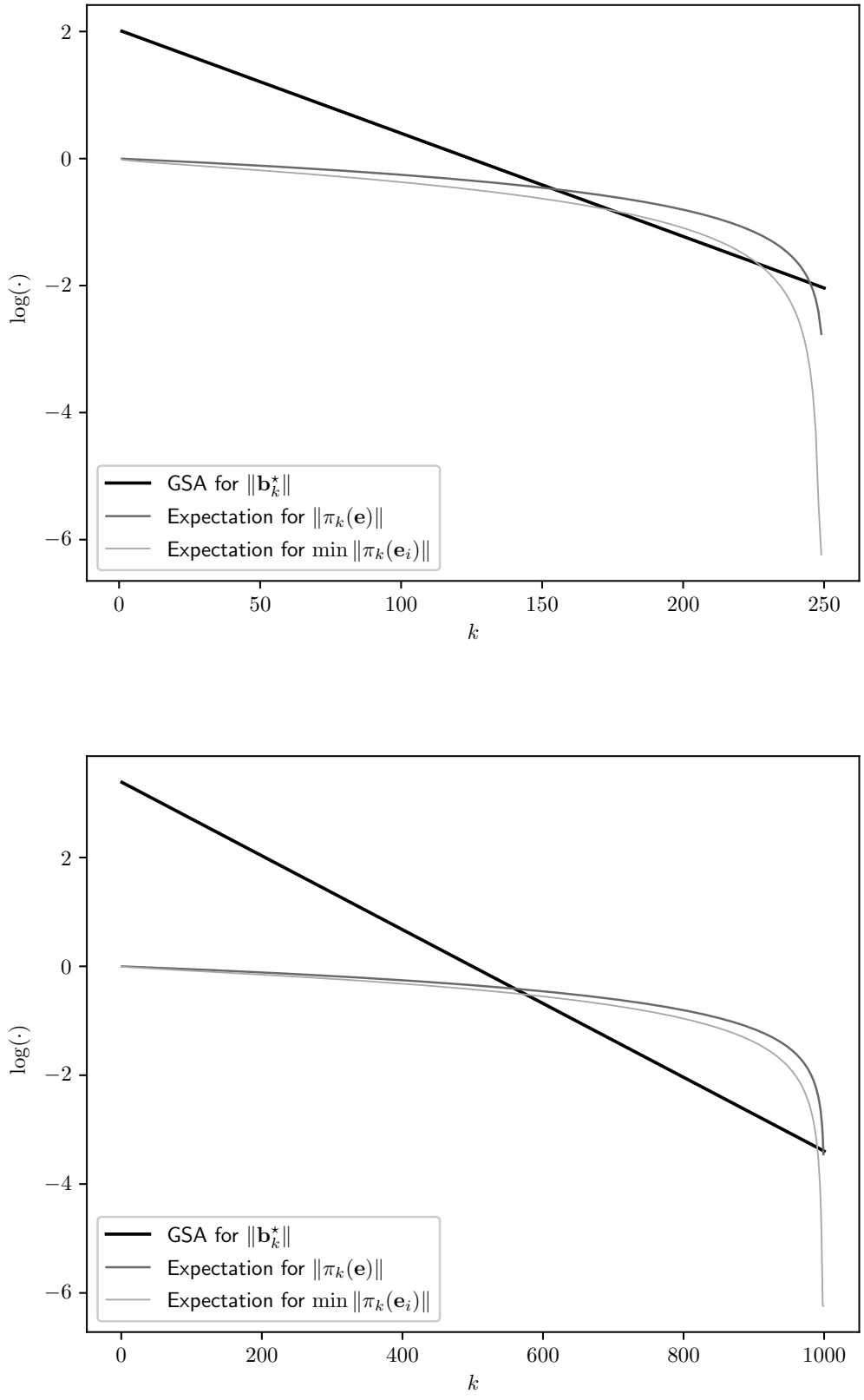


Figure 4.1: Comparing the expected norms of randomised Gram-Schmidt vectors of a basis of $\mathbb{Z}^n$ after BKZ reduction with blocksize $n/2$ with the expected projection norms of one and n unit vectors. $n = 250$ above and $n = 1000$ below.

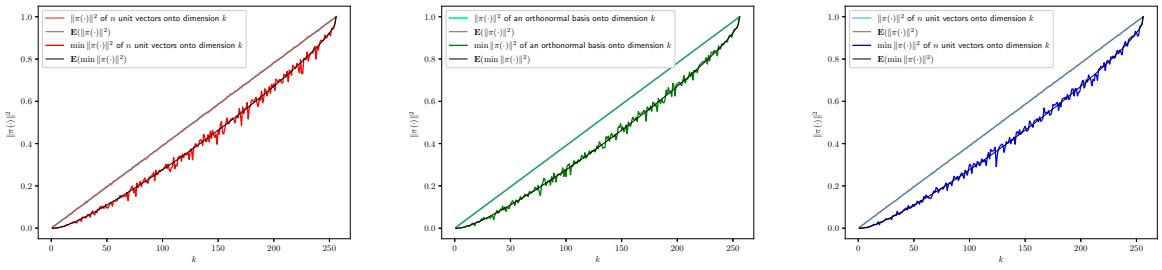
though the χ^2 and the Beta distributions are very good approximations of each other in the small-β context, the difference might become more noticeable for larger blocksizes, so to correct this we choose to work with Beta distributions instead.

We want to emphasise that our framework is not intended for practical use or to supplant existing work. Instead, its purpose is to enhance our comprehension of the components involved in the primary attack. When compared to [DDGR20] it simplifies the situation greatly by not taking into account lifting probabilities, or even more precise Gram-Schmidt norm estimates. It also ignores possible fluctuations in the value of $\|\mathbf{b}^\star_{n-\beta+1}\|$. Estimations for hypercubic lattices obtained by both frameworks are compared in Figure 4.3.

To estimate the expectation of the minimal norm of the projections, we use the following heuristic.

Heuristic 1. Let $0 < k < n$. If a lattice L of rank n contains N vectors $\mathbf{s}_1, \ldots, \mathbf{s}_N$ of equal norm r, then the random variables defined by the squared norms of their projections onto a random dimension k subspace of $\mathbb{R}^n$ are independent.

Heuristic 1 is very close to heuristics used in the study of the dual attack [DP23]. We argue that when N is not too large (we only use $N \leq n$), this heuristic is reasonable for our purposes. See Figure 4.2 for a comparison of the average minimal squared norms of the projections of shortest vectors onto random subspaces in the cases of a random set of unit vectors, an orthonormal basis of $\mathbb{R}^n$, and a cirulant set of n vectors.



(a) Random (b) Orthonormal (c) Structured

Figure 4.2: Comparing average/minimal norms of projections π of sets of n unit vectors onto random k-dimensional subspaces of $\mathbb{R}^n$. $n = 256$, and k ranges from 0 to n. Theoretical expected values are plotted in black. The sets considered are random on the sphere (4.2a), orthonormal basis (4.2b) and structured: all cyclic permutations of a normalised NTRU*-like secret vector (4.2c). Each point correspond to a single random choice of vectors as well as a single random choice of subspace.*

Lemma 4.3.1. *Let $0 < k < n$. Let $\mathbf{s}_1, \ldots, \mathbf{s}_N$ be vectors of norm r in a lattice that satisfies Heuristic 1. Then*

$$\mathbb{E}\left(\min_{1\leq i\leq N} \|\pi(\mathbf{s}_i)\|^2\right) = r^2 \int_0^1 \left(1 - I_x\left(\frac{k}{2}, \frac{n-k}{2}\right)\right)^N dx,$$

where π is the projection onto a random dimension k subspace of $\mathbb{R}^n$, and I is the regularised incomplete beta function.

Proof. All of the $\|\pi(\mathbf{s}_i)\|^2/r^2$ follow the Beta distribution $B(k/2, (n-k)/2)$. Let f denote its probability density function (pdf), and F the associated cumulative distribution function (cdf). Then by independence, the pdf $f_{\min}$ of $\min_{1\leq i\leq N} \|\pi(\mathbf{s}_i)\|^2/r^2$ satisfies $f_{\min}(x) = N(1 - F(x))^{N-1} f(x)$. It follows that

$$\mathbb{E}\left(\min_{1\leq i\leq N} \|\pi(\mathbf{s}_i)\|^2/r^2\right) = \int_0^1 x f_{\min}(x) dx = \int_0^1 (1 - F(x))^N dx,$$

where we used integration by parts. We conclude using the fact that the cdf of the beta function is equal to the regularised incomplete beta function.

□

While Lemma 4.3.1 can be quite practical, we prefer to work with a slightly different quantity that is easier to manipulate.

Lemma 4.3.2. *Let $\tau > 0$. Let $0 < k < n$ and π a projection onto a random dimension k subspace of $\mathbb{R}^n$. Let $\mathbf{s}_1, \ldots, \mathbf{s}_N$ be vectors of norm r in a lattice that satisfies Heuristic 1. Then*

$$\Pr\left(\min_{1\leq i\leq N} \|\pi(\mathbf{s}_i)\| < r\tau\right) = 1 - \left(1 - I_{\tau^2}\left(\frac{k}{2}, \frac{n-k}{2}\right)\right)^N$$

Proof. By independence,

$$\Pr\left(\min_{1\leq i\leq N} \|\pi(\mathbf{s}_i)\| < r\tau\right) = 1 - \prod_{i=1}^{N} \Pr\left(\|\pi(\mathbf{s}_i)\|^2 \geq r^2\tau^2\right).$$

All of the $\|\pi(\mathbf{s}_i)\|^2/r^2$ follow the Beta distribution of parameters $k/2, (n-k)/2$. Each term of the product is exactly the complement to 1 of the cdf of the previous beta function evaluated at τ^2. We conclude by definition of $I_x(a, b)$.

□

In our study we consider blocksizes that are fractions of n. For this reason we will use the notation $\beta = \alpha n$, where $\alpha \in [0, 1]$. Again, we are interested in asymptotic behaviours as n goes to infinity, which means we do not care if β is not integral. In order to get anything meaningful from Lemma 4.3.2, we need a precise estimate of $I_x\left(\frac{\alpha n}{2}, \frac{(1-\alpha)n}{2}\right)$. For this we use a result by Temme [Tem92].

Lemma 4.3.3 (Derived from [Tem92], Sec. 3)**.** *Let $\varepsilon > 0$, $x \in (0, 1)$ and $\alpha \in (\varepsilon, 1-\varepsilon)$. Then*

$$I_x\left(\frac{\alpha n}{2}, \frac{(1-\alpha)n}{2}\right) = \frac{1}{2}\mathrm{erfc}\left(-\frac{\eta\sqrt{n}}{2}\right) + o\left(\mathrm{erfc}\left(-\frac{\eta\sqrt{n}}{2}\right)\right),$$

where $\eta = \mathrm{sign}(x-\alpha)\sqrt{-2\alpha\ln\left(\frac{x}{\alpha}\right) - 2(1-\alpha)\ln\left(\frac{1-x}{1-\alpha}\right)}$, and $\mathrm{erfc} = 1 - \mathrm{erf}$ is the complementary error function.

Proof. We are in the second case studied by [Tem92], where $a = \frac{\alpha n}{2}$ and $b = \frac{(1-\alpha)n}{2}$ are such that $a + b = \frac{n}{2} \to \infty$, and both ratios $\frac{a}{b} = \frac{\alpha}{1-\alpha}$ and $\frac{b}{a} = \frac{1-\alpha}{\alpha}$ are bounded away from 0. The Lemma follows directly from Eq. (3.9) in [Tem92].

□

Lemma 4.3.3 begs the question: how big can η get? By deriving the asymptotic behaviour of η, we can deduce the asymptotic blocksize required by our variant of the primal attack.

Proposition 4.3.4. *Let $0 < p < 1$ be a fixed constant probability. Let $0 < \varepsilon < 1$ and $\varepsilon < \alpha < 1 - \varepsilon$ be a function of n. Let $\pi_{n-\alpha n+1}$ be a projection onto a random dimension αn subspace of $\mathbb{R}^n$. Let $\mathbf{s}_1, \ldots, \mathbf{s}_N$ be $\Theta(n)$ vectors of norm r in a lattice L that satisfies Heuristic 1. Suppose also that $c := \mathrm{vol}(L)^{1/n}/r = \Theta(1)$. Then if the asymptotic identity*

$$\Pr\left(\min_{1\leq i\leq N} \|\pi_{n-\alpha n+1}(\mathbf{s}_i)\| < \delta_{\alpha n}^{(2\alpha-1)n-1}\sqrt{c}\right) = p + o(1) \tag{4.3}$$

holds, then

$$\beta := \alpha n = \frac{n}{2} - \frac{\ln(2c)}{4}\frac{n}{\ln(n)} + o\left(\frac{n}{\ln n}\right).$$

Proof. By Lemma 4.3.2 with $\tau = \delta_{\alpha n}^{(2\alpha-1)n-1}\sqrt{c}$,

$$\Pr\left(\min_{1\leq i\leq N}\|\pi_{n-\alpha n+1}(\mathbf{s}_i)\| < \delta_{\alpha n}^{(2\alpha-1)n-1}\sqrt{c}\right) = 1-\left(1-I_x\left(\frac{\alpha n}{2},\frac{(1-\alpha)n}{2}\right)\right)^N,$$

where $x = \delta_{\alpha n}^{2((2\alpha-1)n-1)}c$ therefore it would suffice to prove that

$$\ln\left(1-I_x\left(\frac{\alpha n}{2},\frac{(1-\alpha)n}{2}\right)\right) \sim \frac{\ln p}{N}. \tag{4.4}$$

Letting $\eta = \operatorname{sign}(x-\alpha)\sqrt{-2\alpha\ln\left(\frac{x}{\alpha}\right)-2(1-\alpha)\ln\left(\frac{1-x}{1-\alpha}\right)}$ as in Lemma 4.3.3 and combining the result of this same Lemma with Equation 4.4, we get

$$-\frac{\ln p}{n} \sim I_x\left(\frac{\alpha n}{2},\frac{(1-\alpha)n}{2}\right) \sim \frac{1}{2}\operatorname{erfc}\left(-\frac{\eta\sqrt{n}}{2}\right).$$

This yields $x<\alpha$ and $\eta^2 \sim 4\frac{\ln n}{n}$ (we used that $N=\Theta(n)$ and the following estimate for large u: $\operatorname{erfc}(u)\sim\pi^{-1/2}u^{-1}e^{-u^2}$. See also [Phi60] for an alternative method). To conclude we look for the most important terms inside of η^2. Looking at

$$4\frac{\ln n}{n} \sim 2\alpha\ln\alpha + 2(1-\alpha)\ln(1-\alpha) - 2\alpha\ln x - 2(1-\alpha)\ln(1-x), \tag{4.5}$$

we deduce that $\alpha = \frac{1}{2}-\frac{\xi}{\ln n}$, where $\xi = O(1)$. By carefully taking care of the little o terms, x can be expressed using

$$\frac{x}{c} = \delta_{\alpha n}^{2((2\alpha-1)n-1)}c = \left(\frac{\alpha n}{2\pi e}(\alpha n\pi)^{1/(\alpha n)}\right)^{\frac{-2\xi n/\ln n-1}{\alpha n-1}} \sim \left(\frac{n}{4\pi e}\right)^{-4\frac{\xi}{\ln n}} \sim e^{-4\xi}.$$

We can now compute the largest contribution K to the right hand side of Equation 4.5:

$$K = \ln\left(\frac{e^{4\xi}}{4c(1-ce^{-4\xi})}\right) = \ln\left(\frac{(e^{4\xi}-2c)^2+4c(e^{4\xi}-c)}{4c(e^{4\xi}-c)}\right).$$

We must have $K+o(K) = 4\frac{\ln n}{n}$, therefore the constant term must be 0, and thus $\xi = \frac{\ln(2c)}{4}+o(1)$, which concludes our proof.

□

Corollary 4.3.5. *Let L be a rank n lattice for which Heuristic 1 holds with vectors $\mathbf{s}_1,\ldots,\mathbf{s}_N$ of norm r such that $N=\Theta(n)$ and $r/\operatorname{vol}(L)^{1/n} := c^{-1/2} = O(1)$. The primal attack framework predicts that applying BKZ reduction with blocksize $\beta = n(1/2-\ln(2c)/4\ln n+o(1/\ln n))$ recovers a vector of norm at most r with high probability. In particular if the heuristic holds for hypercubic and* NTRU *lattices, then so does this result.*

Proof. The main point follows directly from Proposition 4.3.4. For a hypercubic lattice Λ, $\operatorname{vol}(\Lambda) = \|\mathbf{s}\| = 1$. For a NTRU lattice L, $\operatorname{vol}(L)^{1/n} = \sqrt{q} = \Theta(\sqrt{n})$, and $\|\mathbf{s}\| = \Theta(n)$, where $\mathbf{s} = (\mathbf{g},\mathbf{f})$ is the secret key and q is the NTRU modulus.

□

4.3.2 Discussion and Illustration

The results of Proposition 4.2.1 and Proposition 4.3.4 are identical. If we focus uniquely on the primal attack[1], this means that asymptotically, having n short vectors does not offer any advantage over having just one. In fact, we conjecture that for k a constant, if we had a polynomial number N of independent (in the sense of Heuristic 1) equally short vectors, then the following k terms of the expansion of the predicted blocksize assuming the presence of these N vectors would match precisely with the next k terms (of the form $a_i n \ln^{-i}(n)$) derived in the case of a solitary short vector. Indeed, the estimates of Proposition 4.2.1 and Proposition 4.3.4 are not very good in practice, because the convergence rate is very weak (notice that the term in the erfc function is a $\Theta(\sqrt{\ln n})$). This means that the asymptotic regime will only kick in for huge values of n, beyond cryptographic relevance. However, this does not prove that the presence of more short vectors is useless with regards to the primal attack. In fact, the structure of Equation 3.2 indicates that having strictly more short vectors is directly advantageous.

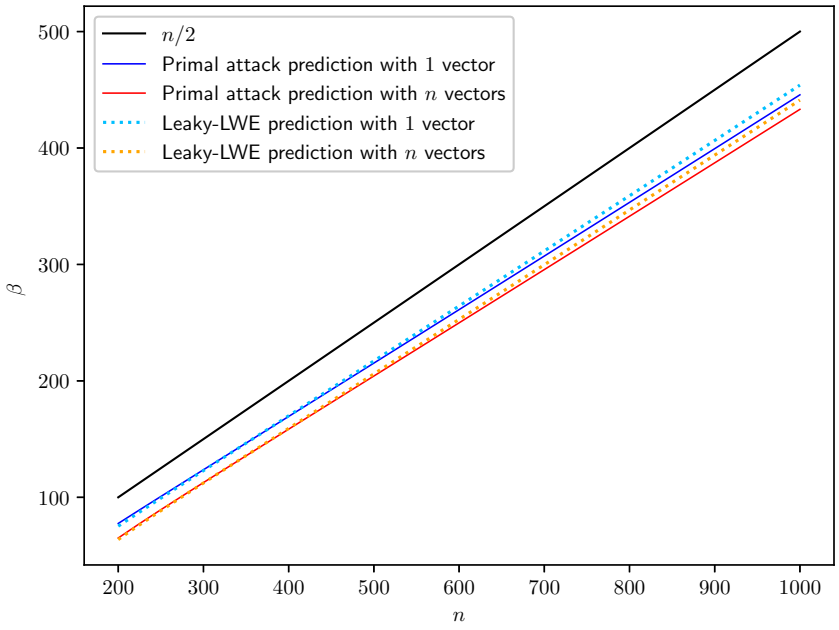


Figure 4.3: Blocksizes required to recover unit vectors in dimension n hypercubic lattices. The predictions in dotted lines were generated using the sage script provided in [DPPvW22]. Our model does not assume progressive-BKZ execution.

Practical alternative

Due to the reasons mentioned above, for practical application of our framework, we recommend directly solving the modified primal attack equation obtained from combining Equation 3.2 and Lemma 4.3.1 numerically. The results for hypercubic lattices are plotted in Figure 4.3, and compared with the predictions of [DDGR20]. In the observed range of dimensions, the heuristic blocksize gain is consistently between 11 and 13, compared to simply evaluating the asymptotic formula. Surprisingly, our naive predictions end up being very close to the more precise predictions of [DDGR20]. We provide a proof of concept sage script at

[1] Dense sublattice attacks can asymptotically outperform generic lattice reduction for NTRU with overstretched parameters [KF17; DvW21], but this is outside the scope of our study.

https://github.com/htmb-bot/NTRU-and-Hypercubic.

4.3.3 How Good is the Geometric Series Assumptions?

Our heuristic estimates heavily rely on our ability to predict the values of the norms of the Gram-Schmidt vectors after lattice reduction. In order to do this we assumed that their log-norms decreased linearly in the indices (GSA), and corrected this in the presence of q-vectors (ZGSA). This GSA heuristic has been discussed at length in [AD21, Section 4] and all references within. The GSA is expected to be a good first order approximation of the Gram-Schmidt norms of an n-dimensional BKZ-β reduced basis for indices that are not small (< 50) nor too close to $n - \beta$ (although this can change in the presence of q-vectors). The most notable difference comes from the fact that the tail (the last β indices) should have the shape of an HKZ-reduced basis. This tail-adapted variant of GSA (TGSA) provides the same prediction for the first indices, and then the last β norms first decrease slower than expected by GSA, and then faster towards the end. The TGSA can be observed through the simulator of [CN11]. Such simulators are inspired and verified by experiments, which means that the estimations are very precise for the dimensions on which the simulators were tested, but it remains unclear if such predictions still hold for cryptographically large dimensions, and how much they depend on specific implementations. The case that is interesting for cryptography: when β is a constant fraction of the dimension n should be studied in more detail. To illustrate, we take lattices of dimensions $n \in \{140, 160, 170, 180, 190, 200\}$ from the SVP Challenge [SG], and run one-tour BKZ progressively from $\beta = 5$ to $\beta = n/2$, using g6k [Alb+19]. We then compare the log-norms of the Gram-Schmidt vectors to those obtained by the simulation of [CN11].

We can clearly see the *spoon* shape predicted by the TGSA on all plots of Figure 4.4. While the prediction in blue is very close to the profile obtained through experiments, it seems to deviate more and more as the dimension n grows. Of course, this could be meaningless or related to a specific implementation, but completing the sequence, one can easily imagine a very big discrepancy at $n = 500$.

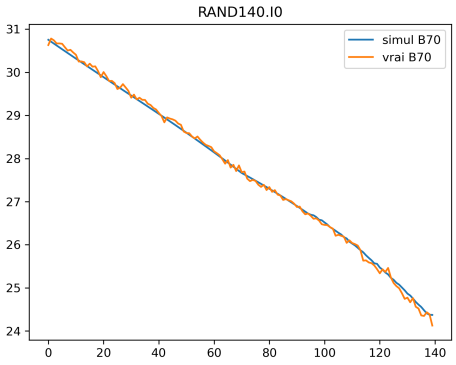
4.4 A Reduction from $\mathbb{Z}$SVP to γ-$\mathbb{Z}$SVP

In this section we focus exclusively on the hypercubic lattice. Recall that a rank n lattice Λ is *hypercubic* if and only if it is generated by an orthonormal basis, and if that is the case, we say that the lattice is isomorphic to $\mathbb{Z}^n$. Recall also that $\mathbb{Z}$SVP denotes the specialisation of the shortest vector problem to lattices that are isomorphic to $\mathbb{Z}^n$, and that for $\gamma > 1$, γ-$\mathbb{Z}$SVP denotes its approximation variant. We focus on the following question.

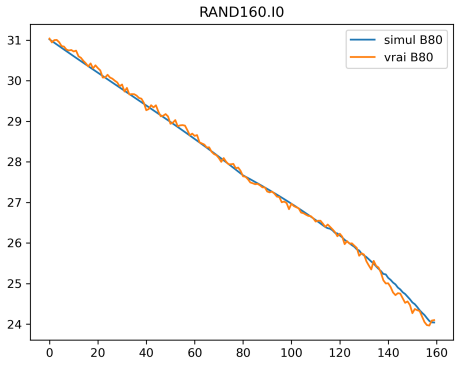
Question 4.4.1. *Let Λ be a hypercubic lattice, and $\gamma > 1$ a real number. Given access to an oracle that outputs vectors in Λ of norm at most γ, how easy is it to recover a unit vector of Λ?*

This question only really makes sense when γ is the square root of an integer, so the first non-trivial instance of Question 4.4.1 is when $\gamma = \sqrt{2}$. This was solved provably using a projection argument by [BGPS23, Theorem 5.1] using at most two queries. The authors claim that the result still holds for $\gamma \in \{\sqrt{3}, 2\}$, although the proofs involve a lot of tedious casework, and the number of queries increases. We were able to solve $\gamma = \sqrt{3}$ using 4 queries, but it is unclear how the number of queries grows with γ, and a proof by casework seems to become unfeasible even for quite small constant values of γ.

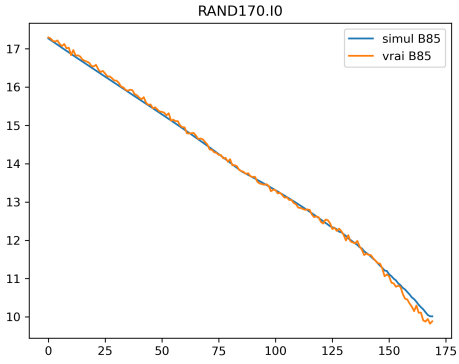
Another approach is [Jia+23, Theorem 3.2], where the authors require $|\Lambda \cap \gamma B_n^{(2)}(\gamma)|$ calls to a randomised γ-$\mathbb{Z}$SVP oracle. As γ grows, this approach remains polynomial in n for constant γ but is not very practical.



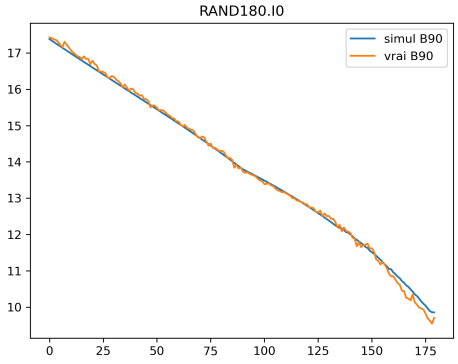
(a) $n = 140$



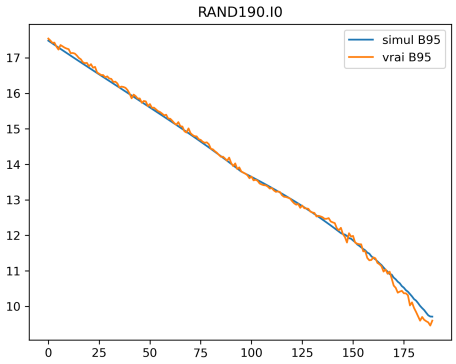
(b) $n = 160$



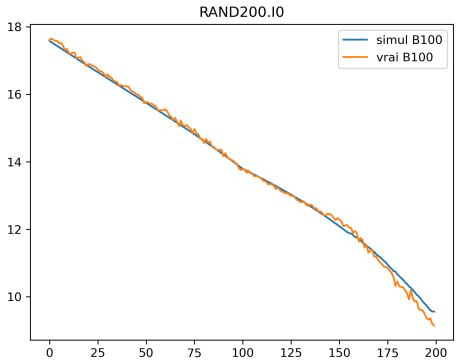
(c) $n = 170$



(d) $n = 180$



(e) $n = 190$



(f) $n = 200$

Figure 4.4: Comparing (progressive) BKZ-n/2 with [CN11] simulation.

4.4.1 Project and Intersect: a New Algorithm

We give a randomised algorithm that solves ℤSVP using $\Theta(n)$ calls to a γ-ℤSVP oracle. Our algorithm is heuristic, but we believe it can be made provable if analysed with care. We leave this for future work.

Proposition 4.4.2. *Let Λ be a hypercubic lattice of rank n. Let L be a primitive sublattice of Λ, of rank k. Define*

$$L_\Lambda^\perp := \pi_{\mathrm{vect}(L)^\perp}(\Lambda) \cap \Lambda,$$

then $L_\Lambda^\perp$ *is a lattice of rank* $n-k$ *and volume* $\mathrm{vol}(L_\Lambda^\perp) = \mathrm{vol}(L)$.

Proof. Because Λ is an integral lattice and L is a primitive sublattice of Λ, [Mar03, Proposition 1.9.6] implies that $L_\Lambda^\perp$ is a lattice. Its volume is then given by [Mar03, Proposition 1.9.8], as Λ is a unimodular lattice ($\Lambda = \Lambda^\times$). □

Using the notations from Proposition 4.4.2, assuming we know a basis $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ for Λ and a basis $C = (\mathbf{c}_1, \dots, \mathbf{c}_k)$ for L, we can efficiently compute a basis of $L_\Lambda^\perp$. For this we follow the procedure described in [NS97], and construct the lattice generated by the following matrix for $\lambda \in \mathbb{R}_{>0}$ (we use row notation).

$$\mathbf{M}_\lambda = \begin{pmatrix} \langle \mathbf{b}_1, \mathbf{c}_1 \rangle & \langle \mathbf{b}_1, \mathbf{c}_2 \rangle & \cdots & \langle \mathbf{b}_1, \mathbf{c}_k \rangle & \lambda & 0 & \cdots & 0 \\ \langle \mathbf{b}_2, \mathbf{c}_1 \rangle & \langle \mathbf{b}_2, \mathbf{c}_2 \rangle & \cdots & \langle \mathbf{b}_2, \mathbf{c}_k \rangle & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{b}_n, \mathbf{c}_1 \rangle & \langle \mathbf{b}_n, \mathbf{c}_2 \rangle & \cdots & \langle \mathbf{b}_n, \mathbf{c}_k \rangle & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

We also directly have $\mathbf{M}_\lambda = \begin{pmatrix} \mathbf{B}^T\mathbf{C} & \lambda \mathbf{I}_n \end{pmatrix}$. We claim that for a small enough value of λ, LLL-reducing $\mathbf{M}_\lambda$ recovers a basis of $L_\Lambda^\perp$. Indeed, if there are integers (x_i) such that $\mathbf{x} = \sum x_i \mathbf{b}_i \in L_\Lambda^\perp$, then for each of the $\mathbf{c}_j$, we have

$$0 = \langle \mathbf{x}, \mathbf{c}_j \rangle = \sum_{i=1}^{n} x_i \langle \mathbf{b_i}, \mathbf{c_j} \rangle,$$

making $(0, \dots, 0, \lambda x_1, \lambda x_2, \dots, \lambda x_n)$ a vector of $\mathcal{L}(\mathbf{M}_\lambda)$. Therefore, a basis for $L_\Lambda^\perp$ embeds in a rank $n-k$ sublattice of $\mathcal{L}(\mathbf{M}_\lambda)$, whose volume is $\lambda^{n-k}\mathrm{vol}(L)$ by Proposition 4.4.2. For a small enough λ, this sublattice is efficiently recovered by the LLL algorithm.

Remark 4.4.3. This procedure does not use the secret orthonormal basis of Λ in any way, only scalar products between lattice vectors. In general for this problem we might only be given Gram matrices, in which case our reasoning still holds. Things can quickly go wrong when conducting experiments directly on $\mathbb{Z}^n$, where one must take extra care not to use the coordinates of vectors in the canonical basis, as those depend directly on the secret basis.

We are now ready to present our randomised algorithm. The main idea is that when a basis vector $\mathbf{b}$ is somewhat short, it contains 0-coordinates in the secret orthonormal basis with high probability. If this is the case, then any secret unit vector $\mathbf{e}_i$ such that $\langle \mathbf{e}_i, \mathbf{b} \rangle = 0$ will be contained in the lattice $\mathcal{L}(\mathbf{b})_\Lambda^\perp$, of rank $n-1$ instead of n, and volume $\|\mathbf{b}\|$ instead of 1. It should be easier to recover a unit vector in this new lattice, although in exchange, the new lattice $\mathcal{L}(\mathbf{b})_\Lambda^\perp$ does not contain n independent unit vectors, but less, which makes the situation unclear, and we will have to resort to precise estimations. If we are lucky, we can project orthogonally to k independent short vectors, and hope that the orthogonal lattice still contains a unit vector, which will now be easier to recover. The procedure we described is reminiscent of May and Silverman's attack from [MS01] in the case of NTRU, although what we are doing is different. The reader can refer to [DDGR20, Section 6.3] for a discussion of the impact of that attack on concrete security estimates of NTRU.

Algorithm 1 A randomised algorithm for ℤSVP.

Require: A basis $\mathbf{B}$ of $\Lambda \simeq \mathbb{Z}^n$. Parameters $k, \beta \in \mathbb{Z}_{>0}$.
Ensure: A unit vector of Λ, or $\perp$.
1: Guess a set of k pairwise distinct indices. $I = \{i_1, \ldots, i_k\} \subseteq [n]$.
2: Compute a basis $\mathbf{B}_1$ of the lattice $\mathcal{L}(\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_k})^{\perp}_{\Lambda}$.
3: Get $\mathbf{B}_2$ by BKZ-β reducing $\mathbf{B}_1$.
4: **if** $\exists \mathbf{e} \in \mathbf{B}_2 : \|\mathbf{e}\| = 1$ **then**
5: Return $\mathbf{e}$
6: **else**
7: Return $\perp$
8: **end if**

Step 2 requires building and LLL-reducing the matrix $\mathbf{M}_\lambda$ as was previously discussed. The key steps are Step 1 (guessing) and Step 3 (lattice reduction). We need parameters k to be carefully selected to allow the guessing step to succeed with high enough probability: this guessing step succeeds when $\lambda_1(L) = 1$, where $L = \mathcal{L}(\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_k})^{\perp}_{\Lambda}$. The blocksize parameter β should also be taken large enough that lattice reduction consistently recovers the unit vector contained in L.

It is clear that Algorithm 1 succeeds in solving ℤSVP with some probability. In order for it to constitute a reduction from ℤSVP to γ-ℤSVP, we need to have a procedure that generates a basis of short vectors, from the oracle only. We admit that this is doable using $\Theta(n)$ calls to a γ-ℤSVP oracle using the randomisation framework of [Jia+23], and will assume for now that we are given an input basis $\mathbf{B}$ of Λ whose vectors have length γ.

4.4.2 Heuristic Analysis

Depending on the value of γ, we will explain how one can select k and β. For the sake of exposition, we will not be concerned about the constraints that k and β and γ^2 must be integers, and we will assume for now that $\Theta(n)$ calls to a γ-ℤSVP oracle allows us to efficiently build a list of n $\mathbb{Z}$-linearly independent vectors $(\mathbf{b}_1, \ldots, \mathbf{b}_n)$, all of norm γ. This is obviously an oversimplification, but it should reflect reality somewhat accurately: we leave more precise statements and proofs for future work.

Estimating the number of restarts

We first turn to Step 1, and estimate its success probability. Recall that we consider this guessing step a success if there exists a unit vector $\mathbf{e}$ of Λ such that $\mathbf{e} \perp \mathbf{b}_i$ for all $i \in I$. We first note that if $\|\mathbf{b}_i\|^2 = \gamma^2$ and γ is small, then $\mathbf{b}_i$ must be orthogonal to many unit vectors of Λ. Indeed, $\mathbf{b}_i$ can be decomposed as

$$\mathbf{b}_i = \sum_{j=1}^{n} x_j \mathbf{e}_j,$$

where the $(\mathbf{e}_j)_j$ are an orthonormal basis of Λ, and all x_j are integers, such that $\sum x_j^2 = \gamma^2$. For non-zero x_j, $x_j^2 \geq 1$, therefore the number of non-zero x_j is at most γ^2, which means that each of the $\mathbf{b}_i$ has at least $n - \gamma^2$ 0-coordinates in the secret basis.

Remark 4.4.4. If γ is large (eg $\gamma = \Theta(\sqrt{n})$), a more precise analysis is required, depending on the distribution. For example if $\mathbf{b}_i$ is sampled via the discrete Gaussian distribution over Λ, then one can easily estimate the number of 0-coordinates. As we focus here on smaller values of γ, no extra care is required.

We now provide a lower bound on the probability that Step 1 succeeds. As each of the $\mathbf{b}_i$ has at least $n - \gamma^2$ 0-coordinates, the sum over all $\mathbf{e}_j$ of the number of $\mathbf{b}_i$ orthogonal to them

is at least $n(n-\gamma^2)$, from which we deduce that there exists a secret unit vector $\mathbf{e} \in \Lambda$ that is orthogonal to at least $n-\gamma^2$ of the $\mathbf{b}_i$. The probability of a guess succeeding for this particular $\mathbf{e}$ is now at least

$$p_k(\gamma) := \binom{n-\gamma^2}{k} \cdot \binom{n}{k}^{-1}. \tag{4.6}$$

Indeed the probability that $\mathbf{e}$ is orthogonal to all k chosen vectors is at least the same probability in the case where exactly $n-\gamma^2$ of the $\mathbf{b}_i$ are orthogonal to $\mathbf{e}$. In this case there are $\binom{n-\gamma^2}{k}$ good choices of I, for $\binom{n}{k}$ total choices.

Remark 4.4.5. With some more work, it should be possible to derive a much tighter bound on the success probability, if not the exact value. For this is might be beneficial to decorrelate the 0-coordinates of the $\mathbf{b}_i$ using the randomisation framework of [Jia+23]. The remaining problem becomes a classical variant of the coupon collector problem. Last but not least, note that here we only use one of the $\mathbf{e}_j$ to get a lower bound on the success probability, but as in Section 4.3, we should take into account all n targets. As in Section 4.3, the impact of using many targets will only lead to logarithmic improvements, which might still be useful for concrete parameters.

Predicting the lattice reduction step

If Step 1 succeeds, then $\mathbf{e} \in L$, where L denotes the lattice $L = \mathcal{L}(\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_k})^{\perp}_{\Lambda}$ obtained at Step 2. L is a sublattice of Λ, here is how they compare:

- $\lambda_1(L) = \lambda_1(\Lambda) = 1$;
- $\mathrm{vol}(\Lambda) = 1$, but $\mathrm{vol}(L) = \mathrm{vol}(\mathcal{L}(\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_k}))$, which is much larger.
- $\mathrm{rank}(\Lambda) = n$ but $\mathrm{rank}(L) = n-k$, which is much smaller.

If the lattice L has smaller rank, larger volume, but has the same shortest vector $\mathbf{e}$ as Λ, then this vector should be much easier to recover using lattice reduction. Heuristically, we argue that

$$\mathrm{vol}(L) = \mathrm{vol}(\mathcal{L}(\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_k})) \approx \prod_{j=1}^{k} \|\mathbf{b}_{i_j}\| \approx \gamma^k.$$

Again, this deserves to be made more rigorous but we choose to keep the analysis heuristic.

We now turn to Step 3. Conditionally on being lucky at Step 1, we can use standard primal attack heuristics to estimate how plausible it is that the shortest vector in the BKZ-β-reduced basis $\mathbf{B}_2$ is $\mathbf{e}$ up to sign (we focus only on $\pm\mathbf{e}$, as for a clever choice of k, it should be extremely unlikely that any other unit vectors of Λ is in L). The 2016 estimate says that $\mathbf{e}$ will be recovered when

$$\sqrt{\frac{n-k-\beta}{n-k}}\|\mathbf{e}\| \leq \|\mathbf{B}^{\star}_{2,n-k-\beta+1}\|. \tag{4.7}$$

Here $\mathbf{B}^{\star}_{2,n-k-\beta+1}$ denotes the $(n-k-\beta+1)$-th vector of the Gram-Schmidt orthogonalisation of $\mathbf{B}_2$. Using the GSA heuristic, we can estimate that

$$\|\mathbf{B}^{\star}_{2,n-k-\beta+1}\| \approx \delta_{\beta}^{n-k-\frac{2n}{n-1}(n-k-\beta)}\mathrm{vol}(L)^{1/(n-k)}, \tag{4.8}$$

where $\delta_\beta = \left(\frac{\beta}{2\pi e}(\pi\beta)^{1/\beta}\right)^{\frac{1}{2(\beta-1)}}$.

Eliminating factors which do not contribute to the main term, the right hand side of the success condition becomes

$$\left(\frac{\beta}{2\pi e}\right)^{1-\frac{n-k}{2\beta}} \cdot \mathrm{vol}(L)^{1/(n-k)} \approx \left(\frac{\beta}{2\pi e}\right)^{1-\frac{n-k}{2\beta}} \cdot \gamma^{\frac{k}{(n-k)}}. \tag{4.9}$$

The constant case

Claim 4.4.6. *Let $\gamma > 1$ be a constant $\gamma = O(1)$. Heuristically, using Algorithm 1 with parameters $k = n - \ln n$ and $\beta = \gamma$ leads to a polynomial-time reduction from ℤSVP to γ-ℤSVP.*

Indeed, with this choice of parameters, the right hand side of Equation 4.9 becomes $\left(\frac{\gamma}{2\pi e}\right)^{1-\frac{\ln n}{2\gamma}} \cdot \gamma^{\frac{n}{\ln n}-1}$. This is much larger than the left hand side of Equation 4.7, which makes the success condition of BKZ-γ always true. Note that we are running BKZ in a lattice of dimension $\ln n$ with constant blocksize. This is polynomial-time, and for practical values of n, LLL would be more than enough to recover $\mathbf{e}$.

It remains to show that $p_k(\gamma)$ is not too small. Using Stirling's formula, a quick computation gives

$$p_k(\gamma) = 1 - \frac{\gamma^2 \ln n}{n} + o\left(\frac{\ln n}{n}\right),$$

which means that asymptotically, a single run of Algorithm 1 is enough to solve ℤSVP.

The logarithmic case

Claim 4.4.7. *Let $d > 0$ be a constant $d = O(1)$ and $\gamma = O((\ln n)^d)$. Heuristically, using Algorithm 1 with parameters $k = n - \sqrt{n}$ and $\beta = \ln n$ leads to a quasi-polynomial time reduction from ℤSVP to γ-ℤSVP.*

Assuming $\beta = \ln n$, the runtime of BKZ-β is polynomial. Following the reasoning from the previous section, for $k = n - \sqrt{n}$, the success condition becomes

$$\sqrt{1 - \frac{\ln n}{\sqrt{n}}} \leq \left(\frac{\ln n}{2\pi e}\right)^{1-\frac{\sqrt{n}}{2\ln n}} \cdot (\ln n)^{d\frac{n-\sqrt{n}}{\sqrt{n}}}.$$

The exponent on $\ln n$ in the RHS above is given by

$$1 - \frac{\sqrt{n}}{2\ln n} + \frac{d(n-\sqrt{n})}{\sqrt{n}} = (1-d) + \sqrt{n}\left(d - \frac{1}{2\ln n}\right),$$

which is positive for large enough values of n, meaning that the inequality is true asymptotically.

We now compute the asymptotic value of $p_k(\gamma)$ from Equation 4.6. For this we use Stirling's formula: $\ln(x!) = x\ln x - x + \frac{1}{2}\ln x + O(1)$. Recall that we ignore any rounding issues.

With $\delta = (\ln n)^{2d}$ and $d = n - k$, we have

$$p_k(\gamma) = \binom{n-\delta}{k} \cdot \binom{n}{k}^{-1} = \frac{(n-\delta)!}{n!} \cdot \frac{(n-k)!}{(n-\delta-k)!} = \frac{(n-\delta)!}{n!} \cdot \frac{d!}{(d-\delta)!}.$$

Now $\ln p_k(\gamma) = \ln\left(\frac{(n-\delta)!}{n!}\right) + \ln\left(\frac{d!}{(d-\delta)!}\right)$, and

$$\begin{aligned}
\ln\left(\frac{(n-\delta)!}{n!}\right) &= \ln((n-\delta)!) - \ln(n!) \\
&= (n-\delta)\ln(n-\delta) - (n-\delta) + \frac{1}{2}\ln(n-\delta) - n\ln n + n - \frac{1}{2}\ln n + O(1) \\
&= n\ln(1-\delta/n) - \delta\ln(n-\delta) + \delta + \frac{1}{2}\ln(1-\delta/n) + O(1) \\
&= -\delta + o(\delta) - \delta\ln n - \delta\ln(1-\delta/n) + \delta + o(\delta) \\
&= -\delta\ln n + o(\delta).
\end{aligned}$$

The same computation gives $\ln\left(\frac{d!}{(d-\delta)!}\right) = \delta \ln d + o(\delta)$. All in all we get

$$\ln p_k(\gamma) = \delta(\ln d - \ln n) + o(\delta) = -\frac{1}{2}(\ln n)^{2d+1} + o\left((\ln n)^{2d}\right),$$

which means that we expect at most $p_k(\gamma)^{-1} = \Theta\left(n^{(\ln n)^{2d}/2}\right)$ runs of the algorithm before $\mathbf{e} \in L$. This is quasi-polynomial in n.

Claim 4.4.7 improves on [Jia+23, Corollary 3.1] applied to $\gamma = O((\ln n)^d)$, however more work should be done to determine the exact improvement.

In practice

Both of our analyses in the cases $\gamma = O(1)$ and $\gamma = O(\ln n)$ were conducted heuristically under the GSA, and lead to asymptotic statements. In practice those asymptotic values do not mean much, and it might be preferable to optimise for k and β numerically. Given k and β, solving the equality case for Equation 4.8 using Equation 4.9 to express the right hand side allows us to build an estimator of the cost of solving $\mathbb{Z}$SVP. This cost is given by the estimated number of restarts, which is roughly[2] $p_k(\gamma)^{-1}/n$, multiplied by the estimated cost of running BKZ-β reduction on a lattice of dimension $n - k$. The cost of BKZ can be estimated in various ways, practical [Alb+18] or theoretical [LN24]. This gives an approximate estimation of the cost of our attack for various parameters, and enables quick and practical selection of the parameters k and n, making it easy to compare our attack with others.

We add that in the same way as [MS01], the guessing step in our attack is inherently parallelisable, whereas this is not the case for lattice reduction in general. Independently, the behaviour of our algorithm that is particular to hypercubic lattices might help towards understanding the threshold phenomenon described in [BGPS23, Section 6.2] and [DPPvW22, Section 4.2].

[2] We divide by n to account for the n possible targets.

Chapter

5 On Provable Reduction of Near-Hypercubic Lattices

Abstract In the previous chapter, we have seen that heuristic attacks on rank n NTRU and hypercubic lattices require subroutines that find short vectors in lattices of dimensions $4n/9$ and $n/2$, respectively. In this chapter, we focus on provable algorithms. We show how blockwise reduction and duality can exploit lattices with special geometric properties, effectively reducing the required blocksize to provably solve the shortest vector problem to half of the lattice's rank, and in the case of the hypercubic lattice $\mathbb{Z}^n$, further relaxing the approximation factor of blocks to $\sqrt{2}$. Remarkably, these near-hypercubic lattices cover Falcon and most concrete instances of the NTRU cryptosystem: this is the first provable result showing that breaking NTRU lattices can be reduced to finding shortest lattice vectors in halved dimension, thereby providing a positive response to a conjecture of Gama, Howgrave-Graham and Nguyen at Eurocrypt 2006.

This chapter is largely based on material presented in the conference paper [BN24].

Chapter content

5.1 Introduction

Gama, Howgrave-Graham and Nguyen [GHN06] conjectured at Eurocrypt 2006 that the reduction of a $2n$-dimensional NTRU lattice could be reduced to that of an αn-dimensional lattice for some $\alpha < 2$. The hypercubic lattice $\mathbb{Z}^n$ was first studied by Szydlo [Szy03] twenty years ago, but it was only shown very recently to be significantly easier to reduce than generic lattices: one can recover an orthonormal basis of $\mathbb{Z}^n$ in time $2^{n/2+o(n)}$ using the algorithm of Bennett, Ganju, Peetathawatchai and Stephens-Davidowitz[1] [BGPS23], or, as shown by Ducas [Duc23] by using polynomially many calls to an oracle for the shortest vector problem (SVP) in dimension $n/2$, which also leads to an asymptotic running time of $2^{n/2+o(n)}$. In other words, solving $\mathbb{Z}$SVP can be reduced to solving SVP in dimension $n/2$.

[1] Note that the *semi-stable* variant of their algorithm [BGPS23, Cor. 5.5] also applies to NTRU lattices.

Our results

We introduce a new blockwise reduction algorithm, which is a variant of Ducas's algorithm [Duc23], itself a variant of Gama-Nguyen's slide reduction [GN08a]. The differences with Ducas's approach are twofold.

First, our algorithm is more general, as it is not restricted to $\mathbb{Z}^n$: it also applies to any lattice L such that the product of its first minimum with that of its dual lattice is small, namely $\lambda_1(L)\lambda_1(L^\vee) < 1 - \frac{1}{\mathrm{poly}(n)}$, where $\lambda_1(\cdot)$ and $L^\vee$ denote respectively the first minimum and the dual lattice. This condition is typically not satisfied for a generic lattice: Minkowski's inequality only implies that $\lambda_1(L)\lambda_1(L^\vee) = O(n)$. But it turns out to be satisfied by most instantiations of NTRU, because the symplectic property of NTRU uncovered by Gama *et al.* [GHN06] implies that $\lambda_1(L)\lambda_1(L^\vee) = \lambda_1(L)^2/q$ where q is the small modulus of the NTRU cryptosystem, and also equal to $\mathrm{vol}(L)^{2/\mathrm{rank}(L)}$. In the recent NTRU-HPS submission [Che+20] to NIST, we have $\lambda_1(L)^2/q < 1/2$ for all three parameter sets proposed, a condition that is necessary to ensure the absence of decryption failures. For the original NTRU [HPS98] from the 90s and for Falcon [Fou+19], this does not hold but can be taken care of by a mild heuristic assumption on the projection of secret vectors over random subspaces related to lattice reduction: similar yet stronger assumptions were made and checked in the context of lattice enumeration [GNR10]. Thus, we show that for the NTRU-HPS submission [Che+20], one can provably find a non-zero lattice vector at least as short as the secret key, by solving the shortest vector problem in a lattice of halved dimension. This is the first rigorous result showing that an NTRU lattice can be solved by working with SVP oracles in a smaller dimension than what is required for a generic lattice. It should not be confused with heuristic security estimates where the blocksize required to break the underlying system is heuristically estimated to be a fraction of the lattice dimension, as studied in Chapter 4.

Second, our algorithm improves that of Ducas in the case of $\mathbb{Z}^n$. Ducas required an exact or nearly-exact algorithm for SVP in dimension $n/2$, whereas our algorithm can tolerate an approximate-SVP algorithm in dimension $n/2$ with an approximation factor essentially $\sqrt{2}$. Intuitively, a factor $\sqrt{2}$ should make the problem easier, and the SVP challenges [SG] suggest that the problem is easier in practice. Eisenbrand and Venzin [EV22] note that the best sieving algorithms give a provable $2^{0.802n+o(n)}$-runtime algorithm for $O(1)$ approximations of the shortest vector, although the constant is larger than $\sqrt{2}$. There is currently no theoretical evidence that approximating SVP to within $\sqrt{2}$ is easier than solving exact SVP, but if ever it is strictly easier, such as solvable in time $2^{\alpha n+o(n)}$ for some $\alpha < 1$, we would immediately obtain an exponentially faster algorithm for the $\mathbb{Z}$LIP, running in time $2^{\alpha n/2+o(n)}$, which would beat all other known attacks, both provable and heuristic.

Technical overview

Our algorithm differs from Ducas's algorithm in two main aspects. First, we distinguish the primal and the dual lattice. Second, we change the termination condition: instead of densifying a certain sublattice until it becomes hypercubic, we check whether our current primal and dual sublattices include a shortest vector.

Ducas's analysis [Duc23] is based on a surprising upper bound $\sqrt{1-1/n}$ on the first minimum of projections of $\mathbb{Z}^n$ over certain subspaces. This upper bound is tight when the subspace is a hyperplane corresponding to the dual root lattice $A_{n-1}^\vee$. However, we show that the upper bound can be improved for certain lower-dimensional subspaces, which might be of independent interest, and allows us to relax the SVP oracle to an approximate-SVP oracle with factor essentially $\sqrt{2}$. More precisely, it is well-known that the expectation of the squared norm of the projection of a unit vector onto a k-dimensional random subspace of $\mathbb{R}^n$ is k/n. We show that the expectation of the squared norm of the projection of a random element of a fixed orthonormal basis of $\mathbb{R}^n$ onto a fixed k-dimensional subspace is also k/n. This allows us to replace the bound $\sqrt{1-1/n}$

by essentially $\sqrt{1/2}$ when $k \approx n/2$.

Roadmap

In Section 5.2, we first describe our blockwise reduction algorithm for near-hypercubic lattices, then specialise its analysis to NTRU. This algorithm uses SVP-oracles in half dimension. We then introduce in Section 5.3 a variation of the algorithm that reduces hypercubic lattices using only approximate SVP-oracles.

5.2 Blockwise Reduction of Near-Hypercubic Lattices

5.2.1 Provable Algorithm

Algorithm 2 Primal/dual reduction with blocksize of halved dimension

Require: A basis $B = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ of a lattice $\Lambda \subseteq \mathbb{Z}^m$, together with two upper bounds r and $r^\vee$ such that $\lambda_1(L) \leq r$ and $\lambda_1(L^\vee) \leq r^\vee$. L (resp. N) is the sublattice spanned by the first $\lfloor n/2 \rfloor$ (resp. $\lfloor n/2 \rfloor + 1$) vectors of B, *i.e.* $L = \mathcal{L}(\mathbf{b}_1, \ldots, \mathbf{b}_{\lfloor n/2 \rfloor})$. Keep in mind that L and N are updated naturally as B evolves.

Ensure: A short non-zero vector in Λ of norm $\leq r$ or a short non-zero vector in the dual $\Lambda^\vee$ of norm $\leq r^\vee$, or a basis B such that $\mathrm{vol}(L)$ is guaranteed to be small.

1: LLL-reduce B.
2: **while** $\mathrm{vol}(L)$ strictly decreases **do**
3: $\quad \mathbf{e} \leftarrow$ SVP-oracle(L) to check for short primal lattice vectors.
4: $\quad$ **if** $\|\mathbf{e}\| \leq r$ **then**
5: $\quad\quad$ Return $\mathbf{e}$.
6: $\quad$ **else**
7: $\quad\quad$ SVP-reduce(Λ/L) to reduce the second half of B modulo its first half.
8: $\quad$ **end if**
9: $\quad \mathbf{e}' \leftarrow$ SVP-oracle$(\Lambda^\vee \cap \mathrm{span}(N)^\perp)$ to check for short dual lattice vectors.
10: $\quad$ **if** $\|\mathbf{e}'\| \leq r^\vee$ **then**
11: $\quad\quad$ Return $\mathbf{e}'$.
12: $\quad$ **else**
13: $\quad\quad$ SVP-reduce$(N^\vee)$ to dual-reduce the first half of B: this is DSVP-reduction of the lattice N.
14: $\quad$ **end if**
15: **end while**
16: Return B.

Algorithm 2 can be viewed as a variant of Gama-Nguyen's slide algorithm [GN08a] and Ducas' algorithm [Duc23, Alg. 1]. However, it differs in a few ways, mainly:

- It is not specialised to $\mathbb{Z}^n$.
- The termination conditions are different: instead of uniquely focusing on the reduction task, our algorithm can also check for the presence of short vectors in the lattice Λ or its dual $\Lambda^\vee$. Indeed, the tests at Lines 4 and 10 are parametrised by values r and $r^\vee$, which will be specified in the case of NTRU lattices and hypercubic lattices. If the user knows that Λ and/or $\Lambda^\vee$ contains a short vector of a prescribed length, then he can change the values of r and $r^\vee$ accordingly, for example by setting $r = \lambda_1(\Lambda)$ and/or $r^\vee = \lambda_1(\Lambda^\vee)$ when the first minima are known.

- Lines 3 and 9 add an extra call to the SVP oracle, which provides a way to prematurely abort if the objective is to find a vector of Λ and/or $\Lambda^\vee$ of norm less than a fixed value. This is especially useful in the case of NTRU and hypercubic lattices where the first minimum is well-known.

- Unlike [GN08a; Duc23], our algorithm assumes no requirement on the parity of n.

We make an important remark on Algorithm 2, which explains why we view this reduction as a primal/dual reduction: Steps 9-14 are dual to Steps 3-8, in the sense that they are exactly Steps 3-8 if we replace the lattice Λ by its dual $\Lambda^\vee$, and the sublattice L by $\Lambda^\vee \cap \mathrm{span}(N)^\perp$.

The efficiency of the algorithm is based on the following key elementary result:

Lemma 5.2.1. *Assume that $\Lambda \subseteq \mathbb{Z}^m$. During a loop iteration, the sublattice L (at the beginning of a loop iteration) is transformed into L', after Step 14. Then:*

$$\frac{\mathrm{vol}(L')}{\mathrm{vol}(L)} = \lambda_1(\Lambda/L)\lambda_1(N^\vee), \tag{5.1}$$

where N is from Step 13. Furthermore, if the exact reduction oracles of Steps 7 and 13 are replaced by approximate-reduction with factor respectively γ and γ', then:

$$\frac{\mathrm{vol}(L')}{\mathrm{vol}(L)} \leq \gamma\gamma'\lambda_1(\Lambda/L)\lambda_1(N^\vee). \tag{5.2}$$

Proof. The sublattice L can only be changed by Step 13, which cannot change the sublattice N. Since $\mathrm{vol}(N) = \mathrm{vol}(L)\|\mathbf{b}_{k+1}^\star\|$, we are interested in $\|\mathbf{b}_{k+1}^\star\|$, which can only be changed by Steps 7 and 13. After Step 7, we have $\|\mathbf{b}_{k+1}^\star\| = \lambda_1(\Lambda/L)$. After Step 13, we have $1/\|\mathbf{b}_{k+1}^\star\| = \lambda_1(N^\vee)$. So $\|\mathbf{b}_{k+1}^\star\|$ changes from $\lambda_1(\Lambda/L)$ to $1/\lambda_1(N^\vee)$, which proves Equation 5.1. The inequality of Equation 5.2 follows from the definition of approximate reduction. □

Theorem 5.2.2. *Let $\Lambda \subseteq \mathbb{Z}^m$ be a rank n lattice. Assume that $\lambda_1(\Lambda)\lambda_1(\Lambda^\vee) < 1 - \varepsilon$ for some $\varepsilon = \frac{1}{\mathrm{poly}(n)}$. Then Algorithm 2 returns a non-zero vector of Λ with norm $\leq r$, or a non-zero vector of its dual $\Lambda^\vee$ with norm $\leq r^\vee$. The number of loop iterations from Step 3 till Step 14 is polynomial in the size of the input basis B and $1/\varepsilon$. The number of SVP oracle queries is linear in the number of loop iterations, and the dimension of the lattice in each oracle query is $\leq \lfloor n/2 \rfloor + 1$.*

Proof. $\Lambda \subseteq \mathbb{Z}^m$ implies that $\mathrm{vol}(L)^2 \in \mathbb{Z}$. $\ln \mathrm{vol}(L)$ is polynomially bounded by the size of the input basis B, and can only decrease with LLL reduction (Step 1). This means that the number of times $\mathrm{vol}(L)$ decreases by $1 - \varepsilon$ is polynomially bounded by the size of the input basis B and $1/\varepsilon$.

If $\|\mathbf{e}\| > r \geq \lambda_1(\Lambda)$, there exists $\mathbf{u} \in \Lambda$ such that $\|\mathbf{u}\| = \lambda_1(\Lambda)$ and $\mathbf{u} \notin L$, therefore $\lambda_1(\Lambda/L) \leq \|\mathbf{u}\| = \lambda_1(\Lambda)$. Similarly, if $\|\mathbf{e}'\| > r^\vee \geq \lambda_1(\Lambda^\vee)$, then $\lambda_1(N^\vee) \leq \lambda_1(\Lambda^\vee)$. Thus, if both $\|\mathbf{e}\| > r$ and $\|\mathbf{e}'\| > r^\vee$, then using our assumption, $\lambda_1(\Lambda/L)\lambda_1(N^\vee) \leq \lambda_1(\Lambda)\lambda_1(\Lambda^\vee) < 1-\varepsilon$. This implies by Lemma 5.2.1 that $\mathrm{vol}(L)$ decreases by at least $1 - \varepsilon$, which can only happen polynomially many times.

Thus, we will find, within polynomially many iterations, some $\mathbf{e} \in L \subseteq \Lambda$ such that $\|\mathbf{e}\| \leq r$ or some $\mathbf{e} \in N^\vee \subseteq \Lambda^\vee$ such that $\|\mathbf{e}'\| \leq r^\vee$.

By definition, each loop iteration makes four calls to an SVP oracle, and the underlying lattice has rank $\in \{\lfloor n/2 \rfloor, n - \lfloor n/2 \rfloor, n - \lfloor n/2 \rfloor - 1, \lfloor n/2 \rfloor + 1\}$. □

5.2.2 Application to NTRU and Falcon

In 2006, Gama, Howgrave-Graham and Nguyen [GHN06] showed that coordinate embedding NTRU lattices from the ring $\mathbb{Z}[X]/(X^n - 1)$ are proportional to symplectic lattices, which is a special case of isodual lattices, *i.e.* there is an isometry between the lattice and its dual. We derive the following property of NTRU lattices:

Theorem 5.2.3. *Let $\mathcal{R}$ be $\mathbb{Z}[X]/(X^n-1)$ or $\mathbb{Z}[X]/(X^n+1)$. Let $(f, g) \in \mathcal{R}^2$ be an* NTRU *secret key corresponding to parameters (q, n) and a lattice Λ obtained from the coefficient embedding. Then there is an explicit bijection $\phi : \Lambda \to q\Lambda^\vee$ which preserves the Euclidean norm, and which can be computed in polynomial time (in both directions). In particular,*

$$\lambda_1(\Lambda^\vee) = \frac{\lambda_1(\Lambda)}{q},$$

where $\lambda_1(\Lambda)^2 \leq \|\mathbf{f}\|^2 + \|\mathbf{g}\|^2$, where $(\mathbf{f}, \mathbf{g})$ is the coefficient embedding of (f, g).

Proof. Using row notation, it is not hard to show that Λ and $\Lambda^\vee$ are respectively generated by the bases B_Λ and $B_{\Lambda^\vee}$, where

$$B_\Lambda = \begin{pmatrix} qI_n & 0 \\ H & I_n \end{pmatrix} \text{ and } B_{\Lambda^\vee} = \begin{pmatrix} \frac{1}{q}I_n & -\frac{1}{q}H^T \\ 0 & I_n \end{pmatrix},$$

where H is circulant (resp. anti-circulant) in the coefficients of $h \in \mathcal{R}$ the public key corresponding to (f, g) if $\mathcal{R} = \mathbb{Z}[X]/(X^n - 1)$ (resp. $\mathcal{R} = \mathbb{Z}[X]/(X^n + 1)$). We claim that

$$\phi : \begin{cases} \Lambda & \to q\Lambda^\vee \\ (\mathbf{u}, \mathbf{v}) & \mapsto (\tilde{\mathbf{v}}, -\tilde{\mathbf{u}}) \end{cases}$$

is the desired isometry, where $\tilde{\mathbf{u}}$ is $\mathbf{u}$ in reverse order. Indeed, because of the circulant or anti-circulant nature of H, the i-th row of h is exactly the same as the $(n + 1 - i)$-th row of H^T in reverse order. The structure of $B_{\Lambda^\vee}$ relatively to B_Λ allows us to conclude that ϕ is a suitable candidate. This map ϕ can clearly be computed in polynomial time, in both directions. Finally, the inequality $\lambda_1(\Lambda)^2 \leq \|\mathbf{f}\|^2 + \|\mathbf{g}\|^2$ follows from the fact that (g, f) is a lattice vector. □

Thus, we can upper bound $\lambda_1(\Lambda^\vee)\lambda_1(\Lambda)$ by $\frac{1}{q}(\|\mathbf{f}\|^2 + \|\mathbf{g}\|^2)$. Table 5.1 gives the explicit value of this upper bound for three types of NTRU lattices: the ones of the NTRU submission to NIST [Che+20], the original NTRU cryptosystem [HPS98], and the NIST signature standard Falcon [Fou+19]. These three types differ from the distribution used for f and g:

- For the first two, f and g have ternary coefficients $\in \{0, \pm 1\}$ but the number of ± 1 differ for each type.
- For Falcon however, f and g no longer have ternary coefficients: instead, its coefficients follow a discrete Gaussian distribution. We used publicly-available key generation software to compute the typical value of $\|\mathbf{f}\|^2 + \|\mathbf{g}\|^2$.

In addition, all of these examples use the coefficient embedding version of NTRU. The first two use the ring $\mathbb{Z}[X]/(X^n - 1)$, and the third uses $\mathbb{Z}[X]/(X^n + 1)$, both of which fall into the scope of Theorem 5.2.3.

In Table 5.1, the green colour indicates that the upper bound is $< 1 - \varepsilon$ for some constant $\varepsilon > 0$, which makes Theorem 5.2.2 applicable: this is the case for all parameter sets of NTRU submission to NIST [Che+20], and for the toy parameter set of the original NTRU [HPS98]. If we run Algorithm 2 with input $r^2 = \|\mathbf{f}\|^2 + \|\mathbf{g}\|^2$ and $r^\vee = \frac{1}{q}\sqrt{\|\mathbf{f}\|^2 + \|\mathbf{g}\|^2}$ (where the exact value may be replaced by a good upper bound): this will return a nonzero vector in the primal

Upper bound on $\lambda_1(L)\lambda_1(L^\vee)$ for various NTRU parameters						
Lattice	N	q	$\|(\mathbf{f},\mathbf{g})\|^2$	$\lambda_1(L)\lambda_1(L^\vee)$	$\frac{1}{2}\lambda_1(L)\lambda_1(L^\vee)$	Approx factor
NIST-1 [Che+20]	509	2048	593	.2897	.1449	2.628
NIST-3 [Che+20]	677	2048	705	.3444	.1722	2.410
NIST-5 [Che+20]	821	2048	1057	.2581	.1291	1.969
Original toy [HPS98]	107	64	53	.8281	.4141	1.554
Original [HPS98]	167	128	161	1.258	.6289	1.261
	263	128	147	1.148	.5742	1.320
	503	256	575	2.246	1.123	N/A
Falcon-512 [Fou+19]	512	12889	16481	1.341	.6706	1.251
Falcon-1024 [Fou+19]	1024	12889	16487	1.342	.6708	1.250

Table 5.1: NTRU *parameters: the two filled-in columns determine whether Theorem 5.2.2 applies, theoretically or heuristically. The last column illustrates by how much we can relax the SVP-reduction used Steps 7 and 13 of Algorithm 2. When* $\|(\mathbf{f},\mathbf{g})\|^2$ *is not fixed by the specifications, we take the experimental median over* 1000 *instances.*

lattice at least as short as the secret key, using only an SVP oracle in halved dimension. Indeed, if ever a dual vector is returned, the isometry of Theorem 5.2.3 allows to transform the short dual vector into a short primal vector. Bare in mind that it is believed that the secret-key vectors are the shortest vectors of the NTRU lattice, but this has not been proved.

We explain the situation of the NTRU submission to NIST [Che+20]. To avoid decryption failures, the generation of $\mathbf{f}$ and $\mathbf{g}$ is such that $\|\mathbf{f}\|^2 + \|\mathbf{g}\|^2 \leq q/2$. In fact, we have $\|\mathbf{f}\|^2 \leq N$ and $\|\mathbf{g}\|^2 = q/8 - 2$. Thus:

$$\lambda_1(\Lambda^\vee)\lambda_1(\Lambda) \leq \frac{1}{q}(\|\mathbf{f}\|^2 + \|\mathbf{g}\|^2) \leq \frac{1}{2}.$$

On the other hand, the historical parameters of NTRU allowed decryption failures, which increased $\|\mathbf{f}\|$ and $\|\mathbf{g}\|$.

The red colour in Table 5.1 shows that the bound is not satisfied. However, there is a way to get around this issue, under a mild assumption, except for the largest parameter of original NTRU [HPS98]. Indeed, Theorem 5.2.2 uses an upper bound on $\lambda_1(\Lambda)\lambda_1(\Lambda^\vee)$ to actually upper bound $\lambda_1(\Lambda/L)$ and $\lambda_1(N^\vee)$, knowing that none of the n short vectors $\mathbf{s}_1, \ldots, \mathbf{s}_n$ related to the secret key, obtained by coefficient embedding of the $(x^i * f, x^i * g)$, belong to the sublattice L (and similarly for the dual, with respect to $N^\vee$). It follows that $\lambda_1(\Lambda/L) \leq \min_{1\leq i\leq n} \|\pi(\mathbf{s}_i)\|$, where π denotes the orthogonal projection over $\mathrm{span}(L)^\perp$. If $\mathrm{span}(L)^\perp$ was a random subspace, the expectation of $\|\pi(\mathbf{s}_i)\|^2$ would be $\|\mathbf{s}_i\|^2 \frac{1}{n} \dim \mathrm{span}(L)^\perp \approx \|\mathbf{s}_i\|^2 \frac{1}{2}$. This suggests to make the mild assumption that:

$$\lambda_1(\Lambda/L)\lambda_1(N^\vee) \leq \frac{\lambda_1(\Lambda)\lambda_1(\Lambda^\vee)}{2}.$$

If this assumption holds at each loop iteration, then the conclusions of Theorem 5.2.2 still hold: we will obtain a nonzero vector in the primal lattice at least as short as the secret key. The second to last column of Table 5.1 therefore shows an upper bound of $\frac{1}{2}\lambda_1(\Lambda)\lambda_1(\Lambda^\vee)$: it turns out that the upper bound is now always $<_{\mathrm{poly}} 1$, except for the largest parameter of original NTRU [HPS98]. If this product is $<_{\mathrm{poly}} 1$, then we can heuristically relax the SVP-reductions used in Steps 7 and 13 of Algorithm 2 to approximate-SVP-reductions with approximation

factor $<_{\text{poly}} \sqrt{\frac{2}{\lambda_1(\Lambda)\lambda_1(\Lambda^\vee)}}$. The rightmost column of Table 5.1 provides explicit values of the best approximation factors.

To summarise, Algorithm 2 provably returns a nonzero lattice vector at least as short as the secret key for all parameter sets of NTRU submission to NIST [Che+20], using only an SVP oracle in halved dimension. And it succeeds heuristically under a mild assumption, for Falcon [Fou+19] and all parameter sets of original NTRU [HPS98] except for one. This gives a positive answer to the conjecture of Gama *et al.* [GHN06]: the reduction of a $2n$-dimensional NTRU lattice can be reduced to that of a n-dimensional lattice[2]. In addition, half of the oracle calls of our algorithm still work with approximate reduction, up to constant approximation factors that increase as $\lambda_1(\Lambda)\lambda_1(\Lambda^\vee)$ decreases.

We provide an additional result regarding the self-dual nature of the NTRU modules, which we believe can be of independent cryptanalytic interest.

Theorem 5.2.4. *Any* NTRU *module is isomorphic to its dual module. Additionally, the canonical embedding* NTRU *lattice is isometric up to a scalar factor to its dual lattice.*

Proof. Let $\mathcal{R} = \mathbb{Z}[X]/P(X)$ for some unitary degree n polynomial $P \in \mathbb{Z}[X]$. Let $h \in \mathcal{R}$ and M_h be a NTRU module as defined in Section 3.3:

$$M_h := \{(u, v) \in \mathcal{R}^2 : hu \equiv v \mod q\mathcal{R}\}.$$

The dual module $M_h^\vee$ is defined as the set of module homomorphisms from M_h to $\mathcal{R}$. We have

$$M_h^\vee = \{(\alpha, \beta) \in (\mathbb{Q}[X]/P(X))^2 : \forall (u, v) \in M_h, \alpha u + \beta v \in \mathcal{R}\}.$$

Let $(\alpha, \beta) \in M_h^\vee$. Observe that $(0, q) \in M_h$. Therefore $q\beta \in \mathcal{R}$, and there exists $\beta' \in \mathcal{R}$ such that $\beta = \frac{1}{q}\beta'$. Now observe that $(1, h) \in M_h$. This gives $\alpha + \frac{1}{q}\beta' h \in \mathcal{R}$, from which we deduce that there also exists $\alpha' \in \mathcal{R}$ such that $\alpha = \frac{1}{q}\alpha'$, and $\frac{1}{q}(\alpha' + \beta' h) \in \mathcal{R}$. Let

$$L_h := \{(x, y) \in \mathcal{R}^2 : hy \equiv -x \mod q\mathcal{R}\},$$

then $(\alpha', \beta') \in L_h$, therefore $qM_h^\vee \subseteq L_h$. But clearly L_h and M_h are isomorphic via the map $\psi : (x, y) \mapsto (y, -x)$, so by examining the index of $qM_h^\vee$ in M_h we can conclude that M_h and $M_h^\vee$ are isomorphic via the map $\frac{1}{q}\psi$. Because the canonical embedding is a ring homomorphism, the second part of the statement follows directly from the shape of the isomorphism. □

Theorem 5.2.4 essentially says that any NTRU lattice can be turned in a self-dual version of itself by a simple change of embedding. Note that this isn't a generalisation of Theorem 5.2.3.

5.3 Reducing Hypercubic Lattices with Approx-SVP Oracles

5.3.1 Provable Algorithm

In this section, we specialise Algorithm 2 to the case of $\mathbb{Z}^n$, and allow to relax the exact-SVP oracle into an approximate-SVP oracle: Ducas [Duc23] was only able to relax his oracle for an approximation factor very close to 1, while we allow an approximation factor close to $\sqrt{2}$. Our improvement also leads to a speculative improvement over the $2^{n/2}$ running time, if approximating SVP to within a factor $\sqrt{2}$ is exponentially faster than solving SVP.

We first present our specialised algorithm: Algorithm 3 is basically Algorithm 2 with $r = r^\vee = 1$ and approximate oracles instead of exact oracles, with a different termination: since we want to obtain an orthonormal basis, we don't stop once a unit vector has been found, we

[2] In this sentence *reduction* and *reduced* have different meanings.

reduce the dimension of Λ by projection, and keep iterating Algorithm 2 until the rank becomes trivial.

Algorithm 3 An algorithm for $\mathbb{Z}$LIP with approximate-SVP oracles in dimension $n/2$.

Require: An approximation factor $\gamma \in [1, \sqrt{2-2/n})$. A basis B of $\Lambda \simeq \mathbb{Z}^n$. L (resp. N) is the lattice spanned by the first $\lfloor n/2 \rfloor$ (resp. $\lfloor n/2 \rfloor + 1$) vectors of B.
Ensure: O an orthonormal basis of Λ.

$O = \{\}$
LLL-reduce B
while $\dim(B) > 0$ **do**
  if γ-SVP-oracle(L) returns a vector $\mathbf{e}$ such that $\|\mathbf{e}\| = 1$ **then**
    $O \leftarrow O \cup \{\mathbf{e}\}$.
    $B \leftarrow \mathrm{LLL}(\pi_{\mathbf{e}^\perp}(B))$ (update L and N accordingly).
  else
    γ-SVP-reduction-oracle(Λ/L) to reduce the second half of B modulo its first half.
  end if
  if γ-SVP-oracle($(\Lambda^\vee/N)^\vee$) returns a vector $\mathbf{e}'$ such that $\|\mathbf{e}'\| = 1$ **then**
    $O \leftarrow O \cup \{\mathbf{e}'\}$.
    $B \leftarrow \mathrm{LLL}(\pi_{\mathbf{e}'^\perp}(B))$ (update L and N accordingly).
  else
    γ-SVP-reduction-oracle($N^\vee$) to dual-reduce the first half of B.
  end if
end while
Return O.

The main result in this subsection is the following:

Theorem 5.3.1. *Given as input a basis B of $\Lambda \simeq \mathbb{Z}^n$ and given access to a γ-SVP approximation oracle in dimension $\lfloor n/2 \rfloor + 1$ where $\gamma \in \left[1, \sqrt{2 - \frac{2}{n}}\right)$, Algorithm 3 returns an orthonormal basis of Λ in polynomial time.*

We briefly compare Algorithm 3 with Ducas's algorithm [Duc23]. Ducas's algorithm restricts to a hypercubic lattice of odd dimension: the algorithm keeps reducing until the "half-sublattice" L (the sublattice generated by the first half of the current basis) generates a hypercubic lattice. Instead, Algorithm 3 checks using an approximate SVP oracle whether the "half-sublattice" L or its dual counterpart contains a unit vector: if not, the first minimum of L is > 1, which allows us to better upper bound the first minimum of Λ/L or its dual counterpart, compared to [Duc23, Lem. 4]. If ever a unit vector is discovered, we can decrement the lattice rank by projection, which also means that our algorithm must not be sensitive to the parity of the rank. The key to our improvement is the following technical result on random projections, which might be of independent interest.

Projecting an orthonormal basis

It is well-known that the expectation of the squared norm of the projection of a unit vector onto a k-dimensional random subspace of $\mathbb{R}^n$ is $\frac{k}{n}$. The following elementary lemma shows that the expectation of the squared norm of the projection of a random element of a fixed orthonormal basis of $\mathbb{R}^n$ onto a fixed k-dimensional subspace is also $\frac{k}{n}$.

Lemma 5.3.2. *Let $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ be an orthonormal basis of $\mathbb{R}^n$. Let π be the orthogonal projection over a k-dimensional subspace F of $\mathbb{R}^n$. Then:*

$$\sum_{i=1}^{n} \|\pi(\mathbf{e}_i)\|^2 = k.$$

Proof. Let $(\mathbf{f}_1, \ldots, \mathbf{f}_k)$ be an orthonormal basis of F. Then for each $1 \leq i \leq n$:

$$\|\pi(\mathbf{e}_i)\|^2 = \sum_{j=1}^{k} \langle \mathbf{e}_i, \mathbf{f}_j \rangle^2.$$

Therefore:

$$\sum_{i=1}^{n} \|\pi(\mathbf{e}_i)\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \mathbf{e}_i, \mathbf{f}_j \rangle^2 = \sum_{j=1}^{k} \sum_{i=1}^{n} \langle \mathbf{e}_i, \mathbf{f}_j \rangle^2 = \sum_{j=1}^{k} 1 = k,$$

because each $\mathbf{f}_j$ is a unit vector and $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ is an orthonormal basis of $\mathbb{R}^n$. □

The previous lemma allows us to upper bound the first minimum of the projection of a hypercubic lattice, as follows:

Corollary 5.3.3. *Let L be a primitive sublattice of rank $1 \leq k < n$ of a full-rank hypercubic lattice Λ of $\mathbb{R}^n$ such that $\lambda_1(L) \geq \sqrt{2}$. Then $\lambda_1(\Lambda/L)^2 \leq 1 - \frac{k}{n}$.*

Proof. L is primitive so Λ/L is a lattice and $\lambda_1(\Lambda/L)$ is well-defined. Let π be the orthogonal projection onto the $(n-k)$-dimensional subspace $L^\perp$. We know that Λ has an orthonormal basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$: this is also an orthonormal basis of $\mathbb{R}^n$ so the lemma shows that

$$\sum_{i=1}^{n} \|\pi(\mathbf{e}_i)\|^2 = n - k$$

Furthermore, note that all the $\pi(\mathbf{e}_i)$'s are nonzero: if $\pi(\mathbf{e}_i) = 0$ for some i, then $\mathbf{e}_i \in L$ because L is primitive, then $\lambda_1(L) \leq 1$, which contradicts $\lambda_1(L) \geq \sqrt{2}$. Therefore there exists an integer $i \in \{1, \ldots, n\}$ such that $0 < \|\pi(\mathbf{e}_i)\|^2 \leq \frac{n-k}{n}$. Hence $\lambda_1(\Lambda/L)^2 \leq 1 - \frac{k}{n}$. □

In other words, under certain conditions over L, we can decrease Ducas [Duc23]'s upper bound $\sqrt{1 - 1/n}$ to $\sqrt{1 - k/n}$, which is better as soon as $k \geq 2$: we note that for $k = 1$, Ducas [Duc23]'s upper bound is actually tight for L spanned by the all-one vector $(1, 1, \ldots, 1)$, which means that Λ/L is the dual root lattice $A_{n-1}^\vee$. We are now ready for the proof of Theorem 5.3.1, which is very similar to that of Theorem 5.2.2: we simply combine Lemma 5.3.2 with (5.2) of Lemma 5.2.1.

of Theorem 5.3.1. $\Lambda \simeq \mathbb{Z}^n$ implies that $\mathrm{vol}(L)^2 \in \mathbb{Z}$. Because $\mathrm{vol}(\Lambda) = 1$ and well-known properties of LLL reduction, Step 2 guarantees $\ln \mathrm{vol}(L) = O(n^2)$. This means that the number of times $\mathrm{vol}(L)$ decreases by a factor $1 - \varepsilon$ (without changing Λ) is $O(n^2/\varepsilon)$.

We have $n = 2k$ or $n = 2k + 1$ where $k = \lfloor n/2 \rfloor$. We let L be the primitive lattice spanned by $(\mathbf{b}_1, \ldots, \mathbf{b}_k)$.

Consider Step 4. if $\|\mathbf{e}\| < \sqrt{2}$, then $\|\mathbf{e}\| = 1$ because Λ has no vector of norm in the interval $(1, \sqrt{2})$. So we recovered a shortest vector $\mathbf{e}$ of Λ, and Step 6 iterates the algorithm, by projecting Λ over the hyperplane orthogonal to $\mathbf{e}$: this is a hypercubic lattice of rank $n - 1$, and we have to recompute an LLL-reduced basis.

Otherwise, $\|\mathbf{e}\| \geq \sqrt{2}$. We deduce that $\lambda_1(L) > 1$, as otherwise $\lambda_1(L) = 1$ because $\lambda_1(\Lambda) = 1$, which would contradict $\|\mathbf{e}\| \leq \gamma$. But $\lambda_1(L) > 1$ implies that $\lambda_1(L) \geq \sqrt{2}$ because Λ has no vector of norm in the interval $(1, \sqrt{2})$. So Corollary 5.3.3 shows that $\lambda_1(\Lambda/L)^2 \leq 1 - \frac{k}{n}$.

The remaining steps are the dual counter part. So if $\|\mathbf{e}'\| \geq \sqrt{2}$ in Step 10, we deduce similarly by applying Corollary 5.3.3 to the sublattice $\Lambda^\vee \cap \mathrm{span}(N)^\perp$ of rank $n - (k+1)$, that $\lambda_1(N^\vee)^2 = \lambda_1(\Lambda^\vee/(\Lambda^\vee \cap \mathrm{span}(N)^\perp))^2 \leq 1 - \frac{n-(k+1)}{n}$. We thus have proved:

$$\lambda_1(\Lambda/L)\lambda_1(N^\vee) \leq \sqrt{1 - \frac{k}{n}}\sqrt{1 - \frac{n-(k+1)}{n}} = \frac{\sqrt{(n-k)(k+1)}}{n}.$$

If $n = 2k$, then:

$$\frac{\sqrt{(n-k)(k+1)}}{n} = \frac{1}{2}\sqrt{1+\frac{2}{n}} = \frac{1}{2}\left(1+\frac{1}{n}-\frac{1}{2n^2}+O\left(\frac{1}{n^3}\right)\right).$$

Otherwise, $n = 2k+1$ and:

$$\frac{\sqrt{(n-k)(k+1)}}{n} = \frac{k+1}{n} = \frac{1}{2}\left(1+\frac{1}{n}\right).$$

Since $\gamma^2 < 2-\frac{2}{n}$, (5.2) of Lemma 5.2.1 implies that, unless we find a unit vector, $\mathrm{vol}(L)$ decreases by at least $\left(1-\frac{1}{n}\right)\left(1+\frac{1}{n}\right) = 1-\frac{1}{n^2}$. Thus, within polynomially many iterations, we will find a unit vector $\mathbf{e}$ or $\mathbf{e}'$. Since there are only n unit vectors, we find all of them within polynomially many iterations. □

A consequence of Theorem 5.3.1 is the following speculative Corollary, that would break the $n/2$ barrier for $\mathbb{Z}$LIP as long as $\sqrt{2}$-approx SVP is exponentially easier than its exact counterpart.

Corollary 5.3.4. *Let $\alpha < 1$ be a constant. If there exists an algorithm for approx-SVP with approximation factor $\sqrt{2-2/n}$ that runs in time $2^{\alpha n+o(n)}$, then there also exists an algorithm for $\mathbb{Z}$LIP that runs in time $2^{\alpha n/2+o(n)}$.*

5.3.2 An Attempt at Dimensions for Free for Approximate-SVP

Aside from being visibly easier in practice, γ-SVP has been shown to be exponentially easier than exact SVP for some larger constant approximation factors (Th. 3.2 of [EV22]), this gives some evidence as to why the premise of Corollary 5.3.4 might be true.

Based on current knowledge, it is widely accepted that the fastest SVP oracles in cryptographically relevant dimensions are obtained via sieving algorithms[NV08; Alb+19]. As sieving algorithms generally produce not only the shortest vector, but roughly all vectors of norm less than $(4/3)^{n/2}\mathrm{GH}(L)$ in a rank n lattice L, a neat trick discovered by Ducas in [Duc18] is to sieve in a projected lattice in dimension $n-d$, hoping that the projection of the shortest vector in L projects to one of the vectors obtained by sieving in the projected lattice. Assuming d is quite small, lifting will recover the shortest vector with high probability. The analysis of [Duc18] concludes that the first order term in the asymptotic expansion of the largest d that works is given by

$$d = \frac{n}{\ln n}\ln 4/3 + o\left(\frac{n}{\ln n}\right). \tag{5.3}$$

In this case, we say that we have gained d *dimensions for free*.

In this subsection, we provide a heuristic asymptotic analysis of the natural generalisation of the *dimensions for free* technique to approximate-SVP oracles with a constant factor. Indeed one can reasonably hope to decrease the sieving dimension further if the target vector length is increased, as there are more valid targets after the lifting step. However we conclude that under our usual heuristic assumption on the distribution of target vectors, there is no asymptotic improvement.

We recall the reasoning made in [Duc18], which we adapt to a γ-SVP oracle for a constant[3] γ. Without loss of generality, we assume our rank n lattice L has volume 1, and assume we know a BKZ-$(n-d)$ reduced basis. This preprocessing is reasonable, as it requires at worst $\mathrm{poly}(n)$ calls to a $(n-d)$-dimensional SVP oracle. We then recall that π_{d+1} denotes the orthogonal projection on $\mathrm{span}(\mathbf{b}_1,\ldots,\mathbf{b}_d)^\perp$.

[3]The non-constant case is also interesting and can be treated in the same way.

We hope that a short vector $\mathbf{s} \in L$ of expected length at most $\gamma \cdot \mathrm{GH}(L)$ projects to a vector recovered by the sieve in $\pi_{d+1}(L)$. As $\|\pi_{d+1}(\mathbf{s})\| \leq \|\mathbf{s}\|$, it is sufficient that

$$\gamma \cdot \mathrm{GH}(L) \leq \sqrt{4/3} \cdot \mathrm{GH}(\pi_{d+1}(L)). \tag{5.4}$$

In fact we expect $\pi_{d+1}(\mathbf{s})$ to be shorter than $\mathbf{s}$ by a factor $\sqrt{\frac{n-d}{n}}$, which could be included in the left hand side of Equation 5.4 to make it less tight.

Computing $\mathrm{GH}(\pi_{d+1}(L))$ requires the volume of the projected lattice, which can be estimated roughly using the GSA and the following expression:

$$\mathrm{vol}(\pi_{d+1}(L)) = \prod_{i=d+1}^{n} \|\mathbf{b}_i^\star\| = \delta_\beta^{\frac{nd(d-n)}{n-1}},$$

where we will assume $\beta = n - d$.

For $\gamma < \sqrt{4/3}$, we derive from Equation 5.4 that the largest d that works satisfies $d \sim \frac{n}{\ln n} \ln \sqrt{4/(3\gamma^2)}$. This in fact is worse than what we obtain for $\gamma = 1$, as the value of d decreased. What went wrong? We forgot to account for the fact that there is not one, but approximately[4] γ^n vectors of norm at most $\gamma \cdot \mathrm{GH}(L)$. Geometrically, we are looking for the probability that at least one of γ^n points of the ball of radius $\gamma \cdot \mathrm{GH}(L)$ projects into the ball of radius $\sqrt{4/3} \cdot \mathrm{GH}(L_d)$. For the sake of continuity, we will assume that all γ^n points are uniform of the sphere of radius $\gamma \cdot \mathrm{GH}(L)$. It would be more accurate to use a uniform distribution on the ball, but we prefer this choice as it enables us to reuse results from Section 4.3.

Proposition 5.3.5. *Let $\mathbf{s}_1, \ldots, \mathbf{s}_N$ be $N = \gamma^n$ uniform points on the n-dimensional space sphere of radius $\gamma \cdot \sqrt{n/(2\pi e)}$. Let π be a projection orthogonal to a fixed random d-dimensional subspace. Then if $\gamma \geq 1$ is a constant and*

$$\Pr\left(\min_{1 \leq i \leq N} \|\pi(\mathbf{s}_i)\| \leq \sqrt{\frac{4}{3}} \cdot \sqrt{\frac{n-d}{2\pi e}} \cdot \left(\frac{n-d}{2\pi e}\right)^{-\frac{d}{2(n-d)}}\right) = \Theta(1), \tag{5.5}$$

then $d = \frac{n}{\ln n} \ln 4/3 + o\left(\frac{n}{\ln n}\right)$.

Proof. The proof reuses the exact same ideas as that of Proposition 4.3.4. We first note from Lemma 4.3.2 that the probability from Equation 5.5 can be written as

$$1 - \left(1 - I_x\left(\frac{n-d}{2}, \frac{d}{2}\right)\right)^N = \Theta(1), \tag{5.6}$$

where x is given by the squared ratio of the radii of the two balls (in the lattice setting, the radii would be $\sqrt{4/3} \cdot \mathrm{GH}(\pi_d(L))$ and $\gamma \cdot \mathrm{GH}(L)$):

$$x = \frac{4}{3\gamma^2} \cdot \frac{n-d}{n} \cdot \left(\frac{n-d}{2\pi e}\right)^{-d/(n-d)}.$$

We can now use Lemma 4.3.3 inside Equation 5.6 to get

$$\frac{c}{\gamma^n} \sim I_x\left(\frac{n-d}{2}, \frac{d}{2}\right) \sim \frac{1}{2}\mathrm{erfc}\left(-\frac{\eta\sqrt{n}}{2}\right),$$

where c is a positive constant, and η is defined as in Lemma 4.3.3, where we use α to denote $\alpha n = \beta = n - d$. From this we are able to derive that $\eta^2 \sim 4 \ln \gamma$, which leads to the following variation of Equation 4.5:

$$2 \ln \gamma \sim \alpha \ln \alpha + (1-\alpha) \ln (1-\alpha) - \alpha \ln x - (1-\alpha) \ln (1-x), \tag{5.7}$$

[4] This uses the Gaussian heuristic, which not true for all lattices.

which leads to $\alpha = 1 - \frac{\xi}{\ln n}$, where $\xi = O(1)$. The exact value of ξ can be derived from Equation 5.7 as $\alpha \ln \alpha \to 0$, $(1-\alpha)\ln(1-\alpha) \to 0$, and $(1-\alpha)\ln(1-x) \to 0$. Therefore we know that $2\ln\gamma \sim -(1-\frac{\xi}{\ln n})\ln(x) \sim \ln(x^{-1})$. Expanding and only keeping the main term we finally get

$$\xi = 2\ln\gamma - \ln(3\gamma^2/4) = \ln(4/3).$$

This concludes our proof. □

Remark 5.3.6. The direct use of Lemma 4.3.3 in the previous proof is in fact flawed, as the value of α that is taken is not bounded away from 1. However, the fact that α is bounded away from 1 is not required by [Tem92] for the first term of the expansion. Deriving the second term would therefore require computations that are even more tedious than what was required for the first one.

From Proposition 5.3.5, we conclude that relaxing the dimensions for free technique to approximate-SVP oracles for a constant approximation should not lead to any major asymptotic gain, as the number of *free* dimensions remains of the order of $\frac{n}{\ln n}\ln 4/3$. However this result is asymptotic and does not exclude a potential practical gain, as was observed in similar situations in the previous chapter.

Part III

Making and Breaking Digital Signatures with Lattices

Chapter

6 On a Signature Scheme of Feussner and Semaev

Abstract If lattice reduction is used today as the main cryptanalytic tool to attack lattice-based schemes, it can also apply to other schemes, for which lattices do not appear explicitly in the construction. We present a key-recovery attack on DEFI, an efficient signature scheme proposed recently by Feussner and Semaev, and based on isotropic quadratic forms, borrowing ideas from both multivariate and lattice cryptography. Our lattice-based attack is partially heuristic, but works on all proposed parameters: experimentally, it recovers the secret key in a few minutes, using less than ten (message, signature) pairs.

This chapter is largely based on material presented in the conference paper [BN25].

Chapter content

6.1 Introduction

The lattice-based signature schemes Dilithium [Duc+21] and Falcon [Fou+19] have been selected by the NIST [NIS22b] as the first standards for post-quantum cryptography. But this post-quantum security comes at a cost: the size of both the public key and the signature of Dilithium and Falcon are significantly bigger than for ECDSA and RSA. It would be useful to have more

efficient post-quantum signature schemes and/or based on different assumptions: this motivated the NIST to open a call for additional digital signature proposals [NIS22a] in 2022. In that call, Feussner and Semaev submitted the lattice-based signature scheme EHTv3v4 [FS23], which currently remains unbroken after a fix. Very recently [FS24a], the same authors proposed a very different and much more efficient scheme, called DEFI, on the NIST pqc mailing list: with a 800-byte public key and a 432-byte signature, DEFI is more efficient than both Dilithium and Falcon, and beats all additional NIST submissions except for SQISign in (public key + signature) size [PQS24]. See Figure 6.1 for an illustration. Even with a non-optimised implementation, DEFI's signature and verification times seem to compare favourably to all proposed signatures [Clo24]. DEFI is a peculiar scheme borrowing from both multivariate cryptography and lattice-based cryptography: its security is based on the hardness of solving systems of quadratic equations over the integers and a polynomial ring R such as $\mathbb{Z}[X]/(X^{64}+1)$. In its general form, this problem is known to be NP-hard, and therefore the authors of DEFI assumed it hard in the worst case, but DEFI uses special instances of the problem, which might be much easier to solve.

More precisely, a DEFI private key is a solution to a small system of quadratic equations over R, determined by the DEFI public key. Because R is a polynomial ring, this small system can be transformed into a large system of quadratic equations over $\mathbb{Z}$, which in general would be an NP-hard problem. DEFI is a hash-and-sign probabilistic scheme: the signature of a hashed message h is simply a randomly-generated solution to a small system of quadratic equations over R, in such a way that the first entry of the solution vector is h, and the other entries depend on the choice of a nonce, a one-time key required for each signature generation. Surprisingly, DEFI does not use modular arithmetic nor finite fields: all operations are in the polynomial ring R, and this was a security argument in [FS24a]. Feussner and Semaev analysed [FS24a] several attacks on DEFI to argue that their scheme DEFI was immune against Gröbner basis attacks and lattice attacks. They proposed a 64-bit numerical challenge, and concrete parameters for which they conjectured a 128-bit security level.

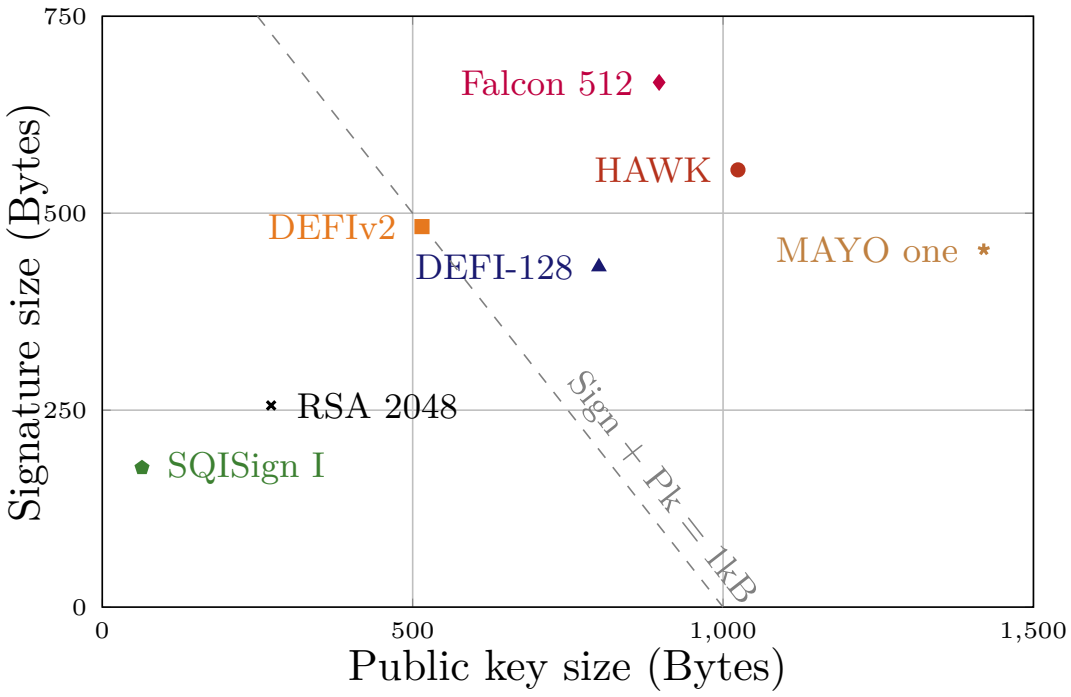


Figure 6.1: Performances of the DEFI *signature scheme.*

Our results

We show that DEFI is completely insecure: experimentally, less than ten (message, signature) pairs are sufficient to recover the secret key in a few minutes, for all parameters proposed in [FS24a], including the 64-bit challenge.

Disclaimer

In reaction to the break of DEFI presented here, the authors of the scheme proposed DEFIv2 [FS24b], a new and improved version of the signature scheme, boasting similarly impressive performances. This work was written before DEFIv2 was announced and thus only focuses on the cryptanalysis of DEFI. The attack presented here does not seem to be directly applicable to DEFIv2.

Technical overview

The starting point of our attack is that each DEFI signature leaks information on the secret key, and we exploit that leakage: each signature provides a linear equation over R, whose solutions are related to the secret key. By collecting enough signatures, we obtain a linear system of equations over R: this gives rise to a lattice whose rank is independent of the number of signatures used, but for which we know an unusually short vector related to the private key and the secret nonces which were used to generate each signature. The more signatures we use, the more unusually short the secret vector becomes, without affecting the rank of the lattice.

By reducing this lattice, we heuristically obtain this unusually short vector. At this point, we cannot yet recover the secret key. However, it allows us to derive a new linear system of equations over R: this gives rise to a second lattice, whose rank depends on the number of signatures. We know that this lattice contains another very short vector, which is directly related to the private and the secret nonces which were used to generate each signature. Again, if we take more signatures into account, then this second very short secret vector becomes even shorter, relatively to what one would expect from a typical lattice. However, the lattice rank increases with the number of signatures used in this second stage, making lattice reduction increasingly expensive. We circumvent this issue by noting the existence of and heuristically recovering an unusually dense sublattice of much smaller rank that contains the targeted second secret short vector. We then reduce the recovered sublattice to obtain the second secret.

Together, the two unusually short vectors that were recovered provide a final system of linear equations over $\mathcal{R}$, whose solutions are exactly the secret key and its rotations. If enough signatures are given, we obtain a linear system over $\mathbb{Z}$ for which there are many more equations than unknowns: this recovers the secret key in polynomial time by linear algebra, provided that the linear system is full-rank. Alternatively, the structure of the two unusually short vectors allows us to recover the secret key efficiently by an ad-hoc process, based on the equation relating the public key and the secret key.

To summarise, there are three stages in the attack: the first two stages use lattice reduction to find extremely short vectors, but the rank of the lattice used in the first stage is independent of the number of signatures used. The final third stage recovers the private key without lattice techniques.

Related work

The authors of DEFI [FS24a] also considered lattice attacks, but showed that their attacks failed. However, their attacks were different from our attack, even though their attacks exploited the same equations that we are using. These attacks failed because of two reasons. The first reason is that [FS24a] used a different lattice, whose dependence on the secret key was much less useful:

in this lattice, there was apparently no unusually short lattice vector related to the secret key. The second reason is that [FS24a] only considered attacks using a single signature.

Roadmap

Section 6.2 introduces a the Quadratic Form Equivalence (QFE) problem as well as its module version, which we view as generalisations of Lattice and module Lattice Isomorphism Problem (LIP). Section 6.3 presents the DEFI signature scheme. We describe our attack in Section 6.4, give some elements of justification in Section 6.5, and present our experimental results in Section 6.6.

6.2 The Quadratic Form Equivalence Problem

Definition 6.2.1. Let K be a field and V a vector space. A quadratic form is an application $Q : V \to K$ such that there exists a symmetric bilinear form $B : V \times V \to K$ such that

$$\forall u \in V,\ \ Q(u) = B(u, u).$$

If V is an n-dimensional real vector space with basis $(\mathbf{e}_i)_{1 \leq i \leq n}$, then Q can be represented in matrix form by the symmetric matrix $\mathbf{Q} \in S_n(K)$, where the entries are defined by $\mathbf{Q}_{ij} = B(\mathbf{e}_i, \mathbf{e}_j)$. In this case, $Q(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x}$.

When using the term *quadratic form*, we will interchangeably speak of applications and symmetric matrices.

Using the notations of Definition 6.2.1 in the setting of real lattices, we have $V = \mathbb{R}^n$ and $K = \mathbb{R}$. The lattice basis $\mathbf{B} \in \mathrm{GL}_n(\mathbb{R})$, naturally defines a quadratic form through its associated Gram Matrix $\mathbf{G} = \mathbf{B}^T\mathbf{B}$.

Quadratic forms associated to lattices are positive definite, which means that for any vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{G}\mathbf{x} \geq 0$, with equality if and only if $\mathbf{x}$ is zero. Two bases $\mathbf{B}_1$ and $\mathbf{B}_2$ represent the same quadratic form $\mathbf{G} = \mathbf{B}_1^T\mathbf{B}_1 = \mathbf{B}_2^T\mathbf{B}_2$ if and only if there exists an linear isometry $\mathbf{O} \in \mathcal{O}_n(\mathbb{R})$ such that $\mathbf{B}_1 = \mathbf{O} \cdot \mathbf{B}_2$. This is reminiscent of the Lattice Isomorphism problem, in which two lattices are considered isomorphic if one is a linear isometry of the other. LIP has an equivalent formulation in terms of integral equivalence of Gram matrices.

Definition 6.2.2 (LIP,[DvW21])**.** Let $\mathbf{Q} \in S_n(\mathbb{Z})$ be a positive definite quadratic form. Given another positive definite quadratic form $\mathbf{Q}'$ that is equivalent to $\mathbf{Q}$, find a unimodular $\mathbf{U} \in \mathrm{GL}_n(\mathbb{Z})$ such that $\mathbf{Q}' = \mathbf{U}^T\mathbf{Q}\mathbf{U}$. We say that two matrices $\mathbf{A} \in S_n(\mathbb{Z})$ and $\mathbf{B} \in S_n(\mathbb{Z})$ are (integrally) equivalent if there exists a $\mathbf{U} \in \mathrm{GL}_n(\mathbb{Z})$ such that $\mathbf{A} = \mathbf{U}^T\mathbf{B}\mathbf{U}$.

Remark 6.2.3. The choice of only considering Gram matrices with integral coordinates is deliberate as working with those matrices on a computer remove the need for complicated precision management, but it is not mathematically motivated.

Definition 6.2.2 only considers *positive definite* quadratic forms, which is natural as they correspond to lattice. In its general version, it is considered a sufficiently hard problem to be considered for post-quantum standardisation. However the classification of quadratic forms includes other types, and we are not aware of any immediate reason why the same problem could not be studied in general.

Definition 6.2.4 (Quadratic Form Equivalence (QFE))**.** Let $\mathbf{Q} \in S_n(\mathbb{Z})$ be a quadratic form. Given another quadratic form $\mathbf{Q}'$ that is equivalent to $\mathbf{Q}$, find a unimodular $\mathbf{U} \in \mathrm{GL}_n(\mathbb{Z})$ such that $\mathbf{Q}' = \mathbf{U}^T\mathbf{Q}\mathbf{U}$.

Clearly, QFE is a generalisation of LIP. It includes the problem of integral equivalence also in cases where the quadratic form is indefinite, *e.g.* when it takes both positive and negative values. If QFE in its general form allows for better performing cryptosystems, then this problem is worth studying, however it might also be the case that there exists a direct attack on QFE in the indefinite case that does not apply to LIP.

If one considers the coefficients of the unknown matrix $\mathbf{U}$ as variables, then $\mathbf{Q}' = \mathbf{U}^T\mathbf{Q}\mathbf{U}$ can be rewritten as a multivariate quadratic system of equations, hinting towards the fact that QFE and MQ[1] are related problems, and some attacks from multivariate cryptography might transpose well to the setting of QFE, that happens to include LIP.

The signature scheme HAWK [DPPvW22] which we have already mentioned in previous chapters relies on a variation of LIP with extra structure: module-LIP. This problem considers quadratic forms with $V = K^r$ and K a number field with complex conjugation, where we are using notations from Definition 6.2.1. The correct notion here is that of Hermitian forms, which are exactly the forms associated to matrices $\mathbf{Q} \in K^{r\times r}$ that are Hermitian, i.e. such that $\overline{\mathbf{Q}}^T = \mathbf{Q}$. If the form is positive definite in the sense that $\mathrm{Tr}(\overline{\mathbf{x}}^T\mathbf{Q}\mathbf{x}) > 0$ for $\mathbf{x} \in K^r\backslash\{\mathbf{0}\}$, then the form corresponds to a (module) lattice. $r \in \mathbb{Z}_{>1}$ is the called the (module) rank. Rank $r = 1$ is excluded as is vulnerable to the Gentry-Szydlo algorithm [GS02].

Definition 6.2.5 (module-LIP, [DPPvW22])**.** Let $r \in \mathbb{Z}_{>1}$ and K be a number field with ring of integers $\mathcal{O}_K$. Let $\mathbf{Q} \in \mathcal{O}_K^{r\times r}$ be a Hermitian matrix such that $\mathrm{Tr}(\overline{\mathbf{x}}^T\mathbf{Q}\mathbf{x}) > 0$ for all $\mathbf{x} \in K^r\backslash\{\mathbf{0}\}$. Given a matrix $\mathbf{Q}' \in \mathcal{O}_K^{r\times r}$ that is equivalent to $\mathbf{Q}$, find a unitary $\mathbf{U} \in \mathrm{GL}_n(\mathcal{O}_K)$ such that $\mathbf{Q}' = \overline{\mathbf{U}}^T\mathbf{Q}\mathbf{U}$. We say that $\mathbf{A} \in \mathcal{O}_K^{r\times r}$ and $\mathbf{B} \in \mathcal{O}_K^{r\times r}$ are (integrally) equivalent if there exists a $\mathbf{U} \in \mathrm{GL}_n(\mathcal{O}_K)$ such that $\mathbf{A} = \overline{\mathbf{U}}^T\mathbf{B}\mathbf{U}$.

Remark 6.2.6. Our definition will be sufficient for the purpose of the chapter. For a better understanding of this problem, we encourage the reader to refer to the definitions given in [MPPW24] and [APvW25]. Indeed in the (free) module case, the Gram matrix is obtained from the module lattice basis via the canonical Hermitian product that is defined using the real and complex embeddings of the number field K. We note that in the case of fields that are neither totally real, nor CM, the problem used in DEFI differs slightly from the one defined in [APvW25], because of the absence of conjugation.

As previously with Definition 6.2.4, module-LIP can be made more general by dropping the condition that the Hermitian form should be positive definite.

Definition 6.2.7 (module-QFE)**.** Let $r \in \mathbb{Z}_{>1}$ and K be a number field with ring of integers $\mathcal{O}_K$. Let $\mathbf{Q} \in \mathcal{O}_K^{r\times r}$ be a Hermitian matrix. Given a matrix $\mathbf{Q}' \in \mathcal{O}_K^{r\times r}$ that is equivalent to $\mathbf{Q}$, find a unitary $\mathbf{U} \in \mathrm{GL}_n(\mathcal{O}_K)$ such that $\mathbf{Q}' = \overline{\mathbf{U}}^T\mathbf{Q}\mathbf{U}$.

6.3 The DEFI Signature Scheme

In [FS24a], Feussner and Semaev propose a new digital signature scheme called DEFI, based on solving systems of quadratic Diophantine equations over the rational integers.

6.3.1 Formal Definition of the Scheme

Let $q \in \mathbb{Z}[X]$ be a monic irreducible polynomial, and $R = \mathbb{Z}[X]/(q)$ its associated polynomial ring, where (q) denotes the ideal of $\mathbb{Z}[X]$ generated by q. Let $\mathbf{J} = \mathrm{diag}(1, 1, -1, -1) \in M_4(R)$. For $\mathbf{A} \in M_n(R)$, we define $f_\mathbf{A}$ by $f_\mathbf{A}(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ for $\mathbf{x} \in R^n$. In [FS24a], the authors seem to consider a wider array of matrices $\mathbf{J} = \mathrm{diag}(\pm1, \ldots, \pm1) \in M_n(R)$, where n can vary, but the instantiation of their scheme heavily relies on the specific choices $n = 4$ and $\mathbf{J} = \mathrm{diag}(1, 1, -1, -1)$.

[1] Although usually, variables in MQ are defined over a field.

Unless design choices are made radically different for another choice of $\mathbf{J}$, an adapted version of our attack would still apply.

Private key

The *private key* is a matrix $\mathbf{B} \in M_4(R)$, defined blockwise as

$$\mathbf{B} = \begin{pmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

where $\mathbf{B}_{21} \in R^3$ and $\mathbf{B}_{22} \in M_3(R)$. $\mathbf{B}$ should be invertible and $\mathbf{B}_{21}, \mathbf{B}_{22}$ and $\mathbf{B}_{22}^{-1}$ are taken with small norm (*i.e.* elements are polynomials of R with small coefficients). Note that invertibility of $\mathbf{B}$ implies it is unimodular. We refer to [FS24a] for the precise generation procedure for $\mathbf{B}$. While the condition on the size of $\mathbf{B}$ in [FS24a] is slightly different, we will assume that $\|\mathbf{B}_{21}\|_\infty < \delta_{\mathbf{B}_{21}}$ and $\|\mathbf{B}_{22}\|_\infty < \delta_{\mathbf{B}_{22}}$, where $\delta_{\mathbf{B}_{21}}$ and $\delta_{\mathbf{B}_{22}}$ are all parameters chosen in Table 6.1. We would like to stress that this is not an important change, as it does not make the scheme less secure and only helps increase the readability of our analysis. We refer to Figure 6.2 for experimental confirmation that all the private keys that were generated from the reference implementation [Feu24] satisfy our bounds.

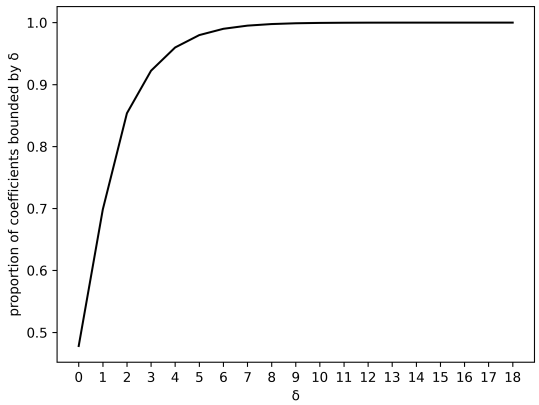


Figure 6.2: Proportion of coefficient embedding *coefficients of* $\mathbf{B}$ *that have absolute value less than an integer* δ*, out of* 1000 DEFI*-128 samples. We note that all coefficients were smaller than* $\delta_{\mathbf{B}_{22}} = 18$.

Public key

The *public key* is the matrix

$$\mathbf{C} = \mathbf{B}^T \mathbf{J} \mathbf{B}.$$

Again, the authors of DEFI choose to reject matrices $\mathbf{C}$ that have large entries, as this allows for shorter public keys. We will not use this fact in our attack.

Signature generation

The following procedure is used to sign a message μ from a private key $\mathbf{B}$. The full pseudo-code is described in Algorithm 5.

1. A message μ is first hashed into $h := H(\mu) \in R$.
2. A special trapdoor procedure constructs a $\mathbf{z} = (h \ || \ \mathbf{z}')$ such that $f_{\mathbf{J}}(\mathbf{z}) = 0$. This step is described in Algorithm 4. It completes the hashed message with a random nonce in such a way that the resulting vector is isotropic with respect to $\mathbf{J}$.

Algorithm 4 GenerateZ($\cdot$)

Require: $z_1 \in R$.
Ensure: $\mathbf{z}' \in R^3$ such that $f_{\mathbf{J}}(z_1 \ || \ \mathbf{z}') = 0$.
1: $u_1, v_2 \leftarrow D_u$ $\{u_1, v_2 \in R\}$
2: $v \leftarrow v_2(1 - u_1^2)$ $\{v \in R\}$
3: $u_2 \leftarrow 2v_2$ $\{u_2 \in R\}$
4: $z_2 \leftarrow v + u_2 u_1^2 - z_1 u_1$ $\{z_2 \in R\}$
5: $z_3 \leftarrow v + z_1 u_1$ $\{z_3 \in R\}$
6: $z_4 \leftarrow u_1 u_2 - z_1$ $\{z_4 \in R\}$
7: $\mathbf{z}' \leftarrow (z_2 \ || \ z_3 \ || \ z_4)$ $\{\mathbf{z}' \in R^3\}$
8: Return $\mathbf{z}'$

Algorithm 5 DEFI signature generation

Require: A message μ and a private key $\mathbf{B} \in M_4(R)$.
Ensure: A valid signature $\mathbf{y} \in R^3$.
1: $h \leftarrow H(\mu)$ $\{h \in R\}$
2: $\mathbf{z}' \leftarrow$ GenerateZ(h) $\{\mathbf{z}' \in R^3\}$
3: $\mathbf{y} \leftarrow \mathbf{B}_{22}^{-1}(\mathbf{z}' - \mathbf{B}_{21}h)$ $\{\mathbf{y} \in R^3\}$
4: Return $\mathbf{y}$

3. The signature is $\mathbf{y} := \mathbf{B}_{22}^{-1}(\mathbf{z}' - \mathbf{B}_{21}h)$.

Signature verification

Verification is described in Algorithm 6. It consists of the following two steps:

1. The message μ is hashed into $h := H(\mu)$.
2. The signature is *accepted* if and only if $f_{\mathbf{C}}((h \ || \ \mathbf{y})) = 0$, where $\mathbf{y}$ is the signature and $\mathbf{C}$ is the public key.

Algorithm 6 DEFI signature verification

Require: A message μ, a signature $\mathbf{y} \in R^3$ and a public key $\mathbf{C} \in M_4(R)$.
Ensure: *Accept* if the signature is correct, *Reject* otherwise.
1: **if** $f_{\mathbf{C}}((h \ || \ \mathbf{y})) = 0$ **then**
2: Return *Accept*
3: **else**
4: Return *Reject*
5: **end if**

6.3.2 Correctness of the Scheme

Let $(\mu, \mathbf{y})$ be a valid signature obtained using the secret key $\mathbf{B}$, $h = H(\mu)$ and $\mathbf{C}$ the associated public key. $\mathbf{y} = \mathbf{B}_{22}^{-1}(\mathbf{z}' - \mathbf{B}_{21}h)$ where $\mathbf{z}'$ is the output of GenerateZ(h). Let $\mathbf{z} = (h \ || \ \mathbf{z}')$, and $\mathbf{x} = (h \ || \ \mathbf{y})$. Then

$$f_{\mathbf{C}}(\mathbf{x}) = \mathbf{x}^T\mathbf{C}\mathbf{x} = (\mathbf{B}\mathbf{x})^T\mathbf{J}(\mathbf{B}\mathbf{x}) = f_{\mathbf{J}}(\mathbf{B}\mathbf{x}) = f_{\mathbf{J}}(\mathbf{z}).$$

Indeed,

$$\mathbf{B}\mathbf{x} = \begin{pmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} h \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} h \\ \mathbf{B}_{21}h + \mathbf{B}_{22}\mathbf{y} \end{pmatrix} = \begin{pmatrix} h \\ \mathbf{z}' \end{pmatrix} = \mathbf{z}. \tag{6.1}$$

	m	λ_u	$\delta_{\mathbf{B}_{21}}$	$\delta_{\mathbf{B}_{22}}$
DEFI-64	32	15	2	15
DEFI-128	64	35	2	18

Table 6.1: Parameters for DEFI

In order to prove the correctness of the scheme, we need to prove that Algorithm 4 produces a vector that is isotropic with respect to $\mathbf{J}$ once is is concatenated with the hash of the message. We use the notations of Algorithm 4 for the coordinates of $\mathbf{z}'$:

$$\begin{aligned} f_{\mathbf{J}}(\mathbf{z}) = \mathbf{z}^T\mathbf{J}\mathbf{z} &= z_1^2 + z_2^2 - z_3^2 - z_4^2 \\ &= (z_1 + z_4)(z_1 - z_4) + (z_2 + z_3)(z_2 - z_3) \\ &= (u_1u_2)(2z_1 - u_1u_2) + (2v + u_2u_1^2)(u_2u_1^2 - 2z_1u_1) \\ &= u_1u_2(2z_1 - u_1u_2) + (u_2(1 - u_1^2) + u_2u_1^2)(u_2u_1^2 - 2z_1u_1) \\ &= u_1u_2(2z_1 - u_1u_2 + u_1u_2 - 2z_1) = 0. \end{aligned}$$

6.3.3 Parameter Choice

DEFI comes in two flavours, a challenge version called DEFI-64, and a reference version DEFI-128 that was claimed to provide 128 bits of security. The ring R is defined according to the parameter m as $\mathbb{Z}[X]/(X^m + 1)$. The distribution D_u samples λ_u non-zero coordinates and uniformly assigns them a number from $\{\pm 1, \pm 2\}$.

6.4 Attacking DEFI

Lattice attacks on DEFI were already discussed in [FS24a, Section *IV.E.*]. The authors of the scheme observe that

$$z_2 + z_3 = u_2 \tag{6.2}$$
$$z_1 + z_4 = u_1u_2, \tag{6.3}$$

where $z_2 = b_{21}h + b_{22}y_2 + b_{23}y_3 + b_{24}y_4$ and $z_3 = b_{31}h + b_{32}y_2 + b_{33}y_3 + b_{34}y_4$, using the notation $\mathbf{B} = (b_{ij})_{i,j\in[4]}$. To exploit Equation 6.2, they argue that recovering the desired vector $\mathbf{b} = (b_{21}, b_{22}, b_{23}, b_{24}, b_{31}, b_{32}, b_{33}, b_{34}, u_2) \in R^9$ as a SVP solution in the lattice

$$L = \{\mathbf{x} \in R^9 : x_1h + x_2y_2 + x_3y_3 + x_4y_4 + x_5h + x_6y_2 + x_7y_3 + x_8y_4 - x_9 = 0\}$$

should be difficult, as the experimental norm of $\mathbf{b}$ is much larger than the Gaussian heuristic for L, implying that *if L behaves like a typical random lattice* then its shortest vector would be much shorter than $\mathbf{b}$. Their analysis is not exhaustive and overlooks a number of things. Most notably, L only exploits the information obtained from a single signature, and focuses on recovering the coefficients of B directly, whereas other more intermediate secrets might be easier to obtain through lattice attacks. No further justification is given to assess that it is sound to use the Gaussian heuristic for the length of the shortest vector here.

In our attack, we assume that the attacker has access to k signatures and use $x^{(i)}$ to denote the value of the parameter x corresponding to the i-th signature. We would like to emphasise that the attack is heuristic (as is often in the case in lattice-based cryptanalysis). Our analysis is presented in Section 6.5. Experiments are later discussed in Section 6.6.

6.4.1 First Step: Recovering u_2

The first step of the attack exploits the fact that each signature contributes to some leakage in order to recover half of the randomness used to generate the isotropic vectors that are used in the signing process (recall that Algorithm 4 samples u_1 and v_1, here we recover $u_2 = 2v_2$), as well as some partial information on coefficients of the secret matrix $\mathbf{B}$. Observation 6.4.1 below summarises exactly what we get from this first step.

Observation 6.4.1 (Informal)**.** *If an attacker has access to a large enough number k of signatures $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k)}$ signed with the same private key $\mathbf{B}$, then it can recover $b_{21} + b_{31}$, $b_{22} + b_{32}$, $b_{23} + b_{33}$, $b_{24} + b_{34}$ and all $u_2^{(i)}$ in time polynomial in $m = \dim(R)$ and the size of the entries.*

We aim to recover this secret information by using lattice reduction, to find a short vector in a lattice. We first define our lattice L_1, show how to construct it efficiently from public information, and finally we write down a short vector $\mathbf{s}_1 \in L_1$.

Definition 6.4.2. Let $k \in \mathbb{Z}_{>0}$. Assuming the vectors $(h^{(i)}, y_2^{(i)}, y_3^{(i)}, y_4^{(i)}) \in R^4$ for $i \in \mathbb{Z}_{>0}$ form a sequence of signatures obtained using the same private key, we define the following lattice

$$L_1 = \left\{ (\alpha, \beta, \gamma, \delta, \varepsilon^{(1)}, \ldots, \varepsilon^{(k)}) \in R^{4+k} : \forall i, \varepsilon^{(i)} = \alpha h^{(i)} + \beta y_2^{(i)} + \gamma y_3^{(i)} + \delta y_4^{(i)} \right\}.$$

We note that L_1 can be seen interchangeably as an R-module or as a Euclidean lattice.

Proposition 6.4.3 (Properties of L_1)**.** *Let $k \in \mathbb{Z}_{>0}$. Then the following statements are true:*

- *L_1 has rank $4m$;*
- *L_1 has ambient dimension $(k+4)m$;*
- *A basis of L_1 can be efficiently computed from the first k signatures.*

Proof. By definition, L_1 is generated by the $4m$ following vectors

$$\begin{gathered}
\left(X^j, 0, 0, 0, X^j h^{(1)}, \ldots, X^j h^{(k)}\right); \\
\left(0, X^j, 0, 0, X^j y_2^{(1)}, \ldots, X^j y_2^{(k)}\right); \\
\left(0, 0, X^j, 0, X^j y_3^{(1)}, \ldots, X^j y_3^{(k)}\right); \\
\left(0, 0, 0, X^j, X^j y_4^{(1)}, \ldots, X^j y_4^{(k)}\right),
\end{gathered}$$

for $j \in [m]$. This gives an efficiently computable generating set of vectors of L_1 of size $4m$. The first four R-coordinates are all linearly independent, therefore L_1 has rank exactly $4m$. □

Let

$$\mathbf{s}_1 = (b_{21} + b_{31}, b_{22} + b_{32}, b_{23} + b_{33}, b_{24} + b_{34}, u_2^{(1)}, \ldots, u_2^{(k)}).$$

Recall that for all $i \in [k]$ we have as in Equation 6.1,

$$\begin{pmatrix} b_{22} & b_{23} & b_{24} \\ b_{32} & b_{33} & b_{34} \end{pmatrix} \begin{pmatrix} y_2^{(i)} \\ y_3^{(i)} \\ y_4^{(i)} \end{pmatrix} = \begin{pmatrix} z_2^{(i)} \\ z_3^{(i)} \end{pmatrix} - h^{(i)} \begin{pmatrix} b_{21} \\ b_{31} \end{pmatrix}.$$

Adding both equations together, and combining the result with Equation 6.2 and the definition of L_1 implies in turn that $\mathbf{s}_1 \in L_1$. Recall that by design, DEFI comes with very short secret key coefficients (which seems inevitable to ensure small public keys), as well as very short coefficients for the nonce $v_1^{(i)}$ (which this time seems inevitable to ensure small signature sizes). This qualitatively justifies the shortness of $\mathbf{s}_1$. In fact, our attack recovers $\mathbf{s}_1$ by using a lattice reduction algorithm with input the basis of L_1 described in Proposition 6.4.3.

Remark 6.4.4. Until now, we have only used a single short vector $\mathbf{s}_1$. Recall that L_1 has R-module structure, and in the case where $q = X^m + 1$, all rotations $X^i \cdot \mathbf{s}_1$ (where $\cdot$ acts on L_1 R-coordinate by R-coordinate) for $i \in [m]$ give linearly independent vectors of L_1 of equal norm. This implies two things:

- If $\mathbf{s}_1$ is unusually short, then so are its $m-1$ other rotations and lattice reduction might recover the wrong one.
- $\mathbf{s}_1$ and its rotations generate an unusually dense rank m sublattice of L_1. Instead of studying the cost of recovering $\mathbf{s}_1$ directly via lattice reduction, it makes more sense to heuristically study the reduction strength needed to recover this special dense sublattice for a given number of signatures.

The first point is in fact not really a problem, as we can simply continue the attack simultaneously with all m rotations. This is only a linear increase in complexity, so a polynomial-time attack would remain polynomial. We will explain later in Section 6.4.4 how we can remove remove the need for this linear increase in complexity. We cover the second point in Section 6.5.1.

6.4.2 Second Step: Recovering u_1

The aim of this Section is to adapt our first step in a way that enables us to use Equation 6.3 to recover the secret information $u_1^{(i)}$. The $u_2^{(i)}$ are different, so as soon as we wish to use more than one signature, an immediate lattice approach does not work. We explain how we can artificially view our problem modulo a large prime p, and how this enables us to view recovering the $u_1^{(i)}$ as yet another lattice problem. Our end result for this step is described in Observation 6.4.5 below.

Observation 6.4.5 (Informal)**.** *If an attacker has access to a large enough number k of signatures $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k)}$ signed with the same private key $\mathbf{B}$, then it can recover b_{41}, b_{42}, b_{43}, b_{44} and all $u_1^{(i)}$ in time polynomial in $m = \dim(R)$ and the size of the entries.*

Lemma 6.4.6. *Let $p \in \mathbb{Z}_{>0}$ be a prime number. Let $q \in \mathbb{Z}[X]$ be a monic irreducible polynomial, and $R_p := (\mathbb{Z}/p\mathbb{Z})[X]/(q)$. Then a polynomial $r \in R_p$ is invertible in R_p if and only if* $\gcd(r, q) = 1$.

In this section we fix a random large prime number p. See Section 6.5.2 for a discussion on the size of p.

Definition 6.4.7. Let $k \in \mathbb{Z}_{>0}$. Assuming the vectors $(h^{(i)}, y_2^{(i)}, y_3^{(i)}, y_4^{(i)}) \in R^4$ for $i \in \mathbb{Z}_{>0}$ form a sequence of signatures obtained using the same private key, we define the following lattice

$$L_2 = \left\{(\alpha, \beta, \gamma, \delta, \varepsilon^{(1)}, \ldots, \varepsilon^{(k)}) \in R^{4+k} : \forall i, \varepsilon^{(i)} u_2^{(i)} \equiv_p \alpha h^{(i)} + \beta y_2^{(i)} + \gamma y_3^{(i)} + \delta y_4^{(i)}\right\}.$$

Proposition 6.4.8 (Properties of L_2)**.** *Let $k \in \mathbb{Z}_{>0}$. The following statements hold:*

- *L_2 has rank $(k+4)m$;*
- *L_2 has ambient dimension $(k+4)m$;*
- *A basis of L_2 can be efficiently computed from the first k signatures;*
- $\mathrm{vol}(L_2) = p^{km}$.

Proof. L_2 is a q-ary lattice whose basis can be written directly in a typical NTRU-like shape:

$$\begin{pmatrix} \mathbf{0} & p\mathbf{I}_{km} \\ \mathbf{I}_{4m} & \mathbf{Y} \end{pmatrix},$$

where the matrix $(\mathbf{I}_{4m} \ \mathbf{Y})$ is derived in the same way as for L_1. Note that although from a distance, L_2 looks just like L_1 with an extra mod p condition, there is an additional subtlety in its definition: an extra multiplication by $u_2^{(i)}$ on the left of the equation defining the lattice. The modulus p was incorporated to enable us to directly write down basis vectors. Indeed by Lemma 6.4.6 all $u_2^{(i)}$ are invertible in R_p and the following vectors for $j \in [k]$ can be efficiently computed:

$$\begin{aligned}
&\left(X^j, 0, 0, 0, X^j (u_2^{(1)})^{-1} h^{(1)}, \ldots, X^j (u_2^{(k)})^{-1} h^{(k)}\right); \\
&\left(0, X^j, 0, 0, X^j (u_2^{(1)})^{-1} y_2^{(1)}, \ldots, X^j (u_2^{(k)})^{-1} y_2^{(k)}\right); \\
&\left(0, 0, X^j, 0, X^j (u_2^{(1)})^{-1} y_3^{(1)}, \ldots, X^j (u_2^{(k)})^{-1} y_3^{(k)}\right); \\
&\left(0, 0, 0, X^j, X^j (u_2^{(1)})^{-1} y_4^{(1)}, \ldots, X^j (u_2^{(k)})^{-1} y_4^{(k)}\right).
\end{aligned}$$

As in Proposition 6.4.3, these $4m$ vectors are linearly independent vectors of L_2. From the previous blockwise representation one can directly read the volume and rank, so this set of linearly independent vectors of L_2 is also generating. Of course in order to freely use the values of $u_2^{(i)}$ we assume that the first step of the attack was already executed successfully. □

We now explain the design of this lattice by exhibiting some of its short vectors. Let

$$\mathbf{s}_2' = (b_{21} + b_{31}, b_{22} + b_{32}, b_{23} + b_{33}, b_{24} + b_{34}, 1, \ldots, 1).$$

For the same reason that $\mathbf{s}_1 \in L_1$, $\mathbf{s}_2' \in L_2$. We now claim that L_2 contains another independent short vector. Let

$$\mathbf{s}_2 = (b_{41} + 1, b_{42}, b_{43}, b_{44}, u_1^{(1)}, \ldots, u_1^{(k)}).$$

For all $i \in [k]$, rewriting the last row of Equation 6.1 we get

$$b_{42} y_2^{(i)} + b_{43} y_3^{(i)} + b_{44} y_4^{(i)} = z_4^{(i)} - h^{(i)} b_{41}.$$

Combining this with Equation 6.3, for which $h = z_1$ proves that $\mathbf{s}_2 \in L_2$.

As with L_1, L_2 has R-module structure. Therefore, all shifts $X^i \cdot \mathbf{s}_2$ and $X^j \cdot \mathbf{s}_2'$ are also vectors of L_2. This lets us define the following sublattice:

$$L_2' := \langle (X^i \cdot \mathbf{s}_2)_i, (X^j \cdot \mathbf{s}_2')_j \rangle_{\mathbb{Z}} \subset L_2.$$

Experimentally, when L_2 is built using enough signatures and p is taken large enough, lattice reduction applied to L_2 happens to recover L_2' in the following way: the first $2m$ vectors of the reduced basis generate L_2' exactly. Note that L_2', by definition is independent of the chosen value of p, as neither $\mathbf{s}_2$ nor $\mathbf{s}_2'$ depend on p.

From there we have a basis for L_2', a lattice of rank $2m \in \{64, 128\}$, and it is quite clear that any lattice reduction operation on L_2' should not be too costly. Morally, this lattice is a compositum of two lattices generated by all cyclic shifts of their respective generators and we need to find a way to act on each half separately. Our goal is to recover $\mathbf{s}_2$ to get access to all the information described in Observation 6.4.5. Because $\|\mathbf{s}_2'\| < \|\mathbf{s}_2\|$, solving an SVP instance on L_2' will do nothing to help. However we notice that the first four (R)-coordinates of $\mathbf{s}_2$ should be smaller than those of $\mathbf{s}_2'$, therefore we can choose a constant value $c(k)$ and use it to skew L_2' by defining

$$L_2'' := \left\{ (c(k)x_1, c(k)x_2, c(k)x_3, c(k)x_4, x_5, \ldots, x_{k+4}) : (x_i)_{i \in [k+4]} \in L_2' \right\}.$$

For a good choice of $c(k)$, *e.g.* a $c(k)$ that skews the lattice enough that the image of $\mathbf{s}_2$ in L_2'' becomes smaller in norm than the image of $\mathbf{s}_2'$, the skewed image of $\mathbf{s}_2$ in L_2'' becomes its shortest vector and can be recovered by lattice reduction in L_2''.

6.4.3 Final Step: Private Key Recovery

We now explain how to recover the full private key after the first two steps.

Observation 6.4.9. *If an attacker has access to $b_{2j} + b_{3j}$ and b_{4j} for $1 \leq j \leq 4$, where the b_{ij} are matrix coefficients of the private key $\mathbf{B}$ associated to $\mathbf{C}$, then it can fully recover $\mathbf{B}$ in time polynomial in $m = \dim(R)$.*

The public key is defined as $\mathbf{C} = \mathbf{B}^T\mathbf{J}\mathbf{B}$, implying the following relations on the diagonal coefficients $c_{jj} \in R$ of $\mathbf{C}$, for $1 \leq j \leq 4$:

$$c_{jj} = b_{1j}^2 + b_{2j}^2 - b_{3j}^2 - b_{4j}^2. \tag{6.4}$$

As would be the case after Observation 6.4.5, we now have access to all b_{4j} for $1 \leq j \leq 4$. All c_{jj} and b_{1j} are known and therefore, Equation 6.4 allows us to recover $b_{2j}^2 - b_{3j}^2 = (b_{2j} - b_{3j})(b_{2j} + b_{3j})$. We also know all $b_{2j} + b_{3j}$ for $1 \leq j \leq 4$ (as would be the case after Observation 6.4.1). The only remaining step is to derive $b_{2j} - b_{3j}$ from $b_{2j}^2 - b_{3j}^2$ and $b_{2j} + b_{3j}$. R being only a ring, it is impossible to simply invert $b_{2j} + b_{3j}$. However, if we pick a large enough prime p as in Section 6.4.2, $b_{2j} + b_{3j}$ can be inverted modulo p without any loss of information.

6.4.4 Exploiting the Ring Choice

In their concrete parameters and for efficiency purposes, DEFI is instantiated using $R = \mathbb{Z}[X]/(X^m + 1)$, where m is a power of two. We explain how this can be used to simplify the attack.

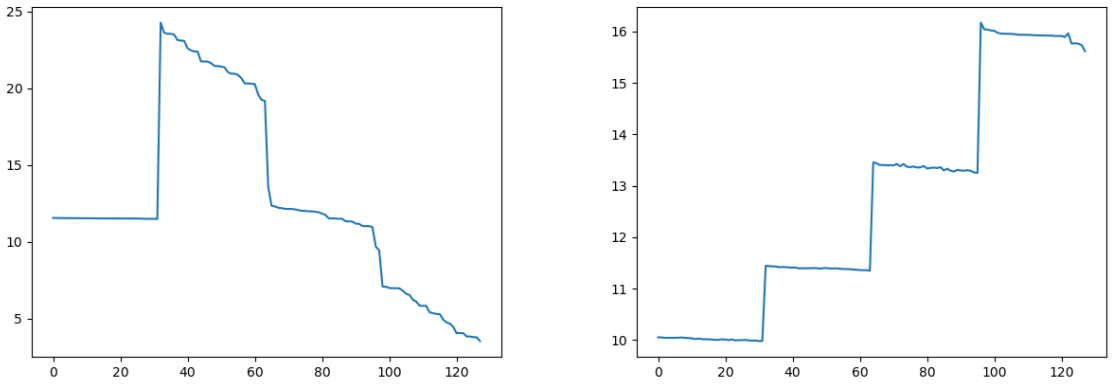
In Remark 6.4.4, we have seen that the first step of our attack might only recover a shift $X^j \cdot \mathbf{s}_1$ instead of our targeted $\mathbf{s}_1$. Instead of running the second step of our attack multiple times with each possible rotation, we notice that L_2 also has R-module structure, so its definition is independent of the shifts of $u_2^{(i)}$ that were obtained in the first step. Therefore the second step of the attack remains valid regardless, and it will output some $X^j \cdot \mathbf{s}_2$ that is a shift of the target secret $\mathbf{s}_2$. Now that we have handled the hardest part of the attack, we can use Equation 6.4 to guess and verify the correct shifts for $\mathbf{s}_1$ and $\mathbf{s}_2$. Indeed, when testing a specific pair of shifts $(X^i \cdot \mathbf{s}_1, X^j \cdot \mathbf{s}_2)$, we can use the procedure described in the proof of Observation 6.4.9 to recover a candidate value b_- for say $b_{22} - b_{32}$, using our guess b_+ for $b_{22} + b_{32}$. If this value b_- is such that $b_- + b_+$ has even and small coordinates, then the fact that a large enough prime p was chosen in the inversion implies that necessarily, the pair of shifts (i, j) has correctly been guessed. This whole procedure of guessing shifts runs at most m^2 times, which makes its runtime negligible compared to the lattice reduction steps.

6.5 Some Elements to Justify the Attack

The description of the attack in Section 6.4 leaves a couple open questions: how can we be sure that a fast lattice reduction algorithm is enough to really recover $\mathbf{s}_1$ from our basis for L_1, as well as $\mathbf{s}_2$ from our basis for L_2? How many signatures are required to mount each step of the attack? Although most of our justification are experimental, we make a few remarks on the shape of the lattices considered. To the best of our knowledge, the behaviour of lattice reduction algorithms on lattices whose geometry resembles that of L_1 or L_2 is poorly documented, making a fully rigorous analysis near-impossible.

6.5.1 Analysing L_1

The first step of the attack relies on recovering $\mathbf{s}_1$ from a poor basis for L_1, a lattice which we generate using k signatures. We give will give precise experimental numbers for the minimal number of signatures k that allow us to mount this first step in Section 6.6. In this Section we study the shape of a reduced basis for L_1, and note that the gap between the expected norm of $\mathbf{s}_1$ and the expected norm of a shortest nonzero vector in L_1 is unusual, in the sense that it is larger than what we would expect from a typical random lattice. This in itself does not explain why polynomial-time lattice reduction algorithms seem to be enough to successfully execute this first step. The estimates from [GN08b] do not apply, as there is no gap between $\lambda_1(L_1)$ and $\lambda_2(L_1)$. In fact, due to the presence of a sublattice of unusually small covolume and of dimension a fraction of the dimension, the situation seems to be closer to that of NTRU where it has been observed in [KF17; DvW21] that when the covolume of the lattice becomes sufficiently large, lattice reduction starts recovering vectors from the dense sublattice earlier than the time we would expect it to recover short vectors. To predict lattice reduction, one might want to simulate the profile of the Gram-Schmidt norms throughout the reduction process. This seems particularly enticing as we would expect a straight and horizontal line for the first quarter of vectors, followed by a sharp increase and then a steadily decreasing line covering the last three quarters, in the style of a Geometric Series Assumption in the presence of q-vectors. However, given that lattice reduction is so effective on L_1 (slight improvements to LLL are essentially already enough to fully reduce the basis in the cases of both DEFI-64 and DEFI-128), it is also practically impossible to simulate the behaviour of the Gram-Schmidt profile for L_1, as the transition happens so quick. If one were to generalise the scheme to larger rings and evaluate its practical security against our attack, then one would need to conduct this analysis thoroughly. In our setting however, the ring size is small enough that we can rely on experiments for this part. The Gram-Schmidt lengths pictured in Figure 6.3 before and after reduction remain an interesting tool to discuss the geometry of L_1. Figure 6.3 shows that the reduced basis separates



(a) Before reduction

(b) After flatter and LLL

Figure 6.3: log Gram-Schmidt norms of L_1 for $R = \mathbb{Z}[X]/(X^{32}+1)$, 10 signatures.

into four projected sublattices with increasing covolumes. The presence of another sublattice of already short vectors on the left of the profile before reduction can only benefit reduction, in the same way as unusually short q-vectors in the input basis naturally aid reduction, as exploited in [DEP23] and asymptotic study of [BN24].

In what follows we bound the gap between $\mathrm{GH}(L_1)$ and $\lambda_1(L_1)$, and show that for a large enough number k of signatures, we can expect $\mathbf{s}_1$ and its rotations to be the shortest non-zero vectors of L_1. This explains how solving an instance of SVP in L_1 allows us to recover a rotation

of $\mathbf{s}_1$.

Lemma 6.5.1. *[MM92, Theorem 4.1.8] If* $\mathbf{A}$ *and* $\mathbf{B}$ *are nonnegative Hermitian square matrices in* $M_n(\mathbb{C})$*, then*

$$\det(\mathbf{A}+\mathbf{B})^{1/n} \geq \det(\mathbf{A})^{1/n} + \det(\mathbf{A})^{1/n}.$$

Proposition 6.5.2 (Volume of L_1)**.** *Let* $k \in \mathbb{Z}_{>0}$ *be an integer divisible by* 4*. If we write the computed basis for* L_1 *blockwise as*

$$\begin{pmatrix} \mathbf{I}_{4m} & || & \mathbf{A}_1 & || & \dots & || & \mathbf{A}_{k/4} \end{pmatrix}$$

with square matrices $\mathbf{A}_i$*, then*

$$\mathrm{vol}(L_1) \geq \left(\frac{k}{4}\right)^{2m} \left(\min_{1\leq i\leq k} \det(\mathbf{A}_i)\right).$$

If we also assume that the $\mathbf{A}_i$ *are independent random variables that follow the same distribution, then*

$$\mathbb{E}\left(\mathrm{vol}(L_1)^{\frac{1}{4m}}\right) \geq \frac{\sqrt{k}}{2}\sqrt{\mathbb{E}(\det(\mathbf{A}_1)^{\frac{2}{4m}})}.$$

Proof. We prove this identity using Lemma 6.5.1, inductively on k. Indeed,

$$\begin{aligned}
\mathrm{vol}(L_1)^{\frac{2}{4m}} &= \det\left(\begin{pmatrix} \mathbf{I}_{4m} & || & \mathbf{A}_1 & || & \dots & || & \mathbf{A}_{k/4} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{I}_{4m} & || & \mathbf{A}_1 & || & \dots & || & \mathbf{A}_{k/4} \end{pmatrix}^T\right)^{\frac{1}{4m}} \\
&= \det\left(\mathbf{I}_{4m} + \mathbf{A}_1\mathbf{A}_1^T + \ldots + \mathbf{A}_{k/4}\mathbf{A}_{k/4}^T\right)^{\frac{1}{4m}} \\
&\geq 1 + \sum_{i=1}^{k/4} \det\left(\mathbf{A}_i\mathbf{A}_i^T\right)^{\frac{1}{4m}} \\
&\geq \frac{k}{4} \min_{1\leq i\leq k} \det(\mathbf{A}_i)^{\frac{2}{4m}},
\end{aligned}$$

where we used the fact that the $\mathbf{A}_i\mathbf{A}_i^T$ are all non-negative, Hermitian and square. The second identity immediately follows from linearity of expectation. □

The following Proposition explains the behaviour of the gap for L_1 generated from asymptotically many signatures.

Proposition 6.5.3. *Let* $k \in \mathbb{Z}_{>0}$ *be an integer divisible by* 4*. Using the same notations as in Proposition 6.5.2, and assuming the* $\mathbf{A}_i$ *are independent random variable that follow the same distribution, then*

$$\frac{\mathrm{GH}(L_1)}{\lambda_1(L_1)} \geq (1+o_{k,m}(1))\sqrt{\frac{m}{32\pi e\lambda_u}} \min \det(\mathbf{A}_i)^{\frac{1}{4m}},$$

and

$$\mathbb{E}\left(\frac{\mathrm{GH}(L_1)}{\|\mathbf{s}_1\|}\right) \leq \sqrt{\frac{m}{4\pi e\lambda_u}}\sqrt{\mathbb{E}(\det(\mathbf{A}_1)^{\frac{2}{4m}})}.$$

Proof. We now show that $\mathbf{s}_1$ is somewhat short. Indeed, the coefficients of the secret key $\mathbf{B}$ are themselves bounded, and the $u_2^{(i)}$ are generated according to a distribution that has short expected norm. More precisely, $u_2^{(i)}$ consists of exactly λ_u coordinates in $\{\pm 2, \pm 4\}$, which means that we have

$$\begin{aligned}
\|\mathbf{s}_1\|^2 &\leq \sum_{j=1}^{4} \|b_{2j} + b_{3j}\|^2 + \sum_{i=1}^{k} \|u_2^{(i)}\|^2 \\
&\leq 16(\delta_{\mathbf{B}_{22}} + \lambda_u k).
\end{aligned}$$

The gap between the Gaussian heuristic on L_1 and the norm of the shortest vector can therefore be bounded using Proposition 6.5.2, where the $o_k(\cdot)$ is taken as $k \to \infty$:

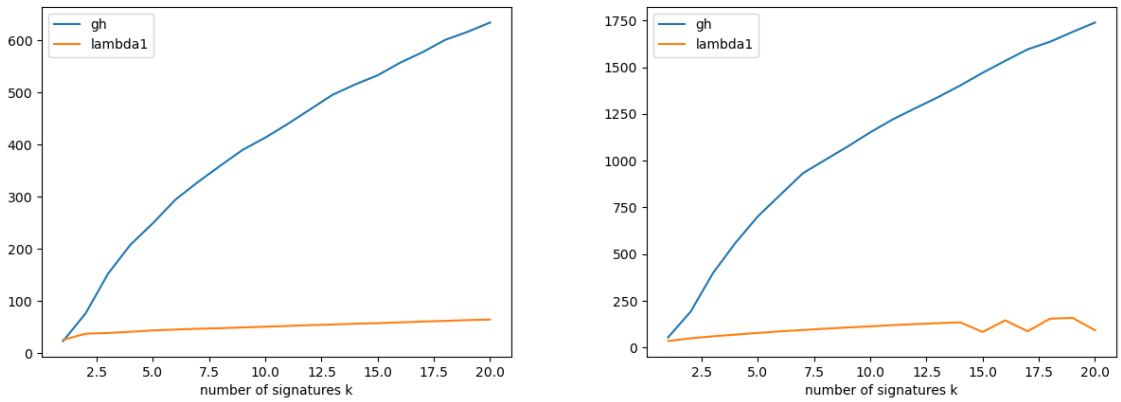
$$\begin{aligned}\frac{\mathrm{GH}(L_1)}{\lambda_1(L_1)} \geq \frac{\mathrm{GH}(L_1)}{\|\mathbf{s}_1\|} &\geq (1+o_m(1))\frac{\sqrt{\frac{4m}{2\pi e}}\mathrm{vol}(L_1)^{\frac{1}{4m}}}{\sqrt{16(\delta_{\mathbf{B}_{22}}+\lambda_u k)}} \\ &\geq (1+o_m(1))\frac{\sqrt{\frac{4m}{2\pi e}}\sqrt{k/4}\min\det(\mathbf{A}_i)^{\frac{1}{4m}}}{\sqrt{16(\delta_{\mathbf{B}_{22}}+\lambda_u k)}} \\ &= (1+o_{k,m}(1))\sqrt{\frac{m}{32\pi e\lambda_u}}\min\det(\mathbf{A}_i)^{\frac{1}{4m}}.\end{aligned}$$

Using the second item of Proposition 6.5.2, and the inequality $\|\mathbf{s}_1\| \geq \sqrt{2k\lambda_u}$ we can now bound

$$\begin{aligned}\mathbb{E}\left(\frac{\mathrm{GH}(L_1)}{\|\mathbf{s}_1\|}\right) &\leq \sqrt{\frac{4m}{2\pi e}}\frac{\sqrt{k/4}}{\sqrt{2k\lambda_u}}\sqrt{\mathbb{E}(\det(\mathbf{A}_1)^{\frac{2}{4m}})} \\ &\leq \sqrt{\frac{m}{4\pi e\lambda_u}}\sqrt{\mathbb{E}(\det(\mathbf{A}_1)^{\frac{2}{4m}})}.\end{aligned}$$

□

Proposition 6.5.3 helps justify two things. First, it is well-known that increasing the gap between a vector's length and the Gaussian heuristic while leaving the covolume to a fixed value equates to increasing the Hermite factor, which heuristically leads to an easier lattice problem. Therefore, as this gap increases with k, it makes sense that more signatures lead to an easier lattice problem. Second, a larger gap between $GH(L_1)$ and $\|\mathbf{s}_1\|$ means that it is more likely (the probability should in fact be overwhelming as in [LN24, Theorem 6]) that $\mathbf{s}_1$ and its rotations are the true shortest vectors in L_1. The dimension of the lattice is 128 for DEFI-64 and 256 for DEFI-128, making running simple lattice reduction algorithms feasible. Figure 6.4 compares the Gaussian heuristic for L_1 with the size of the shortest vector recovered by a run of flatter and LLL in the case of DEFI-64 and DEFI-128.



(a) DEFI-64 (b) DEFI-128

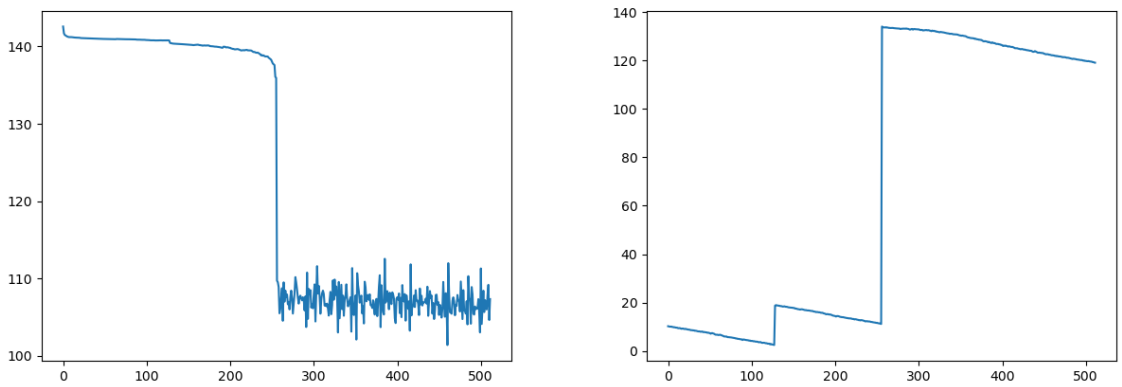
Figure 6.4: Comparing the size of $\mathbf{s}_1$ with the Gaussian heuristic in L_1, for increasing number of signatures k.

Although both curves seem to diverge fast, their main terms are both up to a constant equivalent to $\sqrt{k}$, therefore their ratio converges towards a constant value. Asymptotically, this gap is less than the gap that one would observe in the NTRU lattice [Che+20].

6.5.2 Analysing L_2

Once the first step has successfully been executed, the second step selects a prime modulus p, and then generates a lattice L_2 with the data from k signatures. Contrarily to what happened with L_1, where the number of signatures did not influence the dimension of the lattice, here $\dim(L_2) = (k+4)m$, and therefore as we are going to have to reduce L_2, it is of the utmost importance to limit the number of signatures included into L_2 to the minimum possible.

We start by observing the Gram-Schmidt profiles for L_2, before and after reduction by LLL.



(a) Before reduction

(b) After LLL

Figure 6.5: log Gram-Schmidt norms of L_2 for $R = \mathbb{Z}[X]/(X^{64}+1)$, 4 signatures

Figure 6.5 confirms the observations from Section 6.4.2: if the modulus p is large enough, lattice reduction separates the vectors in L_2 that live in a sublattice independent of p. We notice a first sublattice of dimension $4m$, as well as another of dimension $2m$, which in fact corresponds exactly to the lattice L_2' defined in Section 6.4.2.

To prove that LLL or a stronger form of lattice reduction recovers L_2' would require answering a lot of questions combining sublattice finding with lattice reduction theory. Although some algorithms for the densest sublattice problem [DM13] have been proposed, or studied in the very particular case of NTRU [DvW21], the practical solution to the following question is yet not well understood: given a rank n lattice with the promise that it has an unusually dense sublattice of rank n', how hard is it to recover such a sublattice? This question generalises the more common study of lattice reduction algorithms to solve SVP in the presence of an unusually short vector (or in other words an unusually dense rank 1 sublattice). It is possible to formalise the wording *unusually dense* by comparing the covolume of the sublattice with the appropriate value of the expected covolume of a random rank n' sublattice, as described in [Thu98], but we consider this to be outside of the scope of our account of the proposed attack on DEFI, and choose parameters (number of signatures, size of p, lattice reduction algorithms) that consistently and efficiently recover the sublattice L_2'.

6.5.3 Analysing the Key-Recovery Step

If the first two steps have been performed successfully, then it is possible to prove Observation 6.4.9.

Lemma 6.5.4. *Let $a, b, c \in R$ be ring elements such that $a = bc$, $\gcd(b, q) = 1$. Let p be a prime number such that $p \geq 2\|c\|_\infty$. If for $x \in R$, $\tilde{x}$ denotes the class of x in $R_p = (\mathbb{Z}/p\mathbb{Z})[X]/(q)$, and* $\mathrm{Round}(\tilde{x})$ *is the representative of $\tilde{x}$ in R that has minimal ∞-norm, then*

$$c = \mathrm{Round}(\tilde{b}^{-1}\tilde{a}).$$

Proof. First, $\tilde{b}$ is invertible in R_p thanks to Lemma 6.4.6. Therefore, $\tilde{b}^{-1}\tilde{a} = \tilde{c}$ in R_p. Now $\|\text{Round}(\tilde{c})\|_\infty \leq \frac{p}{2}$ and there is only one such representative of $\tilde{c}$. The bound on $\|c\|_\infty$ implies that $c = \text{Round}(\tilde{c})$. □

Of Observation 6.4.9. As mentioned above, we assume that we can apply Observations 6.4.1 and 6.4.5 to recover the sums $b_{2j} + b_{3j}$ and the b_{4j} for $1 \leq j \leq 4$. Then using Lemma 6.5.4 onto Equation 6.4 we obtain that if $p \geq 2\|b_{2j} - b_{3j}\|_\infty$ is a prime, the smallest representative of $(b_{2j} \tilde{+} b_{3j})^{-1}(\tilde{c_{jj}} - \tilde{b}_{1j}^2 + \tilde{b}_{4j}^2)$ in R is $b_{2j} - b_{3j}$. The half-sums and half-differences give us all remaining coefficients of the secret key $\mathbf{B}$. Finally, note that $2\|b_{2j} - b_{3j}\|_\infty \leq 4\max(\delta_{\mathbf{B}_{21}}, \delta_{\mathbf{B}_{22}})$, so any prime p larger than this value will suffice. □

6.6 Experiments

We ran all experiments on a `intel i7-1065G7 CPU@1.3GHz` processor. Our code for the attack is available online at

https://gitlab.inria.fr/hbambury/defi-nitely-broken.

The full secret-key recovery attack for the challenge instance DEFI-64 runs in less than 20 seconds with 3 signatures. For reference, we give the solution to the challenge in the appendix. We now focus only on the strongest security parameters proposed in [FS24a], DEFI-128.

6.6.1 Running the Attack

Using the code available at [Feu24], we generated 100 DEFI-128 public keys with corresponding signatures, and tested our attack 100 times.

First step

Using the flatter software [RH23] followed by fplll's LLL implementation [The24] as our main lattice reduction tools for reducing L_1 in the first step, we are able to recover the nonces u_2 in less than 50 seconds on average. In all 100 instances, the first step failed with 8 signatures, but was successful with 9. If one is willing to pay the cost of time and run stronger lattice reduction algorithms, then it is likely than one can reduce the number of required signatures. We chose not to explore this path.

Second step

Using flatter combined with LLL on L_2 with 4 signatures and a random 100-bit prime number for p recovered L_2' in all 100 instances, with an average runtime under 180 seconds. 3 signatures were not enough to separate the $2m$-dimensional sublattice from the $4m$-dimensional one (see 6.5b). We were able to recover the nonces u_1 using fplll's implementation of BKZ with blocksize 20 on L_2'' in all 100 instances, with an average runtime under 30 seconds.

Key-recovery

The runtime for the last step is negligible compared to the first two step, and this step is guaranteed to work. We conclude that our attack was successfully able to recover the private key in all 100 of our DEFI-128 challenges using 9 signatures, and in under 5 minutes for each.

6.6.2 Minor Improvements

Lattice weights

It comes as no surprise that the sizes of the coordinates of the short vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ can be roughly predicted from the parameters of the scheme. Indeed, they are directly tied to: on one side the generation of the private key $\mathbf{B}$, and on the other the generation of the nonces u_1 and u_2. In practice, we add some weights to the different columns of the lattices we reduce to ensure that the target vector has balanced coordinates. Even if done very roughly, this allowed us to lower the number of required signatures without the need for stronger lattice reduction.

On the use of flatter

The most expensive part of our attack is by far the lattice reduction step. Lattice reduction can become costly when the lattice dimension is large, or when the size of the input lattice vectors grows. The vectors here are of very reasonable sizes, but the lattices grow in dimension. L_1 is always 256-dimensional, and L_2 with 4 signatures is 512-dimensional, which is far too big for any naïve implementation of LLL. The algorithm of [RH23] enables us to deal with such high dimensions in only a few minutes, we use it as a pre-processing step for LLL[2].

Even lattice intersection

In Section 6.4.1, we aim at recovering a short vector $\mathbf{s}_1$ that contains the elements $u_2^{(i)}$. Because of the trapdoor construction described in Algorithm 4, $u_2^{(i)}$ consists of λ_u coefficients that are all in $\{\pm 2, \pm 4\}$. Consequently,

$$\mathbf{s}_1 \in L_1 \cap L_{even},$$

where $L_{even} \cong \mathbb{Z}^{4m} \times (2\mathbb{Z})^{km}$ is the lattice whose last km coordinates are all even. It might seem natural that using this extra information on the shape of the nonces should help us, especially as this lattice intersection can be efficiently computed through duality. We do not observe any substantial experimental improvement when considering this intersection.

Conclusion: Discussion and Perspectives

We have presented a full key-recovery attack on all proposed parameters of [FS24a], a signature scheme based on an innovative problem involving isotropic vectors of non-definite quadratic forms. Our attack is well motivated, and was shown to work experimentally on every challenge instance we generated.

Sublattice recovery

A point that remained unclear to us in the analysis of our attack is the analysis of the sublattice recovery problem. We show an example of a situation where it would be interesting to understand how, why and when lattice reduction recovers a given unusually dense sublattice. This question having ties with the study of NTRU lattices, we believe it might be of independent interest.

Fixing DEFI?

We see no obvious countermeasure to our attack, other than increasing parameters or radically changing the procedure for generating an isotropic vector. A direct fix by adapting parameter

[2]Using both flatter and an LLL implementation might sound somewhat redundant. Flatter is significantly faster than LLL for lattices in large dimensions, but does not guarantee an LLL-reduced basis. We found that running LLL after flatter improved the basis quality for a minimal overhead.

values would require a careful study of our lattice attack and ensure that more computationally intense lattice reduction as allowed by the desired security requirements does not lead to any exploitable leaks, even when many signatures are available to the attacker. Even if this were possible, it would most certainly increase the parameters for the scheme, making it less competitive, and would still not consist in a sound security proof. A true fix would require changing the procedure for generating an isotropic vector in such a way that the output distribution would become independent of the secret key.

Remark 6.6.1. Such a fix was implemented by the authors in the new scheme [FS24b], achieving similar (if not better) performances. The field choice $\mathbb{Q}(X)/(X^{28}+X+1)$ is different from the one in the first version and somewhat non-standard in cryptography, as it is not a cyclotomic field. This might lead to creative cryptanalysis.

An interesting new assumption

The idea behind the scheme remains new and interesting. The claimed hard problem of recovering a unimodular matrix $\mathbf{B}$ from the public key $\mathbf{C} = \mathbf{B}^T\mathbf{J}\mathbf{B}$, which up to conjugation of the left matrix reformulates as *module-QFE* in Section 6.2 certainly looks a lot like the *(module)-Lattice Isomorphism Problem* [DvW22; BGPS23; DPPvW22], only that it is defined here with $\mathbf{J}$, which is not positive definite, making it different from the classical lattice problem. This Quadratic Form Equivalence problem in the Isotropic case is not well studied from the cryptographic point of view and would benefit from more constructions and direct cryptanalysis using algorithmic ideas from the study of reduction of quadratic forms, as well as mathematical ideas on the classification or decomposition of isotropic forms.

Chapter

7 A New Lattice-Based Signature

Abstract The Fiat-Shamir with Aborts paradigm (FSwA) uses rejection sampling to remove a secret's dependency on a given source distribution. Recent results revealed that unlike the uniform distribution in the hypercube, both the continuous Gaussian and the uniform distribution within the hyperball minimise rejection rates and signature sizes. However, in practice both these distributions suffer from the complexity of their sampler. So far, those three distributions are the only available alternatives, but none of them offer the best of all worlds: competitive proof of knowledge size and rejection rate with a simple sampler.

We introduce a new generic framework for FSwA using polytope based rejection sampling to enable a wider variety of constructions. This framework is the first to generalise these results to integral distributions. To complement the lack of alternatives, we also propose a new polytope construction, whose uniform sampler approaches in simplicity that of the hypercube. At the same time, it provides competitive proof of knowledge size compared to that obtained from the Gaussian distribution. Concurrently, we share some experimental improvements of our construction to further reduce the proof size. Finally, we propose a new signature based on the FSwA paradigm using both our framework and construction: Patronus. We prove it to be competitive with Haetae in signature size and with Dilithium on sampler simplicity, making it the current best FSwA lattice signature that does not rely on Gaussian distributions.

Most of this chapter is based on the conference paper[1] *[BBRS24].*

Chapter content

[1] As this work also appears in Hugo Beguinet's PhD thesis [Beg24], it is customary to highlight our separate contributions. While Hugo supervised the project and designed the signature, I contributed to the mathematical analysis of its components, especially objects related to the sampler.

7.1 Introduction

We have already seen an example of a candidate post-quantum signature scheme in the previous chapter. The *security* of that scheme relied on some untested assumptions, and came with no formal security proofs. Although the underlying quadratic form equivalence problem might be hard, there is no reason to believe that recovering DEFI's private key from its public key implies a solution to said problem in general. Indeed, we saw that the scheme was insecure. In this chapter we focus on a very different class of conjecturally post-quantum signature schemes, whose hardness provably relies on the hardness of well-established lattice problems, that no-one has been able to solve: lattice-based signatures in the Fiat-Shamir with Aborts (FSwA) paradigm.

Lattice-based cryptography offers numerous advantages over traditional number-theoretic public-key cryptography. These advantages span from conjectured resistance to quantum attacks to the capability of performing arbitrary computations on encrypted data, all while maintaining comparable or even superior efficiency. However, a notable challenge persists: the need to reduce the size of transmittable elements, including zero-knowledge proofs of knowledge (ZKPoK).

Even when using algebraic lattices, zero-knowledge proofs still tend to be at least an order of magnitude larger than their traditional counterparts. Consequently, the transition towards this so-called post-quantum cryptography, driven by the release of the first standards, presents a series of challenges. These challenges include a substantial increase in bandwidth consumption. Presently, these issues serve as barriers to the widespread adoption of lattice-based cryptography.

Zero-knowledge There exists a wide variety of lattice-based ZKPoK constructions, starting with [KTX08] and seeing improvements in [LNSW13]. This evolution has led to multiple lines of work, in particular to the birth of the Fiat-Shamir with Aborts [Lyu09] paradigm. In this paradigm, the crucial zero-knowledge step is done through a rejection sampling algorithm in order to remove any sort of secret dependency from the output distribution. This led to a plethora of different improved constructions [BLNS21; LNS20; LNS21a; LNP22], from basic signatures [Duc+21; Che+23] to blind [BLNS23a] and group signatures [dPLS18; LNS21b] as well as anonymous credentials [BLNS23b]. In this chapter we focus on signatures that use the FWsA paradigm.

Fiat-Shamir with aborts Two notable examples of lattice-based digital signatures are

- Dilithium [Duc+21], now standardised by the US standardisation entity under the name ML-DSA[2]. This scheme is already being used by millions and will become the preferred standard for digital signatures from 2035 onwards.
- Haetae [Che+23], winner of the Korean Kpqc competition for post-quantum digital signatures.

Both of these signatures are designed using Lyubashevsky's [Lyu09] scheme as a blueprint, with extra optimisations such as the use of module lattices instead of their unstructured counterpart.

How can we compare digital signatures? In recent years, a lot of new schemes have emerged that aspire to resist against a quantum adversary. This abundance of candidates can be difficult to navigate. Why is scheme A better than scheme B? Why is entity X claiming that scheme B is better than scheme A ? Those questions do not (always) have a clear-cut answer. Indeed digital signatures are one of the most versatile primitives in cryptography, so measuring

[2] Dilithium and ML-DSA are technically not exactly the same, as ML-DSA incorporates extra minor optimisations.

how good a scheme is has to depend on the desired application. In general we care about security as well as efficiency.

- Security depends on the belief that the underlying mathematical problem is computationally intractable, and will remain so in the foreseeable future. This belief is greatly enhanced by security proofs that reduce security properties of the signature (unforgeability and so on) to the hardness of the underlying problem.

- Efficiency mainly has two aspects: the speed at which the algorithm runs on standard hardware, and the size of the elements that need transmitted. In real-world network protocols, signatures and keys might need to fit in packets or groups of packets of data.

There are of course many more criteria, to list just a few, one might need the signature to easily generalise to allow for more sophisticated primitives (can the signature be thresholdised?), to remain secure in other models (eg resistance to side-channel attacks and fault injection), or to be good enough in multiple such metrics.

An ode to simplicity Amidst concerns on efficiency, simplicity of the scheme and its implementation are sometimes overlooked. This was a major selling point for Dilithium compared to other candidates: it does not require the generation of secret randomness from a Gaussian or discrete Gaussian distribution. Generating such samples in a way that resists against side-channel attacks is difficult, see [BHLY16; EFGT17; PBY17; GMRR22; Pre23], and leads to insecure implementations. This seems to be one of the major reason that keeps delaying the publication of the NIST standard for the Falcon signature [Fou+19]. For a standard that will be so widely deployed, one should not rely on the expertise of implementers.

The authors of Haetae use similar techniques as those of Dilithium[3], but they use samples uniformly distributed in a Euclidean ball, and that requires the use of discrete Gaussians. At the cost of simplicity and ease of implementation, they obtain shorter signature sizes for equivalent security parameters.

In the specification document [Duc+21], the authors of the scheme claim the following: "Under the restriction that we avoid discrete Gaussian sampling, to the best of our knowledge, Dilithium has the smallest combination of signature and public key sizes[4] of any post-quantum signature scheme." We aim to improve on this statement.

Question 7.1.1. *Is it possible to build a FSwA signature without the need for Gaussians that relies on the same assumptions as* Dilithium*, but has noticeably shorter signature sizes?*

As an answer to Question 7.1.1, we present an improved Lyubashevsky-like signature scheme called Patronus. The design of Patronus is similar to its peers Dilithium and Haetae, but all three generate secret randomness through the use of different samplers:

- Dilithium requires a uniform sampler on the integral points of a hypercube.

- Haetae requires a uniform sampler on the integral points of a Euclidean ball.

- Patronus requires a uniform sampler on the integral points of a polytope defined by intersecting a hypercube with a cross-polytope. In large dimensions, this intersection approaches that of the cross-polytope.

[3] As well as the extra trick from the BLISS signature scheme of considering bimodal distribution.

[4] by this they mean the sum signature size + public key, although some other applications might need to optimise for other linear combinations of both quantities, *e.g.* six signatures and two public keys for TLS, see [WW21].

Table 7.1: Comparison of both the signature and verification key sizes in bytes for Dilithium, Haetae, *and* Patronus, *where "II", "III" and "V" represent respectively* 120, 180 *and* 260 *bits of classical security.*

Table 7.2: Signature size comparison.

	II	III	V
Haetae	1,463	2,337	2,908
Patronus (this work)	2,070	2,575	3,721
Dilithium	2,420	3,293	4,595

Table 7.3: Verification key size comparison.

	II	III	V
Haetae	992	1,472	2,080
Patronus (this work)	832	1,152	1,632
Dilithium	1,312	1,952	2,592

Satisfyingly, we notice that all three shapes can be associated with the ball of one of the following standard norms: respectively ℓ_∞, ℓ_2 and ℓ_1. Notice also that we always need to sample over integers. Previous analyses of FSwA, such as that of [DFPS22] are conducted exclusively in the continuous setting, leaving a blurry gap between theory (studying continuous shapes) and practice (which requires counting integral points in said shapes). At some point the need to restrict results to integers is necessary, however it appears that a generic construction dealing with this technicality is missing in the literature. This leads to the following questions:

Question 7.1.2. *Can we get a generic FSwA construction that deals directly with distributions over the integers?*

The signature and public key sizes of Patronus are compared with those from Dilithium and Haetae in Table 7.1 and Figure 7.1.

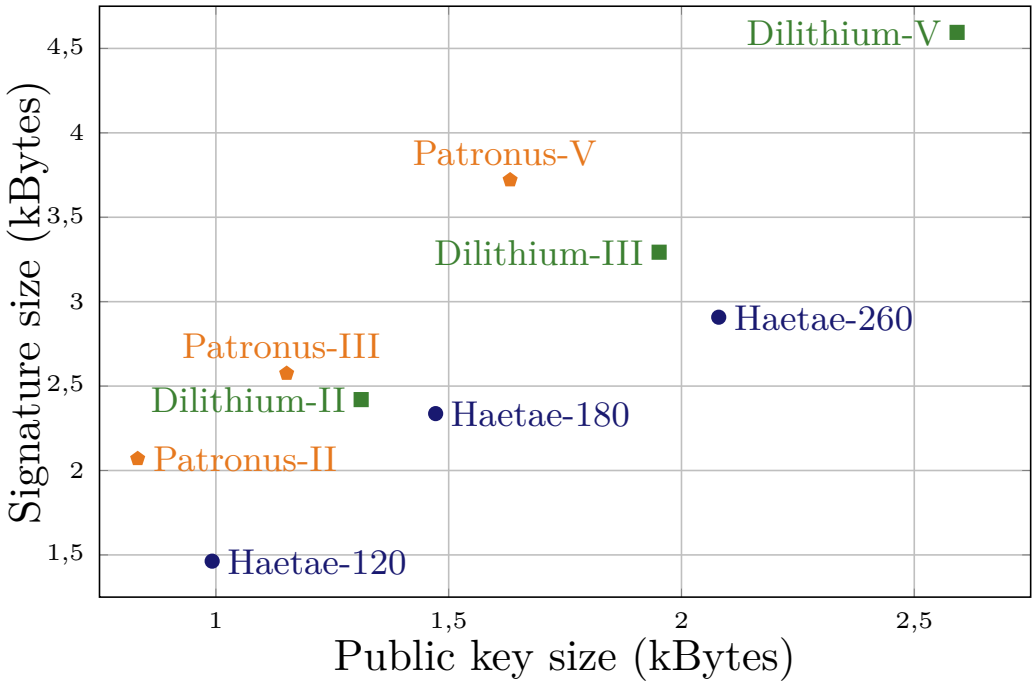


Figure 7.1: Comparing signature performances.

The only drawback that Patronus has compared with Dilithium is the complexity and speed of its sampler, and this is analysed in Table 7.4. Note that a 20-fold loss compared to Dilithium in sampler speed is not as bad as it sounds, because the sampler is only a small brick that accounts for less than 10% of signing time in Dilithium, which makes signing times for Patronus and Dilithium the same order of magnitude. The improvement over Haetae is clear, as Haetae's sampler accounts for roughly 80% of its signing time, even ignoring masking considerations.

Table 7.4: Running time (cycles) and randomness consumption (bytes) for Patronus *and* Dilithium *samplers using SHAKE-256.*

Table 7.5: SampleH *(this work).*

Speed	II	III	V
median	420,721	575,430	1,028,036
average	453,294	594,168	1,111,171
Randomness			
median	16,048	10,064	24,208
average	16,827	11,087	25,221

Table 7.6: ExpandMask *(Dilithium).*

Speed	II	III	V
median	24,152	29,732	42,262
average	24,173	29,943	41,968
Randomness	2,720	3,400	4,760

Desirable specifications In our setting, we use the Fiat-Shamir heuristic to convert an ID scheme into a signature. Recall the notations from the description of the lattice ID scheme in Section 3.4: once the prover receives the challenge c, it samples a mask $\mathbf{y}$ uniformly[5] from a set V_Y, and computes the value of $\mathbf{z} = \mathbf{y} + c\mathbf{s}$. This value must not leak any information regarding the secret-dependant value $c\mathbf{s}$. In order for this to happen, the prover rejects $\mathbf{z}$ and *aborts* if $\mathbf{z} \notin V_Z$. This makes the choice of V_Z crucial to ensure that the protocol yields a zero-knowledge proof of knowledge. In fact, if we denote V_{CS} the set of possible values for $c\mathbf{s}$ (where none of $\mathbf{s}$ and c are fixed), then $\mathbf{z}$ avoids any form of leakage if and only if

$$V_Z \subseteq \bigcap_{\mathbf{x} \in V_{CS}} (V_Y + \mathbf{x}). \tag{7.1}$$

Indeed if $\mathbf{z}$ is contained in all $V_Y + \mathbf{x}$ for $\mathbf{x} \in V_{CS}$, then whatever the secret value $\mathbf{x} = c\mathbf{s}$, there will exist a unique $\mathbf{y} \in V_Y$ that reaches $\mathbf{z}$. As $\mathbf{y}$ is sampled uniformly from V_Y, this makes $\mathbf{z}_1 \in V_Z$ obtained from $\mathbf{x}_1 \in V_{CS}$ statistically indistinguishable from any $\mathbf{z}_2$ obtained from $\mathbf{x}_1 \neq \mathbf{x}_2$.

When designing a scheme, we have to choose sets V_Z, V_Y and V_{CS}. Before explaining our choice, we make a list of properties that we would like them to satisfy:

1. **Integrality:** We require that $V_Y \cup V_{CS} \subset \mathbb{Z}^n$.

2. **Efficient membership testing:** In order to decide when to abort, there must be a straightforward way to check whether $\mathbf{z} \in V_Z$.

3. **Minimising aborts:** If $\mathbf{z} \notin V_Z$, the signing procedure will fail and be forced to restart. Therefore if V_Y and V_{CS} are fixed, then in order to minimise the expected number of restarts, V_Z should be maximal in Equation 7.1, that is

$$V_Z = \bigcap_{\mathbf{x} \in V_{CS}} (V_Y + \mathbf{x}). \tag{7.2}$$

4. **Approximating the ball:** Note that after transforming the ID protocol into a signature through the Fiat-Shamir transform, the signature will consist in both the accepted $\mathbf{z}$ and the challenge c. We are interested in minimising the size of such a tuple of elements. As explained in [DFPS22], c only contributes to a small fraction of the bitsize (the main requirement is that it has enough min-entropy for it to be difficult to guess), and the contribution of $\mathbf{z}$ towards signature length is mostly driven by $\|\mathbf{z}\|$. Therefore we aim to minimise the expected value of $\|\mathbf{z}\|$ when $\mathbf{z}$ is sampled according to the target distribution.

5. **Efficient sampling:** We must know an efficient sampling algorithm in V_Y to sample the mask $\mathbf{y}$.

[5] Uniform sampling is not required, but is a design choice we make.

Remark 7.1.3. Regarding condition 4., the paper [DFPS22] shows that both Gaussian distributions and uniform sampling in the Euclidean ball essentially reach a lower bound for signature size in this setting (this is the design rationale for Haetae), whereas the uniform hypercube sampler from Dilithium does not. In our study we focus only on uniform distributions, and when this is possible we illustrate the quality of our shape with regards to signature size by looking at the ratio between the radii of the circumscribed and inscribed Euclidean balls, with respect to V_Z. Heuristically, the circumradius corresponds to the signature size, whereas the inradius corresponds in some sense to the best possible signature size, for a given security target.

To the extent of our knowledge, none of the previous existing solutions succeeded in tackling all five constraints:

- **Euclidean balls:** This approach excels at point 4. (the ratio would be 1, the best possible result), but fails short of point 5. as efficiently sampling uniform integral points in a Euclidean ball is a notoriously tricky task. The approach used in [Che+23] requires Gaussian sampling and 128-bit fixed-point arithmetic.
- **Hypercubes:** Sampling in a hypercube as in[Duc+21] would be the best solution if not for point 4., the ratio of $\sqrt{n}$ between circum-/in-radius here is essentially the worst possible, resulting in larger signature sizes.
- **Gaussians:** The discrete Gaussian distribution used in [Lyu12] is optimal for point 4., but fails at point 2., as it is not straightforward to simulate such a distribution using rejection sampling.

Polytopes to the rescue We propose a balanced solution to tackles the five desired specifications, by introducing a generic framework for FSwA proofs of knowledge using uniform distributions in polytopes. If $\mathcal{P} \subset \mathbb{R}^n$ is a polytope, we use the notation $\mathcal{P}^n_{r,\mathbf{v}} := (\mathbf{v} + r\mathcal{P})$ to denote a shifted and scaled copy of $\mathcal{P}$. We omit $\mathbf{v}$ in the subscript if $\mathbf{v} = \mathbf{0}$, and add $\mathbb{Z}$ in subscript to denote intersection with $\mathbb{Z}^n$.

Overview In Section 7.2, we introduce a generic result on rejection sampling in the context of FSwA with uniform distributions for polytopes, and carefully prove a variant of this result when the distributions are uniform over the integral points in the polytopes, thereby answering Question 7.1.2. We note here that this framework generalises what has been done in Dilithium as a hypercube is an example of polytope.

In Section 7.3 we introduce and study the polytope

$$\mathcal{H}^n_r = B^{(\infty)}_n(r) \cap B^{(1)}_n(r\sqrt{n}).$$

It is clear that $\mathcal{H}^n_{r,\mathbb{Z}} = \mathcal{H}^n_r \cap \mathbb{Z}^n$ satisfies points 1. and 2. of our wish-list. The result of the previous section shows that it satisfies 3., and we show in Section 7.3 that $\mathcal{H}^n_{r,\mathbb{Z}}$ is a much more suitable candidate than the hypercube for point 4. by reaching a ratio of $\sqrt[4]{n}$. Finally, we study integer points in ℓ_1 balls in order to provide an efficient and isochronous[6] sampler for $\mathcal{H}^n_{r,\mathbb{Z}}$, thereby solving point 5.

In Section 7.4, we argue that one can post-reject values of $\mathbf{z} \in \mathcal{H}^n_{r,\mathbb{Z}}$ such that $\|\mathbf{z}\| \leq \theta r$ for a well-chosen constant value of θ, as this only occurs with a very small probability, because uniform points in an ℓ_1 ball tend to concentrate towards its center. Although the resulting shape $\mathcal{H}^n_{r,\mathbb{Z}} \cap B^{(2)}_n(\theta r)$ is not a polytope and some results from the general framework carry over, while others remain to be adapted.

We wrap all our contributions into one possible application of FSwA by introducing a new post-quantum signature scheme: Patronus. We prove its basic properties such as correctness

[6] By *isochronous* we mean: such that its runtime is independent of secret data, see [HPRR20].

and security properties such as UF-CMA in the QROM using [KLS18; Bar+23; DFPS23]. As can be seen in Table 7.1, this scheme has non-negligibly shorter signatures compared to Dilithium [Duc+21] (around 25% shorter) while providing shorter public keys (around 13%). Sampling is exclusively based on uniform distributions in integral intervals and not fixed-point Gaussian distributions meaning it uses less randomness and should be much easier to protect against side-channel attacks compared to Haetae [Che+23]. This solves Question 7.1.1.

7.2 Rejection Sampling on Polytopes

In this section, we explain how one can instantiate the sets V_Y, V_Z, V_{CS} mentioned in the introduction of this chapter using polytopes. This choice generalises Dilithium and allows for two convenient properties: firstly for well-chosen polytopes, exact characterisations of integral points exist, and secondly polytopes are a versatile tool when attempting to approximate a convex body, *e.g.* for our intents and purposes, the Euclidean ball.

We show that polytopes are a suitable candidate for Equation 7.2, translating the framework of distribution indistinguishability through using intersections of volumes over which probability densities are taken to be homogeneous. Because we only consider uniform distributions, we can conveniently write out everything in terms of intersections of shapes.

This section studies both the continuous and more importantly discrete setting (restricting to $\mathbb{Z}^n$) for FSwA, minimising the rejection rate corresponding to a given target uniform distribution over a polytope. For this, we first prove an important characterisation of a special intersection of polytopes, and then we share our main general theorem on rejection sampling using discrete uniform distributions. $n \in \mathbb{Z}_{>0}$ refers to the dimension of our ambient space throughout.

7.2.1 Convex-Ception: Intersecting Polytopes

Assuming we are given V_Y and V_{CS}, what does $\cap_{\mathbf{x}\in V_{CS}}(V_Y + \mathbf{x})$ look like? We choose to focus on the case where V_{CS} and V_Y are rescaled versions of a symmetric convex set $\mathcal{S}$, *e.g.* balls for the norm $\|\cdot\|_{\mathcal{S}}$ defined by $\|\mathbf{x}\|_{\mathcal{S}} = \inf\{\lambda \geq 0 : \mathbf{x} \in \mathcal{S}_\lambda\}$.

Lemma 7.2.1. *For any symmetric convex set $\mathcal{S}$, let $r, R \in \mathbb{R}_{>0}$ be two radii such that $R > r$. Then:*

$$\mathcal{S}_{R-r} = \bigcap_{\mathbf{c}\in\mathcal{S}_r} \mathcal{S}_{R,\mathbf{c}}.$$

Proof. Starting with the direct inclusion, let $\mathbf{x} \in \mathcal{S}_{R-r}$ and $\mathbf{c} \in \mathcal{S}_r$, we have

$$\|\mathbf{x} - \mathbf{c}\|_{\mathcal{S}} \leq \|\mathbf{x}\|_{\mathcal{S}} + \|\mathbf{c}\|_{\mathcal{S}} \leq (R - r) + r = R.$$

Since $\mathbf{x} - \mathbf{c} \in \mathcal{S}_R$ if and only if $\mathbf{x} \in \mathcal{S}_{R,\mathbf{c}}$, we get $\mathcal{S}_{R-r} \subseteq \cap_{\mathbf{c}\in\mathcal{S}_r}\mathcal{S}_{R,\mathbf{c}}$.

We prove the reverse inclusion by contraposition. Namely, any vector not in $\mathcal{S}_{R-r}$ is not in $\cap_{\mathbf{c}\in\mathcal{S}_r}\mathcal{S}_{R,\mathbf{c}}$. Let $\mathbf{z} \notin \mathcal{S}_{R-r}$, *i.e.* such that $\|\mathbf{z}\|_{\mathcal{S}} > R-r$. Let $\mathbf{x} = \frac{\mathbf{z}}{\|\mathbf{z}\|_{\mathcal{S}}}$ and $\mathbf{c} = -r\mathbf{x}$. By definition $\|\mathbf{x}\|_{\mathcal{S}} = 1$ and by symmetry of $\mathcal{S}$, $\mathbf{c} \in \mathcal{S}_r$. Therefore

$$\|\mathbf{z} - \mathbf{c}\|_{\mathcal{S}} = \|\|\mathbf{z}\|_{\mathcal{S}} \cdot \mathbf{x} + r\mathbf{x}\|_{\mathcal{S}} = \|\mathbf{z}\|_{\mathcal{S}} + r > R,$$

and this concludes our proof. □

Lemma 7.2.1 deals with the continuous case, and remains very general (Note that it includes the ℓ_2 and ℓ_∞ balls studied by [Che+23; Duc+21]). We show that the result still holds with symmetric polytopes that have integer vertices.

Lemma 7.2.2. *Let $\mathcal{P}$ be a convex set, and $\mathbf{a} \in \mathbb{R}^n$ a vector. Then we have*

$$\mathcal{P} \cap (\mathcal{P} + \mathbf{a}) = \bigcap_{t \in [0,1]} (\mathcal{P} + t\mathbf{a}).$$

As a consequence, for any convex set $\mathcal{P}$ and polytope $\mathcal{Q}$:

$$\bigcap_{\mathbf{c} \in \mathcal{Q}} (\mathcal{P} + \mathbf{c}) = \bigcap_{\mathbf{c} \in \partial\mathcal{Q}} (\mathcal{P} + \mathbf{c}) = \bigcap_{\mathbf{c} \in \mathcal{V}(\mathcal{Q})} (\mathcal{P} + \mathbf{c}).$$

Proof. We start with the first equation, *i.e.* $\mathcal{P}$ is convex. Let $\mathbf{x} \in \mathcal{P} \cap (\mathcal{P} + \mathbf{a})$. Then there exists $\mathbf{y} \in \mathcal{P}$ such that $\mathbf{x} = \mathbf{y} + \mathbf{a}$. Let $t \in [0, 1]$. Then $\mathbf{x} = (\mathbf{y} + (1 - t)\mathbf{a}) + t\mathbf{a}$, where $\mathbf{y} + (1 - t)\mathbf{a} = (1 - t)\mathbf{x} + t\mathbf{y}$ and therefore lives in $\mathcal{P}$ by convexity. Thus $\mathbf{x} \in \mathcal{P} + t\mathbf{a}$ and we have proved the direct inclusion. The reverse inclusion is trivial. We now establish the second statement by induction on the number of vertices of $\mathcal{Q}$. The case $|\mathcal{V}(\mathcal{Q})| = 2$ has been dealt with. Now let $\mathcal{P}$ be a convex set and $\mathcal{Q}$ a polytope with $m + 1$ vertices $\mathbf{a}_1, \ldots, \mathbf{a}_{m+1}$, and suppose the result holds for convex hulls with fewer vertices. Let $\mathbf{c} = \sum_{i=1}^{m+1} t_i \mathbf{a}_i \in \mathcal{Q}$, then $\mathbf{c} = (1 - t_{m+1})\mathbf{c}' + t_{m+1}\mathbf{a}_{m+1}$, where $\mathbf{c}' \in \mathcal{Q}'$ and $\mathcal{V}(\mathcal{Q}') = (\mathbf{a}_i)_{i \leq m}$. Reciprocally, any point of $\mathbf{c}' \in \mathcal{Q}'$ gives a segment $[\mathbf{c}', \mathbf{a}_{m+1}] \subseteq \mathcal{Q}$. Now using the first identity multiple times,

$$\begin{aligned}\bigcap_{\mathbf{c} \in \mathcal{Q}} (\mathcal{P} + \mathbf{c}) &= \bigcap_{\mathbf{c}' \in \mathcal{Q}'} \bigcap_{t \in [0,1]} (\mathcal{P} + (1 - t)\mathbf{c}' + t\mathbf{a}_{m+1}) = \bigcap_{\mathbf{c}' \in \mathcal{Q}'} (\mathcal{P} + \mathbf{c}') \cap (\mathcal{P} + \mathbf{a}_{m+1}) \\ &= (\mathcal{P} + \mathbf{a}_{m+1}) \cap \bigcap_{\mathbf{c}' \in \mathcal{Q}'} (\mathcal{P} + \mathbf{c}').\end{aligned}$$

We conclude by our induction hypothesis. Note that the vertices are contained in the boundary so we don't bother with the middle term of the last statement.

□

Proposition 7.2.3 ($\mathcal{P}$-ception: Intersection of polytopes)**.** *Let $\mathcal{P}$ be a symmetric polytope. Let $r, R \in \mathbb{R}$ such that $R > r > 0$. Then:*

$$\bigcap_{\mathbf{c} \in \mathcal{P}_r} \mathcal{P}_{R,\mathbf{c}} = \bigcap_{\mathbf{c} \in \partial\mathcal{P}_r} \mathcal{P}_{R,\mathbf{c}} = \bigcap_{\mathbf{c} \in \mathcal{V}(\mathcal{P}_r)} \mathcal{P}_{R,\mathbf{c}} = \mathcal{P}_{R-r}.$$

In particular, if $\mathcal{V}(\mathcal{P}_r) \subset \mathbb{Z}^n$, $\bigcap_{\mathbf{c} \in \mathcal{P}_{r,\mathbb{Z}}} \mathcal{P}_{R,\mathbf{c},\mathbb{Z}} = \mathcal{P}_{R-r,\mathbb{Z}}$.

Proof. The first statement is a direct application of Lemma 7.2.1, as polytopes are convex by definition. The second statement follows from Lemma 7.2.2, indeed because the vertices of $\mathcal{P}_r$ are integral,

$$\bigcap_{\mathbf{c} \in \mathcal{P}_{r,\mathbb{Z}}} \mathcal{P}_{R-r,\mathbf{c}} = \bigcap_{\mathbf{c} \in \mathcal{V}(\mathcal{P}_r)} \mathcal{P}_{R-r,\mathbf{c}} = \mathcal{P}_{R-r},$$

and intersecting with $\mathbb{Z}^n$ yields the result. □

In general and for practical purposes, it is preferable that the set V_{CS} (corresponding to $\mathcal{P}_{r,\mathbb{Z}}$ in the previous proposition) of points defining the translation of the main polytope lies within a Euclidean ball. As such, we need a stronger result for $\mathcal{P}$-ception. We solve this issue by proving that if each facet of $\mathcal{P}_r$ contains an integral point, Proposition 7.2.3 still holds. When the polytope allows it, this enables us to replace P_r by an inscribed Euclidean ball that is tangent to all of its facets at integral points.

Corollary 7.2.4. *Let $\mathcal{P}$ be a symmetric polytope, and $R > r > 0$ two real radii. For any function $h : \mathcal{F}(\mathcal{P}_r) \to \mathbb{R}^n$ that maps a facet $F \in \mathcal{F}(P)$ to one of its points $\mathbf{h}_F \in \mathcal{F}(\mathcal{P}_r)$,*

$$\bigcap_{F \in \mathcal{F}(\mathcal{P}_r)} (\mathcal{P}_R + \mathbf{h}_F) = \mathcal{P}_{R-r}.$$

Proof. To show the first inclusion, we use the characterisation of polytopes from its facets. Given a facet $F \in \mathcal{F}(\mathcal{P})$, we denote by $H(F)$ the half-space containing zero that is defined by the hyperplane containing F. Notice that $\mathcal{P} = \cap_{F\in\mathcal{F}(\mathcal{P})} H(F)$. Using the same subscript notation that is used for rescaling polytopes in the context of facets, and the fact that $-F_R$ is a facet of $\mathcal{P}_R$, we can write $\mathcal{P}_R + \mathbf{h}_{F_r} \subset H(-F_R) + \mathbf{h}_{F_r} = H(-F_{R-r})$. Inclusion is preserved by intersecting over all facets, therefore

$$\bigcap_{F\in\mathcal{F}(\mathcal{P}_r)} (\mathcal{P}_R + \mathbf{h}_F) = \bigcap_{F\in\mathcal{F}(\mathcal{P}_r)} H(-F_{R-r}) = \bigcap_{F\in\mathcal{F}(\mathcal{P}_r)} H(F_{R-r}) = \mathcal{P}_{R-r}.$$

The reverse inclusion follows by Proposition 7.2.3. □

7.2.2 Fiat-Shamir with Aborts and Polytopes

Rejection sampling will allow us to convert the source distribution (that of $\mathbf{y} + c\mathbf{s}$) into a target distribution that is uniform on the set V_Z which will be defined by the result of Proposition 7.2.3. In this section we formalise rejection sampling in this context.

We use Lemma 3.2.3, which in our case says the following: as long as the Rényi divergence[7] between the uniform distribution on V_Z and the uniform distribution on $V_Y + \mathbf{v}$ for $\mathbf{v} \in V_{CS}$ is smaller than some bound $M > 1$, then perfect rejection sampling is possible: we obtain a distribution that is statistically indistinguishable from the target distribution. M is exactly the expected value for the number of aborts.

Computing the (infinite-)Rényi divergence between uniform distributions in the continuous setting amounts to computing the ratio between the volume of the supports. In general, estimating the volume of a polytope can be a delicate task [DF88; CCF22]. However, due to our special choice of polytopes that are scaled versions of each other, all complications disappear.

Lemma 7.2.5. *Let $\mathcal{P}$ be a symmetric polytope, $\mathbf{v} \in \mathbb{R}^n$ and let $\beta, r, R > 0$ be real numbers such that $R \geq r + \beta$ and $\mathbf{v} \in \mathcal{P}^n_\beta$. Then:*

$$\mathcal{R}_\infty \left[\mathcal{U}\left(\mathcal{P}^n_r\right) \,\|\, \mathcal{U}\left(\mathcal{P}^n_{R,\mathbf{v}}\right) \right] = \left(\frac{R}{r}\right)^n.$$

In particular, if $R = r + \beta$ and $M > 1$, then the inequality $\left(\frac{R}{r}\right)^n \leq M$ holds if and only if r satisfies the condition $r \geq \frac{\beta}{M^{\frac{1}{n}} - 1}$.

Proof. Let $\mathcal{P}$ be a symmetric polytope and $\mathbf{v} \in \mathcal{P}^n_\beta$, by applying Proposition 7.2.3 we have: $\mathcal{P}^n_{R-\beta} = \bigcap_{\mathbf{c}\in\mathcal{P}^n_\beta} \mathcal{P}^n_{R,\mathbf{c}} \subset \mathcal{P}^n_{R,\mathbf{v}}$. With $r = R - \beta$ the Rényi divergence is well-defined. The desired Rényi divergence is then obtained directly. Finally for $M > 1$, by fixing $R = r + \beta$ we can derive the following equivalences:

$$\left(\frac{R}{r}\right)^n \leq M \quad \Leftrightarrow \quad r + \beta \leq r \cdot M^{\frac{1}{n}} \quad \Leftrightarrow \quad r \geq \frac{\beta}{M^{\frac{1}{n}} - 1}.$$

□

Computing the exact number of integral points inside a generic high-dimensional polytope should be at least as difficult as estimating its volume. If the radius is large, a Gaussian heuristic argument tells us that we should expect both the volume and the number of integral points to be very close approximations of each other.

In the discrete setting, the ratio of volumes should be replaced by a ratio of cardinalities, for which the scaling argument used in Lemma 7.2.5 breaks down. However as the argument is *morally* the same, we explain how one can obtain a precise rejection sampling result in the discrete setting, using the following definition.

[7] In our setting, where we only have to deal with uniform distributions, this should read as a fancy word for the ratio of the volume of the supports.

Definition 7.2.6. We define the *defect* of a convex set $\mathcal{S}$ as

$$\varepsilon(\mathcal{S}) := \frac{|\mathcal{S}_{\mathbb{Z}}|}{\mathrm{vol}(\mathcal{S})} - 1.$$

Remark 7.2.7. Note that the defect is not always positive.

- If $\mathcal{S} = [-r, r]^n$ for a positive integer r, $\varepsilon(\mathcal{S}) = \left(1 + \frac{1}{2r}\right)^n - 1$.
- Estimating the defect of $B_n^{(2)}(r)$ is exactly the Gauss circle problem in n dimensions.

Proposition 7.2.8. *Let $\mathcal{P}$ be a symmetric circumscribed polytope, $M > 1$ and $\beta > 0$ be real numbers such that $\mathcal{V}\left(\mathcal{P}_\beta^n\right) \subset \mathbb{Z}^n$, and let $\mathbf{v} \in \mathcal{B}_\beta$ where $\mathcal{B}$ is the Euclidean ball that is tangent to all facets of $\mathcal{P}$. Then if $\mathcal{B}_\beta$ and $\mathcal{P}_\beta^n$ are tangent at integral points only, then for $r \geq \frac{\beta}{M^{1/n}-1}$, $R = r + \beta$, and any $M' \geq \frac{1+\varepsilon(\mathcal{P}_R^n)}{1+\varepsilon(\mathcal{P}_r^n)} \cdot M$,*

$$\mathcal{R}_\infty\left[\,\mathcal{U}\left(\mathcal{P}_{r,\mathbb{Z}}^n\right) \parallel \mathcal{U}\left(\mathcal{P}_{R,\mathbf{v},\mathbb{Z}}^n\right)\,\right] = \left(\frac{R}{r}\right)^n \cdot \frac{1+\varepsilon(\mathcal{P}_R^n)}{1+\varepsilon(\mathcal{P}_r^n)} \leq M'.$$

Proof. As $\mathcal{P}_\beta^n$ is a circumscribed symmetric polytope with integral vertices and tangent points, from Corollary 7.2.4 it is equivalent to consider $\mathbf{v} \in \mathcal{P}_\beta^n$ and $\mathbf{v} \in \mathcal{B}_\beta$. By Proposition 7.2.3 we obtain $\mathcal{P}_{R-\beta,\mathbb{Z}}^n = \bigcap_{\mathbf{c}\in\mathcal{P}_{\beta,\mathbb{Z}}^n} \mathcal{P}_{R,\mathbf{c},\mathbb{Z}}^n \subset \mathcal{P}_{R,\mathbf{v},\mathbb{Z}}^n$. Thus the Rényi divergence in the statement is well-defined, and the equality follows from rewriting:

$$\mathcal{R}_\infty\left[\,\mathcal{U}\left(\mathcal{P}_{r,\mathbb{Z}}^n\right) \parallel \mathcal{U}\left(\mathcal{P}_{R,\mathbf{v},\mathbb{Z}}^n\right)\,\right] = \frac{|\mathcal{P}_{R,\mathbb{Z}}^n|}{|\mathcal{P}_{r,\mathbb{Z}}^n|} = \frac{\mathrm{vol}(\mathcal{P}_R^n)}{\mathrm{vol}(\mathcal{P}_r^n)} \cdot \frac{|\mathcal{P}_{R,\mathbb{Z}}^n|}{\mathrm{vol}(\mathcal{P}_R^n)} \cdot \frac{\mathrm{vol}(\mathcal{P}_r^n)}{|\mathcal{P}_{R,\mathbb{Z}}^n|}.$$

□

Proposition 7.2.3 contributes doubly to Proposition 7.2.8: it first proves the existence of our Rényi divergence by ensuring that the shifted and rescaled polytope contains the support of the target distribution. Second, it helps minimise the expected number of aborts M, which has direct consequences regarding signing speed. Alternatively, one can fixed a value M that corresponds to the desired speed, and use Proposition 7.2.8 to optimise for $R - r$, which affects signature size. We can now state our main theorem, that enables generic FSwA rejection sampling using uniform distributions over polytopes:

Theorem 7.2.9 (Rejection sampling for $\mathcal{U}\left(\mathcal{P}_{\mathbb{Z}}^n\right)$)**.** *Let $\mathcal{P}$ be a symmetric circumscribed polytope, $M > 1$ and $\beta > 0$ be real numbers such that $\mathcal{V}\left(\mathcal{P}_\beta^n\right) \subset \mathbb{Z}^n$, and $\mathcal{B}_\beta$ and $\mathcal{P}_\beta^n$ are tangent at integral points only, where $\mathcal{B}$ is the Euclidean ball that is tangent to all facets of $\mathcal{P}$. Let h be a probability distribution with $\mathsf{Supp}(h) \subseteq \mathcal{B}_\beta$. Let $r \geq \frac{\beta}{M^{1/n}-1}$, $R \geq r + \beta$, and $M' = \frac{1+\varepsilon(\mathcal{P}_R^n)}{1+\varepsilon(\mathcal{P}_r^n)} \cdot M$. Let $\mathbf{v} \in \mathsf{Supp}(h)$ and denote $\rho_{r,\mathbf{v}}^n := \mathcal{U}\left(\mathcal{P}_{r,\mathbf{v},\mathbb{Z}}^n\right)$ with the same subscript convention as usual. If $M' > 1$ then:*

$$\mathcal{R}_\infty\left[\,\mathcal{U}\left(\mathcal{P}_{r,\mathbb{Z}}^n\right) \parallel \mathcal{U}\left(\mathcal{P}_{R,\mathbf{v},\mathbb{Z}}^n\right)\,\right] = \left(\frac{R}{r}\right)^n \cdot \frac{1+\varepsilon(\mathcal{P}_R^n)}{1+\varepsilon(\mathcal{P}_r^n)} = M',$$

and the two algorithms $\mathcal{A}$ and $\mathcal{F}$ below have indistinguishable output distributions:

$\mathcal{A}$	$\mathcal{F}$
$\mathbf{v} \leftarrow\!\!\$\ h$	$\mathbf{v} \leftarrow\!\!\$\ h$
$\mathbf{z} \leftarrow\!\!\$\ \rho_{R,\mathbf{v}}^n$	$\mathbf{z} \leftarrow\!\!\$\ \rho_r^n$
output $(\mathbf{z}, \mathbf{v})$ *if* $\mathbf{z} \in \mathcal{P}_{r,\mathbb{Z}}^n$, *else* $\perp$	*output* $(\mathbf{z}, \mathbf{v})$ *with probability* $1/M'$, *else* $\perp$

Furthermore, $\mathcal{A}$ outputs $(\mathbf{z}, \mathbf{v})$ with probability $1/M'$.

Proof. Given $\mathbf{z} \in \mathcal{P}^n_{R,\mathbf{v},\mathbb{Z}}$, $\min\left(\frac{\rho^n_r(\mathbf{z})}{M' \cdot \rho^n_{R,\mathbf{v}}(\mathbf{z})}, 1\right)$ is either equal to 0 if $\mathbf{z} \notin \mathcal{P}^n_{r,\mathbb{Z}}$ or 1 otherwise. Indeed, we know by Proposition 7.2.8 that:

$$\frac{\rho^n_r(\mathbf{z})}{M' \cdot \rho^n_{R,\mathbf{v}}(\mathbf{z})} = \begin{cases} \geq 1 \text{ if } \rho^n_r(\mathbf{z}) \neq 0 \\ = 0 \text{ if } \rho^n_r(\mathbf{z}) = 0 \end{cases}.$$

Our aim is to apply Lemma 3.2.3 to the distributions $D_s = \{(\mathbf{z}, \mathbf{v}) : \mathbf{v} \hookleftarrow\$ h \wedge \mathbf{z} \hookleftarrow\$ \rho^n_{R,\mathbf{v}}\}$ and $D_t = \{(\mathbf{z}, \mathbf{v}) : \mathbf{v} \hookleftarrow\$ h \wedge \mathbf{z} \hookleftarrow\$ \rho^n_r\}$. First we verify that they both satisfy conditions of this lemma.

Proposition 7.2.3 implies that $\mathsf{Supp}(D_t) \subseteq \mathsf{Supp}(D_s)$. Because we assume $M > 1$, $r \geq \frac{\beta}{M^{\frac{1}{n}} - 1}$ and $R \geq r + \beta$, we get:

$$\begin{aligned} R_\infty(\mathcal{D}_s \parallel \mathcal{D}_t) &= \max_{\mathbf{z} \in \mathcal{P}^n_{r,\mathbb{Z}}, \mathbf{v} \in \mathsf{Supp}(h)} \frac{\mathcal{D}_s((\mathbf{z}, \mathbf{v}))}{\mathcal{D}_t((\mathbf{z}, \mathbf{v}))} = \max_{\mathbf{z} \in \mathcal{P}^n_{r,\mathbb{Z}}, \mathbf{v} \in \mathsf{Supp}(h)} \frac{h(\mathbf{v})\, \rho^n_r(\mathbf{z})}{h(\mathbf{v})\, \rho^n_{R,\mathbf{v}}(\mathbf{z})} \\ &= \max_{\mathbf{v} \in \mathsf{Supp}(h)} \left(\max_{\mathbf{z} \in \mathcal{P}^n_{r,\mathbb{Z}}} \frac{\rho^n_r(\mathbf{z})}{\rho^n_{R,\mathbf{v}}(\mathbf{z})} \right) = \max_{\mathbf{v} \in \mathsf{Supp}(h)} \mathcal{R}_\infty \left[\rho^n_r \parallel \rho^n_{R,\mathbf{v}} \right] \\ &\leq \left(\frac{R}{r}\right)^n \cdot \frac{1 + \varepsilon(\mathcal{P}^n_R)}{1 + \varepsilon(\mathcal{P}^n_r)} \leq M \cdot \frac{1 + \varepsilon(\mathcal{P}^n_R)}{1 + \varepsilon(\mathcal{P}^n_r)} = M', \end{aligned}$$

where the last line uses Proposition 7.2.8 and the definitions of r, R and M'. Now if $M' > 1$, we can apply Lemma 3.2.3 and conclude. □

Remark 7.2.10. Theorem 7.2.9 acts as our main tool to compute rejection rates. It is a rigorous translation of the more straightforward continuous result to the discrete setting, and requires exact computation of the defects $\varepsilon(\mathcal{P}^n_r)$ and $\varepsilon(\mathcal{P}^n_R)$ to prove effective. This pre-computation will only be required a single time. In practice, and when computing the defect is a harder problem than the one underlying the security of the scheme, any pair of bounds on the defects will give a valid M'. Tighter bounds will lead to smaller rejection rates. The tempting approximation $\varepsilon(\mathcal{P}^n_r) = \varepsilon(\mathcal{P}^n_R) = 0$ could either break statistical indistinguishability or lead to a non-optimal rejection rate. In practice, it feels like unless the choice of polytopes and their radii is very unfortunate, computational indistinguishability should still hold, but we won't discuss this further here.

7.3 Introduction to Gemstone Cutting

In this section we choose a polytope that we will later use to instantiate the results of Section 7.2.

Recall that we want a polytope in which we can easily sample (ie without using Gaussians or precision management). We also will need to count its number of integral points, although this step does not need to be efficient, only efficient enough to be pre-computed once.

Crucially for signature size, we also would like the chosen polytope to be a tight approximation of the Euclidean ball, in the sense that we want to minimise the ratio between circumradius and inradius[8]. This ratio is invariant under scaling.

We already know that the Euclidean ball reaches an optimal ratio of 1, and that a hypercube reaches a ratio of $\sqrt{n}$. It is not difficult to prove that the cross-polytope also reaches this same ratio of $\sqrt{n}$ (this is unsurprising as the hypercube and cross-polytope are dual polytopes). Before we discard the cross-polytope as an option, notice (the computation was carried out in Chapter 2) that the expected Euclidean norm of a uniform random point in a hypercube is much larger than the corresponding quantity in the cross-polytope. This means that we would benefit from somehow discarding the points in the cross-polytope $B^{(1)}_n(r)$ that have large ℓ_2 or ℓ_∞ norm.

[8] Assuming the polytope is circumscribed, and this is another requirement.

This leads us to propose a polytope that has better ℓ_2 norm ratio compared to the hypercube ($\sqrt[4]{n}$ instead of $\sqrt{n}$) and a much friendlier sampler compared to the Euclidean ball. We divide this section into four parts: we first define and characterise our special polytope $\mathcal{H}$, and explain why it fits perfectly our previous framework from Section 7.2. We then design an efficient isochronous algorithm for uniform sampling in $\mathcal{H}_{\mathbb{Z}}$. Finally, we provide a worst-case improvement on $\mathcal{H}$, that leads to a constant ℓ_2 norm ratio, under mild heuristic assumptions.

7.3.1 Characterisation of $\mathcal{H}$

Definition 7.3.1. For $n \in \mathbb{Z}_{>0}$ and $r \in \mathbb{R}_{>0}$, we define

$$\mathcal{H}_r^n := B_n^{(\infty)}(r) \cap B_n^{(1)}(r\sqrt{n}).$$

As $\mathcal{H}$ is defined as an intersection of full-rank polytopes, and it contains an open ball, so by Proposition 3.1.5, $\mathcal{H}$ is also a full-rank polytope. The radii of the hypercube and cross-polytope are chosen so that both polytopes share the same circumscribed Euclidean sphere.

Remark 7.3.2. We use the simplified notation $\mathcal{H}$ when both the dimension n and radius r can be omitted without ambiguity. Additionally and in practice, we will restrict r to the positive integers, a distinction that will be useful when using Theorem 7.2.9.

Proposition 7.3.3. *For $n \in \mathbb{Z}_{>0}$ and $r \in \mathbb{R}_{>0}$, $\mathcal{H}_r^n$ is a symmetric inscribed and circumscribed polytope with radius $r\sqrt{\lfloor\sqrt{n}\rfloor + (\sqrt{n} - \lfloor\sqrt{n}\rfloor)^2} \leq r\sqrt[4]{n}$. Equality is achieved when $\sqrt{n} \in \mathbb{Z}$. The vertices of this polytope are all of the form $(\underbrace{r, \ldots, r}_{\lfloor\sqrt{n}\rfloor}, \{\sqrt{n}\}r, 0, \ldots, 0)$ up to signed permutation,*

where $\{\sqrt{n}\} = \sqrt{n} - \lfloor\sqrt{n}\rfloor$.

Proof. From its definition, the polytope $\mathcal{H}_r^n$ is stable under signed permutation of coordinates, in particular it is symmetric. Let $\mathbf{v} \in \mathcal{V}(\mathcal{H}_r^n)$ be a vertex of $\mathcal{H}_r^n$. Without any loss of generality we can assume its coordinates are non-negative and sorted in decreasing order. By Proposition 3.1.6, it is impossible that $\mathbf{v}$ has two coordinates $0 < v_j < v_i < r$ for $i < j$. Indeed if that were the case, we could then write $\mathbf{v} = \frac{\mathbf{v}_1+\mathbf{v}_2}{2}$, where $\mathbf{v}_1 = \mathbf{v}+\varepsilon\mathbf{e}_i-\varepsilon\mathbf{e}_j$ and $\mathbf{v}_2 = \mathbf{v}-\varepsilon\mathbf{e}_i+\varepsilon\mathbf{e}_j$ are both in $\mathcal{H}_r^n$ for a small enough $\varepsilon > 0$ (here $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ denotes the canonical basis for $\mathbb{R}^n$). Using the ℓ_1 norm condition, this shows that $\mathbf{v}$ is of the form $(r, \ldots, r, \{\sqrt{n}\}r, 0, \ldots, 0)$, where $\{\sqrt{n}\} = \sqrt{n} - \lfloor\sqrt{n}\rfloor$ and the first $\lfloor\sqrt{n}\rfloor$ coordinates are equal to r. This proves the first part of the statement, *i.e.* $\mathcal{H}_r^n$ is inscribed with radius $r\sqrt{\lfloor\sqrt{n}\rfloor + \{\sqrt{n}\}^2}$. The inequality follows from the fact that $\{\sqrt{n}\}^2 \leq \{\sqrt{n}\}$, with equality if and only if $\{\sqrt{n}\} = 0$, exactly when n is a perfect square. A little extra effort with Proposition 3.1.6 can also show that all points of the aforementioned form are indeed vertices of $\mathcal{H}_r^n$. By contradiction if wlog $\mathbf{v} = (r, \ldots, r, \{\sqrt{n}\}r, 0, \ldots, 0)$ is not a vertex, then there we can write $\mathbf{v} = t\mathbf{v}_1 + (1-t)\mathbf{v}_2$ for a $t \in (0, 1)$, and $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}_r^n$ not equal to $\mathbf{v}$. The first $\lfloor\sqrt{n}\rfloor$ coordinates of both $\mathbf{v}_1$ and $\mathbf{v}_2$ must be r, otherwise one would escape $B_n^{(\infty)}(r)$. The $(1+\lfloor\sqrt{n}\rfloor)$-th coordinate of $\mathbf{v}_1$ and $\mathbf{v}_2$ must also be $\{\sqrt{n}\}r$, otherwise one of them escapes from $B_n^{(1)}(r\sqrt{n})$. This forces the last coordinates to be 0, and $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{v}$, this is a contradiction. It remains to show that $\mathcal{H}_r^n$ is circumscribed. Indeed, the ball $B_n^{(2)}(r)$ is tangent to all $2n$ facets of $B_n^{(\infty)}(r)$, and also to all 2^n facets of $B_n^{(1)}(r\sqrt{n})$ by taking the dual of the fact that $B_n^{(\infty)}(r)$ is inscribed. The tangency points are all signed permutations of $(r, 0, \ldots, 0)$ and $(r/\sqrt{n}, \ldots, r/\sqrt{n})$, and all of those are in $\mathcal{H}_r^n$. As facet of $\mathcal{H}_r^n$ can only be obtained from facets of $B_n^{(\infty)}(r)$ or $B_n^{(1)}(r\sqrt{n})$, we can conclude that $B_n^{(2)}(r)$ is tangent to all facets of $\mathcal{H}_r^n$. □

In Proposition 7.3.3 we proved different properties of $\mathcal{H}_r^n$:

- All of its vertices lie on a sphere of radius approximately $\sqrt[4]{n}$. In particular, they are integral as soon as $r \in \mathbb{Z}_{>0}$ and $\sqrt{n} \in \frac{1}{r}\mathbb{Z}$, in particular this is the case when $r \in \mathbb{Z}_{>0}$ and n is a perfect square.
- All of its facets are tangent to $B_n^{(2)}(r)$.

Regarding the first point, we need $\{\sqrt{n}\}r$ to be an integer, but this can be circumvented by using the following elementary Lemma on the ℓ_1 norm.

Lemma 7.3.4. *For $n \in \mathbb{Z}_{>0}$ and $r \in \mathbb{R}$, $B_{n,\mathbb{Z}}^{(1)}(r) = B_{n,\mathbb{Z}}^{(1)}(\lfloor r \rfloor)$.*

Now restriction to $\mathbb{Z}^n$ can be handled by Proposition 7.3.3 and Lemma 7.3.4:

Corollary 7.3.5. *Let $n, r \in \mathbb{Z}_{\geq 0}$, and $\{\sqrt{n}\} = \sqrt{n} - \lfloor \sqrt{n} \rfloor$. Then:*

$$\mathcal{H}_{r,\mathbb{Z}}^n = B_{n,\mathbb{Z}}^{(\infty)}(r) \cap B_{n,\mathbb{Z}}^{(1)}(\lfloor \sqrt{n} \rfloor r + \lfloor \{\sqrt{n}\} r \rfloor).$$

In particular by Proposition 7.2.3, for positive integers R and β,

$$\bigcap_{\mathbf{c} \in \mathcal{H}_{\beta,\mathbb{Z}}^n} \mathcal{H}_{R,\mathbf{c},\mathbb{Z}}^n = \mathcal{H}_{R-\beta,\mathbb{Z}}^n,$$

as $B_{n,\mathbb{Z}}^{(1)}(\lfloor \sqrt{n} \rfloor \beta + \lfloor \{\sqrt{n}\} \beta \rfloor)$ is an integral polytope.

In the discrete case, the second point alone is not sufficient to make the Theorem 7.2.9 work, as we would also need the tangency points to all have integral coordinates. This happens exactly when r and $r/\sqrt{n}$ are both integers. See also [Beg24, Lemma 3.3.3] to understand what happens outside of the case where $\sqrt{n}$ divides r.

7.3.2 Rejection Sampling on $\mathcal{H} \cap \mathbb{Z}^n$

We restate our main rejection sampling theorem Theorem 7.2.9 applied directly to $\mathcal{H}_{\mathbb{Z}}$, which has been seen in Section 7.3.1 to satisfy the conditions of the theorem. In this subsection we study the volume and number of integral points of $\mathcal{H}$, which allows us to estimate the defect $\varepsilon(\mathcal{H})$.

Corollary 7.3.6 (Rejection sampling for $\mathcal{U}(\mathcal{H}_{\mathbb{Z}})$)**.** *Assume n is a perfect square. Let $M > 1$ be a real number and $\beta > 0$ an integer such that n divides β^2. Let h be a probability distribution with $\mathsf{Supp}(h) \subseteq B_n^{(2)}(\beta) \cap \mathbb{Z}^n$. Let $r \geq \frac{\beta}{M^{1/n}-1}$, $R = r + \beta$, and $M' = \frac{1+\varepsilon(\mathcal{H}_R^n)}{1+\varepsilon(\mathcal{H}_r^n)} \cdot M$. Let $\mathbf{v} \in \mathsf{Supp}(h)$ and denote $\rho_{r,\mathbf{v}}^n := \mathcal{U}\left(\mathcal{H}_{r,\mathbf{v},\mathbb{Z}}^n\right)$ with the same subscript convention as usual. If $M' > 1$ then:*

$$\mathcal{R}_\infty\left[\mathcal{U}\left(\mathcal{H}_{r,\mathbb{Z}}^n\right) \parallel \mathcal{U}\left(\mathcal{H}_{R,\mathbf{v},\mathbb{Z}}^n\right)\right] = \left(\frac{R}{r}\right)^n \cdot \frac{1+\varepsilon(\mathcal{H}_R^n)}{1+\varepsilon(\mathcal{H}_r^n)} = M',$$

and the two algorithms $\mathcal{A}$ and $\mathcal{F}$ below have indistinguishable output distributions:

$\mathcal{A}$	$\mathcal{F}$
$\mathbf{v} \hookleftarrow h$	$\mathbf{v} \hookleftarrow h$
$\mathbf{z} \hookleftarrow \rho_{R,\mathbf{v}}^n$	$\mathbf{z} \hookleftarrow \rho_r^n$
output $(\mathbf{z}, \mathbf{v})$ *if* $\mathbf{z} \in \mathcal{P}_{r,\mathbb{Z}}^n$, *else* $\perp$	*output* $(\mathbf{z}, \mathbf{v})$ *with probability* $1/M'$, *else* $\perp$

Furthermore, $\mathcal{A}$ outputs $(\mathbf{z}, \mathbf{v})$ with probability $1/M'$.

The existence of M' in Corollary 7.3.6 enables the key zero-knowledge proof of knowledge part of the protocol. In practice, we estimate the expected number of aborts M using the continuous approximation, use it to derive values for r, β and R, and finally we verify that $M' > 1$. Computing the defects can be done in a preprocessing phase. Although we can compute $\mathrm{vol}(\mathcal{H})$ precisely, we are not aware of any closed form formula for $|\mathcal{H}_{\mathbb{Z}}|$.

Lemma 7.3.7 ([Fel71, (I.9 Th.3)])**.** *Let $a, t \in \mathbb{R}_{>0}$ be positive real numbers and $n \geq 2$ an integer. The interval $[0, t]$ is partitioned into n subintervals by choosing independently at random $n-1$ points of division. Then the probability $\varphi_n(t)$ that none of these subintervals is of length exceeding a equals*

$$\varphi_n(t) = \sum_{\nu=0}^{n} (-1)^{\nu} \binom{n}{\nu} \left(1 - \nu \frac{a}{t}\right)_+^{n-1}, \tag{7.3}$$

where $x_+ := \frac{x+|x|}{2}$ is the positive part of x.

Corollary 7.3.8. *Define α_n as the probability that $\mathbf{x}$ belongs to $\mathcal{H}_r^n$ given that $\mathbf{x}$ is uniformly sampled from $B_n^{(1)}(r\sqrt{n})$. Then:*

$$\alpha_n = \sum_{i=0}^{\lfloor \sqrt{n} \rfloor} (-1)^i \binom{n}{i} \left(1 - i \frac{1}{\sqrt{n}}\right)^n. \tag{7.4}$$

Proof. The result is almost a direct application of Lemma 7.3.7, with $a = r$, $t = r\sqrt{n}$ and renaming ν with i. However this time we are dealing with uniformity inside the ball instead of on the sphere. We use the fact that the ball can be peeled into a disjoint union of spheres

$$B_n^{(1)}(r\sqrt{n}) = \coprod_{r'=0}^{r} \mathcal{S}_1^n(r'\sqrt{n}).$$

When uniformly sampling in $B_n^{(1)}(r\sqrt{n})$, the random variable R' corresponding to the layer r' has probability density function $f_{R'}$ obtained by differentiating the ratio of volumes

$$\Pr(R' \leq r') = \Pr(\mathbf{x} \in B_n^{(1)}(r'\sqrt{n})) = \frac{\mathrm{vol}(B_n^{(1)}(r'\sqrt{n}))}{\mathrm{vol}(B_n^{(1)}(r\sqrt{n}))} = \left(\frac{r'}{r}\right)^n.$$

Therefore using Equation 7.3,

$$\begin{aligned}
\alpha_n &= \int_0^r f_{R'}(r') \varphi_n(r'\sqrt{n}) dr' \\
&= \int_0^r \left(\frac{n r'^{n-1}}{r^n}\right) \sum_{i=0}^{n} (-1)^i \binom{n}{i} \left(1 - \frac{ri}{r'\sqrt{n}}\right)_+^{n-1} dr' \\
&= \frac{n}{r^n} \sum_{i=0}^{n} (-1)^i \binom{n}{i} \int_0^r \left(r' - \frac{ri}{\sqrt{n}}\right)_+^{n-1} dr' \\
&= \frac{n}{r^n} \sum_{i=0}^{\lfloor \sqrt{n} \rfloor} (-1)^i \binom{n}{i} \int_{\frac{ri}{\sqrt{n}}}^r \left(r' - \frac{ri}{\sqrt{n}}\right)^{n-1} dr' \\
&= \frac{n}{r^n} \sum_{i=0}^{\lfloor \sqrt{n} \rfloor} (-1)^i \binom{n}{i} \left[\frac{\left(r - \frac{ri}{\sqrt{n}}\right)^n}{n} - \frac{\left(\frac{ri}{\sqrt{n}} - \frac{ri}{\sqrt{n}}\right)^n}{n} \right] \\
&= \sum_{i=0}^{\lfloor \sqrt{n} \rfloor} (-1)^i \binom{n}{i} \left(1 - \frac{i}{\sqrt{n}}\right)^n.
\end{aligned}$$

□

Corollary 7.3.9 (Volume of $\mathcal{H}_r^n$)**.** *For $n \in \mathbb{Z}_{\geq 0}$, $r \in \mathbb{R}_{>0}$ and α_n as defined by Equation 7.4,*

$$\mathrm{vol}(\mathcal{H}_r^n) = \alpha_n \cdot \frac{(2r\sqrt{n})^n}{n!}.$$

Proof. This follows immediately from Corollary 7.3.8, as α_n is the probability that a uniform point in $B_n^{(1)}(r\sqrt{n})$ lies in $\mathcal{H}_r^n$, and $\mathrm{vol}(\mathcal{H}_r^n) = \frac{(2r\sqrt{n})^n}{n!}$. □

We now adapt what we believe to be a standard generating series trick that allows us to count the exact number of points in $\mathcal{H}_{\mathbb{Z}}$. Similar techniques have been used for example in [DEP23, (Section 3.3)].

Lemma 7.3.10. *Let $n, r \in \mathbb{Z}_{\geq 0}$, and $e_{n,r} = \lfloor\sqrt{n}\rfloor r + \lfloor\{\sqrt{n}\}r\rfloor$ using notations from Corollary 7.3.5. Let $(\omega_i)_i$ be the sequence of integers defined by expanding the following polynomial of $\mathbb{Z}[X]$:*

$$\left(1 + 2\sum_{i=1}^{r} X^i\right)^n = \sum_{i=0}^{\infty} \omega_i X^i. \tag{7.5}$$

Then the number of integral points in $\mathcal{H}_r^n$ is given by $|\mathcal{H}_{r,\mathbb{Z}}^n| = \sum_{i=0}^{e_{n,r}} \omega_i$.

Proof. For $\ell \in \mathbb{Z}_{\geq 0}$, the coefficient ω_ℓ counts the number of ways that ℓ can be partitioned in a sum of n integers the interval $[-r, r]$. Indeed it effectively counts partitions whose terms lie in the interval $[0, r]$, with an extra weight of 2 in the count for each non-zero term in the partition. □

Algorithmically, Equation 7.5 should be expanded using a truncated variant of binary exponentiation, as monomials of degree $> e_{n,r}$ do not impact the result. Lemma 7.3.10 does not give a closed form for the cardinality of $\mathcal{H}_{\mathbb{Z}}$ but only an algorithm to compute it. In practice, we use Lemma 7.3.10 to compute the exact value of the defect $\varepsilon(\mathcal{H}_r^n)$ for explicit choices of parameters (n, r).

7.3.3 An Isochronous Sampler on $\mathcal{H} \cap \mathbb{Z}^n$

Recall our aim is to strike a balance between simplicity and optimality. In this subsection, we present an isochronous uniform sampling algorithm for $\mathcal{H}_{r,\mathbb{Z}}^n$ (as detailed in Figure 7.2), which relies solely on uniform sampling without replacement. This approach eliminates the need for Gaussian sampling, albeit at the cost of a low rejection rate.

A sampler is considered (perfectly) isochronous [HPRR20, (Definition 5)] when its running time is independent of any sensitive variable. We establish our main claim in Theorem 7.3.11, demonstrating that our sampler is both uniform in $\mathcal{H}_{r,\mathbb{Z}}^n$ and isochronous.

Our sampler for $\mathcal{H}_{r,\mathbb{Z}}^n$ is based on a uniform sampler in the discrete ℓ_1-ball. We explain how to extend this approach to obtain an isochronous uniform sampler for the discrete ℓ_1-ball of dimension n when we already have one for the discrete ℓ_1-hemisphere of dimension $n+1$.

Theorem 7.3.11. *For $r \in \mathbb{Z}_{>0}$, $\mathsf{SampleH}(n, r)$ is isochronous and uniformly samples from the set $\mathcal{H}_{r,\mathbb{Z}}^n$.*

Proof. Direct consequence of Proposition 7.3.12, Lemma 7.3.13, and Proposition 7.3.14. □

Furthermore, the probabilities of restarting at step 11 of $\mathsf{SampleHemisphere}_1(n, r)$ and step 7 of $\mathsf{SampleH}$ are provided in Proposition 7.3.12 and Proposition 7.3.14, respectively.

To achieve uniform sampling on $\mathcal{H}_{r,\mathbb{Z}}^n$, we can rely on Corollary 7.3.8, which shows that a significant portion of samples from the ℓ_1-ball with an appropriate radius already belong to $\mathcal{H}_{r,\mathbb{Z}}^n$. Hence, we only need to reject samples that are not in the corresponding ℓ_∞-ball.

$\mathsf{SampleHemisphere}_1(n, r)$

1 : $x_0 \leftarrow 0, x_n \leftarrow r + n$
2 : // $\mathcal{S} = \{X \subset [r + n - 1] : \#X = n - 1\}$
3 : $\mathbf{X} \leftarrow\!\!\$\, \mathcal{U}(\mathcal{S})$
4 : $\mathbf{X} \leftarrow \mathbf{X} \cup \{x_0, x_n\}$
5 : $\mathbf{X}.\mathtt{sort}()$
6 : // $x_0, \cdots, x_n$ the ordered elements of X
7 : **for** $i \in [n - 1]$:
8 : $b \leftarrow\!\!\$\, \{0, 1\}$
9 : $y_i \leftarrow (x_i - x_{i-1} - 1)$
10 : **if** $y_i + b = 0$ **then**
11 : **restart**
12 : $y_i \leftarrow (-1)^b y_i$
13 : $y_n \leftarrow (x_n - x_{n-1} - 1)$
14 : **return** $\mathbf{Y} := (y_i)_{1 \le i \le n}$

$\mathsf{SampleBall}_1(n, r)$

1 : $(y_i)_{n+1} \leftarrow\!\!\$$ $\mathsf{SampleHemisphere}_1(n + 1, r)$
2 : **return** $(y_1, \cdots, y_n)$

$\mathsf{SampleH}(n, r)$

1 : $\Delta_n \leftarrow (\sqrt{n} - \lfloor\sqrt{n}\rfloor)$
2 : $r' \leftarrow \lfloor\sqrt{n}\rfloor r + \lfloor\Delta_n r\rfloor$
3 : $\mathbf{Y} \leftarrow \perp$
4 : **while** $\mathbf{Y} = \perp$ **do**
5 : $\mathbf{Y} \leftarrow \mathsf{SampleBall}_1(n, r')$
6 : **if** $\|\mathbf{Y}\|_\infty > r$ **then**
7 : $\mathbf{Y} \leftarrow \perp$
8 : **return** $\mathbf{Y}$

Figure 7.2: Sampling algorithm on $\mathcal{H}_\mathbb{Z}$ from samplers on the ℓ_1-ball and -hemisphere for $r \in \mathbb{Z}_{>0}$.

Proposition 7.3.12. *For $r \in \mathbb{Z}_{>0}$, the sampler $\mathsf{SampleH}(n, r)$ is isochronous and provides uniform samples in $\mathcal{H}^n_{r,\mathbb{Z}}$ if $\mathsf{SampleBall}_1(n, r)$ is isochronous and uniform. Additionally, the probability of $\mathbf{Y} \neq \perp$ in Step 7 of $\mathsf{SampleH}$ is exactly $\alpha_n \frac{1+\varepsilon(B^{(1)}_n(r))}{1+\varepsilon(\mathcal{H}^n_r)}$, where α_n is defined in Equation 7.4 and $\varepsilon(\cdot)$ is the defect, defined in Definition 7.2.6.*

Proof. Given r' in step 2 of $\mathsf{SampleH}(n, r)$, we know that $\mathcal{H}^n_{r,\mathbb{Z}} \subset B^{(1)}_{n,\mathbb{Z}}(r')$. If $\mathsf{SampleBall}_1(n, r)$ is called again due to an abort ($\mathbf{Y} = \perp$ in step 7), we can conclude that $\mathsf{SampleH}(n, r)$ is uniform and isochronous, provided that $\mathsf{SampleBall}_1(n, r)$ is uniform and isochronous. The acceptance rate result follows exactly from Corollary 7.3.8. □

Sampling from the discrete ball in dimension n can be done directly by sampling on the discrete sphere in dimension $n + 1$ and projecting orthogonally to a canonical vector (which can be achieved by removing the last coordinate). However this procedure introduces some bias because if the last of the $n + 1$ coordinates on the sphere is not 0, then projection is 2-to-1, whereas it is a bijection when the last coordinate is 0. For this purpose, we introduce the notation $S^{(1)}_{+,n}$ to denote the ℓ_1-hemisphere, defined as follows:

$$S^{(1)}_{+,n}(R) := S^{(1)}_n(R) \cap \{\mathbf{x} = (x_i)_i \in \mathbb{R}^{n+1} : x_{n+1} \geq 0\}.$$

Projecting away the last coordinate gives a direct bijection between $S^{(1)}_{+,n,\mathbb{Z}}(r\sqrt{n})$ and $B^{(1)}_{n,\mathbb{Z}}(r\sqrt{n})$.

Lemma 7.3.13. *Let $r \in \mathbb{Z}_{>0}$. For all $i \leq n$, let $\mathbf{x} = (x_i)_{1 \le i \le n} \in \mathbb{Z}^n$ and $r > 0$ such that $\|\mathbf{x}\|_1 = r$. If $p_i(x_1, x_2, ..., x_n) = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$ defines the i-th projection, then p_{n+1} is a bijection between $S^{(1)}_{+,n,\mathbb{Z}}(r)$ and $B^{(1)}_{n,\mathbb{Z}}(r)$. Consequently, if $\mathbf{X}$ is a random variable with distribution $\mathcal{U}\left(S^{(1)}_{+,n,\mathbb{Z}}(r)\right)$, then $p_{n+1}(X)$ has distribution $\mathcal{U}\left(B^{(1)}_{n,\mathbb{Z}}(r)\right)$.*

This shows that if $\mathsf{SampleHemisphere}_1(n+1, r)$ is uniform and isochronous, then the sampler $\mathsf{SampleBall}_1(n, r)$ must also have both properties. We thank Jasper Seidensticker for pointing out a subtle flaw in uniformity of the original sampler presented in [BBRS24]. This flaw is corrected above.

Proposition 7.3.14. *For any integer $r \in \mathbb{Z}_{>0}$, the* $\mathsf{SampleBall}_1(n, r)$ *algorithm of Figure 7.2 is both isochronous and uniform in $B^{(1)}_{n,\mathbb{Z}}(r)$. Furthermore, the probability of an abort (triggering the* **restart** *instruction) is equal to:*

$$\beta_{n,r} := \binom{n+r}{n}^{-1} \sum_{j=1}^{n} \binom{n}{j}\binom{r}{n-j}(1-2^{-j}).$$

Proof. All operations within this algorithm, including the uniform selection of $\mathbf{X}$, can be completed in constant time except the sorting algorithm. We claim trivially that even knowing the order of each unknown variable does not help recovering them. The number of aborts is independent of the outputted value since $\mathbf{X}$ is resampled at each restart.

Let's prove that output of $\mathsf{SampleHemisphere}_1(n+1, r)$ follows the uniform distribution in $\mathcal{S}^{(1)}_{+,n,\mathbb{Z}}(r)$. If this is correct, then Lemma 7.3.13 will allow us to conclude. We define:

$$\mathsf{S}_{\mathsf{source}} := \{((b_1, y_1), \ldots, (b_{n+1}, y_{n+1})) \in (\{0,1\} \times [\![0, r]\!])^{n+1} \ : \ \sum_i y_i = r \wedge b_{n+1} = 0\};$$

$$\mathsf{S}_{\mathsf{target}} := \{((b_1, y_1), \ldots, (b_{n+1}, y_{n+1})) \in \mathsf{S}_{\mathsf{source}} \ : \ \forall i \leq n,\ y_i = 0 \Rightarrow b_i = 1\}.$$

A direct analysis reveals that $\mathsf{SampleHemisphere}_1(n+1, r)$ can be reformulated as follows:

$\mathsf{SampleHemisphere}_1(n+1, r)$

1 : $\mathbf{A} = ((b_1, y_1), \ldots, (b_{n+1}, y_{n+1})) \hookleftarrow \mathcal{U}(\mathsf{S}_{\mathsf{source}})$
2 : **if** $\mathbf{A} \notin \mathsf{S}_{\mathsf{target}}$ **then goto** 1
3 : **return** $\left((-1)^{b_1} y_1, \ldots, (-1)^{b_{n+1}} y_{n+1})\right)$

Here, the (b_i, y_i) of the step 1 correspond to the b and y_i values computed in steps 8 and 9 of $\mathsf{SampleHemisphere}_1(n+1, r)$ in Figure 7.2.

Furthermore, we can observe that the mapping:

$$((b_1, y_1), \ldots, (b_{n+1}, y_{n+1})) \to \left((-1)^{b_1} y_1, \ldots, (-1)^{b_{n+1}} y_{n+1}\right),$$

with b_{n+1} always set to 0 is a bijection between $\mathsf{S}_{\mathsf{target}}$ and $S^{(1)}_{+,n,\mathbb{Z}}(r)$. This establishes the uniformity of the sampler.

Finally, the probability of an abort is equal to the probability of sampling an element from $\mathsf{S}_{\mathsf{source}} - \mathsf{S}_{\mathsf{target}}$. An abort happens when $y_i = b_i = 0$ for some $i \leq n$. In fact this happens with probability $1/2$ for each y_i that is zero. Forgetting about signs, there are $\binom{r+n}{n}$ ways to write r as a sum of $n+1$ non-negative integers. If we impose the amount j of y_i that are 0 among the first n coordinates, then there are exactly $\binom{n}{j}\binom{r}{n-j}$ ways to write r as a sum of $n-j$ positive integers and 1 non-negative integer (the $\binom{r}{n-j}$ term follows from the usual stars and bars trick). Therefore the probability $\beta_{n,r}$ of an abort is exactly

$$\binom{n+r}{n}^{-1} \sum_{j=1}^{n} \binom{n}{j}\binom{r}{n-j}(1-2^{-j}),$$

as claimed. □

Theorem 7.3.15. *Let $n, r \in \mathbb{Z}_{>0}$, the total expected number of restarts (steps 11 in* $\mathsf{SampleHemisphere}_1$ *and 7 in* $\mathsf{SampleH}$ *combined) in order to sample a point of $\mathcal{H}^n_{r,\mathbb{Z}}$ using the algorithm of Figure 7.2*

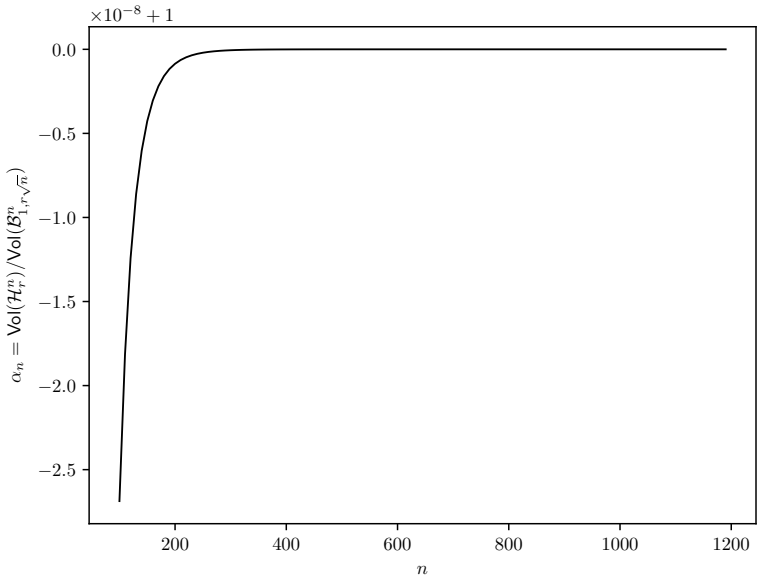


Figure 7.3: Evolution of α_n.

is exactly

$$\frac{(1-\beta_{n,r'})^{-1}}{\alpha_n}\cdot\frac{1+\varepsilon(\mathcal{H}_r^n)}{1+\varepsilon(B_n^{(1)}(r))}=\frac{\left(1-\binom{n+r'}{n}^{-1}\sum_{j=1}^{n}\binom{n}{j}\binom{r'}{n-j}(1-2^{-j})\right)^{-1}}{\sum_{i=0}^{\lfloor\sqrt{n}\rfloor}(-1)^i\binom{n}{i}\left(1-i\frac{1}{\sqrt{n}}\right)^n}\cdot\frac{1+\varepsilon(\mathcal{H}_r^n)}{1+\varepsilon(B_n^{(1)}(r))} \tag{7.6}$$

where $r'=\lfloor\sqrt{n}\rfloor r+\lfloor\{\sqrt{n}\}r\rfloor$.

We end this section with brief note on samplers for the hypercube and the Euclidean ball. The randomness necessary to sample in the hypercube can be obtained directly ($n\ln(2R+1)$ with (n,R) as in Theorem 7.2.9). In this case, the sampling mechanism is easy and direct, without rejections. On the contrary, state-of-the-art samplers in the hyperball are based on projections from a sphere with two extra coordinates and use continuous Gaussian samplers. Simulating continuous Gaussian sampling with discrete Gaussian sampling leads to a large overhead in randomness usage. In addition, to sample inside the integer restriction of the ball, one needs to add specific constraints which lead to rejection (see [Che+23] for more insight). Figure 7.2 allows to uniformly sample in $\mathcal{H}_{\mathbb{Z}}$. It only uses uniform samplers and has small practical rejection rate, which can be computed through Theorem 7.3.15. Combined with Lemma 7.3.10 that allows to pre-compute the defect, this makes $\mathcal{H}$ an ideal candidate for zero-knowledge proofs in the FSwA paradigm.

7.3.4 Why $\mathcal{H}$ Performs Much Better than It Should?

In this subsection, we start from the observation that most of the volume of a ℓ_1 ball is actually inside its inscribed polytope $\mathcal{H}$. We illustrate this below by showing the evolution of α_n (Equation 7.4) as the dimension grows. It can be read on Figure 7.3 that α_n converges swiftly to 1, indicating that the larger the dimension, the closer $\mathcal{H}_r^n$ and $B_n^{(1)}(r\sqrt{n})$ become. Until now, we were measuring performances by considering the ratio between circumradius and inradius, which reached $\sqrt[4]{n}$ in the case of $\mathcal{H}$. In fact this metric for measuring signature sizes is more

concerned with worst-case than it is with average-case. To study average-case, one would need to look at the expected value of the ℓ_2 norm of accepted versions of $\mathbf{z}$. Because the behaviour of $\mathcal{H}$ is close to that of the ℓ_1 ball, a random point of $\mathcal{H}$ is extremely unlikely to have a large ℓ_2 norm, and even reaching $r\sqrt[4]{n}$ is not bound to happen often. This means that intersection with $B_n^{(\infty)}(r)$ is mostly only there for theoretical guarantee that signatures are not unusually large. Pushing this reasoning further leads us to consider the following shape:

Definition 7.3.16. For an integer $n \in \mathbb{Z}_{>0}$ and real numbers $r, \theta \in \mathbb{R}_{>0}$, we define

$$\mathcal{C}_{\theta,r}^n := \mathcal{H}_r^n \cap B_n^{(2)}(\theta r).$$

Remark 7.3.17. Sampling on $\mathcal{C}_{\mathbb{Z}}$ (with the usual abuse of notation) is done by sampling in $\mathcal{H}_{\mathbb{Z}}$ using Figure 7.2 and rejecting if the Euclidean norm is out of bounds. In our application, r is at least 2 and an integer. Recall that for $u > 0$ and $\mathbf{w} \in \mathbb{R}^n$, $\mathcal{C}_{\theta,u,\mathbf{w}}^n = u\mathcal{C}_\theta^n + \mathbf{w}$.

Implications for rejection sampling $\mathcal{C}$ inherits most properties of $\mathcal{H}$ that we need to apply Theorem 7.2.9, except for the it not being a polytope. If the mask $\mathbf{y}$ is sampled from $\mathcal{C}_{\mathbb{Z}}$ directly, Theorem 7.2.9 does not apply, as we are not dealing with a polytope. However, if we first sample $\mathbf{y}$ in $\mathcal{H}_{\mathbb{Z}}$ and then force $\mathbf{z}$ to be in $\mathcal{C}_{\mathbb{Z}}$ by rejecting if needed then we only need to apply Corollary 7.3.6 with $\mathcal{H}_{\mathbb{Z}}$, and no changes are made. This requires that θ is not chosen too aggressively, as if θ is too small, then the amount of last-step rejections will blow up.

Question 7.3.18. *What is a reasonable choice for θ in Definition 7.3.16?*

We justify our choice heuristically, and back it up with experiments. We first give a volumetric argument that explains why intersecting $\mathcal{H}$ with a Euclidean ball of radius θr doesn't lose too many points. Secondly, we use experiments to get an approximation of this ε factor as we only need it to be sufficiently small to not drastically change the rejection rate of our sampler. Computing the exact number of integral points in $\mathcal{C}$ seems much harder than for $\mathcal{H}$, which is why our justification is heuristic.

The next proposition formalises the intuition that because the ℓ_1 ball concentrates towards its centre, cutting the corners of $\mathcal{H}$ with a ℓ_2 ball with radius a constant times larger will not affect its volume too much.

Proposition 7.3.19. *Let $n \in \mathbb{Z}_{>0}$ be a positive integer, and $r \in \mathbb{R}_{>0}$ be a positive real number. Then there exists $c > 0$ such that for any real number $\theta > 1/c$:*

$$1 - \alpha_n^{-1} \exp(-c\theta\sqrt{n}) \leq \mathrm{vol}(\mathcal{C}_{\theta,r}^n)/\mathrm{vol}\mathcal{H}_r^n \leq 1.$$

Proof. The upper bound is clear by definition of $\mathcal{C}_{\theta,r}^n$. The lower bound is more involved. First note that:

$$\mathcal{H}_r^n = (B_n^{(\infty)}(r) \cap B_n^{(1)}(r\sqrt{n}) - B_n^{(2)}(\theta r)) \cup \mathcal{C}_{\theta,r}^n \subset (B_n^{(1)}(r\sqrt{n}) - B_n^{(2)}(\theta r)) \cup \mathcal{C}_{\theta,r}^n.$$

Therefore with volumes:

$$\begin{aligned}\mathrm{vol}(\mathcal{C}_{\theta,r}^n) &\geq \mathrm{vol}(\mathcal{H}_r^n) - \mathrm{vol}(B_n^{(1)}(r\sqrt{n}) - B_n^{(2)}(\theta r)) \\ &\geq \mathrm{vol}(\mathcal{H}_r^n) - \mathrm{vol}(B_n^{(1)}(r\sqrt{n})) + \mathrm{vol}(B_n^{(1)}(r\sqrt{n}) \cap B_n^{(2)}(\theta r)).\end{aligned}$$

The last volume is computed by direct application of a theorem by Schechtman and Zinn [SZ90] restated as theorem 5.1 in [PTT18]. Taking $p = 1$ and $q = 2$, we obtain:

$$\mathrm{vol}(B_n^{(1)}(r\sqrt{n}) \cap B_n^{(2)}(\theta r)) \geq (1 - \exp(-c\theta\sqrt{n}))\mathrm{vol}(B_n^{(1)}(r\sqrt{n})).$$

Using the fact that $\mathrm{vol}(\mathcal{H}_r^n) = \alpha_n \mathrm{vol}(B_n^{(1)}(r\sqrt{n}))$, we conclude. □

Procedures to compute $|\mathcal{C}_{\mathbb{Z}}|$ exist however they are not memory efficient, we present an example of such a procedure using bivariate generating series in the appendix of [BBRS24]. We now comment on the choice of the parameter θ. One natural choice would be to take $\theta = \theta_n$ where we define θ_n in such a way that we obtain $\mathrm{vol}(B_n^{(2)}(\theta_n r)) = \mathrm{vol}(\mathcal{H}_r^n)$. Using Stirling's approximation, this amounts to taking $\theta_n = (\alpha_n/\sqrt{2})^{1/n}\sqrt{\frac{2e}{\pi}} \approx 1.315$.

We provide in Figure 7.4 an estimation of the proportion of rejects added by using $\mathcal{C}_{\mathbb{Z}}$ over $\mathcal{H}_{\mathbb{Z}}$ at fixed dimension with different θ. This experiment show that there is a range of possible θ from 1.35 to 1.5 that enable a trade-off between aggressiveness (smaller proof of knowledge) and additional rejection cost (as the Euclidean norm filter gets tighter). In what follows, we use $\theta = 1.5$ as a conservative choice that still leads to major improvements.

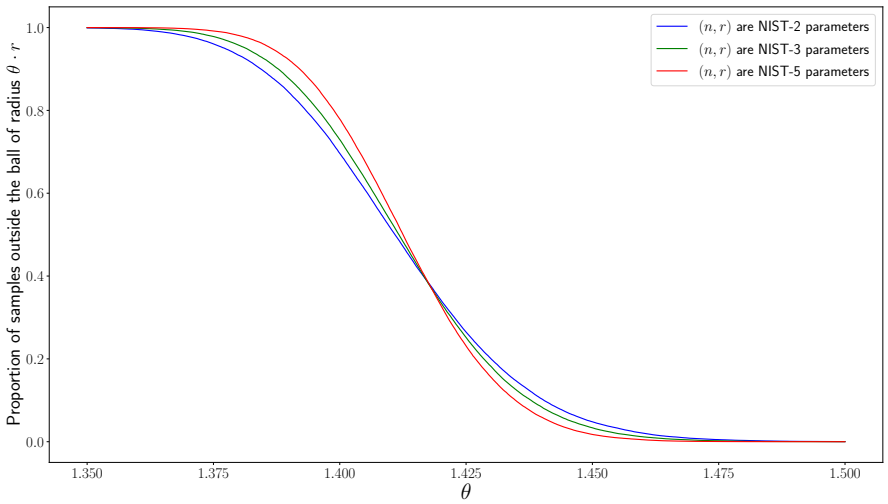


Figure 7.4: Proportion of samples from $\mathcal{H}_{\mathbb{Z}}$ outside of $\mathcal{C}_\theta$ for varying θ. Experiment over 100000 *samples for each parameter set, using our sampler defined in Figure 7.2.*

7.4 An Improved Signature Scheme: Patronus

This section highlights a concrete application of our contributions from Section 7.2, Section 7.3 and Section 7.3.4 through a signature scheme, Patronus, using the FSwA paradigm. We further compare it to Dilithium and Haetae [Duc+21; Che+23], two FSwA signatures. In order to do so, we compare them practically using signature sizes. However, we may study them directly with framework from Lemma 7.2.1 (resp. Corollary 7.2.4) using the hyperball (resp. the hypercube) for Haetae (resp. Dilithium). We describe Patronus, prove its correctness and basic security properties, and propose concrete sets of parameters. As the main change concerns the sampler, our proof follows closely that of Haetae. Note that while this is not used in our analysis, our sets of parameters in Table 7.7 are chosen NTT-friendly.

Security properties and hardness assumptions

As for Dilithium and Haetae, Patronus uses the well-studied lattice-based hardness assumptions MLWE, MSIS and SelfTargetMSIS. The first two were introduced in Section 3.4. To prove our signature Patronus secure, we still need one more common assumption for the specific case of FSwA signatures called Self-Target MSIS (SelfTargetMSIS). For the sake of readability, we omit the explicit mention of the modulus q and the dimension $n = 256$ associated with $\mathcal{R}_q = \mathbb{Z}_q[X]/(X^n + 1)$ when defining the parameters of the problems. We refer to elements of $\mathcal{R}_q^k$ as elements of the $\mathcal{R}_q$-module of rank k and consider them with their embedding in $\mathbb{Z}_q^{k\cdot n}$. We use

the same SelfTargetMSIS problem as Dilithium ([Duc+21, (Section 4.1)]), except that we consider the ℓ_2 norm instead of the ℓ_1 norm:

Definition 7.4.1 (SelfTargetMSIS problem)**.** Suppose that $\mathsf{H} : \{0,1\}^* \times \mathcal{M} \to \mathsf{SetChall} = \{c \in \mathbb{R}_q : \|c\|_1 = \tau \wedge \|c\|_\infty = \tau\}$ is a quantum random hash oracle for some $\tau \in \mathbb{N}$. For positive integers k, l and a positive real number β, the advantage $\mathsf{Adv}^{\mathsf{stmsis}}_{\mathsf{H},k,l,\beta,\mathsf{q_H}}(\mathcal{A})$ against the SelfTargetMSIS problem of an adversary making at most $\mathsf{q_H}$ quantum queries to $|\mathsf{H}\rangle$ is:

$$\Pr\left[\begin{array}{c} 0 < \|\mathbf{y}\|_2 < \beta \\ \wedge \mathsf{H}\left((\mathbf{Id} \mid \mathbf{A})\,\mathbf{y}, M\right) = c \end{array} : \begin{array}{c} \mathbf{A} \leftarrow \mathcal{R}_q^{k\times l} \\ (\mathbf{y} = (\mathbf{r}, c), M) \leftarrow \mathcal{A}^{|\mathsf{H}(\cdot)\rangle}(\mathbf{A}) \\ (\mathbf{y} = (\mathbf{r}, c), M) \in \mathcal{R}_q^{l+k-1} \times \mathsf{SetChall} \times \{0,1\}^* \end{array}\right].$$

In the classical setting, there exists a classical reduction from SelfTargetMSIS to MSIS that exhibits an adversary $\mathcal{B}$ such that $\mathsf{Adv}^{\mathsf{stmsis}}_{\mathsf{H},k,l,\beta,\mathsf{q_H}}(\mathcal{A}) \approx \sqrt{\frac{\mathsf{Adv}^{\mathsf{msis}}_{k,l,2\beta}(\mathcal{B})}{\mathsf{q_H}}}$. More details are given in [KLS18, (Section 4.5.1)].

7.4.1 The Patronus Scheme.

We give a high level description of the different element of the signature from Figure 7.5 and a proof of its correctness. Coefficients of $(\mathbf{e}, \mathbf{s})$ follow centred binomials of parameter $(2, 0.5)$.

An analysis of the entropy is given in the appendix of [BBRS24]. η is a bound on the norm of $\mathbf{s}$ to tailor the Euclidean norm of $\mathbf{s}c$ such that $\|\mathbf{s}c\|_2 \leq \beta = \eta\sqrt{\tau}$ with τ the number of ± 1 in the challenge c. We assert this bound using to the detailed analysis of [Che+23, (Section 3.1)]. This bound ensures that our main theorem on rejection sampling works on $r + \beta$ with r the radius for the target distribution after the rejection step. The parameter r is defined as in Theorem 7.2.9 but with an additional factor $\theta = 1.5$ chosen from Section 7.3.4. ξ in step 5 of Sign is an artificial sample to get sufficient entropy in the challenge c. In step 14 of Sign, we separate $\mathbf{z}$ into its lowbits and highbits part, the lowbits part follows approximately a uniform distribution therefore we do not apply any compression, however we apply the compression from [Dud09] on the highbits part to obtain a compressed signature. Due to the cut constraints we need $q > 2\gamma$ and $\gamma | (q-k)$. Parameter ω has been introduced in Dilithium [Duc+21], it represents the number of carry introduced by $\mathsf{MakeHint}_m$ or alternatively the number of its coefficients equal to 1. This is an interesting information since it allows to obtain a tighter bound in the Patronus security reduction. Differently to Dilithium, in this work we directly use the ℓ_2 bound which implies from the second point in [Duc+21, (Lemma 1)] a tighter bound in the security reduction of Patronus.

We first define necessary primitives to build Patronus.

Definition 7.4.2. Let $r \in \mathbb{Z}$, $d \in \mathbb{N}^*$ and γ a power of two. We define Highbits, Lowbits and Power2round as:

$\mathsf{Power2round}(r, d)$
$r := r \bmod^+ q$
$r_0 := r \bmod^\pm 2^d$
return $((r - r_0)/2^d, r_0)$

$\mathsf{Highbits}(r, \gamma)$
return $\left\lfloor \frac{r}{\gamma} + \frac{1}{2} \right\rfloor$

$\mathsf{Lowbits}(r, \gamma)$
return $r \bmod^\pm \gamma$

Definition 7.4.3. Let $r \in \mathbb{Z}$. Let q be a prime number and $\gamma | (q-k)$ a power of two. Let $m = (q-k)/\gamma$ and $\mathsf{s}(x)$ the function that returns 1 if $x - 1 \geq 0$ and -1 otherwise. Finally, we define $\mathsf{MakeHint}_m$ and $\mathsf{UseHint}_m$ and the subroutines $\mathsf{Highbits}_m$, $\mathsf{Lowbits}_m$ as:

$\mathsf{Highbits}_m(r, \gamma)$

$r_1 \leftarrow \mathsf{Highbits}(r \bmod^+ q, \gamma)$
$r_0 \leftarrow \mathsf{Lowbits}(r \bmod^+ q, \gamma)$
if $r_1 = m$ **then**
  return 0
return r_1

$\mathsf{Lowbits}_m(r, \gamma)$

$r_1 \leftarrow \mathsf{Highbits}(r \bmod^+ q, \gamma)$
$r_0 \leftarrow \mathsf{Lowbits}(r \bmod^+ q, \gamma)$
if $r_1 = m$ **then**
  return $r_0 - k \bmod^\pm \gamma$
return r_0

$\mathsf{UseHint}_m(h, r, \gamma)$

$r_1 \leftarrow \mathsf{Highbits}_m(r, \gamma)$
$r_0 \leftarrow \mathsf{Lowbits}_m(r, \gamma)$
if $h = 1$ **then**
  return $(r_1 + \mathsf{s}(r_0)) \bmod^+ m$
return r_1

$\mathsf{MakeHint}_m(z, r, \gamma)$

$r_1 \leftarrow \mathsf{Highbits}_m(r, \gamma)$
$v_1 \leftarrow \mathsf{Highbits}_m(r + z, \gamma)$
return $[\![r_1 \neq v_1]\!]$

$\mathsf{KeyGen}()$

1 : $\mathbf{A} \leftarrow\$ \mathcal{R}_q^{k \times l}$
2 : $(\mathbf{s}, \mathbf{e}) \leftarrow\$ \mathsf{Binom}_j^{nl} \times \mathsf{Binom}_j^{nk}$
3 : **if** $\mathcal{N}(\mathbf{s}) > n\eta^2$ **then**
4 : **goto** 1
5 : $\mathbf{b} \leftarrow \mathbf{A} \cdot \mathbf{s} + \mathbf{e} \in \mathcal{R}_q^k$
6 : $(\mathbf{b}_\mathsf{H}, \mathbf{b}_\mathsf{L}) \leftarrow \mathsf{Power2round}(\mathbf{b}, d)$
7 : $\mathsf{pk} \leftarrow (\mathbf{A}, \mathbf{b}_\mathsf{H})$
8 : $\mathsf{sk} \leftarrow (\mathbf{s}, \mathbf{b}_\mathsf{L})$
9 : **return** $(\mathsf{pk}, \mathsf{sk})$

$\mathsf{Verify}\,(\mathsf{pk} := (\mathbf{A}, \mathbf{b}_\mathsf{H}), \mu, \sigma := (c, \mathbf{v}))$

1 : $(\bar{\mathbf{z}_\mathsf{H}}, \mathbf{z}_\mathsf{L}, \bar{\mathbf{h}}, \xi) \leftarrow \mathbf{v}$
2 : $\mathbf{z} \leftarrow \alpha_z \cdot \mathsf{Decode}(\bar{\mathbf{z}_\mathsf{H}}) + \mathbf{z}_\mathsf{L}$
3 : $\mathbf{h} \leftarrow \mathsf{Decode}(\bar{\mathbf{h}})$
4 : $\mathbf{w}_\mathsf{H} \leftarrow \mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot 2^d, 2\gamma)$
5 : **return** $[\![c = \mathsf{H}(\mathbf{w}_\mathsf{H}, \xi, \mu) \wedge \|\mathbf{h}\|_1 \leq \omega \wedge \mathbf{z} \in \mathcal{C}^{nl}_{\theta,r,\mathbb{Z}}]\!]$

$\mathsf{Sign}(\mathsf{sk} := (\mathbf{s}, \mathbf{b}_\mathsf{L}), \mu)$

1 : $\mathbf{v} \leftarrow \bot$
2 : **while** $\mathbf{v} = \bot$ **do**
3 : $\mathbf{y} \leftarrow\$ \boxed{\mathcal{H}^{nl}_{r+\beta,\mathbb{Z}}}\ \boxed{\mathcal{C}^{nl}_{\theta,r+\beta,\mathbb{Z}}}$
4 : $\xi \leftarrow\$ \{0,1\}^n$
5 : $\mathbf{w} \leftarrow \mathbf{Ay}$
6 : $\mathbf{w}_\mathsf{H} \leftarrow \mathsf{Highbits}(\mathbf{w}, 2\gamma)$
7 : $c \leftarrow \mathsf{H}(\mathbf{w}_\mathsf{H}, \xi, \mu)$
// $c \in \{\mathbf{x} \in \mathbb{R}_3 : \|\mathbf{x}\|_1 = \tau\}$
8 : $\mathbf{z} \leftarrow \mathbf{y} + \mathbf{s}c$
9 : $\mathbf{r}_0 \leftarrow \mathsf{Lowbits}_m(\mathbf{w} - c\mathbf{e}, 2\gamma)$
10 : **if** $\mathbf{z} \in \mathcal{C}^{nl}_{\theta,r,\mathbb{Z}}$ **and** $\|\mathbf{r}_0\|_\infty < \gamma - \beta'$ **then**
11 : $\mathbf{h} \leftarrow \mathsf{MakeHint}_m(-c\mathbf{b}_\mathsf{L}, \mathbf{w} - c\mathbf{e} + c\mathbf{b}_\mathsf{L}, 2\gamma)$
12 : **if** $\|c\mathbf{b}_\mathsf{L}\|_\infty < \gamma$ **and** $\|\mathbf{h}\|_1 < \omega$ **then**
13 : $\mathbf{v}_1 \leftarrow \mathsf{Encode}(\mathsf{Highbits}(\mathbf{z}, \alpha_z))$
14 : $\mathbf{v}_2 \leftarrow \mathsf{Lowbits}(\mathbf{z}, \alpha_z)$
15 : $\mathbf{v}_3 \leftarrow \mathsf{Encode}(\mathbf{h})$
16 : $\mathbf{v} \leftarrow (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \xi)$
17 : **return** $\sigma := (c, \mathbf{v})$

Figure 7.5: Two variants of the Patronus *signature. The more conservative variant is defined by using $\mathcal{H}$ in the full box while the more aggressive one uses $\mathcal{C}$ in the dashed box. $\mathcal{N}$ is defined as in [Che+23] and $\beta' = \tau$*

. Our concrete parameters are given for the conservative variant.

Correctness is proven using the properties of the functions defined in Definition 7.4.2 and Definition 7.4.3 in a similar way to Dilithium [Duc+21].

The proofs of these lemmas are standard and can be found in the appendix of [BBRS24].

Lemma 7.4.4. *Let $a, b \in \mathbb{Z}$ such that $a \geq 0$ and $b > 0$. It holds that:*

$$a = \left\lfloor \frac{a}{b} + \frac{1}{2} \right\rfloor \cdot b + (a \bmod^\pm b),$$

this form is the unique $a = bq + r$ *with* $r \in \left(-\frac{b}{2}, \frac{b}{2}\right]$.

Lemma 7.4.4 ensures the well definition of Definition 7.4.3 with its existence and unicity. As for the remaining lemmas, they are mandatory milestones to prove Patronus correctness.

Lemma 7.4.5. *Let* $r \in \mathbb{Z}$. *Let* q *a prime,* $\gamma | (q-k)$ *a power of two. Let* $m = (q-k)/\gamma$. *It holds that:*

$$\begin{aligned} &r = \mathsf{Highbits}_m(r,\gamma) \cdot \gamma + \mathsf{Lowbits}_m(r,\gamma) \bmod q \\ &\mathsf{Lowbits}_m(r,\gamma) \in [-\gamma/2, \gamma/2] \\ &\mathsf{Highbits}_m(r) \in [0, m-1]. \end{aligned}$$

Lemma 7.4.6. *Let* $r, z \in \mathbb{Z}_q$ *and* $\|z\|_\infty \leq \gamma/2$*:*

$$\mathsf{UseHint}_m(\mathsf{MakeHint}_m(z,r,\gamma),r,\gamma) = \mathsf{Highbits}_m(r+z,\gamma).$$

Lemma 7.4.7. *If* $\|\mathbf{s}\|_\infty \leq \beta$, $\|\mathsf{Lowbits}_m(\mathbf{r},\gamma)\|_\infty < \gamma/2 - \beta$ *then:*

$$\mathsf{Highbits}_m(\mathbf{r},\gamma) = \mathsf{Highbits}_m(\mathbf{r}+\mathbf{s},\gamma).$$

Lastly we use Lemma 7.4.5, Lemma 7.4.6 and Lemma 7.4.7 to obtain the following correctness proof for Patronus.

Proposition 7.4.8 (Correctness)**.** *Let* $(\mathsf{pk},\mathsf{sk}) \leftarrow \mathsf{KeyGen}()$, $m \in \{0,1\}^*$ *and* $\sigma \leftarrow \mathsf{Sign}(\mathsf{sk}, M)$. *Then,* $\mathsf{Verify}(\mathsf{pk}, M, \sigma) = 1$.

Proof. Let's consider the elements $\mathbf{z}$, $\mathbf{w}$, $\mathbf{w}_\mathsf{H}$, c, ξ, and $\mathbf{v} = (\mathsf{Encode}(\mathsf{Highbits}(\mathbf{z})), \mathsf{Lowbits}(\mathbf{z}), \mathsf{Encode}(\mathbf{h}), \xi)$ computed by Sign.

It is clear by definition of Encode, Decode, Highbits and Lowbits that

$$\mathbf{z} = \alpha_z \cdot \mathsf{Decode}\left(\mathsf{Encode}\left(\mathsf{Highbits}(\mathbf{z})\right)\right) + \mathsf{Lowbits}(\mathbf{z}) \text{ and } \mathbf{h} = \mathsf{Decode}\left(\mathsf{Encode}(\mathbf{h})\right).$$

We thus only need to show that:

$$\mathbf{w}_\mathsf{H} = \mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot 2^d, 2\gamma) \ \wedge \ \|\mathbf{h}\|_p \leq \omega \ \wedge \ \mathbf{z} \in \mathcal{H}^{nl}_{r,\mathbb{Z}}.$$

The two last conditions are trivially verified by definition of the signature algorithm that abort is they are not verified.

Let's show the first equation. The fact, provided by the signature algorithm, that $\|c\mathbf{b}_\mathsf{L}\|_\infty < \gamma$, and $\mathbf{h} = \mathsf{MakeHint}_m(-c\mathbf{b}_\mathsf{L}, \mathbf{w} - c\mathbf{e} + c\mathbf{b}_\mathsf{L}, 2\gamma)$ implies by Lemma 7.4.6 that:

$$\begin{aligned} \mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot 2^d, 2\gamma) &= \mathsf{Highbits}(\mathbf{Az} - c\left(\mathbf{b}_\mathsf{H} \cdot 2^d + \mathbf{b}_\mathsf{L}\right), 2\gamma) \\ &= \mathsf{Highbits}(\mathbf{Az} - c\mathbf{b}, 2\gamma) \quad \text{By Lemma 7.4.5} \\ &= \mathsf{Highbits}(\mathbf{w} - c\mathbf{e}, 2\gamma) \quad \text{by definition of } \mathbf{z} \text{ and } \mathbf{b}. \end{aligned}$$

Finally, the fact that $\|\mathsf{Lowbits}(\mathbf{w} - c\mathbf{e}, 2\gamma)\|_\infty < \gamma - \beta$ (provided by the signature algorithm), and lemma Lemma 7.4.7 allow us to conclude. By assumption $\|c\mathbf{b}_\mathsf{L}\|_\infty \leq \gamma$, by Lemma 7.4.6:

$$\mathsf{UseHint}_m(\mathsf{MakeHint}_m(c\mathbf{b}_\mathsf{L}, \mathbf{w} - c\mathbf{e} + c\mathbf{b}_\mathsf{L}, \gamma), \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot 2^d, 2\gamma) = \mathsf{Highbits}(\mathbf{w}, 2\gamma),$$

because $\|\mathbf{e}\|_\infty \leq (j/2)$ and $\mathsf{Lowbits}(\mathbf{w} - c\mathbf{e}, 2\gamma) < \gamma - \beta'$ with $\beta' = \max \|c\mathbf{e}\|_\infty = \tau(j/2)$. With j being the parameter of the binomial distribution. □

Secret key constraint

The rejection on $\mathbf{s}$ in Figure 7.5 is an important step to define and use properly $c\mathbf{s}$ in the following security proof paragraph. It ensures a theoretical bound directly on $\|c\mathbf{s}\|_2$ following directly the study from [Che+23, (Lemma 6)], which improving drastically the study of the euclidian norm on $c\mathbf{s}$ leading to better signatures.

Lemma 7.4.9 ([Che+23]). *For any $c \in \{0,1\}^n$ with hamming weight τ and a secret $\mathbf{s} \in \mathcal{R}^{k+l}$, $n\,\|c\mathbf{s}\|_2$ is bounded by*

$$\mathcal{N}(\mathbf{s}) = \tau^2 \cdot \sum_{i=1}^{m} \underset{j}{\overset{i\text{-}th}{\max}} \|\mathbf{s}(\omega_j)\|_2^2 + r \cdot \tau \cdot \underset{j}{\overset{(m+1)\text{-}th}{\max}} \|\mathbf{s}(\omega_j)\|_2^2 \|\mathbf{s}(\omega_j)\|_2^2,$$

with $m = \lfloor n/\tau \rfloor$ and $r = n \mod \tau$

Using Lemma 7.4.9, by simply fixing $\mathcal{N}(\mathbf{s}) \leq n\eta^2$ in the rejection rate, we obtain a direct upper bound on $c\mathbf{s}$.

7.4.2 Security of Patronus

We apply [Bar+23, (Theorem 2)] to reduce UF-CMA security to UF-NMA security.

This theorem relies on an analysis of the commitment min-entropy, the property of accepting honest-verifier zero knowledge, and the abort probability inherent in the associated identification protocol, as described in Figure 7.6.

$\mathsf{Com}(\mathsf{sk})$	$\mathsf{Resp}(\mathrm{com}, c, \mathrm{st})$
$\mathbf{s} \leftarrow \mathsf{sk}$	$\mathbf{y} \leftarrow \mathrm{st}$
$\mathbf{v} \leftarrow \perp$	$(\mathbf{w}_\mathsf{H}, \xi) \leftarrow \mathrm{com}$
$\mathbf{y} \leftarrow\!\!\$\ \mathcal{C}^{nl}_{\theta,r+\beta,\mathbb{Z}}$	$\mathbf{z} \leftarrow \mathbf{y} + \mathbf{s}c$
$\xi \leftarrow\!\!\$\ \{0,1\}^n$	$\mathbf{r}_0 \leftarrow \mathsf{Lowbits}(\mathbf{w} - c\mathbf{e},\, 2\gamma)$
$\mathbf{w} \leftarrow \mathbf{A}\mathbf{y}$	$\mathbf{v} \leftarrow \perp$
$\mathbf{w}_\mathsf{H} \leftarrow \mathsf{Highbits}(\mathbf{w}, 2\gamma)$	**if** $\mathbf{z} \in \mathcal{C}^{nl}_{\theta,r+\beta}$ **and** $\|\mathbf{r}_0\|_\infty < \gamma - \beta'$ **then**
$\mathrm{com} \leftarrow (\mathbf{w}_\mathsf{H}, \xi)$	$\quad\mathbf{h} \leftarrow \mathsf{MakeHint}_m(-c\mathbf{b}_\mathsf{L}, \mathbf{w} - c\mathbf{e} + c\mathbf{b}_\mathsf{L},\, 2\gamma)$
$\mathrm{st} \leftarrow \mathbf{y}$	$\quad$**if** $\|c\mathbf{b}_\mathsf{L}\|_\infty < \gamma$ **or** $\|\mathbf{h}\|_1 < \omega$ **then**
return $(\mathrm{com}, \mathrm{st})$	$\quad\quad\mathbf{v} = (\mathsf{Encode}(\mathsf{Highbits}(\mathbf{z})), \mathsf{Lowbits}(\mathbf{z}), \mathsf{Encode}(\mathbf{h}), \xi)$
	return $(c, \mathbf{v})$

Figure 7.6: Identification scheme associated to the signature algorithm.

Zero-knowledge The underlying identification protocol has ε bits of min-entropy when the following condition is met for any $(\mathsf{pk}, \mathsf{sk}) \leftarrow \mathsf{KeyGen}$ and $\mathbf{y} \leftarrow \mathcal{H}_{r_s}$:

$$\forall (\mathbf{w}, \xi), \Pr_{\mathbf{y}}\left[(\mathsf{Highbits}(\mathbf{A}\mathbf{y}, 2\gamma), \varepsilon) = (\mathbf{w}, \xi)\right] \leq 2^{-\varepsilon}.$$

Since ξ represents a uniformly random binary vector of length n, the probability is bounded by 2^{-n}, regardless of the chosen $(\mathsf{pk}, \mathsf{sk})$. Consequently, there is a minimum of 256 bits of entropy.

We must now demonstrate that the underlying Σ-protocol of the signature scheme, in Figure 7.6, satisfies the naHVZK property [Bar+23, (Definition 1)]. This property stipulates that non-aborting transcripts generated by the simulator must have a minimal statistical distance

$\mathsf{Sim}(\mathbf{A}, c)$

$\xi \leftarrow\!\$\ \{0,1\}^n$
$\mathsf{flag} \leftarrow \top$
$\mathbf{z} \leftarrow\!\$\ \mathcal{C}^{nl}_{\theta,r}$
with probability $\dfrac{M-1}{M}$ **return** $\perp$
$\mathbf{r}_0 \leftarrow \mathsf{Lowbits}(\mathbf{Az} - c\mathbf{b},\ 2\gamma)$
// We use the fact that in the signature
// algorithm ,$\mathbf{w} - c\mathbf{e} = \mathbf{Az} - c\mathbf{b}$
if $\|\mathbf{r}_0\|_\infty \geq \gamma - \beta'$ **then return** $\perp$
$\mathbf{h} \leftarrow \mathsf{MakeHint}_m(-c\mathbf{b}_\mathsf{L}, \mathbf{Az} - c\mathbf{b} + c\mathbf{b}_\mathsf{L},\ 2\gamma)$
if $\|c\mathbf{b}_\mathsf{L}\|_\infty \geq \gamma$ **and** $\|\mathbf{h}\|_1 \geq \omega$:
  return $\perp$
$\mathbf{v} = (\mathsf{Encode}(\mathsf{Highbits}(\mathbf{z})), \mathsf{Lowbits}(\mathbf{z}),$
  $\mathsf{Encode}(\mathbf{h}), \xi)$
$\mathbf{w}_\mathsf{H} \leftarrow \mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot \alpha,\ 2\gamma)$
return $((\mathbf{w}_\mathsf{H}, \xi), c, \mathbf{v})$

Figure 7.7: Simulator for the naHVZK *property.*

from real transcripts. In this context, Theorem 7.2.9 implies that the statistical distance is 0, the simulator being described in Figure 7.7.

For a given key pair $(\mathsf{pk}, \mathsf{sk})$, the underlying identification protocol has an abort probability associated to $(\mathsf{pk}, \mathsf{sk})$ equal to:

$$p_{\mathsf{pk},\mathsf{sk}} = \Pr_{(\mathrm{com},\mathrm{st})\leftarrow\mathsf{Com}(\mathsf{sk}),c\leftarrow\mathsf{SetChall}} [\mathsf{Resp}(\mathrm{com}, c, \mathrm{st} = \perp)] .$$

It is imperative to upper-bound every $p_{\mathsf{pk},\mathsf{sk}}$ with a constant p, except possibly for a negligible number of key pairs. Following the approach outlined in [Bar+23, (Theorem 2)], we do not rigorously prove this bound but rather estimate it using a heuristic. We follow the methodology detailed in [Duc+21, (Section 3.2)]. There are three distinct reasons to abort:

- When $\mathbf{z} \notin \mathcal{C}^{nl}_{\theta,r}$, which happens with probability $\frac{M-1}{M}$, as derived from Theorem 7.2.9.
- When $\|\mathsf{Lowbits}(\mathbf{w} - c\mathbf{e},\ 2\gamma)\|_\infty < \gamma - \beta'$. In [Duc+21, (Section 3.2)], it is heuristically assumed that $\mathsf{Lowbits}(\mathbf{w} - c\mathbf{e},\ 2\gamma)$ is uniformly distributed, leading to an estimation of the abort probability as $1 - e^{-256\frac{\beta' k}{\gamma}}$ with $\beta' = \tau$.
- When $\|c\mathbf{b}_\mathsf{L}\|_\infty < \gamma$ or $\|\mathbf{h}\|_1 < \omega$. According to [Duc+21, (Section 3.2)], the parameters suggest an abort probability of less than 0.005 for this scenario.

Based on these considerations, we can estimate that $p \leq \frac{M-1}{M} + 1 - e^{-256\frac{\beta' k}{\gamma}} + \frac{1}{200}$.

UF-NMA security The UF-NMA security is established following a procedure analogous to that of Dilithium or Haetae. The only distinction lies in our usage of a different LWE distribution.

Proposition 7.4.10. *For any quantum adversary $\mathcal{A}$ targeting the UF-NMA security with at most q_H queries to the random oracle $|H(\cdot)\rangle$, we can establish the existence of quantum adversaries $\mathcal{B}$ and $\mathcal{C}$ such that:*

$$\mathsf{Adv}^{\mathsf{ufnma}}_{\mathsf{Patronus}}(\mathcal{A}) \leq \mathsf{Adv}^{\mathsf{d\text{-}mlwe}}_{k,l,\mathcal{U}(\mathsf{Binom}^n)}(\mathcal{B}) + \mathsf{Adv}^{\mathsf{stmsis}}_{\mathsf{H},k+1,l,B_{\mathsf{NMA}},q_\mathsf{H}}(\mathcal{C}) ,$$

where $B_{\mathsf{NMA}} = \max\left(2\gamma + 1 + 2^{d-1}\tau, r\right)$

Proof. We call PatronumUnif the signature scheme Patronus where the vector $\mathbf{b}$ computed in KeyGen is uniformly taken on $\mathcal{R}^k_q$.

We can directly see that there exists an adversary $\mathcal{B}$ such that

$$\left|\mathsf{Adv}^{\mathsf{ufnma}}_{\mathsf{Patronus},q_\mathsf{H}}(\mathcal{A}) - \mathsf{Adv}^{\mathsf{ufnma}}_{\mathsf{PatronumUnif},q_\mathsf{H}}(\mathcal{A})\right| \leq \mathsf{Adv}^{\mathsf{d\text{-}mlwe}}_{k,l,\mathcal{U}(\mathsf{Binom}^n)}(\mathcal{B}) .$$

We now study the UF-NMA security of PatronumUnif.

Let's consider a matrix $\mathbf{M} = (\mathbf{A} \mid \mathbf{b})$ uniformly taken in $\mathcal{R}_q^{k\times(l+1)}$ for the MSIS problem. We compute $(\mathbf{b}_\mathsf{H}, \mathbf{b}_\mathsf{L}) \leftarrow \mathsf{Power2round}(\mathbf{b}, d)$ and set $\mathsf{pk} = (\mathbf{a}, \mathbf{b}_\mathsf{H})$. This pk is indistinguishable from a real public key of PatronumUnif.

Suppose that $\mathcal{A}$ finds a valid signature $\left(c, \mathbf{v} = (\bar{\mathbf{z}_\mathsf{H}}, \mathbf{z}_\mathsf{L}, \bar{\mathbf{h}}, \xi)\right)$ of a message m. We define $\mathbf{z} = \alpha_z \cdot \mathsf{Decode}(\bar{\mathbf{z}_\mathsf{H}}) + \mathbf{z}_\mathsf{L}$ and $\mathbf{h} = \mathsf{Decode}(\bar{\mathbf{h}})$. By definition of Verify, we have $\mathbf{z} \in \mathcal{C}^{nl}_{\theta,r+\beta,\mathbb{Z}}$ and $c = \mathsf{H}\left(\mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot \gamma,\, 2\gamma), \xi, m\right)$. We set $\mathbf{u} = \mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot \gamma,\, 2\gamma) \cdot 2\gamma - \mathbf{Az} - c\mathbf{b}$, so we have

$$c = \mathsf{H}\left((\mathbf{A} \mid \mathbf{b} \mid \mathbf{Id}_{k+1})\,(\mathbf{z}, c, \mathbf{u}),\, \xi, m\right) \;\Leftrightarrow\; c = \mathsf{H}\left((\mathbf{M} \mid \mathbf{Id}_{k+1})\,(\mathbf{z}, c, \mathbf{u}),\, \xi, m\right).$$

Moreover, using properties of $\mathsf{MakeHint}_m$ and [Duc+21, (Lemma 1)], we obtain:

$$\begin{aligned}\|\mathbf{u}\|_\infty &\le \|\mathsf{UseHint}_m(\mathbf{h}, \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot \gamma,\, 2\gamma) - \mathbf{Az} - c\mathbf{b}_\mathsf{H} \cdot \gamma\|_\infty + \|c\,(\mathbf{b} - \mathbf{b}_\mathsf{H} \cdot \gamma)\|_\infty \\ &\le 2\gamma + 1 + 2^{d-1}\tau.\end{aligned}$$

Thus, using $\|\mathbf{z}\|_\infty \le r$ which is implied by $\mathbf{z} \in \mathcal{C}^{nl}_{\theta,r,\mathbb{Z}}$, we have:

$$\|(\mathbf{z}, c, \mathbf{u})\|_\infty \le \max\left(2\gamma + 1 + 2^{d-1}\tau, r\right) = B_{\mathsf{NMA}},$$

□

Parameters and cost of known attacks

To provide concrete parameters and estimate the cost of the best-known attacks against our signature scheme, we adapt the concrete security analysis conduced in Haetae [Che+23] and use an adapted version of the scripts [DS21] and `HAETAE-helper-scripts/HAETAE_security_estimates.py`, included in the Haetae reference implementation version 2023.05.02.v1.0.[9] using the $\|\cdot\|_\infty$ MSIS estimator. Lastly, to obtain shorter signatures we use [Dud09] as compression algorithm.

In Table 7.7, we propose concrete parameters for Patronus and present estimated security levels and sizes. These parameters aim for a similar rejection rate M to Dilithium, but the rejection rates for Patronus are slightly larger to improve public key sizes. Then, in Table 7.1, we compare Patronus to Dilithium and Haetae, demonstrating that it offers a compelling trade-off in terms of signature size between these two constructions.

We follow the established *core-SVP* methodology as in Haetae [Che+23] to estimate the number of gates required to solve MLWE, MSIS and SelfTargetMSIS. Since we do not currently know of any way of exploiting the ring structure to solve MLWE and MSIS problems, we are simply viewing these problems as LWE and SIS problems. We consider the *primal* and *dual* attacks against LWE, and *plain BKZ* attacks for SIS and SelfTargetSIS, and note that the recent work [DEL25] does not impact our concrete security. Replacing vectors $\mathbf{v}$ with $\mathrm{vec}(\mathbf{v})$ the vector obtained by concatenating the coefficients of its coordinates, and matrix entries $\mathbf{a}_{ij} \in \mathcal{R}_q$ by the 256×256 matrix whose i-th column is $\mathrm{vec}(\mathbf{x}^{i-1} \cdot \mathbf{a}_{ij})$.

The security of $\mathsf{SelfTargetMSIS}_{\mathsf{H},k,l,B_{\mathsf{NMA}},\mathsf{q_H}}$ is estimated based on the security of $\mathsf{MSIS}_{k,l,B_{\mathsf{NMA}}}$ with the same bound B_{NMA}, following the analysis in [Duc+21, (Section 6.2.1)]. While it could have been possible to use the non-tight reduction from $\mathsf{SelfTargetMSIS}_{\mathsf{H},k,l,B_{\mathsf{NMA}},\mathsf{q_H}}$ to $\mathsf{MSIS}_{k,l,2B_{\mathsf{NMA}}}$, as described under the definition of SelfTargetMSIS, we note that this choice aligns with neither Dilithium nor Haetae's approaches for the security property UF-CMA that we consider.

We recall that a brief comparison with Dilithium and Haetae is provided in Table 7.1.

[9] Accessible at https://www.kpqc.cryptolab.co.kr/haetae.

Table 7.7: Patronus *parameter sets for NIST security levels II, III and V.*

Security target	120	180	260
n	256	256	256
(k, l)	(5,4)	(7,5)	(10,7)
q	523,777	523,777	1,047,041
η	31	34	37
j	2	2	2
τ	39	49	60
r	180,350	210,424	467,345
M	3	4.25	3
$\beta = \eta\sqrt{\tau}$	194	238	287
γ	104,755	104,755	349,013
d	14	14	15
ω	70	80	50
Forgery			
BKZ block-size b	412	625	899
Classical hardness	120	182	262
Quantum hardness	105	160	231
Key Recovery			
BKZ block-size b	475	630	902
Classical hardness	138	184	263
Quantum hardness	122	161	231
Size			
vk (with seed)	832	1,152	1,632
sign	2,070	2,575	3,721

Sampler implementation

To demonstrate the practical viability of our sampler, we developed an unoptimised, isochronous, and portable implementation in C for SampleH and conducted experiments using the parameter sets presented in Table 7.7. The source code is publicly available at:

https://github.com/patronus-signature/patronus-signature.

In Table 7.4, we present the competitive performance of our sampler, tested on an i5-1021U CPU. We utilised the same SHAKE-256 code as provided in Dilithium's reference implementation[10] to generate the pseudo-randomness. One should note that, on average, the sampling time constitutes less than 10% of the total signature generation time for the Dilithium reference implementation.

A bimodal variant of Patronus?

In this section, we study the feasibility of a bimodal version of our scheme. Using the same notations as in the introduction, recall that in the FSwA paradigm, an element $\mathbf{z} = \mathbf{y} + c\mathbf{s}$ is rejected when it leaks information about $c\mathbf{s}$. It was shown in [DDLL13] in the case of Gaussian distributions that if one generates $\mathbf{z}$ as $\mathbf{z} = \mathbf{y} + bc\mathbf{s}$ instead, where b is a secret uniform random bit $b \in \{\pm 1\}$, then one can rescale the initial target distribution for $\mathbf{z}$ and obtain substantially shorter sizes. This technique has been widely reused in modern signature schemes and with various distributions, such as Haetae with uniform distributions over hyperballs.

[10] Accessible at https://github.com/pq-crystals/dilithium/tree/master/ref.

We use the following proposition to argue that a direct application of this technique does not improve the results of this paper.

Proposition 7.4.11. *Let $R, r \in \mathbb{R}_{>0}$ be two radii with $R > r$. Let $n \geq 4$ denote the ambient dimension. Then:*

$$\max\left\{\rho \in \mathbb{R}_{\geq 0} : \mathcal{H}_{\rho}^{n} \subseteq \bigcap_{\mathbf{c}\in\mathcal{H}_{r}^{n}} \left(\mathcal{H}_{R,\mathbf{c}}^{n} \cup \mathcal{H}_{R,-\mathbf{c}}^{n}\right)\right\} = R - r.$$

Proof. Cleary the bimodal intersection contains the unimodal intersection so $\mathcal{H}_{R-r}^{n}$ is contained in both. Therefore $\rho \geq R-r$. Assume by contradiction that there exists a radius $r' > R-r$ such that $\mathcal{H}_{r'}^{n}$ is included in $\bigcap_{\mathbf{c}\in\mathcal{H}_{r}^{n}} \left(\mathcal{H}_{R,\mathbf{c}}^{n} \cup \mathcal{H}_{R,-\mathbf{c}}^{n}\right)$. Alternatively, for all $\mathbf{z} \in \mathcal{H}_{r'}^{n}$ and all $\mathbf{c} \in \mathcal{H}_{r}^{n}$, at least one of $\mathbf{z}+\mathbf{c}$ and $\mathbf{z}-\mathbf{c}$ should be an element of $\mathcal{H}_{R}^{n}$. Recall the notation $\{\sqrt{n}\} = \sqrt{n} - \lfloor\sqrt{n}\rfloor$, and take:

$$\mathbf{c} = (0, \ldots, 0, \underbrace{r, \ldots, r}_{\lfloor\sqrt{n}\rfloor}, \{\sqrt{n}\}r) \qquad \mathbf{z} = (\underbrace{r', \ldots, r'}_{\lfloor\sqrt{n}\rfloor}, \{\sqrt{n}\}r', 0, \ldots, 0).$$

$\mathbf{z}$ is one of the vertices of $\mathcal{H}_{r'}^{n}$ and $\mathbf{c}$ is a vertex of $\mathcal{H}_{r}^{n}$. We have $\|\mathbf{z}+\mathbf{c}\|_1 = \|\mathbf{z}-\mathbf{c}\|_1 = \|\mathbf{z}\|_1 + \|\mathbf{c}\|_1$ as both have disjoint coordinate support because $n \geq 4$. This leads to $\|\mathbf{z}+\mathbf{c}\|_1 = \|\mathbf{z}-\mathbf{c}\|_1 = \sqrt{n}\cdot(r+r')$. Since $r' > R-r$, $r'+r > R$, which means that neither $\mathbf{z}+\mathbf{c}$ nor $\mathbf{z}-\mathbf{c}$ are in $\mathcal{H}_{R}^{n}$. We get our contradiction and this proves that $\rho \leq R-r$, which concludes. □

Proposition 7.4.11 shows that in the case of uniform distribution rejection sampling, the largest version of $\mathcal{H}$ contained in the set of possible values of $\mathbf{z}$ that leak no information on the secret is the same as that obtained in the unimodal case, and therefore using a bimodal variant would not improve our results. A similar -although slightly different- limitation appears in the case of Dilithium, with hypercubes. Indeed in our case, there might still exist another polytope between $\mathcal{H}_{R-r}$ and the bimodal intersection. However, it is not hard to see that this bimodal intersection is not convex, so a convenient polytope is unlikely to exist, would complicate the analysis, and would lead to imperfect rejection sampling, which falls out of the scope of our analysis.

Part IV

Mathematical Properties of Cryptography-Related Objects

Chapter

8 On the First Minimum of a Random Ideal Lattice

Abstract The notion of a random lattice was introduced by Siegel in 1945, and the theory of random lattices enables precise estimates for quantities such as the expected value of the first minimum $\lambda_1(\cdot)$. What can be said of a random ideal lattice? In this chapter, we first recall some results on random real lattices, and then focus our attention to random ideal lattices. In the case of quadratic fields, we explain how the exact values of the moments of $\lambda_1(\cdot)$ can be computed. In an attempt to generalise Siegel's formula to random ideal lattices, we then derive a formula that allows us to compute the expected number of ideal lattice points in an origin-centred ball for number fields of small degree and discriminant. While working on the same problem, Gargava and Viazovska [GV24] successfully obtained an analogue of Siegel's formula for random ideal lattices that shows asymptotic insight for large degree cyclotomic fields. We find that the study of random ideal/module lattices and their geometry seems to connect deeply to many advanced branches of mathematics, including sphere packings, the study of modular and automorphic forms, ergodic theory and homogeneous dynamics. We hope that the nice problems in this topic can act as a bridge between mathematically inclined cryptographers and cryptographically inclined mathematicians.

The results of this chapter are unpublished. Section 8.4 is joint work with Seungki Kim.

Chapter content

8.1 Introduction

The length of the shortest non-zero vector $\lambda_1(\Lambda)$ in a lattice Λ is an extremely important quantity in cryptography, as it governs the hardness of lattice problems such as SVP. This makes understanding its properties a crucial step in understanding attacks on lattice-based

cryptography. When the behaviour of $\lambda_1(\Lambda)$ is too difficult to predict, cryptographers often make the heuristic assumption that the lattice Λ behaves as if it were random. For better or for worse, the number of lattice points in a given ball is often approximated by the volume of the ball. This Gaussian heuristic holds when the radius of the ball goes to infinity, but it also holds on average, *i.e.* for Haar-random lattices. This gives more weight to heuristics that are used to analyse lattice algorithms like BKZ when the lattices at play look random enough.

As we have now seen multiple times by now, cryptographic schemes use special lattices with additional structure, such as ideal (resp. module) lattices: lattices who can be seen as rank 1 (resp. r) $\mathcal{O}_K$-modules, where $\mathcal{O}_K$ is the ring of integers of a number field K, most of the time cyclotomic. The study of *random* lattices in this setting is very new to cryptology, but was relevant in the study of dense lattice packings in high dimensions. In 2013, Venkatesh [Ven13] proved the (at the time) best lower bounds for high-dimensional sphere packings. His argument follows the probabilistic method. He considers a space of rank 2 module-lattices over cyclotomic fields, and is able to show that their first minima will be on average slightly larger than what one would expect for random lattices without structure. Therefore there must exist a lattice in this class with large first minimum, or equivalently a dense lattice packing. A lot of recent progress was made towards lower bounds in lattice and sphere packings, see [CJMS23; Kla25].

The recent work of Gargava, Serban, Viazovzka and Viglino [GSV24; GSVV24] studies the moments of the expected number of lattice points in a ball where the lattices are random module lattices, but for technical reasons their results do not cover the smaller ranks that arise in cryptology. In a concurrent work, [GV24] give a formula for the first moment of the expected number of lattice points in a ball in the rank 1 ideal lattice case.

In this chapter we first recall classic theory by Siegel and Rogers in Section 8.2, and show how to get probabilistic results on both λ_1 and the smoothing parameter η_ε of random (real) lattices. In Section 8.3, we study and visualise the structure of ideal lattices in real quadratic fields, and this leads to an algorithm that computes the expected value of λ_1 exactly. In Section 8.4, we show an analogue of Siegel's formula for ideal lattices, which arises naturally when looking at the problem through an adelic perspective. It allows us to precisely compute the expected number of lattice points in a ball for random ideal lattices in fields of small degree and discriminant. We conclude by comparing our work to the brilliant formula of [GV24].

8.2 Random Real Lattices

Recall that the space of full-rank n-dimensional lattices of covolume 1 is defined as $X_n = \mathrm{SL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{Z})$, equipped with the unique $\mathrm{SL}_n(\mathbb{Z})$-invariant probability measure μ_n. For a regular enough function f, Siegel proved in [Sie45] that the average value of the *Siegel transform* $\sum_{\mathbf{x}\in\Lambda} f(\mathbf{x})$ over lattices Λ sampled from μ can be obtained by simply integrating f over the full space $\mathbb{R}^n$:

Theorem 8.2.1 (Siegel's mean value formula)**.** *Let $n \in \mathbb{Z}_{>0}$ be an integer and $f : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ a Schwartz function, then*

$$\int_{\Lambda\in X_n} \sum_{\mathbf{x}\in\Lambda\setminus\{\mathbf{0}\}} f(\mathbf{x}) d\mu_n = \int_{\mathbb{R}^n} f(\mathbf{x}) d\mathbf{x}.$$

Ten years later, Rogers gave a formula in [Rog55] for higher moments of the Siegel transform. We only restate his formula for the second moment.

Theorem 8.2.2 (Rogers's second moment formula)**.** *Let $n \in \mathbb{Z}_{>2}$ be an integer, and $f : \mathbb{R}^n \mapsto \mathbb{R}_{\geq 0}$ be a Schwartz function, then*

$$\int_{\Lambda\in X_n} \left(\sum_{\mathbf{x}\in\Lambda\setminus\{\mathbf{0}\}} f(\mathbf{x}) \right)^2 d\mu_n = \left(\int_{\mathbb{R}^n} f(\mathbf{x}) d\mathbf{x} \right)^2 + \sum_{\substack{k\in\mathbb{Z}_{>0}, q\in\mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \int_{\mathbb{R}^n} f(k\mathbf{x}) f(q\mathbf{x}) d\mathbf{x}.$$

Note that this formula is only defined for $n > 2$. The reader can refer to [Sch60] for the two-dimensional analogue. We chose to exclude the trivial vector $\mathbf{0}$ from the integral, as this reduces the number of terms in the sum.

If for a given f, we know both the first and second moments of the Siegel transform, then probabilistic arguments such as Chebyshev's inequality give access to probabilistic results on the concentration of the Siegel transform around its mean. Two natural choices for f relate to useful quantities in cryptography:

- If f is the indicator function $\mathbf{1}_S$ of a measurable set S, then the Siegel transform counts the number of lattice points in S. A natural choice for S is the origin-centred ball $B_n(R)$, as the number $\#(\Lambda \cap B_n(R))$ of points of the lattice Λ is strictly less than 3 if and only if $R \geq \lambda_1(\Lambda)$. This choice of function gives probabilistic results on $\lambda_1(\Lambda)$, which were used to prove lower bounds on the density of lattice packings.

- If f is a Gaussian mass function $\rho_s(\mathbf{x}) = \exp\left(-\pi\|\mathbf{x}\|^2/s^2\right)$, then the Siegel transform gives the *Gaussian energy* of the lattice. This quantity relates to the smoothing parameter of a lattice Λ, defined for $\varepsilon > 0$ by $\eta_\varepsilon(\Lambda) := \min\{s > 0 : \rho_{1/s}(\Lambda^\vee) \leq 1 + \varepsilon\}$. The smoothing parameter was introduced in [MR07] for cryptographic applications, to quantify when a Gaussian error looks uniform modulo Λ.

We now study the first and second moments for both choices of functions. The case of the first function is classical and was known to Rogers, and used in [AEN18] to derive probabilistic bounds on $\lambda_1(\cdot)$. The case of the second function feels like it should be known already by mathematicians and physicists but was only derived in the context of cryptography independently to us by Pouly and Shen in [PS24].

8.2.1 Gaussian Heuristic

The fact that the number of points of a lattice Λ in a well-behaved set S should roughly be equal to $\mathrm{vol}(S)/\mathrm{vol}(\Lambda)$ is commonly referred to as the Gaussian heuristic. Applying Theorem 8.2.1 to the function $\mathbf{1}_S$ formalises the Gaussian heuristic *on average*. Computing Theorem 8.2.2 with $f_R = \mathbf{1}_{B_n(R)}$ leads to average results on λ_1.

Theorem 8.2.3. *Let $n \in \mathbb{Z}_{>0}$ be an integer and $R \in \mathbb{R}_{\geq 0}$ be a real number, then*

$$\int_{X_n} \#((\Lambda\backslash\{\mathbf{0}\}) \cap B_n(R)) d\mu_n = \mathrm{vol}(B_n(R)),$$

where $B_n(R)$ is the n-dimensional Euclidean ball of radius R.

Proof. By introducing f_R the indicator function of $B_n(R)$, we can use Theorem 8.2.1 to get

$$\begin{aligned}\int_{X_n} \#((\Lambda\backslash\{\mathbf{0}\}) \cap B_n(R)) d\mu_n &= \int_{X_n} \sum_{\mathbf{x}\in\Lambda\backslash\{\mathbf{0}\}} f_R(\mathbf{x}) d\mu_n \\ &= \int_{\mathbb{R}^n} f_R(\mathbf{x}) d\mathbf{x} \\ &= \mathrm{vol}(B_n(R)).\end{aligned}$$

□

This statement confirms that the Gaussian heuristic holds for random lattices, as long as we exclude the zero vector. The second moment can also be computed.

Theorem 8.2.4. *Let $n \in \mathbb{Z}_{>2}$ be an integer and $R \in \mathbb{R}_{\geq 0}$ be a real number, then*

$$\int_{X_n} \#((\Lambda \backslash \{\mathbf{0}\}) \cap B_n(R))^2 d\mu_n = \mathrm{vol}(B_n(R))^2 + 2\left(\frac{2\zeta(n-1)}{\zeta(n)} - 1\right)\mathrm{vol}(B_n(R)),$$

where $B_n(R)$ is the n-dimensional Euclidean ball of radius R.

Proof. Again, we use the indicator function f_R of $B_n(R)$ to count lattice points in the ball, and use Theorem 8.2.2 to get

$$\begin{aligned}
\int_{X_n} \#((\Lambda \backslash \{\mathbf{0}\}) \cap B_n(R))^2 d\mu_n &= \int_{X_n} \left(\sum_{\mathbf{x} \in \Lambda \backslash \{\mathbf{0}\}} f_R(\mathbf{x})\right)^2 d\mu_n \\
&= \left(\int_{\mathbb{R}^n} f_R(\mathbf{x}) d\mathbf{x}\right)^2 + \sum_{\substack{k \in \mathbb{Z}_{>0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \int_{\mathbb{R}^n} f_R(k\mathbf{x}) f_R(q\mathbf{x}) d\mathbf{x} \\
&= \mathrm{vol}(B_n(R))^2 + \sum_{\substack{k \in \mathbb{Z}_{>0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \int_{\mathbb{R}^n} \mathbf{1}\left(\max(k, |q|) \cdot \|\mathbf{x}\| \leq R\right) d\mathbf{x} \\
&= \mathrm{vol}(B_n(R))^2 + \sum_{\substack{k \in \mathbb{Z}_{>0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \frac{\mathrm{vol}(B_n(R))}{\max(k, |q|)^n} \\
&= \mathrm{vol}(B_n(R))^2 + \mathrm{vol}(B_n(R)) \left(\sum_{\substack{k \in \mathbb{Z}_{\geq 0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \frac{1}{\max(k, |q|)^n} - 2\right) \\
&= \mathrm{vol}(B_n(R))^2 - 2\mathrm{vol}(B_n(R)) + \frac{1}{2}\mathrm{vol}(B_n(R)) \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|_\infty^{-n},
\end{aligned}$$

where $P(\mathbb{Z}^2)$ denotes the set of primitive vectors of $\mathbb{Z}^2$. The sum can be written over non-primitive vectors. Indeed

$$\begin{aligned}
\sum_{\mathbf{v} \in \mathbb{Z}^2 \backslash \{\mathbf{0}\}} \|\mathbf{v}\|_\infty^{-n} &= \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \sum_{q \geq 1} \|q\mathbf{v}\|_\infty^{-n} = \sum_{q \geq 1} q^{-n} \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|_\infty^{-n} \\
&= \zeta(n) \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|_\infty^{-n}.
\end{aligned}$$

The remaining sum can be computed as follows:

$$\sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|_\infty^{-n} = \sum_{N=1}^{\infty} \sum_{\|\mathbf{v}\|_\infty = N} \frac{1}{N^n} = \sum_{N=1}^{\infty} \sum_{\|\mathbf{v}\|_\infty = N} \frac{8N}{N^n} = 8\zeta(n-1).$$

Combining the previous identities proves the claim. □

A commonly used extension of the Gaussian heuristic is the fact that a *typical* full-rank lattice of $\mathbb{R}^n$ should satisfy $\lambda_1(\Lambda) \approx \sqrt{\frac{n}{2\pi e}}\mathrm{vol}(\Lambda)^{1/n}$. This can be formalised by applying Chebyshev's inequality to bound

$$\Pr(\lambda_1(\Lambda) \leq R) = \Pr(\#(\Lambda \backslash \{\mathbf{0}\} \cap B_n(R)) \geq 2)$$

for a well-chosen R. More precise computations can be found in [AEN18; PS24].

8.2.2 Gaussian Energy

Of interest to cryptography is the smoothing parameter. We note that it is possible to get probabilistic bounds by studying the average Gaussian mass of the dual of a random lattice. The Haar-measure μ_n is invariant under duality, so Theorem 8.2.1 and Theorem 8.2.2 can be applied to get the first two moments of the random Gaussian mass.

Theorem 8.2.5. *Let $n > 0$ and $s > 0$. Then*

$$\int_{\Lambda \in X_n} \rho_{1/s}(\Lambda^\vee \backslash \{\mathbf{0}\}) d\mu_n = \frac{1}{s^n}.$$

Proof. The space X_n is invariant under duality. Using Theorem 8.2.1, we can write

$$\begin{aligned}
\int_{\Lambda \in X_n} \rho_{1/s}(\Lambda^\vee \backslash \{\mathbf{0}\}) d\mu_n &= \int_{\Lambda \in X_n} \rho_{1/s}(\Lambda \backslash \{\mathbf{0}\}) d\mu_n \\
&= \int_{\mathbb{R}^n} \rho_{1/s}(\mathbf{x}) d\mathbf{x} \\
&= \left(\int_{-\infty}^{\infty} e^{-\pi s^2 x^2} dx \right)^n \\
&= 1/s^n.
\end{aligned}$$

□

Theorem 8.2.6. *Let $s > 0$. Then*

$$\int_{\Lambda \in X_n} \rho_{1/s}(\Lambda^\vee \backslash \{\mathbf{0}\})^2 d\mu_n = \frac{1}{s^{2n}} + \left(\frac{\zeta(n/2)\beta(n/2)}{\zeta(n)} - 1 \right) \cdot \frac{2}{s^n},$$

where $\zeta(\cdot)$ and $\beta(\cdot)$ respectively denote the Riemann zeta function and the Dirichlet beta function.

Proof. Using Theorem 8.2.2 with the Gaussian mass function, we get

$$\begin{aligned}
\int_{\Lambda \in X_n} \rho_{1/s}(\Lambda^\vee \backslash \{\mathbf{0}\})^2 d\mu_n &= \int_{\Lambda \in X_n} \rho_{1/s}(\Lambda \backslash \{\mathbf{0}\})^2 d\mu_n \\
&= \left(\int_{\mathbb{R}^n} \rho_{1/s}(\mathbf{x}) d\mathbf{x} \right)^2 + \sum_{\substack{k \in \mathbb{Z}_{>0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \int_{\mathbb{R}^n} e^{-\pi s^2 (k^2+q^2)\|\mathbf{x}\|^2} d\mathbf{x} \\
&= \frac{1}{s^{2n}} + \sum_{\substack{k \in \mathbb{Z}_{>0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \left(\frac{1}{s\sqrt{k^2+q^2}} \right)^n \\
&= \frac{1}{s^{2n}} - \frac{2}{s^n} + \sum_{\substack{k \in \mathbb{Z}_{\geq 0}, q \in \mathbb{Z}_{\neq 0} \\ \gcd(k,q)=1}} \left(\frac{1}{s\sqrt{k^2+q^2}} \right)^n \\
&= \frac{1}{s^{2n}} - \frac{2}{s^n} + \frac{1}{2s^n} \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|^{-n},
\end{aligned}$$

where $P(\mathbb{Z}^2)$ denotes the set of primitive vectors of $\mathbb{Z}^2$.

The remaining work is the following: rewrite the sum as an Epstein zeta sum over non-primitive vectors. Indeed

$$\begin{aligned}
E_2(\mathbb{Z}^2, n/2) &:= \sum_{\mathbf{v} \in \mathbb{Z}^2 \backslash \{\mathbf{0}\}} \|\mathbf{v}\|^{-n} = \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \sum_{q \geq 1} \|q\mathbf{v}\|^{-n} = \sum_{q \geq 1} q^{-n} \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|^{-n} \\
&= \zeta(n) \sum_{\mathbf{v} \in P(\mathbb{Z}^2)} \|\mathbf{v}\|^{-n}.
\end{aligned}$$

A standard computation (see also [W, Equation 38]), gives

$$E_2(\mathbb{Z}^2, n/2) = 4\zeta(n/2)\beta(n/2).$$

Stitching the equations together yields the claimed result. □

Similar to the previous section, we note that for $\varepsilon, s > 0$,

$$\Pr(\eta_\varepsilon(\Lambda) > s) = \Pr(\rho_{1/s}(\Lambda^\vee \backslash \{\mathbf{0}\}) > \varepsilon),$$

and this quantity can be bounded above by Chebychev's inequality using the first and second moment, directly leading to probabilistic bounds on $\eta_\varepsilon(\Lambda)$. Again, see [PS24] for the details of this step.

Remark 8.2.7. We note that similar computation with the general version of Rogers formula will give expressions for the higher moments, but the resulting sums will become much more difficult to compute exactly and might need to be approximated. Some of the sums that appear are also useful in chemistry for the study of crystals [BGMWZ13], and some can be decomposed as L-functions, hinting to more advanced number theory.

8.3 Short Vectors in Real Quadratic Fields

The aim of this section explores the following question. We use it to illustrate why the Arakelov class group (introduced by [dBDPW20] for cryptography) is the correct way to view ideal lattices.

Question 8.3.1. *Is is possible to compute the expected value and possibly higher moments of $\lambda_1(\Lambda)$, where Λ is a random ideal lattice of $\tilde{\mathrm{Cl}}(K)$, where K is a quadratic field?*

For this section we let $d > 0$ be a non-zero squarefree integer, and $K = \mathbb{Q}(\sqrt{d})$ the associated real quadratic field. If $d \equiv 1 \pmod 4$, then K has discriminant $\Delta_K = d$, and $\Delta_K = 4d$ otherwise. We denote by $\mathrm{Cl}(K)$ the class group of K, and h_K its order, the class number. We recall that elements in K are of the form $a+b\sqrt{d}$ for $a, b \in \mathbb{Q}$, and that its ring of integers $\mathcal{O}_K$ is $\mathbb{Z}[\sqrt{d}]$ when $d \equiv 1 \pmod 4$, and $\mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right]$ otherwise. The units $\mathcal{O}_K^\times$ are solutions to the Pell equation defined by d. $\mathcal{O}_K^\times$ is finitely generated and one-dimensional, which allows us to define the fundamental unit of $\mathcal{O}_K$ as the unique $\varepsilon_0 > 1$ such that $\mathcal{O}_K = \pm\varepsilon_0^{\mathbb{Z}}$ (the only roots of unity in real quadratic fields are ± 1). The regulator of K can be expressed using the fundamental unit as $R_K = \ln \varepsilon_0$. K embeds into $K_\mathbb{R} = K \otimes \mathbb{R} \cong \mathbb{R}^2$ canonically through the embedding

$$\sigma : \begin{cases} K & \to K_\mathbb{R} \cong \mathbb{R}^2 \\ \alpha = a + b\sqrt{d} & \mapsto (\sigma_1(\alpha), \sigma_2(\alpha)) = (a + b\sqrt{d}, a - b\sqrt{d}) \end{cases}.$$

We recall that σ sends ideals $I \subseteq \mathcal{O}_K$ to the lattice $\sigma(I) \subseteq K_\mathbb{R}$ of volume $\mathrm{vol}(\sigma(I)) = \mathcal{N}(I)\sqrt{\Delta_K}$. Recall that we assume that lattices are always normalised to have covolume 1, and that the distribution over ideal lattices stems from the uniform measure on the Arakelov class group of K (or norm 1 idèle class group). The Arakelov class group of K can be seen as a product of h_K circles, as we have the following exact sequence:

$$0 \to \mathbb{R}/(R_K\mathbb{Z}) \to \tilde{\mathrm{Cl}}(K) \xrightarrow{\phi} \mathrm{Cl}(K) \to 0. \tag{8.1}$$

An element $x \cdot \mathfrak{a} \in \tilde{\mathrm{Cl}}(K)$ maps through ϕ to the ideal class $[\mathfrak{a}] \in \mathcal{O}_K$, obtained by forgetting the information at the infinite places.

Remark 8.3.2. We focus our study on real quadratic fields for the following reason: imaginary quadratic fields only have a finite number of ideal lattices, one for each class in the class group. Statistics of λ_1 over each class become trivial.

8.3.1 Lattices as Elements of the Upper Half-Plane $\mathbb{H}$

The hyperbolic upper half-plane $\mathbb{H} = \{x+iy \in \mathbb{C} : y > 0\}$ equipped with the hyperbolic measure $ds = \frac{dxdy}{y}$ allows us to represent a two dimensional lattice $\Lambda = \mathbf{v}_1\mathbb{Z} + \mathbf{v}_2\mathbb{Z}$ by the element

$$\tau_{(\mathbf{v}_1,\mathbf{v}_2)} = \frac{\phi(\mathbf{v}_2)}{\phi(\mathbf{v}_1)},$$

where $\phi(\mathbf{v}) = v_1 + iv_2$ when $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$. If the imaginary part of $\tau_{(\mathbf{v}_1,\mathbf{v}_2)}$ as defined above were negative, then we simply assume that the basis was given in reverse order, so that our definition forces $\tau_{(\mathbf{v}_1,\mathbf{v}_2)} \in \mathbb{H}$.

The point $\tau_{(\mathbf{v}_1,\mathbf{v}_2)}$ corresponds in fact to the complex number representing the position of the second basis vector after Λ is rescaled, rotated (and possibly flipped) so that the new position of $\mathbf{v}_1$ is the unit vector $(1, 0)$.

The modular group $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ acts on lattice bases in $\mathbb{H}$ through the operation

$$\tau \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \tau = \frac{a\tau + b}{c\tau + d}.$$

The same lattice Λ is represented by an infinite orbit of points of $\mathbb{H}$, of which there exists always a unique (Lagrange-reduced) representative in the fundamental domain

$$\mathcal{F} = \left\{ z \in \mathbb{H} : |z| \geq 1 \text{ and } -\frac{1}{2} < \Re(z) \leq \frac{1}{2} \right\}.$$

Reciprocally, all points of $\mathcal{F}$ represent a lattice, such that we can identify $\mathcal{F}$ and $\mathbb{H}/\mathrm{PSL}_2(\mathbb{Z})$.

In plot 8.1a, we represent random lattices in $\mathcal{F}$. Notice that after rescaling to volume 1, the shortest non-zero vector of a lattice represented by τ in the fundamental domain is given by $\Im(\tau)^{-1/2}$. Therefore following [AEN18, Theorem 13], we get that the expected value of λ_1 by integrating against ds

$$\mathbb{E}_{X_2}(\lambda_1(L)) = \frac{1}{N_{\mathcal{F}}} \int_{\mathcal{F}} y^{-1/2} \frac{dxdy}{y^2} = \frac{2}{\pi} \int_{-1/2}^{1/2} \frac{dx}{(1-x^2)^{3/4}} \approx 0.6826,$$

where $N_{\mathcal{F}}$ is a normalisation term corresponding to the total measure of $\mathcal{F}$. Squaring gives a nicer expression:

$$\mathbb{E}_{X_2}(\lambda_1(L)^2) = \frac{3\ln 3}{2\pi} \approx 0.5245.$$

8.3.2 Ideals Lattices in $\mathbb{H}$: Picturing the Arakelov Class Group

The exact sequence of Equation 8.1 can be visualised by looking at reduced representatives of ideal lattices in $\mathbb{H}$.

Let $I \subseteq \mathcal{O}_K$ be an ideal, with integral basis $(\omega_1, \omega_2) \in K^2$, such that $I = \omega_1\mathbb{Z} + \omega_2\mathbb{Z}$. $\sigma(I)$ is a rank 2 lattice generated by the basis $(\sigma(\omega_1), \sigma(\omega_2))$. What can be said of $\tau_{(\sigma(\omega_1),\sigma(\omega_2))} \in \mathbb{H}$?

Theorem 8.3.3. *All ideals I of the same ideal class have a representative of their embedded lattice $\sigma(I)$ that lies on a circle in $\mathbb{H}$.*

Proof. If I has basis ω_1, ω_2, then $\sigma(I)$ has basis $(\sigma(\omega_1), \sigma(\omega_2))$. Using ω' as a notation for the Galois conjugate $\sigma_2(\omega)$, we have up to reordering of the basis,

$$\tau_{(\sigma(\omega_1),\sigma(\omega_2))} = \frac{\omega_2 + i\omega_2'}{\omega_1 + i\omega_1'}.$$

For notational convenience, we denote as z_I the above quantity. Now let $J \in [I]$ be an ideal in the same class, then there exists an $\alpha \in K^\times$ such that $J = (\alpha)I$. Using the multiplicativity of σ, $(\sigma(\alpha)\sigma(\omega_1), \sigma(\alpha)\sigma(\omega_2))$ is a basis of $\sigma(J)$, and with the same conventions as above, we get

$$z_J = \frac{\alpha\omega_2 + i\alpha'\omega_2'}{\alpha\omega_1 + i\alpha'\omega_1'}.$$

Rearranging gives

$$\frac{z_J - \omega_2'/\omega_1'}{z_J - \omega_2/\omega_1} = -i\frac{\alpha}{\alpha'} \cdot \frac{\omega_1}{\omega_1'}. \tag{8.2}$$

Note that α' is not zero otherwise we would have $\mathcal{N}(\alpha) = \alpha\alpha' = 0$. In particular, because the right hand side is purely imaginary, the lines through $(\omega_2'/\omega_1', z_J)$ and $(\omega_2/\omega_1, z_J)$ must be orthogonal, which equivalently says that z_J lies on the (semi)-circle of diameter given by the real points ω_2/ω_1 and ω_2'/ω_1'. □

Another way to see this using notations from the above proof is that the transformation that gives the locus of points z_J as α varies in $K^\times$ depends only on the value of $c = \alpha'/\alpha$. As α describes $K^\times$, c traces a dense subset of the real line, which is then mapped through the Möbius transform $c \mapsto \frac{\omega_2 + ic\omega_2'}{\omega_1 + ic\omega_1'}$ to a circle.

Corollary 8.3.4. *If I is an ideal with $\mathbb{Z}$-basis (ω_1, ω_2), then all ideals of the same class have a representative in $\mathbb{H}$ on the circle with centre $\frac{\omega_2}{2\omega_1} + \frac{\omega_2'}{2\omega_1'}$, and radius $\left|\frac{\omega_2}{2\omega_1} - \frac{\omega_2'}{2\omega_1'}\right|$.*

The ideals form a dense subset of the semi-circle, which is transformed into the whole circle when also considering distortions at the infinite places. Interestingly, when α is a taken to be a unit $\varepsilon \in \mathcal{O}_K^\times$, we get an equal ideal, whose representative is further along the circle. Therefore if ε_0 is a fundamental unit of the number field, then the following arc represents all ideals in the ideal class:

$$\mathcal{A} = \left\{ z \in \mathbb{H} : z = \frac{c\omega_2 + i\omega_2'}{c\omega_1 + i\omega_1'}, c \in [1, \varepsilon_0^2) \right\},$$

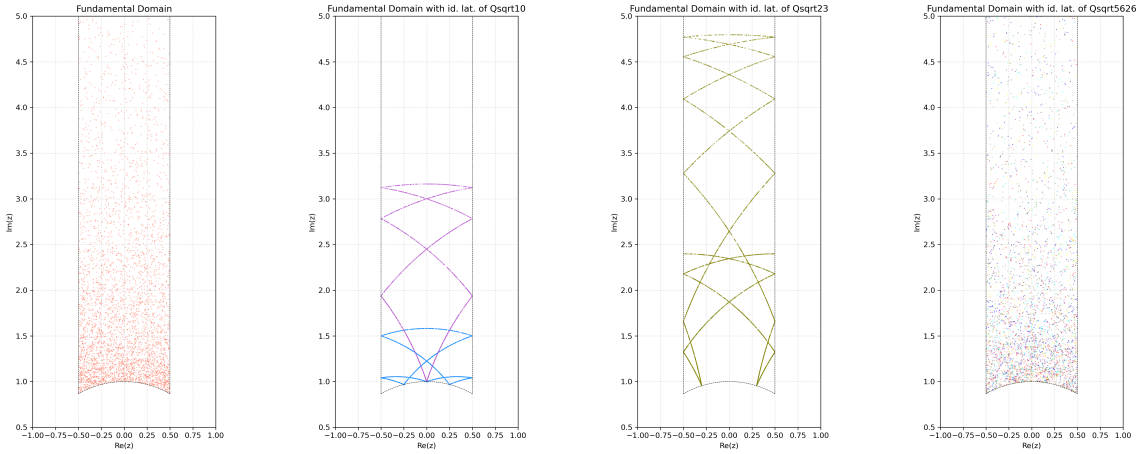
as it takes all values of c obtained by picking α between 1 and ε_0. We used that $\varepsilon_0/\varepsilon_0' = \varepsilon_0^2$ as it is a unit so $\varepsilon_0\varepsilon_0' = 1$.

The Haar measure on this circle is the hyperbolic measure, as it is invariant under the action of $\mathrm{PSL}_2(\mathbb{Z})$, and guarantees that all fundamental arcs have the same measure. The length of the arc $\mathcal{A}$ can be computed by integrating over $\mathcal{A}$ directly, but this is cumbersome and instead we first map $\mathcal{A}$ to a vertical line through the measure-preserving Möbius transform $\mu : z \mapsto \frac{z - \omega_2'/\omega_1'}{z - \omega_2/\omega_1}$ that maps $\mathcal{A}$ to $[i, i\varepsilon_0^2]$:

$$\int_{\mathcal{A}} ds = \int_i^{i\varepsilon_0^2} ds = \int_1^{\varepsilon_0^2} \frac{dy}{y} = 2\ln\varepsilon_0 = 2R_K.$$

This is not a surprise as $\tilde{\mathrm{Cl}}(K)$ can be seen as h_K copies of a circle of length R_K, where the extra factor 2 comes from the fact that ideal lattices that are symmetric across the vertical axis are considered the same because they are isometric.

We illustrate the behaviour of random two-dimensional ideal lattices in Figure 8.1. The first plot 8.1a corresponds to random real lattices. The second plot 8.1b corresponds to ideal lattices of $\mathbb{Q}(\sqrt{10})$, whose class number is 2. Visually, we can immediately see two circles, which correspond directly to what the theory predicts for the Arakelov class group. Notice that we only represent points of $\mathbb{H}$ that lie in the fundamental domain. An arc that exits the fundamental domain through the right boundary should be seen as entering again through the left boundary and vice versa. The situation is more complicated when the arc exits through the lower boundary, as it ends up being reflected via inversion $z \mapsto -1/\overline{z}$. The behaviour at



(a) $\mathbb{Q}$ (b) $\mathbb{Q}(\sqrt{10})$ (c) $\mathbb{Q}(\sqrt{23})$ (d) $\mathbb{Q}(\sqrt{5626})$

Figure 8.1: Comparing reduced random lattices and random ideal lattices in dimension 2, 5000 samples using the random walk strategy of [dBDPW20] for each field indicated. We only plot $\Im(z) \leq 5$. Different ideal classes are represented with different colours.

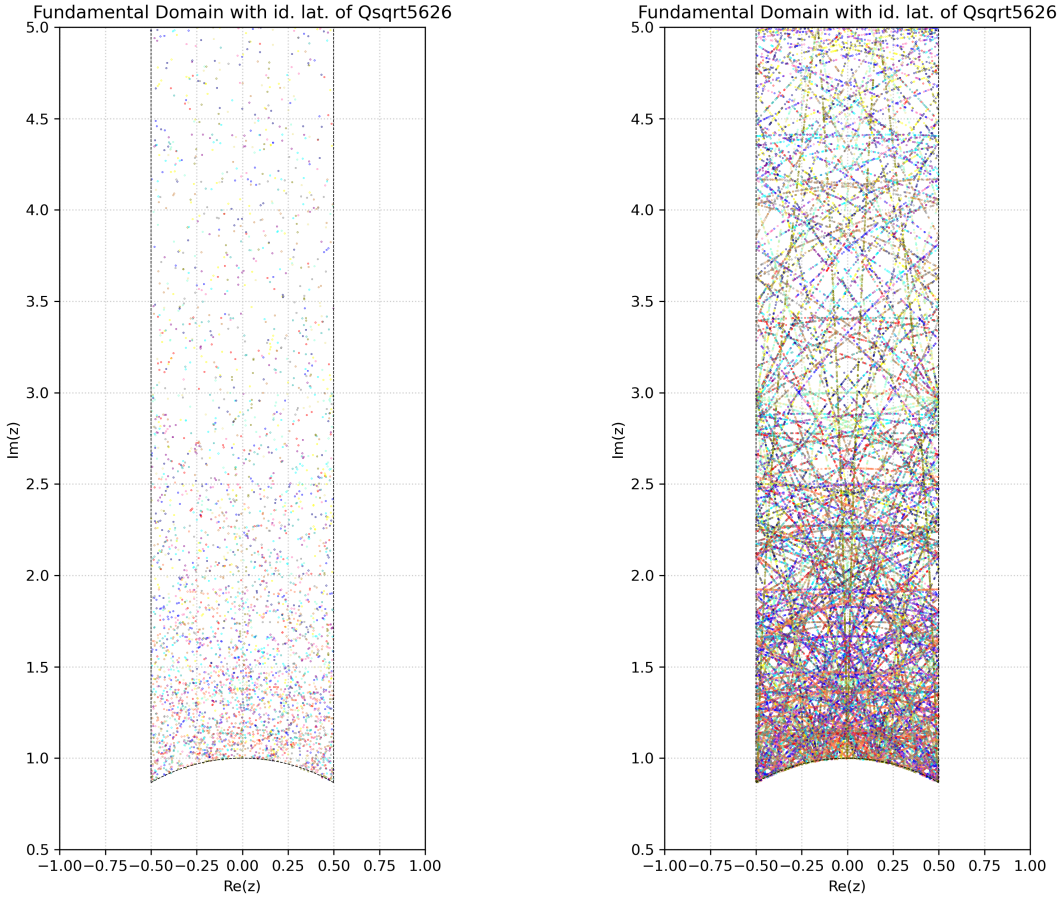
the intersections between boundaries is slightly more complicated, and requires both translation and inversion. The cycles obtained using this process are called closed geodesics. Such geodesics that trace out their image once are called primitive or prime, and share distribution laws with the prime numbers, see [Sar80]. Plot 8.1c shows an example where the class number is one: $\mathbb{Q}(\sqrt{23})$. Lastly, plot 8.1d shows what happens when the discriminant increases. This last field $\mathbb{Q}(\sqrt{5626})$ was specially chosen to have a large class number: 28. It illustrates a very deep result called Duke's theorem, of which a proof was given by Einsiedler, Lindenstrauss, Michel, and Venkatesh in [ELMV12]. This theorem states that as Δ_K grows to infinity, the distribution converges towards the distribution that one would expect from lattices without structure. Note that [ELMV12, Figure 2] plots exactly the same type of object as we do in Figure 8.1. While plot 8.1d might seem like it looks exactly the same as plot 8.1a, it remains quite structured, as can be seen in Figure 8.2 by increasing the number of samples. Equidistribution for ideal lattices is also known for cubic fields [ELMV11], but seems to remain open for higher degrees.

8.3.3 An Algorithm that Computes Moments of λ_1

We can compute statistics on the shortest vector for ideals lattices in a given class using the closed geodesics presented in the previous section. For this we let $f : \mathcal{F} \to \mathbb{R}$ be the function we want to average. For a given ideal lattice $x \in \mathrm{Cl}(K)$ we denote by $\mathcal{G}_x$ the projection of the fundamental arc $\mathcal{A}$ associated to x onto $\mathcal{F}$. Each point on $\mathcal{G}_x$ corresponds to a coset of the quotient of the norm 1 idèle class group (or Arakelov class group) $\mathbb{I}_K^1/K^\times$ by $\mathrm{Cl}(K)$. Because the action of Γ preserves measure, the correct measure on $\mathcal{G}_x$ from taking a uniform idèle class is the hyperbolic measure.

For each $x \in \mathrm{Cl}(K)$, we explain how to compute $\mathbb{E}_{\tau \in \mathcal{G}_x}(f(\tau))$. The first observation that follows from the previous section is the fact that $\mathcal{G}_x$ is obtained by projecting the arc $\mathcal{A}$ into the fundamental domain $\mathcal{F}$. This process is in fact equivalent to lattice or quadratic form reduction, and consists in applying elements of Γ to the arc $\mathcal{A}$ that send it to the closed geodesic $\mathcal{G}_x \subseteq \mathcal{F}$. Because the action of Γ preserves the measure and circle arcs, it follows that $\mathcal{G}_x$ can be decomposed into a finite chain of circle arcs $\mathcal{G}_{x,k}$. If N_x denotes the number of such arcs, we write

$$\mathcal{G}_x = \bigcup_{k \in \mathbb{Z}/N_x\mathbb{Z}} \mathcal{G}_{x,k}. \tag{8.3}$$



(a) 5000 samples

(b) 10^5 samples

Figure 8.2: Random ideal lattices in $\mathbb{Q}(\sqrt{5626})$, sampled using the random walk strategy of [dBDPW20]. We only plot $\Im(z) \leq 5$. Different ideal classes are represented with different colours.

For a given class $x \in \mathrm{Cl}(K)$, how can we compute the decomposition above? We use the following strategy:

1. Identify one of the arcs $\mathcal{G}_{x,i_0}$ by examinating reduced quadratic forms associated with x.
2. Iteratively compute $\mathcal{G}_{x,i+1}$ from $\mathcal{G}_{x,i}$ by following the arc $\mathcal{G}_{x,i}$ in a chosen direction until it hits the boundary of $\mathcal{F}$, and then identify the transformation of Γ required to keep the arc inside $\mathcal{F}$.
3. This algorithm will end when the new arc is the initial one.

For point (1.) we pick an ideal $\mathfrak{a}$ such that $[\mathfrak{a}] = x$, find a $\mathbb{Z}$-basis (ω_1, ω_2) of $\mathfrak{a}$, and (Lagrange) reduce it. This gives a new basis (ω_1', ω_2'), whose image in the upper half plane is in $\mathcal{F}$. Using Corollary 8.3.4, we can find the centre and radius of the circle. Intersecting it with $\mathcal{F}$ gives our initial arc $\mathcal{G}_{x,i_0}$.

Point (2.) is classical, and reminiscent of dynamical billiard problems where a particle bounces inside a closed table. For the purpose of the explanation, we follow an imaginary particle through the different arcs. Each arc has exactly two intersections with the boundary of $\mathcal{F}$. Except for the first step, one of the intersections corresponds to the previous rebound, so there is only one remaining intersection that makes sense. We use (R_i) and (z_i) to denote the sequences of the radii and centres of the arcs $\mathcal{G}_{x,i}$. The boundary of $\mathcal{F}$ can have three different behaviours.

- If this intersection is on the left boundary of $\mathcal{F}$, *i.e.* $x = -1/2$ and $x^2 + y^2 > 1$, then the particle *bounces* to the other (right) boundary, and the next arc will have the same radius $R_{i+1} = R_i$ and centre $z_{i+1} = z_i + 1$. The same happens when the particle hits the right boundary defined by $x = 1/2$ and $x^2 + y^2 > 1$, where $R_{i+1} = R_i$ and $z_{i+1} = z_i - 1$. This is justified by the fact that $z \mapsto z+1$ and $z \mapsto z-1$ are both elements of Γ, so in our setting the right and left boundaries are identified, the particle is in fact not bouncing at all!
- If the second intersection is on the lower boundary defined by $x^2 + y^2 = 1$ and $|x| < 1/2$, then the circle is transformed through inversion $z \mapsto -1/\overline{z}$, another element of Γ, and we have
$$\begin{cases} R_{i+1} &= \frac{R_i}{|z_i - R_i^2|} \\ z_{i+1} &= \frac{z_i}{z_i^2 - R^2} \end{cases}.$$
- The third case is slightly more delicate, and happens only when one of the arcs goes through one of the points of extra symmetry $e^{2i\pi/3}$ or $e^{2i\pi/6}$ (note that they are in fact the same point in $\mathbb{H}/\Gamma$). In this case the circle must undergo first inversion, then translation by ± 1 depending on the direction of incidence, and then a second inversion.

If the computation is carried out formally, or with sufficient precision, it is easy to then check if an arc is equal to the initial arc by comparing centres and radii, so this part of the algorithm terminates. This settles point (3.).

From this computation, it is now possible to compute averages of f over $\mathcal{G}_x$ by piecewise integration, as long as f is integrable on each arc:

$$\mathbb{E}_{\tau \in \mathcal{G}_x}(f(\tau)) = \frac{1}{N_{\mathcal{G}_x}} \sum_{k \in \mathbb{Z}/N_x\mathbb{Z}} \int_{\tau \in \mathcal{G}_{x,k}} f(\tau) \frac{dxdy}{y^2}, \tag{8.4}$$

where $N_{\mathcal{G}_x}$ is a normalisation factor. In fact this normalisation term is independent of x. Indeed it corresponds to integrating $f = 1$ over the same geodesic, and we have already seen previously that for functions of τ that do not depend on $\Im(\tau)$, the integral over different arcs is invariant by action of Γ, and so $N_{\mathcal{G}_x} = 2R_K$ for every class $x \in \mathrm{Cl}(K)$. This is a nice sanity check when implementing Equation 8.4 numerically.

Remark 8.3.5. The computation can also be carried out by restricting $\mathcal{F}$ to $\Re(\tau) \geq 0$, as the integrals can be grouped into pairs of equal value. The normalisation term would then only be the regulator.

As in the case of real lattices, for $k \in \mathbb{Z}_{\geq 0}$, we can obtain the k-th moment of λ_1 on normalised ideal lattices in a given class by plugging in $f(\tau) = \Im(\tau)^{-k/2}$ into Equation 8.4. The value $\Im(\tau)^{-1/2}$ comes from the fact that the lattice with basis $(1, \tau)$ has volume $\Im(\tau)$ (for example as a product of the corresponding Gram-Schmidt norms), and we rescale the lattice so that it has volume 1. The shortest non-zero vector used to be 1, it then becomes $\Im(\tau)^{-1/2}$.

We can remove the ideal class condition by simply averaging over all classes of $\mathrm{Cl}(K)$, by taking exactly one representative per ideal class.

Remark 8.3.6. We note that the function f used to compute the average λ_1 depends on $\Im(z)$, and therefore *unfolding* the geodesic to recover a circle will not agree with the integral. In some very special cases, for example $\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(\sqrt{13})$ or the principal ideal class of $\mathbb{Q}(\sqrt{10})$: we start at $\tau = i$ and there are no inversions. The other translations preserve the imaginary part, and therefore we can unfold the path to be a clean-cut single arc. This allows us to give very simple formulae for the expected value of λ_1^k as a single real integral in such cases. Likewise to the real case, such integrals can be computed if $k = 2$.

Our experimental results in Figure 8.3 confirm that as Δ_K increases, the expected behaviour of λ_1 over ideal lattices converges towards what we expect from a random real lattice.

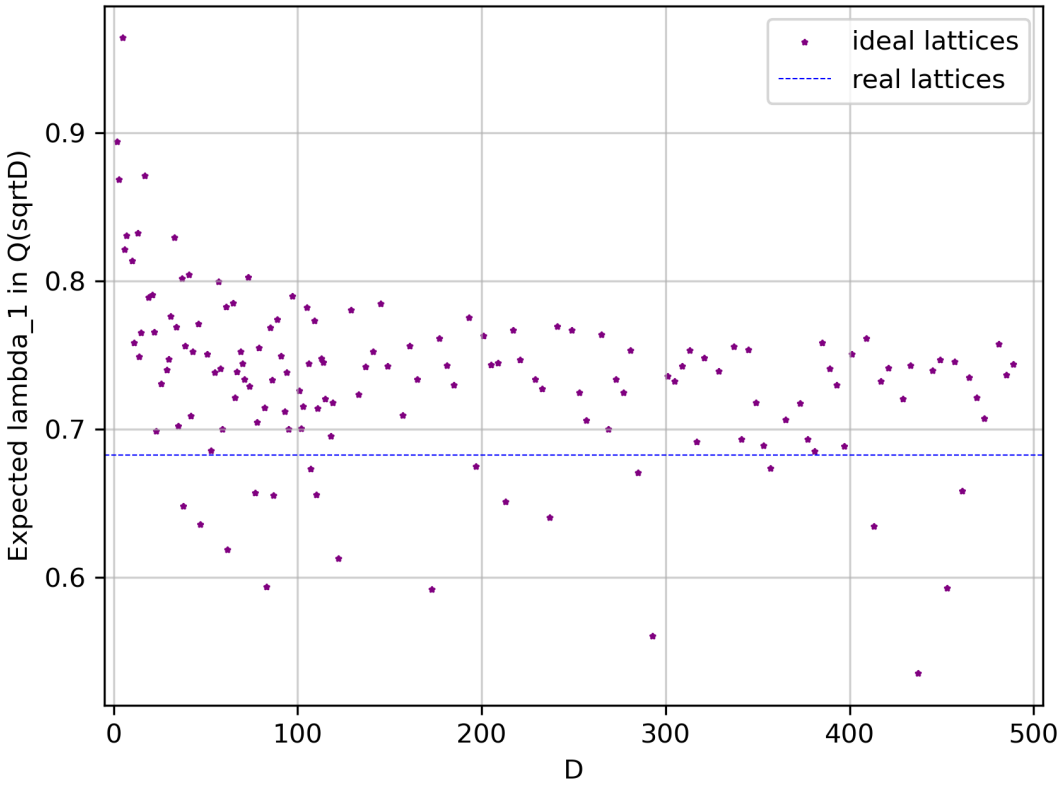


Figure 8.3: The expected value of λ_1 for random ideal lattices of the real quadratic field $\mathbb{Q}(\sqrt{D})$.

8.4 The Expected Number of Ideal Lattice Points in a Ball

The method that we gave in the previous section exploits properties of X_2. Generalising it might be possible but would at the very least require explicit descriptions of fundamental domains of moduli spaces in higher dimensions. This is possible, but very cumbersome as the dimension of the fundamental domain for X_n grows quadratically in n. Equidistribution results such as Duke's theorem are very deep and are not known for fields of degree larger than four. By analogy to the study of λ_1 for random real lattices, we focus on what should be an easier task: counting the average number of ideal lattice points in a ball of radius R. In this section we explain how one can reduce this problem to computing a finite sum of Euclidean integrals. This method requires both enumerating integral ideals of norm bounded by $\sqrt{|\Delta_K|}\cdot(R/\sqrt{n})^n$, and computing n-dimensional integrals. For both of these reasons our method leads to an exact formula that can only be computed in fields with both small discriminant and small degree.

8.4.1 The Siegel-Arakelov Formula

The following proposition links averaging the Siegel transform of f over ideal lattices to an integral over the norm one idèle class group $\mathbb{I}_K^1/K^\times$.

Proposition 8.4.1. *Let K be a number field of degree n, and let f be a real-valued Schwartz function on $K_\mathbb{R}$. Then the average over ideal lattices Λ normalised to have* $\mathrm{vol}(\Lambda) = \sqrt{|\Delta_K|}$ *of*

the Siegel transform $\sum_{\mathbf{v}\in\Lambda\setminus\{\mathbf{0}\}} f(\mathbf{v})$ *is equal to*

$$\frac{1}{\mu(\mathbb{I}_K^1/K^\times)}\int_{\mathbb{I}_K^1/K^\times}\sum_{\alpha\in K^\times} f(\alpha x)d\mu,$$

where f *is extended to* $\mathbb{A}_K$ *by fixing* $f_\nu = \mathbf{1}_{\mathcal{O}_\nu}$ *at every finite place* $\nu \nmid \infty$*, and* f_∞ *to be the previous* f*.*

Proof. Recall that the space of ideal lattices can be identified with the quotient $\mathbb{I}_K/K^\times$, or Arakelov class group. For an idèle $x \in \mathbb{I}_K^1$, the condition $\|x\|_K = 1$ imposes that the corresponding ideal lattice has volume $\sqrt{|\Delta_K|}$. For an ideal lattice Λ_x corresponding to an idèle $x \in \mathbb{I}_K^1$, the non-zero lattice points are exactly given by the set

$$\{\alpha x_\infty : \alpha \in K^\times, (\alpha x)_\nu \in \mathcal{O}_\nu \ \ \forall \nu \nmid \infty\},$$

which means that the extension of f to $\mathbb{A}_K$ ensures precisely that

$$\sum_{\mathbf{v}\in\Lambda_x\setminus\{\mathbf{0}\}} f(\mathbf{v}) = \sum_{\alpha\in K^\times} f(\alpha x).$$

The sum is well-defined over the quotient as replacing x by some γx for $\gamma \in K^\times$ simply permutes the terms in the sum. Averaging over the Haar measure μ inherited from the Tamagawa measure on $\mathbb{I}_K$ (see Definition 3.5.4) gives the desired result. □

The same reasoning can be applied to higher moments of the Siegel transform, for example in the case of the second moment, the correct average of the square of the Siegel transform of f over ideal lattices is

$$\frac{1}{\mu(\mathbb{I}_K^1/K^\times)}\int_{\mathbb{I}_K^1/K^\times}\left(\sum_{\alpha\in K^\times} f(\alpha x)\right)^2 d\mu,$$

where f is extended to $\mathbb{A}_K$ in the same way as in Proposition 8.4.1. Recall that $\mu(\mathbb{I}_K^1/K^\times)$ is equal to the residue of the Dedekind zeta function ζ_K at 1.

By unfolding, we obtain a first analogue of Siegel's formula for averaging over ideal lattices. It says that averaging the Siegel transform of f over ideal lattices is the same as integrating f over the space of norm one idèles.

Theorem 8.4.2 (Siegel-Arakelov formula)**.** *Let* K *be a number field of degree* n*, and let* f *be a real-valued Schwartz function on* $K_\mathbb{R}$*. Then*

$$\int_{\tilde{\mathrm{Cl}}(K)}\sum_{\mathbf{v}\in\Lambda_x\setminus\{\mathbf{0}\}} f(\mathbf{v})d\mu = \frac{1}{\mu(\mathbb{I}_K^1/K^\times)}\int_{\mathbb{I}_K^1} f(x)d\mu,$$

where on the right f *is extended to* $\mathbb{A}_K$ *by fixing* $f_\nu = \mathbf{1}_{\mathcal{O}_\nu}$ *at every finite place* $\nu \nmid \infty$*, and* f_∞ *to be the previous* f*.*

Proof. We call $I(f)$ the integral on the left. Letting $\mathcal{F} \subset \mathbb{I}_K^1$ be a fundamental domain for the

action of $K^\times$, by Proposition 8.4.1 we have

$$\begin{aligned}\mu(\mathbb{I}_K^1/K^\times)\cdot I(f) &= \int_{\mathbb{I}_K^1/K^\times}\sum_{\alpha\in K^\times} f(\alpha x)d\mu\\ &= \int_{\mathcal{F}}\sum_{\alpha\in K^\times} f(\alpha x)d\mu\\ &= \sum_{\alpha\in K^\times}\int_{\mathcal{F}} f(\alpha x)d\mu\\ &= \sum_{\alpha\in K^\times}\int_{\alpha\mathcal{F}} f(y)\|\alpha\|_K d\mu\\ &= \int_{\bigcup_{\alpha\in K^\times}\alpha\mathcal{F}} f(y)d\mu = \int_{\mathbb{I}_K^1} f(y)d\mu,\end{aligned}$$

where we used Fubini's theorem on the absolutely integrable function f, and then the change of variable $y=\alpha x$. Note that this change of variable changes the measure into $\|\alpha\|_K d\mu$, but $\|\alpha\|_K = 1$ by the product formula. □

We can use a similar reasoning for higher moments, to obtain some analogue of Rogers's formula. For the second moment we get the following result.

Theorem 8.4.3. *Let K be a number field of degree n, and let f be a real-valued Schwartz function on $K_\mathbb{R}$. Then*

$$\int_{\tilde{\mathrm{Cl}}(K)}\left(\sum_{\mathbf{v}\in\Lambda_x\setminus\{\mathbf{0}\}} f(\mathbf{v})\right)^2 d\mu = \frac{1}{\mu(\mathbb{I}_K^1/K^\times)}\sum_{\alpha\in K^\times}\int_{\mathbb{I}_K^1} f(x)f(\alpha x)d\mu,$$

where on the right f is extended to $\mathbb{A}_K$ by fixing $f_\nu = \mathbf{1}_{\mathcal{O}_\nu}$ at every finite place $\nu\nmid\infty$, and f_∞ to be the previous f.

Proof. The proof follows the same reasoning as the proof of Theorem 8.4.2. We want to compute

$$\frac{1}{\mu(\mathbb{I}_K^1/K^\times)}\int_{\mathbb{I}_K^1/K^\times}\left(\sum_{\alpha\in K^\times} f(\alpha x)\right)^2 d\mu,$$

and for this we compute

$$\begin{aligned}\int_{\mathbb{I}_K^1/K^\times}\left(\sum_{\alpha\in K^\times} f(\alpha x)\right)^2 d\mu &= \int_{\mathbb{I}_K^1/K^\times}\sum_{\alpha_1\in K^\times}\sum_{\alpha_2\in K^\times} f(\alpha_1 x)f(\alpha_2 x)d\mu\\ &= \int_{\mathbb{I}_K^1/K^\times}\sum_{\beta\in K^\times}\sum_{\gamma\in K^\times} f(\beta\gamma x)f(\gamma x)d\mu\\ &= \sum_{\beta\in K^\times}\int_{\mathbb{I}_K^1/K^\times}\sum_{\gamma\in K^\times} F_\beta(\gamma x)d\mu\\ &= \sum_{\beta\in K^\times}\int_{\mathbb{I}_K^1} F_\beta(x)d\mu = \sum_{\beta\in K^\times}\int_{\mathbb{I}_K^1} f(\beta x)f(x)d\mu,\end{aligned}$$

where we used the same unfolding trick from the proof of Theorem 8.4.2 on the uniformly integrable function $F_\beta(x) = f(\beta x)f(x)$. □

8.4.2 Application to Point-Counting

In this section we fix a radius $R \in \mathbb{R}_{>0}$, and take f to be $\mathbf{1}_{B_n(R)}$ on $K_\mathbb{R}$, and $\mathbf{1}_{\mathcal{O}_\nu}$ at finite places $\nu \nmid \infty$ when extended to $\mathbb{A}_K$.

Proposition 8.4.4. *Let K be a number field of degree n, $R \in \mathbb{R}_{>0}$ and f be the indicator function described above. Then*

$$\int_{\mathbb{I}_K^1} f(x)d\mu = \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \int_{N(x_\infty)=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{B_n(R)}(x_\infty)d\mu,$$

where x_∞ denotes the infinite part of the idèle x, the norm $\mathcal{N}$ is the algebraic norm of integrals, and the norm N in $K_\mathbb{R}$ is obtained by multiplying absolute values at all infinite places.

Proof. In the integral $\int_{\mathbb{I}_K^1} f(x)d\mu$, $f(x)$ is 1 only if the infinite part $x_\infty \in K_\mathbb{R}$ of x is in the ball $B_n(R)$, and its finite parts are integral $x_\nu \in \mathcal{O}_\nu$ for all $\nu \nmid \infty$. Therefore the integral can be seen as the μ-measure of the set of idèles that satisfy those conditions. We partition $\mathbb{I}_K^1$ according to the unique integral ideal $\phi(x) = \mathfrak{a}$ generated by the finite part of x. This ideal is indeed integral, as saying $x_\nu \in \mathcal{O}_\nu$ is the same as saying that the ν-adic valuation of the (a priori fractional) ideal $\phi(x)$ is non-negative. Therefore

$$\int_{\mathbb{I}_K^1} f(x)d\mu = \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \mu\left(\{x \in \mathbb{I}_K^1 : x_\infty \in B_n(R) \text{ and } \phi(x) = \mathfrak{a}\}\right).$$

The norm of the finite part of the idèle x corresponding to the integral ideal $\mathfrak{a}$ is exactly $\mathcal{N}(\mathfrak{a})^{-1}$, so with the normalisation that we chose (see the Remark after Definition 3.5.4), the integral over all idèles with $x_\nu \in \mathcal{O}_\nu$ for all $\nu \nmid \infty$, and that map to the same ideal $\mathfrak{a}$ is exactly 1. Additionally, if x_∞ denotes the infinite part of x, the norm 1 condition of x translates as $N(x_\infty)\cdot\mathcal{N}(\mathfrak{a})^{-1} = 1$. Therefore,

$$\int_{\mathbb{I}_K^1} f(x)d\mu = \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \int_{N(x_\infty)=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{B_n(R)}(x_\infty)d\mu,$$

and this completes the proof. □

While this sum might seem much more complicated than the integral, it has become simply a sum of integrals in $K_\mathbb{R}$ over a finite number of terms, which is simpler to deal with, and makes it computable if the integral can be expressed in a nice way, and ideals of $\mathcal{O}_K$ enumerated under a certain bound. Indeed, for the integration domain to be non-zero above, the hyperbola $\|x_\infty\| = \mathcal{N}(\mathfrak{a})$ must intersect $B_n(R)$. This fails as soon as $\mathcal{N}(\mathfrak{a}) > (R/\sqrt{n})^n$ by AM-GM.

Proposition 8.4.4 can also be generalised to the second moment, but the sum becomes more complicated (although it can be shown to remain finite). In what follows we only focus on the first moment.

Application to imaginary quadratic fields

We start with the least interesting case and specialise Proposition 8.4.4 to K an imaginary quadratic field. The main challenge is computing the integral for a given field K, as the measure μ inherited from the Tamagawa measure does not directly agree with the Euclidean measure.

Proposition 8.4.5. *Let $K = \mathbb{Q}(\sqrt{-D})$ with $D \in \mathbb{Z}_{>0}$ squarefree be an imaginary quadratic field, and $R \in \mathbb{R}_{>0}$ then*

$$\int_{\tilde{\mathrm{Cl}}(K)} \#\left(\Lambda_x\backslash\{\mathbf{0}\} \cap B_2\left(R\right)\right) d\mu = \frac{w_K}{h_K} \cdot \sqrt{|\Delta_K|} \cdot I_{R^2},$$

where w_K is the number of roots of unity in K, h_K is the class number, and I_{R^2} counts the number of non-zero integral ideals of $\mathcal{O}_K$ of norm less than or equal to R^2.

Proof. We use Theorem 8.4.2 and Proposition 8.4.4 with f defined as the indicator of a ball of radius R. So we have to compute

$$\frac{1}{\mu(\mathbb{I}_K^1/K^\times)} \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \int_{N(x_\infty)=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{B_2(R)}(x_\infty) d\mu.$$

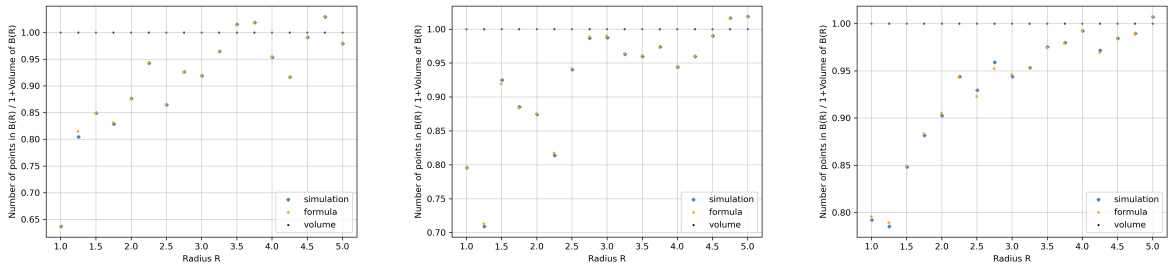
K has are no real embeddings and a single complex embedding, therefore $K_\mathbb{R} \cong \mathbb{C}$ and for $z \in K_\mathbb{R}$ we have $N(z) = |z|^2$. Therefore for a given $\mathfrak{a} \in \mathcal{O}_K$ we must compute the integral over all $z \in \mathbb{C}$ such that $|z| = \sqrt{\mathcal{N}(\mathfrak{a})}$, for which $\mathcal{N}(\mathfrak{a}) \leq R^2$. The measure $d\mu$ is derived from the Tamagawa measure $d^\times z$, which is twice the Lebesgue measure on $\mathbb{C}^\times$ divided by $|z|^2$, which in polar coordinates can be written $\frac{2rdrd\theta}{r^2}$. We change coordinates to (β, θ), where $\beta = r^2$ is defined via the norm map N. Now we compute the 2×2 Jacobian to get $drd\theta = \frac{d\beta d\theta}{2\sqrt{\beta}}$, which in turn from the decomposition $d^\times z = d\mu \cdot d\beta/\beta$ leads to $d\mu = d\theta$. So the integral is 2π when $\mathcal{N}(\mathfrak{a}) \leq R^2$, and 0 otherwise. Because for imaginary quadratic fields we get

$$\mu(\mathbb{I}_K^1/K^\times) = \frac{2\pi h_K}{w_K\sqrt{|\Delta_K|}},$$

the result follows. □

Remark 8.4.6. The normalisation chosen on the left in Proposition 8.4.5 is naturally chosen such that $\mathrm{vol}(\Lambda_x) = \sqrt{|\Delta_K|}$. To obtain a point-counting formula for ideal lattices of volume 1, we must remove the $\sqrt{|\Delta_K|}$ term on the right, and appropriately rescale R.

We validate Proposition 8.4.5 experimentally in Figure 8.4 by sampling many random ideal lattices of imaginary quadratic fields K, and enumerating their non-zero points in an origin-centred ball. Because the Gaussian heuristic must hold for large values of R, we focus on small values of R, which are more relevant to estimating the expected value of λ_1.



(a) $\mathbb{Q}(\sqrt{-5})$ (b) $\mathbb{Q}(\sqrt{-17})$ (c) $\mathbb{Q}(\sqrt{-161})$

Figure 8.4: Comparing the (simulated) enumerated average number of lattice points in balls of varying radii for 1000 random ideal lattices in different imaginary quadratic number fields with our exact formula.

Application to real quadratic fields

The case of real quadratic fields $K = \mathbb{Q}(\sqrt{D})$ is more interesting, as ideal lattices are not h_K isolated points, but have the structure of h_K circles.

Proposition 8.4.7. *Let $K = \mathbb{Q}(\sqrt{D})$ with $D \in \mathbb{Z}_{>0}$ squarefree be an real quadratic field, and $R \in \mathbb{R}_{>0}$ then*

$$\int_{\tilde{\mathrm{Cl}}(K)} \#\left(\Lambda_x \backslash \{\mathbf{0}\} \cap B_2(R)\right) d\mu = \frac{4}{\mathrm{Res}_{s=1}\,\zeta_K(s)} \sum_{\substack{\mathfrak{a}\subseteq\mathcal{O}_K \\ \mathcal{N}(\mathfrak{a})<R^2/2}} \sqrt{\mathcal{N}(\mathfrak{a})}\,\mathrm{arccosh}\left(\frac{R^2}{2\mathcal{N}(\mathfrak{a})}\right).$$

Proof. We again use Theorem 8.4.2 and Proposition 8.4.4 with f defined as the indicator of a ball of radius R. So we have to compute

$$\frac{1}{\mu(\mathbb{I}_K^1/K^\times)} \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \int_{N(x_\infty)=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{B_2(R)}(x_\infty) d\mu.$$

Here $K_\mathbb{R} \cong \mathbb{R}^2$ as we have two real embeddings and no complex embeddings. So an element x_∞ is a pair $(x, y) \in \mathbb{R}^2$, and the norm is $N(x_\infty) = |xy|$, so the integrals are over the four branches of the hyperbolae defined by $|xy| = \mathcal{N}(\mathfrak{a})$, intersected with the disc $B_2(R)$. By noticing that the points on such hyperbolae that are closest to the origin have $|x| = |y|$, we easily get that the intersection between hyperbola and ball is empty if and only if $\mathcal{N}(\mathfrak{a}) > R^2/2$ (although the integral is still 0 for $\mathcal{N}(\mathfrak{a}) = R^2/2$). Again we must find the correct measure $d\mu$. The Tamagawa measure on $(\mathbb{R}^\times)^2$ is given by $\frac{dx}{|x|} \cdot \frac{dy}{|y|}$. A change of variable $\alpha = (|x|/|y|)^{1/2}$ and $\beta = |xy|$ gives $\frac{dxdy}{|xy|} = \frac{d\alpha d\beta}{|\alpha\beta|}$, which then allows us to conclude that the correct measure is $d\mu = \alpha^{-1}d\alpha$. Therefore

$$\int_{\tilde{\mathrm{Cl}}(K)} \#\left(\Lambda_x\backslash\{\mathbf{0}\} \cap B_2\left(R\right)\right) d\mu = \frac{1}{\mu(\mathbb{I}_K^1/K^\times)} \sum_{\substack{\mathfrak{a}\subseteq\mathcal{O}_K \\ \mathcal{N}(\mathfrak{a})<R^2/2}} \int_{|xy|=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{x^2+y^2\leq R^2} \frac{d\alpha}{\alpha}.$$

For every $\mathfrak{a} \subseteq \mathcal{O}_K$ of norm bounded by $R^2/2$ we have

$$\begin{aligned}\int_{|xy|=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{x^2+y^2\leq R^2} \frac{d\alpha}{\alpha} &= 4 \int_{\substack{xy=\mathcal{N}(\mathfrak{a}) \\ x,y>0}} \mathbf{1}_{x^2+y^2\leq R^2} \frac{d\alpha}{(x/y)^{1/2}} \\ &= 4 \int_{x_{\min}}^{x_{\max}} \mathbf{1}_{x^2+(\mathcal{N}(\mathfrak{a})/x)^2\leq R^2} \sqrt{\mathcal{N}(\mathfrak{a})} \frac{dx}{x},\end{aligned}$$

where $x_{\min}$ and $x_{\max}$ are the ordered positive roots of $x^2 + (\mathcal{N}(\mathfrak{a})/x)^2 = R^2$. Solving the quadratic equation gives

$$x_{\min} = \sqrt{\frac{R^2 - \sqrt{R^4 - 4\mathcal{N}(\mathfrak{a})}}{2}} \text{ and } x_{\max} = \sqrt{\frac{R^2 + \sqrt{R^4 - 4\mathcal{N}(\mathfrak{a})}}{2}},$$

which in turn allows us to compute

$$\int_{x_{\min}}^{x_{\max}} \mathbf{1}_{x^2+(\mathcal{N}(\mathfrak{a})/x)^2\leq R^2} \sqrt{\mathcal{N}(\mathfrak{a})} \frac{dx}{x} = \sqrt{\mathcal{N}(\mathfrak{a})} \operatorname{arccosh}\left(\frac{R^2}{2\mathcal{N}(\mathfrak{a})}\right),$$

which concludes, after noticing that $\mu(\mathbb{I}_K^1/K^\times) = \operatorname{Res}_{s=1} \zeta_K(s)$. □

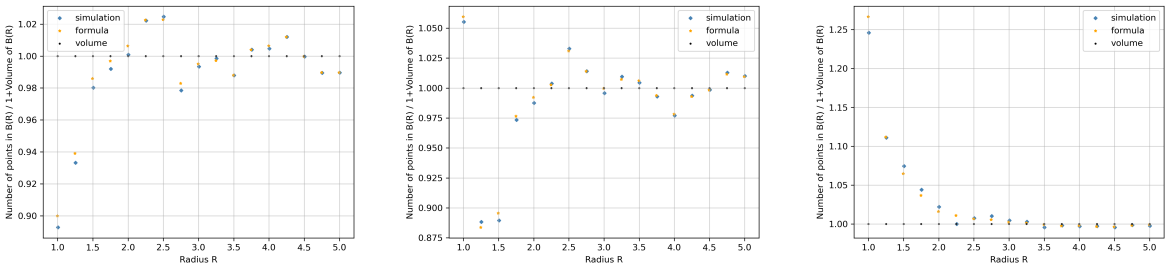
We validate Proposition 8.4.7 experimentally in Figure 8.5 by sampling many random ideal lattices of real quadratic fields K, and enumerating their non-zero points in an origin-centred ball. Because the Gaussian heuristic must hold for large values of R, we focus on small values of R, which are more relevant to estimating the expected value of λ_1.

Application to cyclotomic fields

Last but not least, we focus on fields that are relevant to cryptology: cyclotomic fields.

Proposition 8.4.8. *Let $K = \mathbb{Q}(\zeta)$ be a cyclotomic field of degree n. If $R \in \mathbb{R}_{>0}$ then*

$$\int_{\tilde{\mathrm{Cl}}(K)} \#\left(\Lambda_x\backslash\{\mathbf{0}\} \cap B_n\left(R\right)\right) d\mu = \frac{(2\pi)^{n/2}}{\operatorname{Res}_{s=1} \zeta_K(s)} \sum_{\substack{\mathfrak{a}\subseteq\mathcal{O}_K \\ \mathcal{N}(\mathfrak{a})<(R/\sqrt{n})^n}} f_\infty(\mathcal{N}(\mathfrak{a})),$$



(a) $\mathbb{Q}(\sqrt{10})$ (b) $\mathbb{Q}(\sqrt{23})$ (c) $\mathbb{Q}(\sqrt{5626})$

Figure 8.5: Comparing the (simulated) enumerated average number of lattice points in balls of varying radii for 1000 random ideal lattices in different real quadratic number fields with our exact formula.

where

$$f_\infty(r) = \int_{\gamma_i>0} \mathbf{1}_{\left(\prod_{i\geq 2}\gamma_i^{-1}+\sum_{i\geq 2}\gamma_i \leq \frac{R^2}{2r^{2/n}}\right)} \prod_{i\geq 2}\frac{d\gamma_i}{\gamma_i},$$

for $r > 0$, where sums and products are indexed from 2 to $n/2$.

Proof. Again, as previously, we apply Theorem 8.4.2 and Proposition 8.4.4 with f defined as the indicator of a ball of radius R. Because K has only $n/2$ pairs of complex embeddings (note that n has to be even), $K_\mathbb{R} \cong \mathbb{C}^{n/2}$, and the norm on infinite parts is given by

$$N(x_\infty) = \prod_{i=1}^{n/2} |\sigma_i(x_\infty)|^2,$$

where the σ_i are the complex embeddings. The measure $d\mu$ is obtained by writing

$$\prod_{i=1}^{n/2} \frac{2dz_i}{|z_i|^2} = d\mu \cdot \frac{d\beta}{\beta},$$

where the left hand side corresponds to the Tamagawa measure on $(\mathbb{C}^\times)^{n/2}$ and β correspond to the norm N on infinite parts. This can be handled as in the imaginary quadratic case through polar coordinates by noticing that for $z_i = \sqrt{\beta_i}e^{i\theta_i}$ we get

$$\prod_{i=1}^{n/2} \frac{2dz_i}{|z_i|^2} = \prod_{i=1}^{n/2} d\theta_i \frac{d\beta_i}{\beta_i}.$$

We make the change of variable

$$\gamma_1 = \prod_{i=1}^{n/2} \beta_i, \gamma_i = \beta_i \gamma_1^{-2/n} \text{ for } i > 1$$

and compute the Jacobian

$$\left(\frac{\partial \beta_i}{\partial \gamma_j}\right) = \begin{pmatrix} \frac{2}{n}\beta_1\gamma_1^{-1} & -\beta_1\gamma_2^{-1} & \cdots & -\beta_1\gamma_{n/2}^{-1} \\ \frac{2}{n}\gamma_1^{\frac{2}{n}-1}\gamma_2 & \gamma_1^{\frac{2}{n}} & & \\ \vdots & & \ddots & \\ \frac{2}{n}\gamma_1^{\frac{2}{n}-1}\gamma_2 & & & \gamma_1^{\frac{2}{n}} \end{pmatrix}.$$

Its discriminant is $\gamma_2^{-1} \cdots \gamma_{n/2}^{-1}$. Thus

$$\prod_{i=1}^{n/2} \frac{d\beta_i}{\beta_i} = \prod_{i=1}^{n/2} \frac{d\gamma_i}{\gamma_i},$$

and from $d\gamma_1/\gamma_1 = d\beta/\beta$ we get

$$d\mu = \prod_{i=1}^{n/2} d\theta_i \prod_{i=2}^{n/2} \frac{d\gamma_i}{\gamma_i}.$$

We can now compute the integral for each $\mathfrak{a} \subseteq \mathcal{O}_K$:

$$\begin{aligned}\int_{N(x_\infty)=\mathcal{N}(\mathfrak{a})} \mathbf{1}_{x_\infty \in B_n(R)} d\mu &= (2\pi)^{n/2} \int_{\gamma_i>0} \mathbf{1}_{(\sum_i \beta_i \le R^2/2)} \prod_{i\ge 2} \frac{d\gamma_i}{\gamma_i} \\ &= (2\pi)^{n/2} \int_{\gamma_i>0} \mathbf{1}_{\left(\prod_{i\ge 2} \gamma_i^{-1} + \sum_{i\ge 2} \gamma_i \le \frac{R^2}{2\mathcal{N}(\mathfrak{a})^{2/n}}\right)} \prod_{i\ge 2} \frac{d\gamma_i}{\gamma_i},\end{aligned}$$

where at the end we have used that $\beta_i = \gamma_i \cdot \mathcal{N}(\mathfrak{a})^{2/n}$ for $i > 1$. □

The integrals of Proposition 8.4.8 measure the volume of the intersection of a high-dimensional hyperbola and a ball, and are increasingly difficult to evaluate as the degree of the number field increases. Luckily in small dimensions, this integral is computable.

Proposition 8.4.9. *Let $K = \mathbb{Q}(\zeta_5)$. If $R \in \mathbb{R}_{>0}$ then*

$$\int_{\tilde{\mathrm{Cl}}(K)} \#\left(\Lambda_x \backslash \{\mathbf{0}\} \cap B_4(R)\right) d\mu = \frac{(2\pi)^2}{\mathrm{Res}_{s=1}\, \zeta_K(s)} \sum_{\substack{\mathfrak{a} \subseteq \mathcal{O}_K \\ \mathcal{N}(\mathfrak{a}) < R^4/4}} 2 \operatorname{arccosh}\left(\frac{R^2}{4\sqrt{\mathcal{N}(\mathfrak{a})}}\right).$$

Proof. We simply apply Proposition 8.4.8 to $n = 4$. For each $r = \mathcal{N}(\mathfrak{a})$, we have

$$\begin{aligned} f_\infty(r) &= \int_{\gamma>0} \mathbf{1}_{\left(\gamma^{-1} + \gamma \le \frac{R^2}{2r^{1/2}}\right)} \frac{d\gamma}{\gamma} \\ &= 2 \operatorname{arccosh}\left(\frac{R^2}{4\sqrt{r}}\right), \end{aligned}$$

where we used the same integration as for the real quadratic case. □

We validate Proposition 8.4.9 experimentally in Figure 8.6.

8.4.3 The Formula of Gargava and Viazovska

At the same time as we were exploring the question of counting the average number of ideal lattice points in a ball, motivated by understanding the geometry of ideal lattices in the context of cryptography, Gargava and Viazovska [GV24] worked on exactly the same question, but motivated by a different goal: improving lower bounds for the sphere packing problem in high dimension. Their main result for cyclotomic fields is the following:

Theorem 8.4.10 (Theorem 1 of [GV24])**.** *Let K be a cyclotomic number field. Let $B \subseteq K_{\mathbb{R}}$ be the ball of volume V. Then as $\deg K \to \infty$, we have for some $\eta > 0$*

$$\mathbb{E}_{\text{Unit covolume } \Lambda \in \tilde{\mathrm{Cl}}(K)}(\#B \cap \Lambda \backslash \{\mathbf{0}\}) = V + \varepsilon(V, K),$$

where the error term $\varepsilon(V, K)$ behaves as $\deg K \to \infty$ like

$$|\varepsilon(V, K)| \ll \sqrt{V} |\Delta_K|^{-\eta}.$$

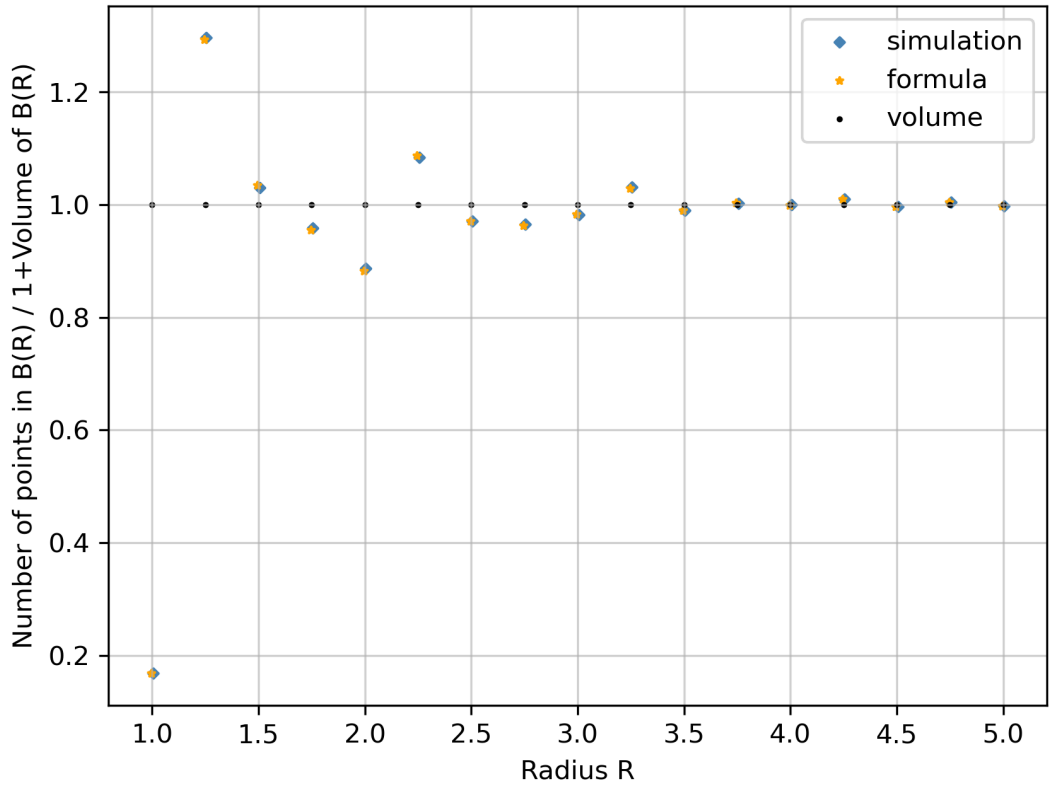


Figure 8.6: Comparing the (simulated) enumerated average number of lattice points in balls of varying radii for 1000 random ideal lattices in $\mathbb{Q}(\zeta_5)$ with our exact formula.

Up to normalisation of the ideals, Theorem 8.4.10 computes the same quantity as Proposition 8.4.8, but acts as a nicer variant of Theorem 8.2.3, because the main volume term directly appears in the expression, which says that as the degree of the cyclotomic field increases, the expected number of ideal lattice points gets closer and closer to the expected number of lattice points for random real lattices, confirming the Gaussian heuristic for large dimensional cyclotomic ideal lattices. The proof of Theorem 8.4.10 starts as our proof of the Siegel-Arakelov formula, but then uses more involved ideas similar to those used in the proof of Duke's theorem for cubic fields [ELMV11]. The key idea is to use a formula of Hecke to relate the average over the space of all ideal lattices of an Epstein zeta function to the Dedekind zeta function of K. This relates back to an average of the point counting function through an inverse Mellin transform, via a contour shifting argument.

Because cryptographically relevant fields use dimensions that are large concrete values, it is tempting to model the behaviour using the asymptotic regime, but because of possible hidden constants (or for other reasons), we might want to evaluate the error term of Theorem 8.4.10 precisely. In order to do this we need the following statement of [GV24].

Theorem 8.4.11 (Corollary 21 of [GV24])**.** *Let K be a cyclotomic number field of degree $n \geq 8$. Let $R \in \mathbb{R}_{>0}$. Then*

$$\mathbb{E}_{\text{Unit covolume } \Lambda \in \tilde{\mathrm{Cl}}(K)}(\#\Lambda \backslash \{\mathbf{0}\} \cap B_n(R)) = \mathrm{vol}(B_n(R)) + \varepsilon(R, K),$$

where the error term is

$$\varepsilon(R,K) = \frac{(2\pi)^{n/2}}{\pi i\sqrt{|\Delta_K|}\cdot \operatorname{Res}_{s=1}\zeta_K(s)} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} |\Delta_K|^{s/2}\zeta_K(s)\frac{R^{ns}}{ns}\left(\frac{\Gamma(s)^{n/2}}{2^{ns/2}\Gamma(ns/2)}\right) ds.$$

While all the terms are computable as long as $|s|$ is not too large, the integral oscillates very fast, making the computation quite expensive. A clever change of variable could help minimising the number of terms one would need to compute, for example by following the Double Exponential method of [TM73], later refined for precise integration of L-functions by Molin in [Mol10]. Unfortunately the integrand does not decrease fast enough along the vertical line to apply the $\sinh(\sinh(\cdot))$ change of variable. By shifting the integration path and adapting the change of variable, some progress might be made towards an explicit result.

Because of technical constraints related to rapid decay of the inverse Mellin transform in the proof of Theorem 8.4.11, the formula for the error term is only valid for $n \geq 8$, which means our efforts in the previous section were not in vain.

Conclusions

While Theorem 8.4.2 provides a variant of Siegel's mean formula for ideal lattices, it is challenging to compute the integral over the norm one idèles. For the problem of counting points in a ball, our method gives an exact formula that can be computed by computing Euclidean integrals in $\deg K$ dimensions, as well as enumerating integral ideals of bounded norm. The cost of both tasks explodes as the degree and discriminant of the number field grow. The formula of [GV24] is much better suited for asymptotic results as $\deg K$ goes to infinity for cyclotomic fields. However precisely computing the error term of Theorem 8.4.11 remains a challenging problem, and only works for cyclotomic fields of large enough degree.

Regarding extensions of both formulae to the second moment, an important step in bounding the average value of λ_1 for ideal lattices, we can say the following:

- We have seen that the Siegel-Arakelov formula generalises to the second moment, in a manner that is similar to Roger's formula. With enough work, it should allow for exact computations on fields of small degree and discriminant, however the computations will be more expensive as for the first moment.
- The next step is to get asymptotic results on the second moment through similar techniques as those used in [GV24] is to relate the integral from the previous point to the study of a $\mathrm{GL}_2(\mathbb{A}_K)$ Eisenstein series, which promises to be a more difficult problem involving the theory of automorphic forms.

Chapter

9 On the Ordinary Isogeny Graph

Abstract In this final chapter, we take a break from lattice-based cryptography and investigate another structure that is not unrelated to cryptography: the isogeny graph. We restrict our study to prime fields $\mathbb{F}_p$ and ordinary (as opposed to supersingular) elliptic curves. We give a detailed presentation of ℓ-isogeny graphs associated with ordinary elliptic curves defined over $\mathbb{F}_p$. We then focus on the following inverse problem: given an abstract (graph-theoretic) volcano V, can one always find primes $\ell, p \in \mathbb{N}$ such that the ordinary ℓ-isogeny graph over $\mathbb{F}_p$ contains V as a connected component? We provide an answer to both this and a stronger version of the question, where all the vertices of the crater of the abstract volcano are required to correspond to elliptic curves with CM by a maximal order $\mathcal{O}$. Interestingly, viewed as elliptic curves over $\mathbb{C}$, such elliptic curves are in fact lattices with $\mathcal{O}$-module structure. To this aim, we generalise a classical theorem of Yamamoto on orders of prime ideals in the class group of imaginary quadratic fields.

Most of this chapter is based on the journal paper [BCP24].

Chapter content

9.1 Introduction

Given a finite field $\mathbb{F}$ of characteristic p and a prime number $\ell \neq p$, one can consider the so-called *ℓ-isogeny graph over* $\mathbb{F}$. Roughly speaking, one can construct this graph $\mathcal{G}$ by taking as vertices the elements of $\mathbb{F}$, which in this context are seen as the j-invariants of elliptic curves defined over $\mathbb{F}$, and by drawing an oriented edge from one vertex to another if there is a geometric isogeny of degree ℓ between the corresponding elliptic curves. The graph $\mathcal{G}$ naturally decomposes into two subgraphs $\mathcal{G}_{\text{ord}}$ and $\mathcal{G}_{\text{ss}}$, which are respectively induced by the elements of $\mathbb{F}$ that are j-invariants of ordinary and supersingular elliptic curves. The structure of these two subgraphs is very different.

The supersingular isogeny graph $\mathcal{G}_{\text{ss}}$ has been studied in [Piz90] as an example of large Ramanujan graph. Due to their complicated structure, supersingular isogeny graphs have found several cryptographic applications, see for instance [CGL09; DJP14] and more recently the signature scheme SQIsign [Aar+25]. On the other hand, the ordinary isogeny graph $\mathcal{G}_{\text{ord}}$ has a much more regular structure which has been studied in the seminal work of Kohel [Koh96] and subsequently by Fouquet and Morain [FM02]. In particular, in [FM02] the authors coined the term *isogeny volcano* to denote the connected components of $\mathcal{G}_{\text{ord}}$. This terminology is justified by the fact that the connected components of ordinary isogeny graphs often appear as a cycle, the *crater*, whose vertices are roots of isomorphic trees, the *lava flows*. Because of this satisfyingly regular structure, ordinary isogeny graphs have attracted a lot of attention, finding applications both in computational number theory (see for instance [Sut12; Sut13]) and in cryptography (see for instance [MSTTV07; IJ10]).

We focus here mainly on ordinary isogeny graphs over $\mathbb{F} = \mathbb{F}_p$ and we start by a review of known results on the structure of $\mathcal{G}_{\text{ord}}$, which we rename $\mathcal{G}_\ell(\mathbb{F}_p)$ from now on. More precisely, we gather in one place the terminology that appeared in different works and provide detailed proofs when needed.

Within what we call the *volcano park*, we thus identify the *cordilleras* in Section 9.2.1, the *belts* in Section 9.2.2, and finally the *volcanoes* in Section 9.2.3. We also treat in full details the pathological cases corresponding to the j-invariants 0 and 1728 (see for instance Proposition 9.2.19 and Proposition 9.2.20). The geological lexicon we use is sometimes new and sometimes borrowed from previous works; for instance, the term “cordillera" first appeared in [MSTTV07]. We give an illustration of such a *volcano park* below, in Figure 9.1.

The connected components of $\mathcal{G}_\ell(\mathbb{F}_p)$ seem to mostly be isolated vertices or pairs of vertices connected with a single edge. However other special structures appear, in the form of isomorphic trees glued along a central ring. In fact, all components have volcano-shape (a notion that is made precise in Definition 9.3.1, such a graph will be called an abstract volcano). Now one may wonder to what extent a graph with the shape of a volcano can be realised as such a connected component. More precisely, we ask the following question:

Question 9.1.1 (Inverse Volcano Problem over $\mathbb{F}_p$)**.** *Let V be an abstract volcano as in Definition 9.3.1. Can we find primes $p, \ell \in \mathbb{Z}$ with $p \neq \ell$ such that V is a connected component in the isogeny graph $\mathcal{G}_\ell(\mathbb{F}_p)$?*

We provide an affirmative and explicit answer to Question 9.1.1. We can even characterise the cases when there exists a solution where the elliptic curves associated with the crater of V have complex multiplication by a maximal order. If the target abstract volcano is given only as a crater with no lava flows, the answer to Question 9.1.1 is easier since we have a lot of freedom in the choice of ℓ and p. We prove the following result (see Section 9.2 and again Definition 9.3.1 for terminology):

Theorem 9.1.2. *Let V be an abstract volcano of depth 0. Then there exist infinitely many distinct primes $p, \ell \in \mathbb{Z}$ such that V is a connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$ and the vertices on V correspond to elliptic curves with complex multiplication by a maximal order.*

On the other hand, if V has lava flows then the inverse volcano problem becomes more difficult, since the theory of isogeny volcanoes now fixes ℓ uniquely (we speak of *ℓ-volcano* in this case) and we thus only have freedom on the choice of p. We prove nonetheless:

Theorem 9.1.3. *Let $\ell \in \mathbb{Z}$ be a prime number and let V be an abstract ℓ-volcano of depth $d > 0$. Then the following holds:*

1. *There exist infinitely many primes $p \in \mathbb{Z}$ such that V is a connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$.*

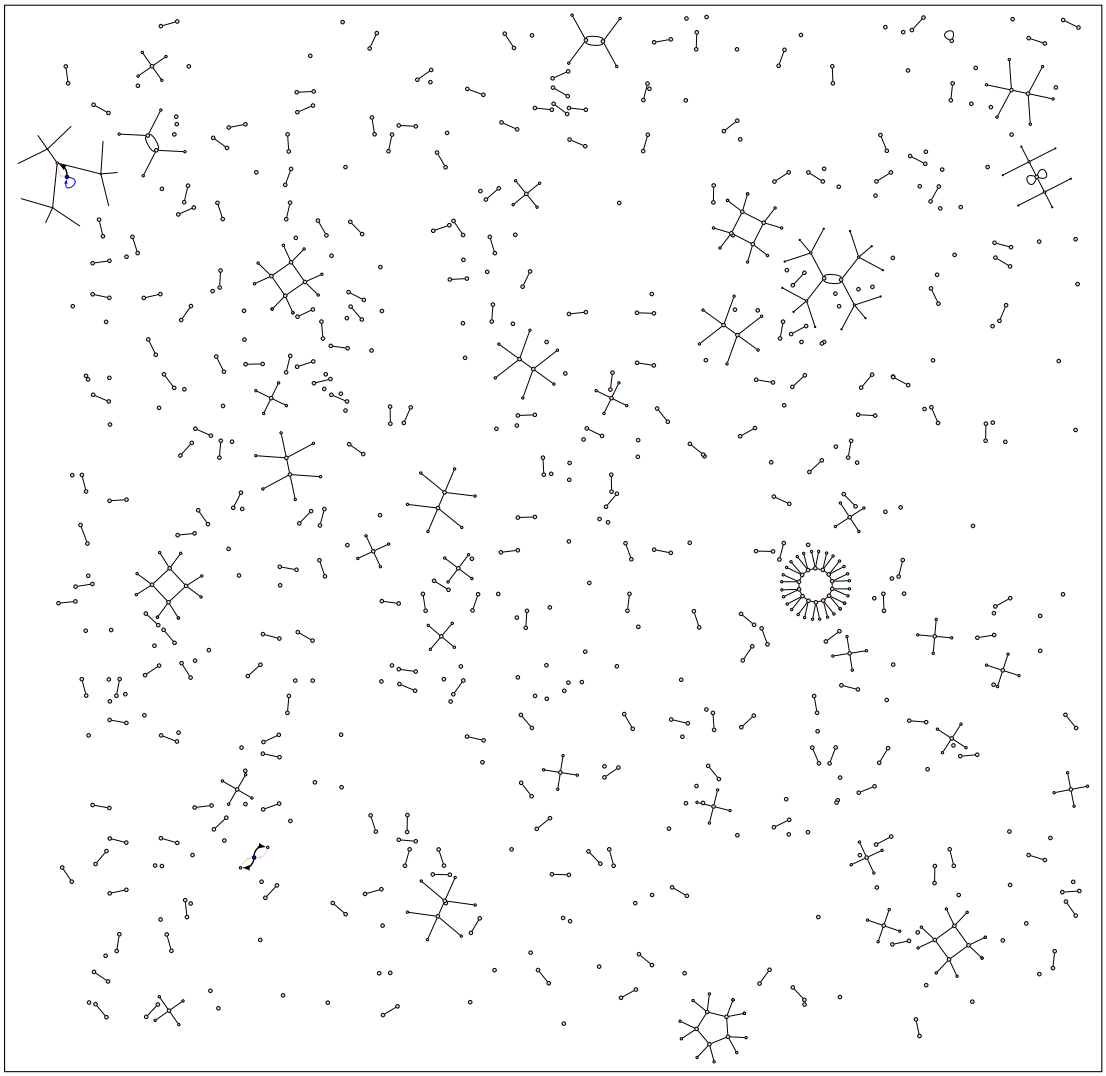
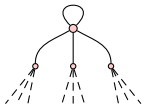


Figure 9.1: The volcano park $\mathcal{G}_3(1009)$.

2. *There exists a prime $p \in \mathbb{Z}$ such that V is a connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$ and the vertices on the crater of V correspond to elliptic curves with complex multiplication by a maximal order if and only if the graph induced by the first two levels of V is not isomorphic to*



When such a prime p exists, there exist in fact infinitely many of them.

The proofs of Theorem 9.1.2 and Theorem 9.1.3 feature a study of elements in the class group of imaginary quadratic fields: in a nutshell, volcanoes with crater size n exist because one can find ideal classes in well-chosen imaginary quadratic fields with order n. To study these questions on orders of elements in class groups, we are lead to study and explicitly solve some Diophantine equations. Using variations of arguments of Nagell, Mahler, Pell, we are able to prove the following generalisation of the classical work of Yamamoto [Yam70]:

Theorem 9.1.4. *The following properties hold.*

1. *Let $n \neq 4$ be a positive integer and let $K = \mathbb{Q}(\sqrt{1-2^{n+2}})$. Then in $\mathcal{O}_K$ the prime 2 splits into two prime ideals whose corresponding classes in $\mathrm{Cl}(\mathcal{O}_K)$ have order n.*

2. *Let $K = \mathbb{Q}(\sqrt{-39})$. Then in $\mathcal{O}_K$ the prime 2 splits into two prime ideals whose corresponding classes in $\mathrm{Cl}(\mathcal{O}_K)$ have order 4.*

3. *Let $\ell \in \mathbb{Z}$ be an odd prime and let $n \in \mathbb{Z}_{>0}$. Define $K_1 := \mathbb{Q}(\sqrt{1-\ell^n})$ and $K_2 := \mathbb{Q}(\sqrt{1-4\ell^n})$. Then either in $\mathcal{O}_{K_1}$ or in $\mathcal{O}_{K_2}$ the prime ℓ splits into two prime ideals whose corresponding classes in $\mathrm{Cl}(\mathcal{O}_{K_i})$ have order n.*

This result is interesting on its own, and plays a key role in our solution of the inverse volcano problem. To obtain a proof of Theorem 9.1.4, we prove Proposition 9.3.6, Proposition 9.3.7, and Proposition 9.3.8. We remark that the third item of Theorem 9.1.4 cannot be strengthened by considering only one among K_1 and K_2: for example, the prime 3 splits into two principal ideals in $\mathbb{Q}(\sqrt{1-3^2}) = \mathbb{Q}(\sqrt{-2})$ and the prime 5 splits into two principal ideals in $\mathbb{Q}(\sqrt{1-4\cdot 5^2}) = \mathbb{Q}(\sqrt{-11})$. On the other hand, in Proposition 9.3.7 and Proposition 9.3.8 we are in fact more precise and we are able to decide which of K_1 or K_2 is the correct field to consider, for a given pair (ℓ, n).

In Section 9.4 we note that our families of explicit solution to the inverse problems are far from being minimal, both in terms of the size of the characteristic p and of the size of the associated discriminant. We give a somewhat naïve algorithm that computes the minimal characteristic p in which a given abstract volcano exists, and provide explicit values for all cycloalkane-shaped volcanoes.

In Section 9.5, we show that the inverse volcano problem over more general finite fields does not always have a solution. We give in Proposition 9.5.1 an example of abstract 2-volcano that is not a connected component of any ordinary isogeny graph over $\mathbb{F}_{p^2}$ for any prime $p \neq 2$.

9.2 Ordinary Isogeny Graphs Over $\mathbb{F}_p$

Let $p \geq 5$ be a prime number and denote by $\mathbb{F}_p$ the prime field of characteristic p, with algebraic closure $\overline{\mathbb{F}}_p$. The condition on p is not restrictive for our study of the inverse volcano problem and, besides, it allows us to simplify the exposition of the theory in this section (cfr. for instance the use of Waterhouse's [Wat69, Theorem 4.1] in the proof of Lemma 9.2.4). We want to define and study ordinary isogeny graphs over $\mathbb{F}_p$. All the constructions appearing in this section can

be performed, *mutatis mutandis*, over $\mathbb{F}_{p^s}$ for any $s > 1$. We decided to focus only on prime fields, over which we will formulate and solve the inverse volcano problem.

Let $\mathcal{V}$ be the subset of $j \in \mathbb{F}_p$ such that the elliptic curves $E/\overline{\mathbb{F}}_p$ with $j(E) = j$ are ordinary. For any $j \in \mathcal{V}$, let $\mathcal{E}_j$ be the set of $\mathbb{F}_{p^2}$-isomorphism classes of elliptic curves $E/\mathbb{F}_p$ with $j(E) = j$. In other words, two elliptic curves E, E' over $\mathbb{F}_p$ represent the same class in $\mathcal{E}_j$ if and only if $j(E) = j(E') = j$ and E is a quadratic twist of E'. If $j \neq 0, 1728$, the set $\mathcal{E}_j$ has cardinality 1 (see [Sil09, Proposition 5.4]), and we fix a representative $E_j/\mathbb{F}_p$. If $j = 0$ then $\#\mathcal{E}_0 = 3$ and we fix $\{E_0^{(1)}, E_0^{(2)}, E_0^{(3)}\}$ to be three representatives of the different isomorphism classes. Similarly, if $j = 1728$ then $\#\mathcal{E}_{1728} = 2$ and we fix $\{E_{1728}^{(1)}, E_{1728}^{(2)}\}$ to be two representatives of the different isomorphism classes. In these two special cases, we generically use E_0 (resp. E_{1728}) to denote any of the three elliptic curves $\{E_0^{(1)}, E_0^{(2)}, E_0^{(3)}\}$ (resp. $\{E_{1728}^{(1)}, E_{1728}^{(2)}\}$).

Definition 9.2.1. For a prime $\ell \neq p$, the ordinary *ℓ-isogeny graph* over $\mathbb{F}_p$ is the directed graph $\mathcal{G}_\ell(\mathbb{F}_p)$ whose vertex set equals $\mathcal{V}$ and such that, for every $j, j' \in \mathcal{V}$, there are as many directed edges from j to j' as there are equivalence classes of separable cyclic isogenies of degree ℓ $\varphi : E_j \to E_{j'}$ defined over $\overline{\mathbb{F}}_p$, where $E_j/\mathbb{F}_p$ denotes any elliptic curve with j-invariant j.

Remark 9.2.2. The directed edges from one node j_1 to another node j_2 in $\mathcal{G}_\ell(\mathbb{F}_p)$ represent the non-equivalent classes of cyclic isogenies of degree ℓ between E_{j_1} and E_{j_2}. If there is a directed edge between j_1 and j_2 with $j_1 \neq j_2$, and if $j_1, j_2 \notin \{0, 1728\}$, then there is a unique edge from j_2 to j_1 corresponding to the dual isogeny. By convention, in this situation we only represent one undirected edge in our figures. The case $j_1 = j_2$ can be more subtle, as an elliptic curve may have a self ℓ-isogeny φ with dual $\widehat{\varphi}$ satisfying $\ker\varphi \neq \ker\widehat{\varphi}$. In this case, the graphic representations display both φ and $\widehat{\varphi}$.

We begin our systematic study of the structure of $\mathcal{G}_\ell(\mathbb{F}_p)$. Moving from general to specific, we divide the isogeny graph into progressively smaller subgraphs: cordilleras, belts and volcanoes. We follow the geological terminology from [FM02; MSTTV07].

Before leaving on a mountain hike, we fix some notations: for each $j \in \mathcal{V}$, the elliptic curve E_j has complex multiplication by an order $\mathcal{O}_j := \mathrm{End}_{\overline{\mathbb{F}}_p}(E_j)$ in an imaginary quadratic field K. We write $D(\mathcal{O})$ for the discriminant of a general order $\mathcal{O}$. For imaginary quadratic orders, it is a negative integer congruent to 0 or 1 mod 4.

9.2.1 Cordilleras

For every $j \in \mathcal{V}$ let us denote by $\mathrm{Tr} : \mathrm{End}_{\overline{\mathbb{F}}_p}(E_j) \otimes_{\mathbb{Z}} \mathbb{Q} \to \mathbb{Q}$ the trace map and by $\pi_j \in \mathrm{End}_{\overline{\mathbb{F}}_p}(E_j)$ the Frobenius endomorphism of the elliptic curve E_j. By [Sil09, Chapter V, Theorem 2.3.1] we have that $\mathrm{Tr}(\pi_j) = p + 1 - \#E_j(\mathbb{F}_p)$ so that $\mathbb{Z}[\pi_j]$ is isomorphic to the imaginary quadratic order of discriminant $D(\mathbb{Z}[\pi_j]) = \mathrm{Tr}(\pi_j)^2 - 4p$.

Definition 9.2.3. Let $t \in \mathbb{Z}$. The *t-cordillera* in $\mathcal{G}_\ell(\mathbb{F}_p)$ is the subgraph of $\mathcal{G}_\ell(\mathbb{F}_p)$ induced by the subset of vertices

$$\mathcal{V}_t := \{j \in \mathbb{F}_p : \mathrm{Tr}(\pi_j) = \pm t\} \subseteq \mathcal{V}.$$

By definition, if $j, j' \in \mathcal{V}_t$ we have $D(\mathbb{Z}[\pi_j]) = D(\mathbb{Z}[\pi_{j'}])$, so that both $\mathrm{End}_{\overline{\mathbb{F}}_p}(E_j)$ and $\mathrm{End}_{\overline{\mathbb{F}}_p}(E_{j'})$ contain an order isomorphic to the imaginary quadratic order of discriminant $t^2 - 4p$. In particular, E_j and $E_{j'}$ have complex multiplication by an order inside the *same* imaginary quadratic field K. We call K the *field associated with the t-cordillera* and, if $D(\mathcal{O}_K) < -4$, we say that the cordillera is *regular*. Cordilleras partition the vertex set of the isogeny graph. Note that their vertices are defined independently of the isogeny degree ℓ.

Lemma 9.2.4. *For $t \in \mathbb{Z}_{\neq 0}$ the following holds:*

1. *The set $\mathcal{V}_t$ is non-empty if and only if $-2\sqrt{p} < t < 2\sqrt{p}$;*

2. *If $\mathcal{V}_t$ is non-empty then for every order $\mathcal{O}$ containing the order of discriminant t^2-4p there exists an elliptic curve $E/\mathbb{F}_p$ such that $\mathrm{End}_{\overline{\mathbb{F}}_p}(E) \cong \mathcal{O}$ and $j(E)$ is in $\mathcal{V}_t$;*

3. *If $j(E) \in \mathcal{V}_t$ and if $\mathrm{End}_{\overline{\mathbb{F}}_p}(E) \cong \mathcal{O}$, then all the $h(\mathcal{O})$ j-invariants corresponding to curves with CM by $\mathcal{O}$ are also in $\mathcal{V}_t$.*

Proof. The first part is given in [Wat69, Theorem 4.1]. We now prove the second part: let $j(E)$ be an element in $\mathcal{V}_t$. Let $\mathcal{O}$ be the endomorphism ring of E. It contains the order $\mathbb{Z}[\pi_E]$ generated by the Frobenius π_E, which has discriminant t^2-4p. Each order containing an order isomorphic to $\mathbb{Z}[\pi_E]$ can be realised as the endomorphism ring of an elliptic curve defined over $\mathbb{F}_p$ and $\mathbb{F}_p$-isogenous to E, see [Wat69, Theorem 4.2, point (2)]. In particular, as these curves are all $\mathbb{F}_p$-isogenous to E, they have the same trace t and this proves the second part. For the third part, by Corollary 3.5.18, all the j-invariants corresponding to curves with CM by $\mathcal{O}$ are rational over $\mathbb{F}_p$. If $j(E)=0$ or $j(E)=1728$, the class number of $\mathcal{O}$ is one and the result is immediate. If $j(E) \notin \{0, 1728\}$, consider a j-invariant $j(E') \in \mathbb{F}_p$ of an elliptic curve E' with complex multiplication by $\mathcal{O}$. Since E and E' have complex multiplication by the same order, they are geometrically isogenous. Let $k/\mathbb{F}_p$ be a degree r field extension over which this isogeny is defined. The Frobenius endomorphisms π_E and $\pi_{E'}$ of the two elliptic curves, seen as elements of the same imaginary quadratic field $K = \mathrm{Frac}(\mathcal{O})$, satisfy $\pi_E^r = \pi_{E'}^r$ or $\pi_E^r = \overline{\pi}_{E'}^r$ by Tate's isogeny theorem [Tat66, Theorem 1 (c4)]. This means that there exists a root of unity $\zeta \in K$ such that

$$\pi_E = \zeta \pi_{E'} \quad \text{or} \quad \pi_E = \zeta \overline{\pi}_{E'}.$$

If $\zeta = \pm 1$, it follows immediately that $j(E') \in \mathcal{V}_t$. Otherwise, $\zeta \in \{\pm i, \pm\zeta_3, \pm\zeta_3^2\}$, where i is a primitive fourth root of unity and ζ_3 is a primitive third root of unity. In this case, one easily deduces that $j(E) = j(E') \in \{0, 1728\}$, a contradiction. There are $h(\mathcal{O})$ possibilities for $j(E')$, and that gives the claim. □

Fix a nonzero integer $t \in (-2\sqrt{p}, 2\sqrt{p})$ and let K be the field associated with the t-cordillera. All the j-invariants of elliptic curves with complex multiplication by the maximal order in K belong to $\mathcal{V}_t$ by Lemma 9.2.4 (3). Fix such a singular invariant j and let E_j be a corresponding elliptic curve over $\mathbb{F}_p$. If $D(\mathcal{O}_K) < -4$, then E_j is determined only up to quadratic twisting, and the Frobenius endomorphism of any of its quadratic twists has, up to sign, the same trace. This in particular implies, using Lemma 9.2.4 (2) that any imaginary quadratic number field K with discriminant $D(\mathcal{O}_K) < -4$ can be associated *to at most one cordillera* (with positive trace). One can formulate this statement in terms of norm equations as follows.

Lemma 9.2.5. *Let K be an imaginary quadratic field of discriminant $D(\mathcal{O}_K) < -4$. Then the equation*

$$4p = t^2 - v^2 D(\mathcal{O}_K) \tag{9.1}$$

has at most one solution $(t, v) \in \mathbb{N}^2$ with $p \nmid t$.

If $D(\mathcal{O}_K) \in \{-3, -4\}$ the situation is more delicate because the vertices $j=0$ and $j=1728$ are represented by 3 and 2 curves respectively, and the squared traces of the Frobenius endomorphisms of these curves may be different. These vertices may thus belong to different cordilleras at the same time. This is indeed always the case, as the following lemma shows.

Lemma 9.2.6. *The following holds:*

(i) The equation

$$4p = t^2 + 3v^2 \tag{9.2}$$

has no integer solutions if $p \equiv 2 \bmod 3$, and exactly three solutions (t, v) with $t > 0$ and $v > 0$ if $p \equiv 1 \bmod 3$.

(ii) The equation

$$4p = t^2 + 4v^2 \tag{9.3}$$

has no solutions if $p \equiv 3 \bmod 4$, and exactly two solutions (x, y) with $x > 0$ and $y > 0$ if $p \equiv 1 \bmod 4$.

Proof. Both statements reduce to solving a norm equation in the ring $\mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$ and $\mathbb{Z}[\sqrt{-1}]$ respectively. □

In particular, Lemma 9.2.6 shows that if $j = 0$ is an ordinary invariant then $K = \mathbb{Q}(\sqrt{-3})$ is associated with three t-cordilleras, and if $j = 1728$ is an ordinary invariant then $K = \mathbb{Q}(\sqrt{-1})$ is associated with two t-cordilleras (t positive).

9.2.2 Belts

In this section, we work inside a fixed non-empty t-cordillera $\mathcal{C}_t$ with vertex set $\mathcal{V}_t$. Let K be the imaginary quadratic number field associated with this cordillera and denote by $\mathcal{O}_K$ its ring of integers. Choose $\pi \in K$ to be any element such that

$$\pi^2 - t\pi + p = 0$$

so that, if $v := [\mathcal{O}_K : \mathbb{Z}[\pi]]$ denotes the conductor of the order $\mathbb{Z}[\pi]$ in K, then

$$4p - t^2 = -v^2 D(\mathcal{O}_K)$$

where $D(\mathcal{O}_K)$ is the discriminant of K. Denoting by $v_\ell(\cdot)$ the usual ℓ-adic valuation on $\mathbb{Q}$, we let $d := v_\ell(v)$ and $v' := v \cdot \ell^{-d}$.

Definition 9.2.7. An order $\mathcal{O}$ such that $\mathbb{Z}[\pi] \subset \mathcal{O} \subset \mathcal{O}_K$ is said to be *ℓ-saturated* if $v_\ell([\mathcal{O}_K : \mathcal{O}]) = d$. It is said to be *$\ell$-dry* if $v_\ell([\mathcal{O}_K : \mathcal{O}]) = 0$.

For every order $\mathbb{Z}[\pi] \subset \mathcal{O} \subset \mathcal{O}_K$ we define its *saturation* $\widetilde{\mathcal{O}}$ as the unique order in K such that

$$[\mathcal{O}_K : \widetilde{\mathcal{O}}] = [\mathcal{O}_K : \mathcal{O}] \cdot \ell^{d - v_\ell([\mathcal{O}_K : \mathcal{O}])}.$$

Note that $\widetilde{\mathcal{O}}$ still contains $\mathbb{Z}[\pi]$. We are now ready to state the main definition of this section.

Definition 9.2.8. For a positive integer $m \mid v'$, the *t-belt* of index m, denoted by $\mathcal{B}_{t,m}$, is the subgraph of $\mathcal{G}_\ell(\mathbb{F}_p)$ induced by the subset

$$\mathcal{V}_{t,m} = \{j \in \mathcal{V}_t : \widetilde{\mathrm{End}_{\mathbb{F}_p}(E_j)} \cong \mathbb{Z} + m\ell^d \mathcal{O}_K\} \subseteq \mathcal{V}_t.$$

Lemma 9.2.9. *The following holds:*

1. *The cut-set determined by the vertex partition $(\mathcal{V}_{t,m}, \mathcal{V} \setminus \mathcal{V}_{t,m})$ is empty.*
2. *Different t-belts form disjoint subgraphs of $\mathcal{G}_\ell(\mathbb{F}_p)$.*

Proof. Both points follow immediately from Proposition 3.5.10 and the fact that the endomorphism ring of every elliptic E with $j(E) \in \mathcal{V}_t$ must contain a subring isomorphic to $\mathbb{Z}[\pi]$. □

Note that two belts $\mathcal{B}_{t,m_1}$ and $\mathcal{B}_{t,m_2}$ for $m_1 \neq m_2$ cannot contain j-invariants of elliptic curves that have isomorphic endomorphism ring. Moreover, a belt does not have to be connected.

9.2.3 Isogeny Volcanoes

Isogeny volcanoes are the connected components of $\mathcal{G}_\ell(\mathbb{F}_p)$. The vertices of these graphs come endowed with a natural "stratification" by level, in the sense of the following definition.

Definition 9.2.10. An ordinary elliptic curve $E/\mathbb{F}_p$, its j-invariant $j(E)$, and its associated order $\mathrm{End}(E) = \mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$ are said to lie at *level* n in the volcano containing $j(E)$ if $v_\ell(f) = n$.

The subgraph induced by the subset of level-n vertices in a volcano V is denoted by V_n. The graph V_0 is called the *crater* of V while we say that the vertices in $V \setminus V_0$ are on the *lava flows* of the volcano V. The *depth* of V is the maximal level on which vertices $j \in V$ lie.

Lemma 9.2.11. *Let k be a finite field of characteristic p and $\ell \neq p$ a prime number. Let E and E' be two elliptic curves over k with complex multiplication by an order $\mathcal{O}$ of conductor f in an imaginary quadratic field K.*

1. *If $\mathfrak{L} \subseteq \mathcal{O}$ is an invertible ideal of norm ℓ, then $E/E[\mathfrak{L}]$ has complex multiplication by $\mathcal{O}$ and the quotient map $E \to E/E[\mathfrak{L}]$ has degree ℓ;*
2. *If $\varphi : E \to E'$ is an isogeny of degree ℓ, then there exists an invertible ideal $\mathfrak{L} \subseteq \mathcal{O}$ of norm ℓ such that $E' \cong E/E[\mathfrak{L}]$;*
3. *If $\mathfrak{L} \subseteq \mathcal{O}$ is a non-invertible ideal of norm ℓ, then $\ell \mid f$ and $E/E[\mathfrak{L}]$ has complex multiplication by the unique quadratic order containing $\mathcal{O}$ with index ℓ.*

Proof. For an ideal $I \subseteq \mathcal{O}$ we denote by $\mathcal{O}(I) := \{x \in K : xI \subseteq I\}$ the order associated with I. We always have $\mathcal{O} \subseteq \mathcal{O}(I)$ and $\mathcal{O}(I)$ is the smallest quadratic order $\mathcal{O}' \supseteq \mathcal{O}$ such that $I\mathcal{O}'$ is invertible in $\mathcal{O}'$, as can be deduced from [JT15, Proposition 5.2].

Part (1) follows from [Wat69, Proposition 3.9] and the fact that $\#E[\mathfrak{L}] = \#(\mathcal{O}/\mathfrak{L})$, since $\ell \neq p$.

We now prove part (2). Consider the ideal

$$\mathfrak{L} := \{f \in \mathrm{End}_{\overline{k}}(E) : f(\ker\varphi) = 0\} \subseteq \mathrm{End}_{\overline{k}}(E) = \mathcal{O}.$$

We clearly have $\ker\varphi \subseteq E[\mathfrak{L}]$. On the other hand, it is not difficult to see, using a finite fields analogue of [Sil94, II, Corollary 1.1.1] and the fact that E and E' have complex multiplication by the same order, that $\ker\varphi$ is a cyclic $\mathcal{O}$-module which is then isomorphic to $\mathcal{O}/\mathrm{Ann}_{\mathcal{O}}(\ker\varphi) = \mathcal{O}/\mathfrak{L}$, where $\mathrm{Ann}_{\mathcal{O}}(\ker\varphi)$ denotes the annihilator ideal of $\ker\varphi$. Since $\ell \neq p$ we have

$$\#E[\mathfrak{L}] = \#\mathcal{O}/\mathfrak{L} = \#\ker\varphi = \ell$$

and we then obtain $\ker\varphi = E[\mathfrak{L}]$. Hence $E' \cong E/\ker\varphi = E/E[\mathfrak{L}]$. Using [Wat69, Proposition 3.9] (which can be applied thanks to [Wat69, Theorem 4.5]) we see that $\mathcal{O}(\mathfrak{L}) = \mathcal{O}$ and so $\mathfrak{L}$ is invertible in $\mathcal{O}$.

We finally prove part (3). Note that the existence of a non-invertible ideal of norm ℓ implies that the conductor of $\mathcal{O}$ is divisible by ℓ. Using [Wat69, Proposition 3.9] again, we need to prove that $\mathcal{O}(\mathfrak{L})$ is equal to the unique order $\mathcal{O}'$ containing $\mathcal{O}$ with index ℓ. Set $\mathcal{O}' = \mathbb{Z}[\omega]$. Then we have $\mathcal{O} = \mathbb{Z}[\ell\omega]$ and $\mathfrak{L} = (\ell, \ell\omega)$. Hence $\mathfrak{L}\mathcal{O}'$ is principal, generated by ℓ, and in particular invertible. We deduce that $\mathcal{O}(\mathfrak{L}) = \mathcal{O}'$ and the proof is concluded. □

Remark 9.2.12. If $\mathcal{O}$ is an imaginary quadratic order of discriminant $D(\mathcal{O})$ and ℓ is a prime, then there are exactly $1 + \left(\frac{D(\mathcal{O})}{\ell}\right)$ prime ideals in $\mathcal{O}$ with norm ℓ. If $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$, this follows from [Cox13, Proposition 7.20]. In this case, the primes of norm ℓ are always invertible. On the other hand, if $\ell | [\mathcal{O}_K : \mathcal{O}]$, the ideal $\mathfrak{L} = \mathbb{Z}\ell + [\mathcal{O}_K : \mathcal{O}]\mathcal{O}_K$ in $\mathcal{O}$ is the unique prime ideal of norm ℓ and is not invertible. Uniqueness follows from the following argument: if $\mathfrak{L}'$ is another such ideal, then $\mathfrak{L}^2 \subseteq \ell\mathcal{O} \subseteq \mathfrak{L}'$, hence $\mathfrak{L} \subseteq \mathfrak{L}'$, which implies $\mathfrak{L} = \mathfrak{L}'$ by maximality.

From a volcanic perspective, Lemma 9.2.11 and Remark 9.2.12 translate in the following way.

Corollary 9.2.13. *Horizontal isogenies in $\mathcal{G}_\ell(\mathbb{F}_p)$ can only occur between vertices at level 0 in the volcanoes belonging to the same belt. More precisely, for every non-zero $t \in \mathbb{Z}$, for every t-cordillera, if K is the field associated with the t-cordillera, then for every m, for every belt $\mathcal{B}_{t,m}$, for every volcano V in $\mathcal{B}_{t,m}$, there are exactly*

$$1 + \left(\frac{D(\mathcal{O}_K)}{\ell}\right) = \begin{cases} 0 & \text{if } \ell \text{ is inert in } K, \\ 1 & \text{if } \ell \text{ is ramified in } K, \\ 2 & \text{if } \ell \text{ splits in } K, \end{cases}$$

distinct edges of $\mathcal{G}_\ell(\mathbb{F}_p)$ from $j \in V_0$ to other vertices in V_0.

We will now describe the structure of an isogeny volcano by analysing crater and lava flow separately. We start by analysing crater structure, then focus on a single belt (that is, looking at a fixed endomorphism structure), and finally we let the lava flow down, looking at class numbers at every level of the volcanoes. In other words, there is a simultaneous eruption in all volcanoes belonging to the same belt, and class number arithmetic constrains the shape to which lava solidifies!

The crater

The following proposition presents the possible craters that can occur.

Proposition 9.2.14. *Let $\mathcal{B}_{t,m}$ be a belt in the isogeny graph $\mathcal{G}_\ell(\mathbb{F}_p)$ and let $\mathcal{C}_{t,m}$ be the subgraph of $\mathcal{B}_{t,m}$ induced by the vertices at level 0. Let $\mathcal{O}$ be the CM order associated with any vertex in $\mathcal{C}_{t,m}$ and let $\mathfrak{L} \subseteq \mathcal{O}$ be a prime ideal dividing ℓ. Then each connected component of $\mathcal{C}_{t,m}$ is a crater of a volcano in $\mathcal{B}_{t,m}$ and consists of:*

1. *A single vertex if ℓ is inert in K;*
2. *A single vertex with a self-loop if ℓ is ramified in K and $\mathfrak{L}$ is principal;*
3. *A single vertex with two self-loops if ℓ splits in K and $\mathfrak{L}$ is principal;*
4. *Two vertices connected with a single edge if ℓ ramifies in K and $\mathfrak{L}$ is not principal;*
5. *Two vertices connected with a double edge if ℓ splits in K and $\mathfrak{L}$ has order 2 in the class group $\mathrm{Cl}(\mathcal{O})$;*
6. *More generally, a cycle of size the order of $[\mathfrak{L}]$ in the class group $\mathrm{Cl}(\mathcal{O})$ if ℓ splits in K and we are not in any of the previous cases.*

Proof. Fix $j(E) \in \mathcal{C}_{t,m}$. We want to analyse the connected component of $\mathcal{C}_{t,m}$ containing $j(E)$. Using Lemma 9.2.11, there is an edge from $j(E)$ to $j(E/E[\mathfrak{L}])$. As the action described in Theorem 3.5.14 is faithful and transitive, by iteration we obtain a cycle of length the order of the class of $\mathfrak{L}$ in $\mathrm{Cl}(\mathcal{O})$.

A case by case analysis concludes the proof, for instance: In the case where ℓ is ramified in K, and the unique ideal $\mathfrak{L}$ lying above ℓ is principal, then $\mathfrak{L} = \overline{\mathfrak{L}}$, and $\mathfrak{L}$, $\overline{\mathfrak{L}}$ correspond to mutually dual isogenies with the same kernel. This situation is represented by a unique self-loop. In the case where ℓ splits into principal ideals $\mathfrak{L}, \overline{\mathfrak{L}}$ in K, and the ideal $\mathfrak{L}$ lying above ℓ is principal, then $\mathfrak{L} \neq \overline{\mathfrak{L}}$. These ideals correspond to two self-isogenies with distinct kernels. This situation is represented by two self-loops in the graph (cfr. Remark 9.2.2). □

Remark 9.2.15. Case (6) in the above proposition in fact includes also case (5). However, we decided to separate the two cases in order to underline the difference between case (4) and case (5).

See Figure 9.2 for a view of all possible subgraphs at level 0. The rightmost graph in Figure 9.2 indicates a generic c-cycle, for any integer $c \geq 5$.

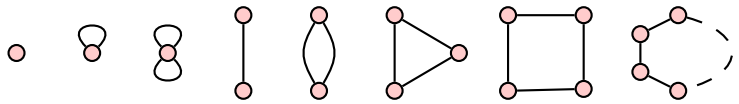


Figure 9.2: All possible craters.

The lava flows

We now explain what the other levels of a volcano look like.

Lemma 9.2.16. *Let E/k be an elliptic curve. The number of non-equivalent ℓ-isogenies from E to any other curve with k-rational j-invariant is $0, 1, 2$ or $\ell + 1$.*

Proof. The kernel of an isogeny from E must be a subgroup of $E[\ell] \cong \mathbb{Z}/\ell\mathbb{Z} \times \mathbb{Z}/\ell\mathbb{Z}$. For it to be an ℓ-isogeny the subgroup must have order ℓ. In order for the target curve to be k-rational, the kernel of the isogeny must be invariant under the action of the Galois group $G = \mathrm{Gal}(\overline{k}/k)$. There are $\ell + 1$ subgroups of order ℓ in $E[\ell]$, that can be seen as lines in an $\mathbb{F}_\ell$-vector space. G acts linearly, and if it fixes three or more such lines, then it acts as an homothety and hence fixes every subgroup. □

Proposition 9.2.17. *Let V be a volcano of depth d in the isogeny graph $\mathcal{G}_\ell(\mathbb{F}_p)$. Assume that the crater V_0 contains j-invariants corresponding to elliptic curves with complex multiplication by an order $\mathcal{O}$ of discriminant $D(\mathcal{O})$ with $D(\mathcal{O}) \neq -3, -4$. Then:*

- *For every $j \in V_0$, the elliptic curve E_j has, up to equivalence, 0 descending isogenies if $d = 0$ and $\left(\ell - \left(\frac{D}{\ell}\right)\right)$ descending isogenies otherwise;*
- *For every $n > 0$ and every $j \in V_n$, the elliptic curve E_j has, up to equivalence, ℓ descending isogenies if $n < d$ and 0 descending isogenies otherwise;*
- *For every $n > 0$ and every $j \in V_n$, the elliptic curve E_j has, up to equivalence, exactly 1 ascending isogeny.*

Proof. The proof of this proposition is essentially contained in the second part of the proof of [Sut13, Lemma 6]. We sketch the argument here.

Let $\mathcal{B}_{t,m}$ be the belt containing the volcano V. We determine the vertical isogenies between the elliptic curves corresponding to the vertices in $\mathcal{B}_{t,m}$ by induction on the level n of the vertices themselves.

If $n = 0$, all the elliptic curves corresponding to the vertices at level 0 of $\mathcal{B}_{t,m}$ have, by Corollary 9.2.13, exactly $1 + \left(\frac{D(\mathcal{O}_K)}{\ell}\right)$ horizontal isogenies. Note that, since the order of discriminant D must be ℓ-dry by the level-0 assumption, we have $\left(\frac{D(\mathcal{O}_K)}{\ell}\right) = \left(\frac{D(\mathcal{O})}{\ell}\right)$. The maximal possible number of non-equivalent isogenies from every E_j is $\ell + 1$, so the maximal possible number of non-equivalent descending isogenies from every E_j is

$$(\ell + 1) - \left(1 + \left(\frac{D(\mathcal{O})}{\ell}\right)\right) = \ell - \left(\frac{D(\mathcal{O})}{\ell}\right).$$

If $d = 0$, there are none. If $d > 0$, then by Corollary 3.5.18 there are exactly $h(\mathcal{O}')$ vertices in $\mathcal{B}_{t,m}$ at level 1, where $\mathcal{O}'$ is the unique order contained in $\mathcal{O}$ with index ℓ. Using [Cox13, Corollary 7.28] and the fact that $D \neq -3, -4$ we have

$$h(\mathcal{O}') = h(\mathcal{O})\left(\ell - \left(\frac{D(\mathcal{O})}{\ell}\right)\right).$$

By Lemma 9.2.11, all of the $h(\mathcal{O})\left(\ell - \left(\frac{D(\mathcal{O})}{\ell}\right)\right)$ elliptic curves with CM by $\mathcal{O}'$ admit a vertical isogeny. Since, by duality and the fact that $D(\mathcal{O}) \neq -3, -4$, the number of ascending isogenies from level 1 to level 0 is the same as the number of descending isogenies from level 0 to level 1, a counting argument shows that every elliptic curve E_j at level 0 has, up to equivalence, $\left(\ell - \left(\frac{D(\mathcal{O})}{\ell}\right)\right)$ descending isogenies and every elliptic curve E_j at level 1 has, up to equivalence, precisely 1 vertical isogeny.

If $n = 1$, there are no horizontal isogenies by Corollary 9.2.13. Hence, all elliptic curves corresponding to the vertices at level 1 of $\mathcal{B}_{t,m}$ have one ascending isogeny (by the previous discussion) and $(\ell + 1) - 1 = \ell$ descending isogenies if $d > 1$, no descending isogenies otherwise. Suppose then that $d > 1$. Then by Corollary 3.5.18 there are exactly $h(\mathcal{O}'')$ vertices in $\mathcal{B}_{t,m}$ at level 2, where $\mathcal{O}''$ is the unique order contained in $\mathcal{O}'$ with index ℓ. Using [Cox13, Corollary 7.28] we see that

$$h(\mathcal{O}'') = \ell h(\mathcal{O}').$$

By Lemma 9.2.11, all of the $\ell h(\mathcal{O}')$ elliptic curves with CM by $\mathcal{O}''$ admit a vertical isogeny. As above we conclude that every elliptic curve E_j at level 1 has, up to equivalence, ℓ descending isogenies and every elliptic curve E_j at level 2 has, up to equivalence, precisely 1 vertical isogeny. An easy induction now leads to the claim. □

It is now time to deal with the pathological volcanoes containing 0 and 1728. We warm up with the following lemma.

Lemma 9.2.18. *Suppose that $0 \in \mathbb{F}_p$ or $1728 \in \mathbb{F}_p$ is the j-invariant of an ordinary elliptic curve E and let $\ell \neq p$ be a prime. Suppose that there exists a cyclic subgroup $H \subseteq E(\overline{\mathbb{F}}_p)$ of order ℓ which is fixed by all the automorphisms of E. Then there exists an ideal $\mathfrak{L} \subseteq \mathrm{End}_{\overline{\mathbb{F}}_p}(E)$ of norm ℓ such that $H = E[\mathfrak{L}]$.*

Proof. We give the details for 0, the proof is similar for 1728. There is an isomorphism $\mathcal{O} := \mathrm{End}_{\overline{\mathbb{F}}_p}(E) \cong \mathbb{Z}[\zeta_6]$ where $\zeta_6 \in \overline{\mathbb{Q}}$ is a primitive 6-th root of unity. In particular, $\mathrm{Aut}_{\overline{\mathbb{F}}_p}(E) \cong \langle \zeta_6 \rangle$. Hence, if H is fixed by all the automorphisms of E, then it is in fact fixed by all its endomorphisms. We deduce that H is a cyclic $\mathcal{O}$-module of order ℓ. We then have an $\mathcal{O}$-module isomorphism

$$\mathcal{O}/\mathrm{Ann}_{\mathcal{O}}(H) \cong H$$

where $\mathrm{Ann}_{\mathcal{O}}(H)$ is the annihilator of H in $\mathcal{O}$. The isomorphism above shows that $\mathfrak{L} := \mathrm{Ann}_{\mathcal{O}}(H)$ is an ideal of $\mathcal{O}$ of norm ℓ and $H \subseteq E[\mathfrak{L}]$. Since these two groups have the same cardinality, we deduce that $H = E[\mathfrak{L}]$, as wanted. □

We now describe the directed neighbourhood of the vertex 0.

Proposition 9.2.19. *Suppose that $0 \in \mathbb{F}_p$ is the j-invariant of an ordinary elliptic curve E and let $\ell \neq p$ be a prime. In the volcano containing 0, the directed subgraph induced by the neighbours of 0 at level zero and level 1 can be described as:*

- *If $\ell = 3$: the vertex 0 has one self-loop, three descending isogenies towards a unique vertex j at level 1, and there is one unique isogeny from j to 0;*

- *If $\ell \equiv 1 \mod 3$: the vertex 0 has two self-loops. Level 1 has either zero or $(\ell-1)/3$ vertices. In the latter case, all of these vertices receive three descending isogenies from 0 and send one ascending isogeny to 0;*
- *If $\ell \equiv 2 \mod 3$, the vertex 0 has no self-loop. Level 1 has either zero or $(\ell+1)/3$ vertices. In the latter case, all of these vertices receive three descending isogenies from 0, and send one ascending isogeny to 0.*

Proof. The structure of the crater is already described in Proposition 9.2.14. If level 1 is empty we are done. We then assume that there is at least one j-invariant at level 1.

Let μ_6 be the group of $\overline{\mathbb{F}}_p$-automorphisms of E. It acts on the set S of subgroups of order ℓ in E. The orbits of this action can be described by means of Lemma 9.2.18: if $H \in S$ is of the form $H = E[\mathfrak{L}]$ for some ideal $\mathfrak{L} \subseteq \mathbb{Z}[\zeta_3]$ of norm ℓ then its orbit is a singleton, otherwise the orbit of H contains three elements. Each singleton orbit gives rise to a self-loop around 0 thanks to Proposition 9.2.14, while each orbit of cardinality 3 gives rise to three directed edges (descending isogenies) from 0 towards the same j-invariant in level 1 by Lemma 9.2.11. Note also that there is a directed edge from each j-invariant at level 1 to 0 by Lemma 9.2.11. Dualising this isogeny, we see that there is at least one, hence three, directed edges from 0 to each j-invariant at level 1. Let $\mathcal{O}$ be the order corresponding to the vertices at level 1. The number of vertices on this level is given by the class number $h(\mathcal{O})$ (see Corollary 3.5.18), which in turn is equal to

$$h(\mathcal{O}) = \frac{1}{3}\left(\ell - \left(\frac{-3}{\ell}\right)\right)$$

by [Cox13, Corollary 7.28]. On the other hand, the total number of oriented edges from level 0 to level 1 is precisely $\ell - \left(\frac{-3}{\ell}\right)$. This, together with the discussion above, implies that for each vertex j at level 1 there are exactly three directed edges from 0 to j in the isogeny graph. By counting, one also easily sees that there is exactly one ascending isogeny from each vertex at level 1 to 0.

□

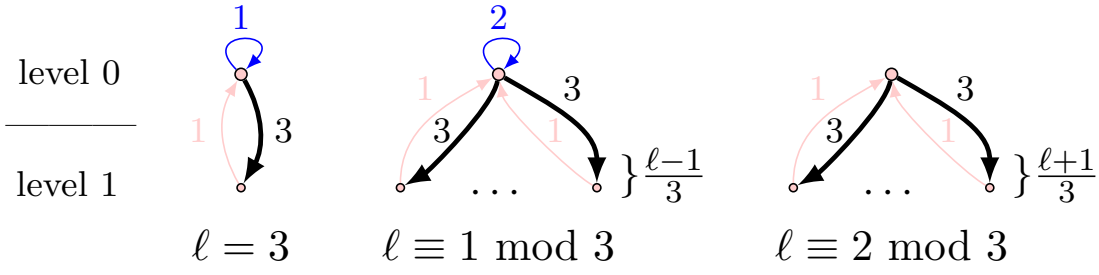


Figure 9.3: All possible neighbourhoods of depth 1 around 0. Numbers next to directed edges represent multiplicities.

Proposition 9.2.20. *Suppose that $1728 \in \mathbb{F}_p$ is the j-invariant of an ordinary elliptic curve E and let $\ell \neq p$ be a prime. In the volcano containing 0, the directed subgraph induced by the neighbours of 0 at level zero and level 1 can be described as:*

- *If $\ell = 2$: the vertex 1728 has one self-loop, two descending isogenies towards a unique vertex j at level 1, and there is one unique isogeny from j to 1728;*
- *If $\ell \equiv 1 \mod 4$: the vertex 1728 has two self-loops. Level 1 has either zero or $(\ell-1)/2$ vertices. In the latter case, all of these vertices receive two descending isogenies from 1728 and send one ascending isogeny to 1728;*

- *If $\ell \equiv 3 \bmod 4$, the vertex 1728 has no self-loop. Level 1 has either zero or $(\ell+1)/2$ vertices. In the latter case, all of these vertices receive two descending isogenies from 1728, and send one ascending isogeny to 1728.*

Proof. The proof goes along the same lines as the proof of Proposition 9.2.19. □

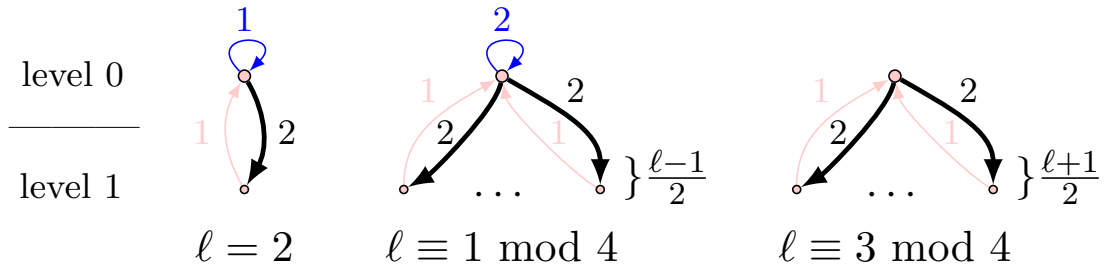


Figure 9.4: All possible neighbourhoods of depth 1 around 1728. Numbers next to directed edges represent multiplicities.

9.2.4 Mapping the Territory: Counting Structures

In this paragraph, we count the number of volcanic structures.

Lemma 9.2.21. *For $t \in \mathbb{Z}_{>0}$, the number of non-empty t-cordilleras in $\mathcal{G}_\ell(\mathbb{F}_p)$ is $\lfloor 2\sqrt{p} \rfloor$.*

Proof. The number of cordilleras is the number of possible traces up to sign. By Waterhouse [Wat69, Theorem 4.1 page 536], each trace in the Hasse interval $[-2\sqrt{p}, 2\sqrt{p}]$ is attained over $\mathbb{F}_p$. This gives $\lfloor 2\sqrt{p} \rfloor$ possibilities for non-zero traces up to sign. It is independent of ℓ. □

Lemma 9.2.22. *Let $t \in [1, 2\sqrt{p}]$ be an integer. Let K be the associated field $\mathbb{Q}(\sqrt{t^2-4p})$. Let $D(\mathcal{O}_K)$ denote its discriminant. Let $v = \sqrt{(t^2-4p)/D(\mathcal{O}_K)}$. Let $d = v_\ell(v)$. The number of belts in the t-cordillera is $\omega(v\ell^{-d})$, where $\omega(k)$ is the number of positive divisors of the integer k.*

Proof. For every positive divisor m of $v\ell^{-d}$, we have the belt $B_{t,m}$ induced by the following set of vertices:

$$\mathcal{V}_{t,m} = \{j \in \mathcal{V}_t \,|\, \widetilde{\mathrm{End}(E_j)} \cong \mathbb{Z} + m\ell^d\mathcal{O}_K\}.$$

□

Lemma 9.2.23. *Let $\mathcal{B}_{t,m}$ be a belt associated with a trace $t \neq 0$ and integer m. Let K be the associated field. Let $\mathcal{O} = \mathbb{Z} + m\mathcal{O}_K$.*

(i) If ℓ is inert, the number of volcanoes in $\mathcal{B}_{t,m}$ is $h(\mathcal{O})$;

(r) If ℓ is ramified, if $\ell\mathcal{O} = \mathcal{L}^2$ with $\mathcal{L}$ principal, then the number of volcanoes in $\mathcal{B}_{t,m}$ is $h(\mathcal{O})$. If $\ell\mathcal{O} = \mathcal{L}^2$ with $\mathcal{L}$ not principal, then the number of volcanoes in $\mathcal{B}_{t,m}$ is $h(\mathcal{O})/2$;

(s) If ℓ is split and $\ell\mathcal{O} = \mathfrak{L}\overline{\mathfrak{L}}$, let r be the order of $\mathcal{L}$ in $\mathrm{Cl}(\mathcal{O})$, then the number of volcanoes in $\mathcal{B}_{t,m}$ is $h(\mathcal{O})/r$.

Proof. The number of volcanoes is the number of craters, and the craters are isomorphic within a belt. The total number of vertices of depth 0 in $\mathcal{B}_{t,m}$ is $h(\mathcal{O})$. The size of each crater is given by Proposition 9.2.14. □

Remark 9.2.24. All volcanoes on the same cordillera have the same depth. All volcanoes on the same belt have the exact same shape. The total number of vertices at level 0 in a belt with endomorphism order $\mathcal{O}$ is $h(\mathcal{O})$.

Lemma 9.2.25. *Let V be a volcano in the belt $\mathcal{B}_{t,m}$, where t corresponds to a cordillera of depth d. Let K be the associated field and $\mathcal{O} = \mathbb{Z} + m\mathcal{O}_K$. Let c be the number of vertices in the crater of V. The number of vertices in the volcano V is*

$$c + \frac{2c}{\#\mathcal{O}^\times}\left(\left(\ell - \left(\frac{D(\mathcal{O})}{\ell}\right)\right)\frac{\ell^d - 1}{\ell - 1}\right).$$

Proof. Use Proposition 9.2.17 in the regular case, and Proposition 9.2.19 and Proposition 9.2.20 in the non-regular case. One can compute c by looking at Proposition 9.2.14. □

Remark 9.2.26. There are exactly p j-invariants in $\mathbb{F}_p$. Therefore, by taking into account those that correspond to supersingular curves (for which there is a nice formula), we can partition p and obtain a mass formula. The interested reader can find a full study in the case $p = 1009$ and $\ell = 3$ in the appendix of [BCP24].

9.3 The Inverse Volcano Problem

9.3.1 Abstract volcanoes

Inspired by the structure of the connected components of $\mathcal{G}_\ell(\mathbb{F}_p)$, we give the following definition.

Definition 9.3.1. An *abstract volcano* $V = (\mathcal{V}, \mathcal{E})$ of *depth* $d \geq 0$ is a connected undirected graph together with a distinguished subset $\mathcal{V}_0 \subseteq \mathcal{V}$ such that the subgraph V_0 induced by $\mathcal{V}_0$ is one of the graphs described in Proposition 9.2.14. Moreover, $d > 0$ if and only if $\mathcal{V} \setminus \mathcal{V}_0 \neq \emptyset$, in which case there exists a partition

$$\mathcal{V} \setminus \mathcal{V}_0 = \bigcup_{i=1}^{d} \mathcal{V}_i$$

and a prime number $\ell \in \mathbb{Z}$ such that, denoting by V_i the subgraph induced by $\mathcal{V}_i$, the following holds:

1. All vertices in $\mathcal{V}_0 \cup ... \cup \mathcal{V}_{d-1}$ have degree $\ell + 1$ and all vertices in $\mathcal{V}_d$ have degree 1;
2. If $v \in \mathcal{V}_r$ and $v' \in \mathcal{V}_k$ are connected by an edge, then $|r - k| \leq 1$;
3. For all $0 < r \leq d$, the graph V_r is totally disconnected;
4. For $0 < r \leq d$, each vertex in V_r has exactly one edge to a vertex in V_{r-1};

We call V_0 the *crater* of V and we say that the vertices in $\mathcal{V}_r$ lie at *level* r.

If the depth d of an abstract volcano V is strictly positive, then the prime ℓ appearing in Definition 9.3.1 is uniquely determined by condition (1) above. In this case, we will also say that V is an (abstract) ℓ-volcano.

It follows from the discussion in Section 9.2 that a connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$ not containing 0 or 1728 is an abstract volcano in the sense above. One may wonder whether the converse is also true. We restate Question 9.1.1 from the introduction: let V be an abstract volcano as in Definition 9.3.1. Can we find primes $p, \ell \in \mathbb{Z}$ with $p \neq \ell$ such that V is a connected component in the isogeny graph $\mathcal{G}_\ell(\mathbb{F}_p)$?

Roughly speaking, one can study this question by dividing it into two subcases:

1. The volcano V has depth $d = 0$ i.e. $V = V_0$;

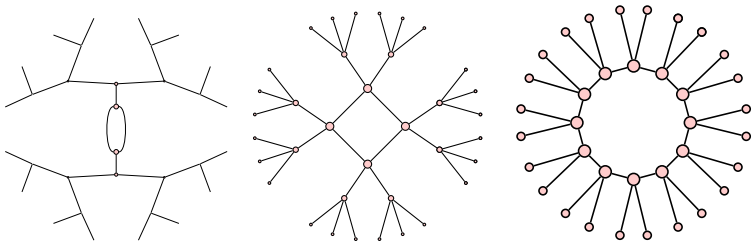


Figure 9.5: Three abstract volcanoes.

2. The volcano V has depth $d > 0$ i.e. $V \neq V_0$.

These two cases are fundamentally different in nature. Indeed, in the first case we can try to find (and we will find) the volcano V as connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$ where both ℓ and p are allowed to vary. However, in the second case the prime ℓ is fixed, since it must be one less than the degree of any vertex in V_0. Hence, in this second setting we have less freedom of choice and the inverse volcano problem becomes more difficult.

To study Question 9.1.1, it is useful to introduce a couple of definitions. First of all note that, letting V_0 be one of the graphs described in Proposition 9.2.14, ℓ a prime number and $d \in \mathbb{Z}_{>0}$, there exists a unique abstract ℓ-volcano V with crater V_0 and depth d. We call $V(V_0, \ell, d)$ the *volcano induced by the triple* (V_0, ℓ, d). We also give the following definition of abstract crater associated with a prime ideal.

Definition 9.3.2. Let $\mathcal{O}$ be an imaginary quadratic order and $\ell \in \mathbb{Z}_{>0}$ a prime number not dividing the conductor of $\mathcal{O}$. Choose a prime ideal $\mathfrak{L} \subseteq \mathcal{O}$ lying above ℓ. The *abstract crater* V_0 *associated with* $\mathfrak{L}$ is the graph consisting of:

- A single vertex if ℓ is inert in $\mathcal{O}$;
- A single vertex with a self-loop if ℓ is ramified in $\mathcal{O}$ and $\mathfrak{L}$ is principal;
- A single vertex with two self-loops if ℓ splits in $\mathcal{O}$ and $\mathfrak{L}$ is principal;
- Two vertices connected with a single edge if ℓ ramifies in $\mathcal{O}$ and $\mathfrak{L}$ is not principal;
- A cycle of size the order of $[\mathfrak{L}]$ in the class group $\mathrm{Cl}(\mathcal{O})$ if ℓ splits in $\mathcal{O}$ and we are not in any of the previous cases.

The relationship between Definition 9.3.2, Proposition 9.2.14 and Question 9.1.1 is clear: if we want to realise an ℓ-volcano V as the connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$ for some prime p, it seems natural to begin by realising its crater V_0. By Proposition 9.2.14, a good start would be to find an imaginary quadratic field K whose ring of integers contains an ideal $\mathfrak{L} \mid \ell$ such that V_0 is equal to the abstract crater associated with $\mathfrak{L}$ (here we are evidently hoping to view the vertices of V_0 as j-invariants of CM elliptic curves with complex multiplication by the *maximal order* in K).

In fact, we will now prove that, in order to solve the inverse volcano problem, realising V_0 as a crater is the key step.

Proposition 9.3.3. *Let $\mathcal{O}$ be an order of discriminant $D(\mathcal{O}) < -4$ in an imaginary quadratic field K and let $\ell \in \mathbb{Z}_{>0}$ be a prime not dividing the conductor of $\mathcal{O}$. Let $\mathfrak{L} \subseteq \mathcal{O}$ be a prime lying above ℓ and let V_0 be the abstract crater associated with $\mathfrak{L}$. Then for every $d > 0$ there exist infinitely many primes $p = p(d) \in \mathbb{Z}_{>0}$ such that $\mathcal{G}_\ell(\mathbb{F}_p)$ contains the volcano $V(V_0, \ell, d)$ as connected component. Moreover, if $\ell \neq 2$ the same is true with $d = 0$.*

Proof. Given a positive integer d, our goal is to find infinitely many primes $p \in \mathbb{Z}_{>0}$ and pairs $(t, v) \in \mathbb{Z}^2$ such that $t \neq 0$, the prime power ℓ^d divides exactly v and $4p = t^2 - v^2 D(\mathcal{O})$. Indeed, suppose this is the case. Then of course we must have $-2\sqrt{p} \leq t \leq 2\sqrt{p}$ so by Lemma 9.2.4 the t-cordillera in $\mathcal{G}_\ell(\mathbb{F}_p)$ is non-empty. Since the order $\mathcal{O}$ certainly contains the order of discriminant $v^2 D(\mathcal{O})$ and ℓ does not divide the conductor of $\mathcal{O}$, by Proposition 9.2.17 combined again with Lemma 9.2.4 the t-cordillera possesses a connected component isomorphic to the volcano $V(V_0, \ell, d)$.

To find primes p and corresponding couples (t, v) as above, we proceed as follows. For every $k \in \mathbb{N}$, denote by H_k the ring class field of the order $\mathbb{Z}[\ell^k \sqrt{D(\mathcal{O})}] \subseteq \mathcal{O}$. We have that $H_k \subsetneq H_{k+1}$ for all $k \in \mathbb{N}$. To see this, one can simply use [Cox13, Corollary 7.28] to compute $[H_{k+1} : H_k]$: if $\ell \neq 2$ one has $[H_{k+1} : H_k] \in \{\ell - 1, \ell, \ell + 1\}$ for all $k \geq 0$, while in the case $\ell = 2$ one has $[H_{k+1} : H_k] = 2$ for all $k \geq 0$ since $\mathbb{Z}[2^k \sqrt{D(\mathcal{O})}]$ has always even discriminant.

In particular $H_d \subsetneq H_{d+1}$, so by the Chebotarëv density theorem there are infinitely many primes $p \nmid 2\ell D(\mathcal{O})$ that split completely in H_d but do not split completely in H_{d+1}. Using [Cox13, Theorem 9.4], we see that there exist $x, y \in \mathbb{Z}$ such that $p = x^2 - D(\mathcal{O})\ell^{2d} y^2$. Moreover, we also have $\ell \nmid y$, since otherwise there would exist $\tilde{y} \in \mathbb{Z}$ such that $p = x^2 - D(\mathcal{O})\ell^{2d+2}\tilde{y}^2$ and then p would split completely (again by [Cox13, Theorem 9.4]) in H_{d+1}, contradicting our assumptions. Now such a p satisfies the norm equation

$$4p = t^2 - v^2 D(\mathcal{O})$$

where $t = 2x$ and $v = 2\ell^d y$. We certainly have $t \neq 0$ since p is split in K by our choices. Moreover, if $\ell \neq 2$, the power ℓ^d divides exactly v so in this case the theorem is proved.

If $\ell = 2$ then the same arguments work by considering primes splitting completely in H_{d-1} but not in H_d (here we use the fact that $d > 0$). This concludes the proof. □

Remark 9.3.4. In the above proof we have chosen a down-to-earth approach, proving the existence of the volcano induced by (V_0, ℓ, d) in $\mathcal{G}_\ell(\mathbb{F}_p)$ by solving an equation of the form $4p = t^2 - v^2\ell^{2d} D(\mathcal{O})$ with $t, v \in \mathbb{Z}$. We could have been more sophisticated and argued as follows: if one manages to find a prime that splits completely in the ring class field of $\mathbb{Z} + \ell^d\mathcal{O}$ but not in the ring class field of $\mathbb{Z} + \ell^{d+1}\mathcal{O}$, then the residue field at p will certainly contain all j-invariants of elliptic curves with CM by $\mathbb{Z} + \ell^d\mathcal{O}$ but no j-invariant of elliptic curves with CM by $\mathbb{Z} + \ell^{d+1}\mathcal{O}$ (cfr. Corollary 3.5.18) and this would ensure the existence of the desired volcano. By the Chebotarëv density theorem, this can be achieved if and only if the two aforementioned ring class fields are distinct, which certainly happens if $\ell \neq 2$ or $\ell = 2$ and $d > 0$.

Ultimately, the difference in the two proofs lies in the following fact: in the first proof we have solved the equation

$$4p = t^2 - v^2\ell^{2d} D(\mathcal{O}) \tag{9.4}$$

by first solving the auxiliary equation

$$p = x^2 - y^2\ell^{2d} D(\mathcal{O}) \tag{9.5}$$

and then multiplying by 2 the solutions x, y. However, (9.4) may have a solution even if (9.5) does not. For instance, the prime 3 is not of the form $x^2 + 11y^2$, but we have $4 \cdot 3 = 1^2 + 11 \cdot 1^2$. The proof appearing in this remark directly solves equation (9.4) without expressing p itself in the form $t^2 - v^2\ell^{2d} D(\mathcal{O})$. This may be useful in view of explicit computations, since the smallest prime p solving (9.4) is smaller or equal than the smallest prime solving (9.5).

9.3.2 Depth $d = 0$

Let us now analyse more closely the first instance of the inverse volcano problem, that is, the case when our given volcano V coincides with its crater V_0. We will now answer Question 9.1.1 in this case.

Proof of Theorem 9.1.2. We refer to the possible shapes of $V = V_0$ as described in Proposition 9.2.14. For each of the cases appearing in the proposition, we first want to find an imaginary quadratic field K and a prime $\mathfrak{L} \subseteq \mathcal{O}_K$ such that V is the abstract crater associated with $\mathfrak{L}$. In cases $(1) - (5)$ it is easy to find such a field K of discriminant $D(\mathcal{O}_K) < -4$ and a prime ideal $\mathfrak{L}$ with odd residue characteristic.

To deal with the case where V is a cycle of length $n \geq 3$ we appeal to [Yam70, Theorem 2], which ensures the existence of an imaginary quadratic field K of discriminant < -4 whose class group contains an element of order n. Since by [Cox13, Theorem 9.12] every ideal class contains infinitely many prime ideals, we deduce that there exists a prime $\ell > 2$ that splits in $\mathcal{O}_K$ into two prime ideals of order n in the class group.

Applying Proposition 9.3.3 now allows us to conclude. □

9.3.3 Depth $d > 0$

We now turn to the second, more difficult instance of the inverse volcano problem *i.e.* the case when V has depth $d > 0$. Given an ℓ-volcano V of depth $d > 0$, realising its crater V_0 as an abstract crater amounts to finding an imaginary quadratic order of conductor coprime to ℓ where there exists a prime ideal $\mathfrak{L} \supseteq \ell$ satisfying the condition in Proposition 9.2.14 corresponding to V_0. In order to apply Proposition 9.3.3 we may also want the discriminant of the order to be smaller than -4. So let us fix ℓ prime and see if we manage to find imaginary quadratic orders that realise the conditions expressed in Proposition 9.2.14:

1. By Dirichlet's theorem on primes in arithmetic progression, there are infinitely many imaginary quadratic fields where ℓ is inert;
2. If $\ell \neq 3$, then ℓ ramifies in a principal ideal in the ring of integers of $\mathbb{Q}(\sqrt{-\ell})$ and the latter has discriminant < -4. If $\ell = 3$ the same is true if we consider the order $\mathbb{Z}[\sqrt{-3}]$ of conductor 2 in $\mathbb{Q}(\sqrt{-3})$;
3. In the imaginary quadratic field $K = \mathbb{Q}(\sqrt{1-4\ell})$ the integral element $\alpha = \frac{1+\sqrt{1-4\ell}}{2}$ has norm ℓ. We deduce that ℓ splits in $\mathcal{O}_K$ into the two principal prime ideals generated by α and $\overline{\alpha}$;
4. If q is a prime that is sufficiently large with respect to ℓ then in the ring of integers of $K = \mathbb{Q}(\sqrt{-\ell q})$ the prime ℓ is ramified into a non-principal ideal. This follows from the fact that every element $\alpha \in \mathcal{O}_K \setminus \mathbb{Z}$ has norm $N_{K/\mathbb{Q}}(\alpha) \geq (1 + \ell q)/4$.

Conditions (5) and (6) in Proposition 9.2.14 are more obscure, as they require to construct an imaginary quadratic field K where ℓ splits into two prime ideals whose class in the ideal class group of K has prescribed order n. We now prove that it is always possible to find such a field. Our construction is inspired by the techniques used by Nagell in [Nag55] and very much depends on whether $\ell = 2$ or $\ell > 2$. We will use in particular a theorem of Mahler [Nag55, Theorem 16], which we recall here for convenience.

Theorem 9.3.5. *Let $D > 1$ be a positive integer which is not a perfect square, and let C be a square-free divisor of $2D$, with $|C| \neq 1$ and $C \neq D$. Let U, V be two natural integers satisfying the equation*

$$U^2 - DV^2 = C. \tag{9.6}$$

If all prime factors of V divide D, we have either

$$U = U_1, \quad V = V_1, \tag{9.7}$$

or

$$U = \frac{U_1^3 + 3U_1V_1^2D}{|C|}, \quad V = \frac{3U_1^2V_1 + DV_1^3}{|C|}, \tag{9.8}$$

where U_1, V_1 denote the least positive solutions in integers of (Equation 9.6). The numbers U and V in Equation 9.8 are determined by the formula

$$\frac{U + V\sqrt{D}}{\sqrt{|C|}} = \left(\frac{U_1 + V_1\sqrt{D}}{\sqrt{|C|}}\right)^3. \tag{9.9}$$

We now begin by treating the case $\ell = 2$.

Proposition 9.3.6. *Let $n \neq 4$ be a positive integer and let $K = \mathbb{Q}(\sqrt{1-2^{n+2}})$. Then in $\mathcal{O}_K$ the prime 2 splits into two prime ideals whose corresponding classes in $\mathrm{Cl}(\mathcal{O}_K)$ have order n.*

Proof. Some preliminary remarks: write $\sqrt{1-2^{n+2}} = x \cdot \sqrt{-A}$ with $x, A \in \mathbb{Z}$ and $A > 0$ square-free. In particular, we can write

$$Ax^2 + 1 = 2^{n+2}. \tag{9.10}$$

Moreover, since $n \geq 1$ we know that x is odd and we thus also have

$$1 \equiv 1 - 2^{n+2} \equiv x^2 \cdot (-A) \equiv -A \bmod 8 \tag{9.11}$$

so that $D(\mathcal{O}_K) = -A$ and $2\mathcal{O}_K$ splits into two distinct conjugate prime ideals $\mathfrak{p}_2$ and $\overline{\mathfrak{p}}_2$.

Consider now the two conjugate principal ideals $\mathfrak{a} := \left(\frac{1+x\sqrt{-A}}{2}\right)$ and $\overline{\mathfrak{a}}$. We have

$$\mathfrak{a} \cdot \overline{\mathfrak{a}} = N_{K/\mathbb{Q}}\left(\frac{1+x\sqrt{-A}}{2}\right) = \frac{Ax^2+1}{4} = 2^n$$

hence these two ideals can be divisible only by $\mathfrak{p}_2$ and $\overline{\mathfrak{p}}_2$. Moreover, the ideals $\mathfrak{a}$ and $\overline{\mathfrak{a}}$ are coprime, as one can see by adding their generators. Hence, we can assume without loss of generality that the prime ideal factorization of $\mathfrak{a}$ is

$$\mathfrak{a} = \left(\frac{1+x\sqrt{-A}}{2}\right) = \mathfrak{p}_2^n$$

and in particular, the class of $\mathfrak{p}_2$ in $\mathrm{Cl}(\mathcal{O}_K)$ has order dividing n. Our goal for the rest of this proof is to prove that this order is precisely n.

Assume by contradiction that this is not the case. Then there exists a prime q and $\tilde{u}, \tilde{v} \in \mathbb{Z}$ such that the following equality of ideals holds:

$$\left(\frac{\tilde{u}+\tilde{v}\sqrt{-A}}{2}\right)^q = \left(\frac{1+x\sqrt{-A}}{2}\right). \tag{9.12}$$

We now distinguish two cases.

First case: q odd. We begin by noting that $A \neq 3$ since by (9.11) we have $-A \equiv 1 \bmod 8$. This implies that $\mathcal{O}_K^\times = \{\pm 1\}$. Since q is odd, we obtain that all the units in $\mathcal{O}_K$ are q-th powers. Hence, there exist $u, v \in \mathbb{Z}$ with $u \equiv v \bmod 2$ such that the following equality *of elements of K* holds:

$$\left(\frac{u+v\sqrt{-A}}{2}\right)^q = \frac{1+x\sqrt{-A}}{2}. \tag{9.13}$$

Expanding the left-hand side and collecting the rational terms together, we reach the equality

$$u^q - \binom{q}{2}u^{q-2}v^2A + \ldots + \binom{q}{q-1}uv^{q-1}(-A)^{\frac{q-1}{2}} = 2^{q-1}. \tag{9.14}$$

Clearly, (9.14) implies that u divides 2^{q-1}. Suppose initially that u is even *i.e.* $u = 2c$ for some $c \in \mathbb{Z}$. Since $u \equiv v \bmod 2$ we can also write $v = 2d$ for some $d \in \mathbb{Z}$. Plugging $2c$ and $2d$ in place of u and v in equation (9.14), one obtains that 2^q divides 2^{q-1}, a contradiction.

Hence, u must be odd, *i.e.* $u = \pm 1$. Reducing (9.14) modulo q we obtain

$$(\pm 1)^q \equiv 2^{q-1} \equiv 1 \bmod q,$$

so, since q is odd, we in fact have $u = 1$. In particular, equality (9.13) now reads

$$\left(\frac{1 + v\sqrt{-A}}{2}\right)^q = \frac{1 + x\sqrt{-A}}{2}$$

and taking norms we obtain

$$\left(\frac{1 + v^2 A}{4}\right)^q = \frac{Ax^2 + 1}{4}.$$

After some manipulations, this can be rewritten as follows:

$$A^2x^2 - (A + v^2A^2)\left[\left(\frac{1 + v^2A}{4}\right)^{\frac{q-1}{2}}\right]^2 = -A$$

so we reach an equation of the form

$$U^2 - DV^2 = -A \tag{9.15}$$

where $U = Ax$, $D = A + v^2A^2$ and $V = [(1+v^2A)/4]^{(q-1)/2}$. We now wish to apply Theorem 9.3.5, since all the prime factors of V divide D. Let us verify that the hypotheses of the theorem are satisfied:

- We certainly have $A \neq 1$ since $-A \equiv 1 \bmod 8$. This also implies that $A \neq D$. Moreover A is square-free by hypothesis;
- By the previous bullet point we have $D = A(1 + v^2A) > 1$. We know that $v \neq 0$ because $A \neq D$. As $A > 1$ is square-free, then $D > 1$ is not a perfect square: indeed if it was, then a prime dividing A would divide D with even exponent, hence would divide both A and $1 + v^2A$, which is impossible;

Hence, Theorem 9.3.5 implies that the only positive solutions to (9.15) can be given by the least positive solution $(U, V) = (U_1, V_1)$ and by

$$U = \frac{U_1^3 + 3U_1V_1^2D}{A}, \qquad V = \frac{3U_1^2V_1 + DV_1^3}{A}.$$

In our case, the fundamental solution is $U = vA$ and $V = 1$. This latter equality yields in particular $v^2A = 3$, hence $A = 3$, which is not possible.

By looking at the V of the non-fundamental solution, we find

$$\left(\frac{1 + v^2A}{4}\right)^{\frac{q-1}{2}} = V = \frac{3v^2A^2 + D}{A} = 1 + 4v^2A = 4(1 + v^2A) - 3.$$

In order to conclude, note that since v is odd, we have

$$1 + v^2A \equiv 1 + 1 \cdot 7 \equiv 0 \bmod 8$$

so in particular the left-hand side of the above equality is even. However, the right-hand side of the equality is odd, contradiction. This concludes the proof in this case.

Second case: $q = 2$. By (9.11) we have $A \neq 1, 3$ and so $\mathcal{O}_K^\times = \{\pm 1\}$. Hence the equality of ideals (9.12) yields an equality of elements of K

$$\left(\frac{u + v\sqrt{-A}}{2}\right)^2 = \pm\frac{1 + x\sqrt{-A}}{2}$$

where $u, v \in \mathbb{Z}$ are such that $u \equiv v \bmod 2$ (here we have simply set $u = \tilde{u}$ and $v = \tilde{v}$). Expanding and looking at rational/irrational parts gives

$$\begin{cases} u^2 - v^2 A & = \pm 2 \\ 2uv & = \pm 2x. \end{cases}$$

By substituting $v = \pm x/u$ in the first equation and expanding we get

$$u^4 \mp 2u^2 - Ax^2 = 0,$$

and solving this quadratic equation gives, after using (9.10)

$$u^2 = 2^{n/2+1} \pm 1.$$

Looking modulo 4 one sees that $u^2 = 2^{n/2+1} - 1$ cannot hold, as $n \geq 2$. Hence we must have $u^2 = 2^{n/2+1} + 1$, or otherwise written $(u-1)(u+1) = 2^{n/2+1}$. This implies $u = \pm 3$ and $n = 4$. However, this case is excluded by our assumptions and the theorem is proved. □

One can directly verify that in $\mathbb{Q}(\sqrt{-39})$ the prime 2 splits into two prime ideals having order 4 in the class group. This observation and the previous proposition imply that for every $n \in \mathbb{Z}_{>0}$ there exists an imaginary quadratic field K where 2 splits into two prime ideals having order n in $\mathrm{Cl}(\mathcal{O}_K)$.

We now treat the case when ℓ is odd. We will prove that for every $n \in \mathbb{Z}_{>0}$ at least one among the imaginary quadratic fields $\mathbb{Q}(\sqrt{1-\ell^n})$ and $\mathbb{Q}(\sqrt{1-4\ell^n})$, call it K, has the property that the prime ℓ splits in $\mathcal{O}_K$ into two prime ideals having order n in $\mathrm{Cl}(\mathcal{O}_K)$. Let us begin by studying these fields separately.

Proposition 9.3.7. *Let $\ell \in \mathbb{N}$ be an odd prime and let $n \in \mathbb{Z}_{>0}$. Define $K := \mathbb{Q}(\sqrt{1-\ell^n})$. Suppose that:*

1. *Either $n \geq 3$ is odd and $(\ell, n) \neq (3, 5)$;*
2. *Or n is even and neither $\frac{\ell^{n/2}+1}{2}$ nor $\frac{\ell^{n/2}-1}{2}$ is a square;*

Then in $\mathcal{O}_K$ the prime ℓ splits into two prime ideals whose corresponding classes in $\mathrm{Cl}(\mathcal{O}_K)$ have order n.

Proof. We proceed as in the proof of Proposition 9.3.6. Write $\sqrt{1-\ell^n} = x \cdot \sqrt{-A}$ with $x, A \in \mathbb{Z}$ and $A > 0$ square-free, so that we have

$$Ax^2 + 1 = \ell^n. \tag{9.16}$$

This equation implies in particular that $-A$ is a square modulo ℓ, so that $\ell\mathcal{O}_K = \mathfrak{p}_\ell\overline{\mathfrak{p}}_\ell$ with $\mathfrak{p}_\ell, \overline{\mathfrak{p}}_\ell \subseteq \mathcal{O}_K$ distinct prime ideals. Consider the two conjugate principal ideals $\mathfrak{a} := (1 + x\sqrt{-A})$ and $\overline{\mathfrak{a}}$. We have

$$\mathfrak{a} \cdot \overline{\mathfrak{a}} = N_{K/\mathbb{Q}}(1 + x\sqrt{-A}) = \ell^n.$$

Since ℓ is odd, the ideals $\mathfrak{a}$ and $\overline{\mathfrak{a}}$ are coprime, so we have, without loss of generality, $\mathfrak{a} = \mathfrak{p}_\ell^n$. In particular, the class of $\mathfrak{p}_\ell$ in $\mathrm{Cl}(\mathcal{O}_K)$ has order dividing n. If this order is not precisely n, then

there exists a prime q and $u, v \in \mathbb{Z}$ with $u \equiv v \bmod 2$ such that the following equality of ideals holds:

$$(1 + x\sqrt{-A}) = \left(\frac{u + v\sqrt{-A}}{2}\right)^q. \tag{9.17}$$

Suppose first that q is odd. Then, the proof of [Nag55, Theorem 25] shows that we must have $x = 11$, $A = 2$, $\ell = 3$ and $q = 5$. From (9.16), we deduce that $q = n = 5$, which contradicts assumption (1).

Hence, we must have $q = 2$ and, in particular, n is even. Reducing equation (9.16) modulo 4 and using that A is square-free, we see that x is even, say $x = 2y$ for $y \in \mathbb{Z}$. Now we can write

$$Ay^2 = \left(\frac{\ell^{n/2} - 1}{2}\right)\left(\frac{\ell^{n/2} + 1}{2}\right). \tag{9.18}$$

The factors on the right hand side are consecutive, hence coprime, integers. By assumption (2) neither of them can be a square, and we deduce that A must be divisible by at least two different primes. In particular, $A > 3$ and $\mathcal{O}_K^\times = \{\pm 1\}$.

Now from (9.17) we get the following equality of elements of K:

$$\left(\frac{u + v\sqrt{-A}}{2}\right)^2 = \pm(1 + x\sqrt{-A})$$

Expanding and looking at rational/irrational parts gives

$$\begin{cases} u^2 - v^2 A & = \pm 4 \\ 2uv & = \pm 4x. \end{cases}$$

From the second equation we get $u \equiv v \equiv x \equiv 0 \bmod 2$. So writing $(u', v', y) = (\frac{u}{2}, \frac{v}{2}, \frac{x}{2})$ and proceeding as in the second case of Proposition 9.3.6, we get

$$u^2 = \frac{\pm 1 + \sqrt{1 + 4y^2 A}}{2} = \frac{\pm 1 + \ell^{n/2}}{2},$$

which is excluded by our hypotheses. This concludes the proof. □

Proposition 9.3.8. *Let $\ell \in \mathbb{N}$ be an odd prime, and n an even positive integer. Define $K := \mathbb{Q}(\sqrt{1 - 4\ell^n})$. If $\ell^{n/2}$ is not the sum two consecutive squares, then ℓ splits in $\mathcal{O}_K$ into two prime ideals whose classes in $\mathrm{Cl}(K)$ have order n.*

Proof. Write $\sqrt{1 - 4\ell^n} = x \cdot \sqrt{-A}$ with $x, A \in \mathbb{Z}$ and $A > 0$ square-free. In particular,

$$Ax^2 + 1 = 4\ell^n. \tag{9.19}$$

We have that x is odd as it divides $4\ell^n - 1$. Thus $A \equiv 3 \bmod 8$ and so $A \neq 1$. Suppose we have $A = 3$. Then (9.19) becomes

$$3x^2 = (2\ell^{n/2} - 1)(2\ell^{n/2} + 1).$$

Both factors on the right hand side are consecutive odd integers, so they are coprime. Hence one of them must be a perfect square and the other three times a perfect square. Suppose that $2\ell^{n/2} - 1$ is a square. Since it is odd, we would have $2\ell^{n/2} - 1 = (2j + 1)^2$ which is equivalent to $\ell^{n/2} = j^2 + (j + 1)^2$, contradicting our assumptions. This means that there exists $k \in \mathbb{Z}_{>0}$ such that $2\ell^{n/2} + 1 = k^2$. However, reducing this equality modulo 4 shows that this cannot happen either and we conclude that $A \neq 3$. In particular, $\mathcal{O}_K^\times = \{\pm 1\}$.

Equality (9.19) implies that ℓ splits in $\mathcal{O}_K$ into distinct conjugate prime ideals $\mathfrak{p}_\ell$ and $\overline{\mathfrak{p}}_\ell$. Now consider the conjugate principal ideals $\mathfrak{a} = \left(\frac{1+x\sqrt{-A}}{2}\right)$ and $\overline{\mathfrak{a}}$. We have

$$\mathfrak{a}\overline{\mathfrak{a}} = N_{K/\mathbb{Q}}\left(\frac{1+x\sqrt{-A}}{2}\right) = \frac{1+Ax^2}{4} = \ell^n,$$

and $\mathfrak{a} + \overline{\mathfrak{a}} = \mathcal{O}_K$, so we can assume without loss of generality that $\mathfrak{a} = \mathfrak{p}_\ell^n$. Once again the class of $\mathfrak{p}_\ell$ in $\mathrm{Cl}(K)$ has order dividing n and we want to prove this order is exactly n. Assume by contradiction it is not the case. Then there exists a prime divisor q of m and $u, v \in \mathbb{Z}$ with $u \equiv v \bmod 2$ such that the following equality of ideals holds:

$$\left(\frac{u+v\sqrt{-A}}{2}\right)^q = \left(\frac{1+x\sqrt{-A}}{2}\right).$$

If q is odd we use the same argument as in the first part of the proof of Proposition 9.3.6, using that $A \neq 1, 3$ to rule out this case. The argument goes through unchanged up until the final part, when we reach the equality

$$\left(\frac{1+v^2A}{4}\right)^{\frac{q-1}{2}} = 4(1+v^2A) - 3$$

with $v \in \mathbb{Z}$ odd. In Proposition 9.3.6 here we concluded by using the fact that $A \equiv 7 \bmod 8$ in that setting. In the current setting however, the congruence $A \equiv 3 \bmod 8$ does not yield any contradiction. Instead to conclude one can notice that after setting $z = \frac{1+v^2A}{4}$, the above equation becomes

$$z^{\frac{q-1}{2}} = 16z - 3$$

which does not have any integral solution.

Hence, we can assume that $q = 2$. Again using the fact that $\mathcal{O}_K^\times = \{\pm 1\}$, we have the following equality of elements of K:

$$\left(\frac{u+v\sqrt{-A}}{2}\right)^2 = \pm\frac{1+x\sqrt{-A}}{2}$$

where $u, v \in \mathbb{Z}$ are such that $u \equiv v \bmod 2$. With the usual arguments we arrive at

$$u^2 = \frac{\pm 2 + \sqrt{4+4x^2A}}{2} = \pm 1 + \sqrt{1+x^2A} = \pm 1 + 2\ell^{m/2}.$$

As we showed above this is impossible and the proof is concluded. □

One can directly verify that in $\mathbb{Q}(\sqrt{-971}) = \mathbb{Q}(\sqrt{1-4\cdot 3^5})$ the prime 3 splits into two prime ideals of order 5 in the class group. This observation and the two previous propositions imply that for every $n \in \mathbb{Z}_{>0}$ and every odd prime ℓ there exists an imaginary quadratic field K where ℓ splits into two prime ideals having order n in $\mathrm{Cl}(\mathcal{O}_K)$.

We are now ready to prove Theorem 9.1.4.

Proof of Theorem 9.1.4. By Proposition 9.3.6, Proposition 9.3.7 and Proposition 9.3.8 and the comments in between, it suffices to show that for every even $n \in \mathbb{N}$ the two sets

$$E_1(n) := \left\{\ell > 2 \text{ prime} : \frac{\ell^{n/2}-1}{2} = j^2 \text{ or } \frac{\ell^{n/2}+1}{2} = j^2 \text{ for some } j \in \mathbb{N}\right\},$$

$$E_2(n) := \left\{\ell > 2 \text{ prime} : \ell^{n/2} = j^2 + (j+1)^2 \text{ for some } j \in \mathbb{N}\right\}$$

have empty intersection. Let $\ell \in E_1(n) \cap E_2(n)$. Then there exists $j \in \mathbb{N}$ such that $\ell^{n/2} = j^2 + (j+1)^2$. We have

$$\frac{\ell^{n/2} - 1}{2} = j^2 + j,$$

and

$$\frac{\ell^{n/2} + 1}{2} = j^2 + j + 1.$$

Suppose there exists $k \in \mathbb{N}$ such that $k^2 = \frac{\ell^{n/2}-1}{2}$. Then $(k-j)(k+j) = j$, so $k+j$ divides j, which is impossible as $j, k > 0$. Hence, since $\ell \in E_1(n)$, there exists $k \in \mathbb{N}$ such that $k^2 = \frac{\ell^{n/2}+1}{2}$, that is, $(k-j)(k+j) = j+1$. Now we must have $k+j$ divides $j+1$ yielding $k = 1$ and $j = 0$. This is impossible, and the corollary follows. □

We can now prove Theorem 9.1.3, which in particular shows that the inverse volcano problem over $\mathbb{F}_p$ is always solvable.

Proof of Theorem 9.1.3. The first part of the theorem follows by combining Proposition 9.3.3 and Theorem 9.1.4. The abstract 3-volcanoes whose crater consists of a single vertex with a self loop cannot be realised in such a way that the vertices on their crater correspond to elliptic curves with complex multiplication by maximal orders. This is because the only imaginary quadratic field where 3 ramifies into a principal ideal is $K = \mathbb{Q}(\sqrt{-3})$ and any elliptic curve with complex multiplication by $\mathcal{O}_K$ has j-invariant 0, which we have excluded from our considerations. On the other hand, as explained in the discussion preceding Theorem 9.3.5, this is the only case where this realisation is not possible. This concludes the proof. □

9.4 Minimal Characteristic Volcanoes and How to Find Them

We have shown that for any given abstract volcano V it is possible to explicitly find some ordinary isogeny graph $\mathcal{G}_\ell(\mathbb{F}_p)$ that contains V as a connected component. But this does not give any guarantee on the size of p that works, or on the discriminant of the endomorphism ring of curves at the crater of such a volcano. For instance, by Proposition 9.3.8 the prime 3 splits into two prime ideals having order 5 in the ideal class group of $\mathbb{Q}(\sqrt{-971})$, where $-971 = 1 - 4 \cdot 3^5$. However, the same is true for the field $\mathbb{Q}(\sqrt{-47})$, whose discriminant is more than 20 times smaller in absolute value.

Question 9.4.1. *For a given abstract volcano V, what is the minimal prime p such that V is realised as a connected component in a $\mathcal{G}_\ell(\mathbb{F}_p)$?*

A first very naïve algorithm to solve the minimal prime question is to list all roots of Φ_ℓ mod p for increasing values of p, thereby constructing $\mathcal{G}_\ell(\mathbb{F}_p)$ in its entirety. For each value of p, one would need to exclude all supersingular j-invariants, and then proceed to use a depth first search or equivalent algorithm to detect cycles of length c, where c is the crater size.

Question 9.4.1 does not seem to have an easy answer, as it in fact asks to make the Chebotarev density theorem effective, which in general in a difficult task in analytic number theory. For this reason we focus instead on the following amusing computational question[1] that helps illustrate how one would find the minimal primes in practice.

Question 9.4.2. *An abstract 3-volcano is called a* cycloalkane *if it has depth 1 and its crater is a cycle of size $c \geq 3$ (we then use the notation C_cH_{2c}). For a given integer $c \geq 3$, what is the minimal prime p such that the cycloalkane C_cH_{2c} is realised as a connected component in $\mathcal{G}_3(\mathbb{F}_p)$?*

[1] This question is due to Fabien Pazuki

To help answer this question, we prove the following two easy lemmas.

Lemma 9.4.3. *Let ℓ be a prime and $n \in \mathbb{Z}_{\geq 1}$ a fixed integer. Suppose that $\mathcal{O}$ is an imaginary quadratic order of conductor coprime to ℓ where ℓ splits into two prime ideals having order n in $\mathrm{Cl}(\mathcal{O})$. Then $|D(\mathcal{O})| \leq 4\ell^n - 1$.*

Proof. Set $K := \mathrm{Frac}(\mathcal{O})$ and let $\mathfrak{L}$ and $\overline{\mathfrak{L}}$ be the two distinct prime ideals lying above ℓ. Then by assumption there exists $\alpha \in \mathcal{O}$ such that $\mathfrak{L}^n = (\alpha)$. We have that $\alpha \notin \mathbb{Z}$ since otherwise n would be even and $\alpha = \pm\ell^{n/2}$, implying that $\mathfrak{L} = \overline{\mathfrak{L}}$. The lemma now follows from the fact that every element $\beta \in \mathcal{O} \setminus \mathbb{Z}$ satisfies $N_{K/\mathbb{Q}}(\beta) \geq \frac{1+|D(\mathcal{O})|}{4}$. □

As in our proof of the inverse problem, our strategy is to first find an appropriate discriminant (or equivalently an appropriate imaginary quadratic order), which we then use to derive a prime number that satisfies the problem condition. Heuristically, one would expect small primes to arise only when the discriminant associated to the imaginary quadratic order is also small. The above lemma allows us to limit the search to a bounded set of possible discriminants.

In some cases, Lemma 9.4.3 sometimes leads to smaller (*i.e.* with associated imaginary quadratic field that has smaller discriminant, or with smaller characteristic p) solutions than the ones provided by Theorem 9.1.4.

Example 9.4.4. *Let V_0 be a cycle of length 6 and let V be the volcano induced by $(V_0, 2, 1)$, as in Figure 9.6.*

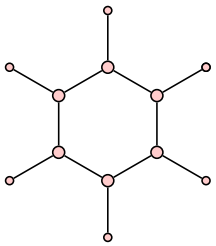


Figure 9.6: The abstract volcano V.

We claim that V is realisable as a connected component of $\mathcal{G}_2(\mathbb{F}_{103})$ (more precisely, as a connected component of the 8-cordillera in this graph). Indeed, using Lemma 9.4.3 one finds that $K = \mathbb{Q}(\sqrt{-87})$ is an imaginary quadratic field where $\ell = 2$ splits into two ideals having order 6 in $\mathrm{Cl}(\mathcal{O}_K)$. The prime $p = 103$ splits completely in the ring class field of $\mathbb{Z}[\sqrt{-87}]$ but does not split completely in the ring class field of $\mathbb{Z}[2\sqrt{-87}]$. We have

$$4 \cdot 103 = 8^2 + 87 \cdot 2^2$$

and the results explained in Section 9.3 now imply the result.

The second lemma below shows that if we have found a prime p that works, then this allows us to bound the candidate discriminants.

Lemma 9.4.5. *Let $p \in \mathbb{Z}_{>0}$ be a prime such that the abstract volcano $V = V(c, \ell, d)$ is a connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$ for some value of c. Then for any prime $p' < p$ such that V is a connected component of $\mathcal{G}_\ell(\mathbb{F}_{p'})$,*

$$|D(\mathcal{O})| < \frac{4p-1}{\ell^{2d}},$$

where $\mathcal{O}$ is the endomorphism ring at the crater of the volcano of $\mathcal{G}_\ell(\mathbb{F}_{p'})$ realising V.

Proof. If p' is a solution to the inverse problem for V, then the associated order $\mathcal{O}$ must satisfy the following equation:

$$4p' = t^2 - \ell^{2d} D(\mathcal{O}),$$

where t is the trace of the associated Frobenius endomorphism. $t \neq 0$ as the elliptic curves are not supersingular, which means that both t^2 and v^2 are at least 1, which in turn implies the result. □

From both lemmas we deduce the following algorithm:

Algorithm 7 Finding the minimal prime

Require: An abstract volcano $V = V(c, \ell, d)$, where $c \geq 2$ and $d \geq 1$. A bound B.
Ensure: The minimal prime realising V, if it is smaller than B.

1: $p_{min} \leftarrow B$; $b \leftarrow 0$
2: **for** Discriminants $D \equiv 0, 1 \pmod 4$ between -7 and $1 - 4\ell^c$ **do**
3: **if** $|D| > \frac{4p_{min}-1}{\ell^{2d}}$ **then**
4: **Return** $b \cdot p_{min}$
5: **else**
6: Compute the order $\mathcal{O}$ such that $D(\mathcal{O}) = D$
7: **if** a prime factor $\mathfrak{L}$ of (ℓ) in $\mathcal{O}$ has order c in $\mathrm{Cl}(\mathcal{O})$ **then**
8: Compute Hilbert class polynomial $H_1 = H_{\mathbb{Z}+\ell^{2d}\mathcal{O}}(X)$
9: Compute Hilbert class polynomial $H_2 = H_{\mathbb{Z}+\ell^{2(d+1)}\mathcal{O}}(X)$
10: **for** primes $p < p_{min}$ **do**
11: **if** $X^2 - D$ and H_1 split in $\mathbb{F}_p[X]$ but H_2 does not split **then**
12: $p_{min} \leftarrow p$; $b \leftarrow 1$
13: Exit inner **for** loop
14: **end if**
15: **end for**
16: **end if**
17: **end if**
18: **end for**
19: **Return** 0

Algorithm 7 relies on a guess B on the size of the minimal prime. If the guess it too low, the algorithm clearly fails and returns 0. Otherwise, its runtime will depend on how close the guess is.

Theorem 9.4.6. *Let p be the minimal prime realising the abstract volcano V, then if $p < B$, then Algorithm 7 terminates and returns p.*

Proof. Termination of the algorithm follows from Lemma 9.4.3 and Lemma 9.4.5. Indeed, let $\mathcal{O}$ be a winning order. Its discriminant $D(\mathcal{O})$ is no more than -7 as V is in a regular cordillera, and at least $1 - 4\ell^c$ by Lemma 9.4.3. If $p < B$, then by Lemma 9.4.5, the condition on line 3 will never be satisfied before p has been found. Indeed the variable p_{min} is always either B (in which case the condition is not satisfied), or a valid solution that triggers the inequality in the lemma. Therefore, one of the checked discriminants is the correct one. For such a D, the if condition on line 7 will trigger, and enumerating over all primes less than B will include the solution p. □

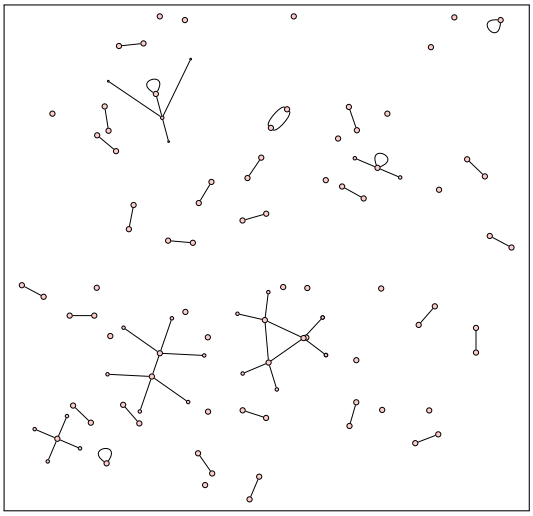
Remark 9.4.7. In the case where we want to compute multiple minimal primes for various abstract volcanoes, it will be useful to note the following:

- If the volcanoes share the same ℓ then the order of $\mathfrak{L}$ in $\mathrm{Cl}(\mathcal{O})$ can be precomputed or memoïsed;
- If the depth is also the same across all volcanoes, then the Hilbert class polynomial information can be precomputed or memoïsed.

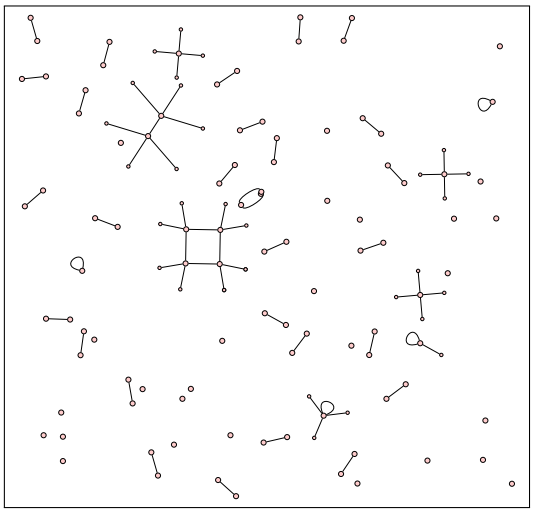
As we have no answer to Question 9.4.1 in general, we will not analyse the complexity of our algorithm in great detail. In fact, the complexity of each iteration of the main loop of Algorithm 7 is divided into two main contributions: the class group related computation on line 7, which is sub-exponential in $|D|$ using the Hafner-McCurley algorithm, and the loop on line 10 which runs in B times expected polynomial time using a randomised factoring algorithm for polynomials mod p. Therefore the overall runtime will heavily depend on the value of the minimal solution, which we do not estimate.

Example 9.4.8. *We encourage the reader to use the algorithm to work out that the three abstract volcanoes appearing in Figure 9.5 are connected components, respectively, of the isogeny graphs* $\mathcal{G}_2(\mathbb{F}_{1009})$, $\mathcal{G}_3(\mathbb{F}_{1303})$ *and* $\mathcal{G}_3(\mathbb{F}_{997})$.

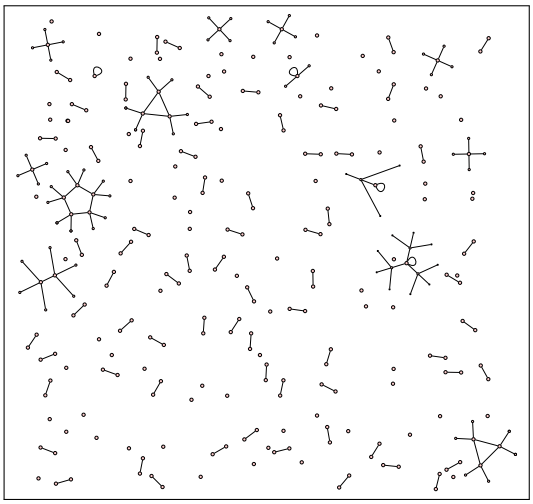
Minimal characteristic for cycloalkanes



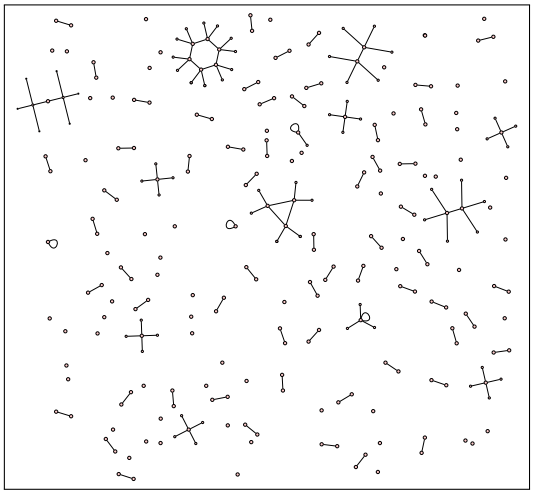
(a) The volcano park $\mathcal{G}_3(103)$



(b) The volcano park $\mathcal{G}_3(127)$



(c) The volcano park $\mathcal{G}_3(307)$



(d) The volcano park $\mathcal{G}_3(277)$

Figure 9.7: Minimal characteristic appearances of cycloalkanes for $c \in \{3, 4, 5, 6\}$.

To solve Question 9.4.2 for small[2] values of c, we run Algorithm 7 and get the following

[2]Cycloalkanes with larger values of c might not occur in nature.

results (we choose to include $c = 2$):

Table 9.1: Minimal primes realizing cycloalkanes C_cH_{2c}.

c	p	c	p	c	p	c	p
2	61	10	823	18	2293	26	5023
3	103	11	1483	19	2803	27	3967
4	127	12	997	20	1879	28	3637
5	307	13	1723	21	3943	29	6619
6	277	14	1009	22	2437	30	4597
7	571	15	1279	23	4231		
8	373	16	1597	24	3229		
9	547	17	2467	25	4327		

Remark 9.4.9. It might be interesting to note that not all values of the minimal p reach volcanoes that have maximal orders as crater endomorphism rings. Indeed for crater sizes $c \in \{3, 9, 15, 16, 18, 24, 27, 28\}$ we noted that this was not the case.

As an illustration, we represent below the full volcano park for cyclopropane, cyclobutane, cyclopentane and cyclohexane.

9.5 The Inverse Volcano Problem over $\mathbb{F}_{p^s}$ with $s > 1$

The inverse volcano problem over $\mathbb{F}_{p^s}$ with $s > 1$ does not always have a solution. An example where $s = 2$ is provided by the next proposition.

Proposition 9.5.1. *Let V_0 be a cycle of length 2 and let V be the abstract volcano induced by $(V_0, 2, 1)$, as in Figure 9.8. For every prime $p \neq 2$, the volcano V is not a connected component of $\mathcal{G}_2(\mathbb{F}_{p^2})$.*

Figure 9.8: The abstract volcano induced by $(V_0, 2, 1)$.

Proof. Let us first notice that the ring of integers $\mathcal{O}_K$ of $K := \mathbb{Q}(\sqrt{-15})$ is the only imaginary quadratic order where 2 splits into two ideals having order 2 in its class group (for example this can be seen using Lemma 9.4.3). Hence, if V were an isogeny volcano in characteristic p, the elliptic curves corresponding to the vertices on its crater would have necessarily complex multiplication by $\mathcal{O}_K$ and p would be split in it. Let H be the ring class field relative to the order $\mathcal{O} := \mathbb{Z}[2\sqrt{-15}]$. The natural exact sequence (see [Neu99b, Chapter I, Proposition 12.9])

$$1 \to (\mathcal{O}_K/4\mathcal{O}_K)^\times/\{\pm 1\} \to \mathrm{Cl}(\mathcal{O}) \to \mathrm{Cl}(\mathcal{O}_K) \to 1$$

splits, so we have $\mathrm{Gal}(H/K) \cong (\mathbb{Z}/2\mathbb{Z})^2$. As K is a quadratic field of small discriminant, this can be verified either by using class field theory or directly by explicitly computing $\mathrm{Cl}(\mathcal{O})$. In particular, by [Cox13, Lemma 9.3]

$$\mathrm{Gal}(H/\mathbb{Q}) \cong (\mathbb{Z}/2\mathbb{Z})^2 \rtimes \mathbb{Z}/2\mathbb{Z} \cong (\mathbb{Z}/2\mathbb{Z})^3$$

and this implies that every prime $\mathfrak{p} \subseteq \mathcal{O}_H$ has residue degree bounded by 2. Hence, for every prime $p \neq 2$ and split in $\mathcal{O}_K$ the field $\mathbb{F}_{p^2}$ contains the j-invariants of elliptic curves with complex multiplication by $\mathcal{O}$. The connected component of $\mathcal{G}_2(\mathbb{F}_{p^2})$ containing them must necessarily be a volcano of depth ≥ 2 whose crater vertices correspond to all the elliptic curves with complex multiplication by $\mathcal{O}_K$. This proves the proposition. □

This fact triggers many questions: what are the obstructions to a possible solution? Are there infinitely many abstract volcanoes that are not connected components of ordinary isogeny graphs over $\mathbb{F}_{p^s}$, for fixed $s > 1$? If there are only finitely many counter-examples, how many of them?

Part V

Conclusions

Chapter

10 Conclusion: Open Questions

Throughout the course of this manuscript, we have come across many interesting questions in the area of post-quantum cryptography, some of which could be answered, while others remain interesting avenues for future research.

Our work have explored different kinds of families of lattices, each with their own behaviour relative to lattice reduction algorithms. We briefly list our main results and the questions that remain to be explored.

Questions from Part II In Chapter 4, we studied heuristic and asymptotic blocksizes for the primal attack, on NTRU and hypercubic lattices. Both kinds of lattices have unusually short vectors, which decreases the asymptotic heuristic blocksize to a fraction of the dimension n: $n/2$ for $\mathbb{Z}^n$ and $4n/9$ for NTRU, the difference occurs only because of the presence of somewhat short q-vectors in NTRU bases. Our estimations assume that BKZ-reduced bases follow the GSA or Z-GSA, and rely on the 2016 estimate for the primal attack. We argue that asymptotic results provide a cleaner way of comparing the hardness of SVP over different lattices. The predictions of the primal attack, often at the core of parameter estimation for deployed schemes, rely strongly on a simple equation, and simulations of the Gram-Schmidt profile of reduced bases. While such heuristics have been verified experimentally in small dimensions, but our understanding of lattice simulators might be biased towards behaviour observable in dimensions we are able to experiment in, making precise statements in cryptographically relevant dimensions impossible.

Question 10.1. *How can we increase our confidence in lattice simulators, especially in dimensions that remain computationally infeasible? Is it possible to understand precisely how and when the 2016 condition kicks in during lattice reduction?*

We have also investigated the asymptotic impact of the presence of multiple shortest vectors (as in NTRU and $\mathbb{Z}^n$), and showed that a linear number of shortest vectors do not change the first terms in the expansion of the asymptotic blocksize. Because the vectors are not distributed independently, such methods cannot be extended to much larger sets of target vectors, a situation that could be handy in the process of analysing algorithms for approximate problems.

Question 10.2. *Can we use a model similar to ours to predict asymptotic behaviour of the primal attack for approximate-SVP?*

The presence of multiple short vectors is not always useless, as we have seen with $\mathbb{Z}^n$: they allow for guessing-style attacks that exploit the orthogonality. We hope to make this project-and-intersect reduction provable in future work.

We have seen in Chapter 5 that $\sqrt{2}$-approximate SVP is an interesting problem, as our generalisation of Ducas's primal-dual algorithm only requires such oracles in dimension $n/2$ to provably reduce $\mathbb{Z}^n$. An improvement on algorithms for $\sqrt{2}$-SVP might lead to faster attacks for both unstructured and structured LIP.

Question 10.3. *Is $\sqrt{2}$-SVP easier than SVP? If so, by how much?*

While $n/2$ in asymptotic blocksize shows that provable algorithms are essentially as good as heuristic algorithms in the case of hypercubic lattices, this is not the case for NTRU. We have shown that with SVP oracles in dimension $n/2$, it is also possible to provably reduce most NTRU instances, leaving a gap with the heuristic $4n/9$.

Question 10.4. *Can we find an algorithm that provably reduces* NTRU*, using only polynomially many SVP oracles in dimension* $4n/9$*? Is* $4n/9$ *even the best constant?*

We note that NTRU is unimodular after rescaling. However our primal-dual algorithm only uses a condition on the size of the product $\lambda_1(\Lambda) \cdot \lambda_1(\Lambda^\vee)$. It feels like we are not using the full potential of duality.

Question 10.5. *Can duality be used more naturally to speed up primal-dual style lattice reduction? Will this lead to a better constant for* NTRU*?*

There is also some work to do relating to the practicality of primal-dual reduction algorithms, where making the dual steps as efficient as the primal steps can be a nice implementation challenge. In the case of $\mathbb{Z}^n$ where this algorithm is asymptotically as good as its heuristic counterparts, primal-dual reduction might even be competitive.

Questions from Part III In Chapter 6, we have provided a full cryptanalysis of DEFI, a very efficient signature scheme. Its updated version has not yet been attacked, and boasts similar performance. We believe that such a scheme is interesting, and should be tied through a security reduction to the module-Quadratic Form Equivalence problem that we introduced.

Question 10.6. *How secure is* DEFI*v2? Is it possible to design a version of* DEFI *whose security relies on the hardness of (module)-QFE?*

Question 10.7. *How does the hardness of QFE with non-necessarily positive definite quadratic forms compare with the hardness of LIP? Is it inherently a bad idea, or can this new assumption be used to make efficient and post-quantum signature schemes? What about the structured versions?*

Reduction algorithms for indefinite quadratic forms are much weaker than their counterparts for definite quadratic forms, which might lead to think that the best algorithms for LIP should outperform the best algorithms for general QFE. However it is not unreasonable to think that there could exist a reduction from practical indefinite QFE problems to multiple definite instances of LIP in smaller dimension. This possibility makes the answer to the first part of the previous question unclear, as QFE could be hard in the worst case, but not necessarily in the average case.

Our new signature scheme Patronus, from Chapter 7 provides a new perspective on distributions one can use in the Fiat-Shamir with Aborts paradigm. This generalisation of Dilithium to polytopes successfully solves the question of providing a signature scheme based on module-lattices that is more compact and still does not rely on Gaussian distributions.

Question 10.8. *In the context of FSwA, is there more appropriate choice for a polytope than the cross-polytope?*

Logical next steps for Patronus would be a fully optimised implementation of the full scheme, as well as side-channel protection.

Questions from Part IV In Chapter 8, we have studied average Gaussian heuristic-style behaviour of random ideal lattices, uncovering questions of unsuspected mathematical beauty. Our approach enabled us to precisely compute the expected value of λ_1 in quadratic number fields, as well as the first moment of the expected number of ideal lattice points in a ball for fields of small degree and discriminant. Our approach complements the formula of Gargava and Viazovska that gives asymptotic insight on the same quantity. Practical computations require evaluating of many integrals, whose complexity increases with the degree of the field.

Question 10.9. *Can we efficiently compute the integrals that appear in our formulae for the first moment? What about the error term in the formula of Gargava and Viazovska?*

Natural extensions of this work includes questions about higher moments. A result on the second moment would allow for probabilistic bounds on the average value of λ_1 in random ideal lattices.

Question 10.10. *Can we compute the second moment of the expected number of ideal lattice points in a ball? Practically and in small degree through a generalisation of our formula and integration of complicated shapes in Euclidean space, or through asymptotical results following [GV24]?*

Solving the last part of the previous question will likely require the machinery of automorphic forms.

Both our experimental results in small dimension and the formula of Gargava and Viazovska point towards the following important idea: while the short vectors of ideal lattices might be constrained in small dimensions, ideal lattices seem to exhibit very deep equidistribution properties, both as the size of the discriminant and the degree of the field increase. This is confirmed in dimensions 2 and 3 by Duke's theorem. The answer to the following question is probably yes.

Question 10.11. *For ideal lattices of cyclotomic fields, how fast does the asymptotic regime kick in? Is the geometric behaviour of cryptographically sized random ideal lattices already very close to the behaviour of random real lattices?*

These results assert that ideal lattices in large dimension behave nicely, which adds weight to the reasoning that lattice reduction should behave the same way on ideal lattices and on general lattices. This hints that possible weaknesses of structured lattices are more likely to lie on the algebraic side. Indeed lattices such as the log-unit and log-S-unit lattices seem to behave very differently. Ultimately, it would be nice to have a precise idea of the size of the short vector in module lattices that are used in cryptographic standards.

Question 10.12. *Can similar results be obtained for rank* 2, 3 *or* 4 *module lattices?*

Regarding Chapter 9, we have proved that all possible abstract volcanoes arise as a connected component of infinitely many ordinary isogeny graphs $\mathcal{G}_\ell(\mathbb{F}_p)$. We list a few interesting questions.

Question 10.13. *What is the minimal prime p such that a given abstract volcano appears as a connected component of $\mathcal{G}_\ell(\mathbb{F}_p)$? What happens if we ask for multiple abstract volcanoes to appear in the same graph?*

The analogous inverse problem for *supersingular* isogeny graphs does not really make sense, as for a given p, the graph is connected, so there is only a single connected component. However a question with similar flavour can be asked:

Question 10.14. *Does there exist a supersingular isogeny graph with no (graph-theoretic) cycles of length a power of two?*

A positive answer to the previous question would provide a counterexample to a famous unsolved conjecture of Erdős and Gyárfás on graphs with a minimum degree of 3.

Bibliography

[Aar+25] Marius A. Aardal, Gora Adj, Diego F. Aranha, Andrea Basso, Isaac Andrés Canales Martínez, Jorge Chávez-Saab, Maria Corte-Real Santos, Pierrick Dartois, Luca De Feo, Max Duparc, Jonathan Komada Eriksen, Tako Boris Fouotsa, Décio Luiz Gazzoni Filho, Basil Hess, David Kohel, Antonin Leroux, Patrick Longa, Luciano Maino, Michael Meyer, Kohei Nakagawa, Hiroshi Onuki, Lorenz Panny, Sikhar Patranabis, Christophe Petit, Giacomo Pope, Krijn Reijnders, Damien Robert, Francisco Rodríguez-Henríquez, Sina Schaeffler, and Benjamin Wesolowski. *SQIsign*. Tech. rep. National Institute of Standards and Technology, 2025. URL: https://sqisign.org (cit. on p. 168).

[ADS15] Divesh Aggarwal, Daniel Dadush, and Noah Stephens-Davidowitz. "Solving the Closest Vector Problem in 2^n Time - The Discrete Gaussian Strikes Again!" In: *Proc. IEEE 56th FOCS*. 2015, pp. 563–582 (cit. on p. 50).

[ALNS20] Divesh Aggarwal, Jianwei Li, Phong Q. Nguyen, and Noah Stephens-Davidowitz. "Slide Reduction, Revisited - Filling the Gaps in SVP Approximation". In: *Advances in Cryptology - Proc. CRYPTO 2020, Part II*. Ed. by Daniele Micciancio and Thomas Ristenpart. Vol. 12171. Lecture Notes in Computer Science. Springer, 2020, pp. 274–295 (cit. on p. 50).

[AS18] Divesh Aggarwal and Noah Stephens-Davidowitz. "Just take the average! An embarrassingly simple 2^n-time algorithm for SVP (and CVP)". In: *SOSA*. 2018. URL: http://arxiv.org/abs/1709.01535 (cit. on p. 50).

[Ajt96] Miklós. Ajtai. "Generating hard instances of lattice problems (extended abstract)". In: *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*. STOC '96. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 1996, pp. 99–108. ISBN: 0897917855. DOI: 10.1145/237814.237838. URL: https://doi.org/10.1145/237814.237838 (cit. on pp. 16, 37).

[Alb+18] Martin R. Albrecht, Benjamin R. Curtis, Amit Deo, Alex Davidson, Rachel Player, Eamonn W. Postlethwaite, Fernando Virdia, and Thomas Wunderer. "Estimate All the {LWE, NTRU} Schemes!" In: *Security and Cryptography for Networks*. Ed. by Dario Catalano and Roberto De Prisco. Cham: Springer International Publishing, 2018, pp. 351–367. ISBN: 978-3-319-98113-0 (cit. on p. 80).

[AD21] Martin R. Albrecht and Léo Ducas. "Lattice Attacks on NTRU and LWE: A History of Refinements". In: *Computational Cryptography: Algorithmic Aspects of Cryptology*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2021, pp. 15–40 (cit. on pp. 50, 74).

[Alb+19] Martin R. Albrecht, Léo Ducas, Gottfried Herold, Elena Kirshanova, Eamonn W. Postlethwaite, and Marc Stevens. "The General Sieve Kernel and New Records in Lattice Reduction". In: *Advances in Cryptology – EUROCRYPT 2019*. Ed. by Yuval Ishai and Vincent Rijmen. Cham: Springer International Publishing, 2019, pp. 717–746. ISBN: 978-3-030-17656-3 (cit. on pp. 74, 90).

[AGVW17] Martin R. Albrecht, Florian Göpfert, Fernando Virdia, and Thomas Wunderer. "Revisiting the Expected Cost of Solving uSVP and Applications to LWE". In: *Proc. ASIACRYPT 2017, Part I*. Vol. 10624. Lecture Notes in Computer Science. Springer, 2017, pp. 297–322 (cit. on pp. 51, 64, 68).

[ADPS16] Erdem Alkim, Léo Ducas, Thomas Pöppelmann, and Peter Schwabe. "Post-quantum Key Exchange - A New Hope". In: *Proc. 25th USENIX*. USENIX, 2016, pp. 327–343 (cit. on pp. 51, 64).

[APvW25] Bill Allombert, Alice Pellet-Mary, and Wessel van Woerden. "Cryptanalysis of Rank-2 Module-LIP: A Single Real Embedding Is All It Takes". In: *Advances in Cryptology – EUROCRYPT 2025*. Ed. by Serge Fehr and Pierre-Alain Fouque. Cham: Springer Nature Switzerland, 2025, pp. 184–212. ISBN: 978-3-031-91124-8 (cit. on p. 99).

[AEN18] Yoshinori Aono, Thomas Espitau, and Phong Q. Nguyen. *Random Lattices: Theory and Practice*. Unpublished. 2018. URL: https://api.semanticscholar.org/CorpusID:271090054 (cit. on pp. 147, 148, 151).

[Ava+19] Roberto Avanzi, Joppe Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, John M. Schanck, Peter Schwabe, Gregor Seiler, and Damien Stehlé. *CRYSTALS-Kyber (version 2.0) – Submission to round 2 of the NIST post-quantum project*. Mar. 2019 (cit. on p. 64).

[Bam22] Henry Bambury. "S-unit attacks on structured lattice-based cryptosystems". Available at https://hbambury.github.io/HBmfocs.pdf. Dissertation. Oxford,UK: University of Oxford, Sept. 2022 (cit. on pp. 17, 38).

[BBRS24] Henry Bambury, Hugo Beguinet, Thomas Ricosset, and Éric Sageloli. "Polytopes in the Fiat-Shamir with Aborts Paradigm". In: *Advances in Cryptology – CRYPTO 2024*. Ed. by Leonid Reyzin and Douglas Stebila. Cham: Springer Nature Switzerland, 2024, pp. 339–372. ISBN: 978-3-031-68376-3 (cit. on pp. 115, 130, 134–136).

[BCP24] Henry Bambury, Francesco Campagna, and Fabien Pazuki. "Ordinary isogeny graphs over $\mathbb{F}_p$: the inverse volcano problem". In: *Annali della Scuola Normale di Pisa* (2024), p. 33. DOI: 10.2422/2036-2145.202310_022 (cit. on pp. 60, 167, 180).

[BN24] Henry Bambury and Phong Q. Nguyen. "Improved Provable Reduction of NTRU and Hypercubic Lattices". In: *Post-Quantum Cryptography - 15th International Workshop, PQCrypto 2024, Part I*. Ed. by Markku-Juhani Saarinen and Daniel Smith-Tone. Oxford, UK: Springer, Cham, Switzerland, June 2024, pp. 343–370. DOI: 10.1007/978-3-031-62743-9_12 (cit. on pp. 63, 64, 81, 107).

[BN25] Henry Bambury and Phong Q. Nguyen. "Cryptanalysis of an Efficient Signature Based on Isotropic Quadratic Forms". In: *Post-Quantum Cryptography*. Ed. by Ruben Niederhagen and Markku-Juhani O. Saarinen. Cham: Springer Nature Switzerland, 2025, pp. 153–175. ISBN: 978-3-031-86602-9 (cit. on p. 95).

[Bar+23] Manuel Barbosa, Gilles Barthe, Christian Doczkal, Jelle Don, Serge Fehr, Benjamin Grégoire, Yu-Hsuan Huang, Andreas Hülsing, Yi Lee, and Xiaodi Wu. *Fixing and Mechanizing the Security Proof of Fiat-Shamir with Aborts and Dilithium*. Cryptology ePrint Archive, Report 2023/246. https://eprint.iacr.org/2023/246. 2023 (cit. on pp. 53, 121, 138, 139).

[Beg24] Hugo Beguinet. "Quantum-resistant Authentification from Lattice Assumptions". PhD thesis. ENS-PSL, 2024 (cit. on pp. 115, 127).

[BF14] Karim Belabas and Eduardo Friedman. “Computing the residue of the Dedekind zeta function”. en. In: *Mathematics of Computation* 84.291 (May 2014), pp. 357–369. ISSN: 0025-5718, 1088-6842. DOI: 10.1090/S0025-5718-2014-02843-3. URL: https://www.ams.org/mcom/2015-84-291/S0025-5718-2014-02843-3/ (visited on 02/25/2025) (cit. on p. 57).

[BGPS23] Huck Bennett, Atul Ganju, Pura Peetathawatchai, and Noah Stephens-Davidowitz. “Just How Hard Are Rotations of $\mathbb{Z}^n$? Algorithms and Cryptography with the Simplest Lattice”. In: *Advances in Cryptology – EUROCRYPT 2023, Part V*. Ed. by Carmit Hazay and Martijn Stam. Vol. 14008. Lecture Notes in Computer Science. Lyon, France: Springer, Cham, Switzerland, Apr. 2023, pp. 252–281. DOI: 10.1007/978-3-031-30589-4_9 (cit. on pp. 50, 64, 74, 80, 81, 113).

[BLNR21] Olivier Bernard, Andrea Lesavourey, Tuong-Huy Nguyen, and Adeline Roux-Langlois. *Log-S-unit lattices using Explicit Stickelberger Generators to solve Approx Ideal-SVP*. Cryptology ePrint Archive, Paper 2021/1384. https://eprint.iacr.org/2021/1384. 2021. URL: https://eprint.iacr.org/2021/1384 (cit. on pp. 17, 38).

[BR20] Olivier Bernard and Adeline Roux-Langlois. “Twisted-PHS: Using the Product Formula to Solve Approx-SVP in Ideal Lattices”. In: *Advances in Cryptology – ASIACRYPT 2020*. Ed. by Shiho Moriai and Huaxiong Wang. Cham: Springer International Publishing, 2020, pp. 349–380. ISBN: 978-3-030-64834-3 (cit. on pp. 17, 38, 64).

[Ber24] Daniel J. Bernstein. *Asymptotics for the standard block size in primal lattice attacks: second order, formally verified*. Cryptology ePrint Archive, Paper 2024/592. https://eprint.iacr.org/2024/592. 2024. URL: https://eprint.iacr.org/2024/592 (cit. on pp. 64, 68).

[BCLvV16] Daniel J. Bernstein, Chitchanok Chuengsatiansup, Tanja Lange, and Christine van Vredendaal. *NTRU Prime: reducing attack surface at low cost*. Cryptology ePrint Archive, Paper 2016/461. 2016 (cit. on p. 52).

[BLNS23a] Ward Beullens, Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. *Lattice-Based Blind Signatures: Short, Efficient, and Round-Optimal*. Cryptology ePrint Archive, Report 2023/077. https://eprint.iacr.org/2023/077. 2023 (cit. on p. 116).

[BEFGK17] Jean-Franois Biasse, Thomas Espitau, Pierre-Alain Fouque, Alexandre Gélin, and Paul Kirchner. “Computing generator in cyclotomic integer rings”. In: *Advances in Cryptology-EUROCRYPT 2017*. Vol. 10210. Lecture Notes in Computer Science. Paris, France, Apr. 2017, pp. 60–88. DOI: 10.1007/978-3-319-56620-7_3. URL: https://hal.archives-ouvertes.fr/hal-01518438 (cit. on pp. 17, 38).

[BS16] Jean-Franois Biasse and Fang Song. “Efficient quantum algorithms for computing class groups and solving the principal ideal problem in arbitrary degree number fields”. en. In: *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, Jan. 2016, pp. 893–902. ISBN: 9781611974331. DOI: 10.1137/1.9781611974331.ch64. URL: http://epubs.siam.org/doi/10.1137/1.9781611974331.ch64 (visited on 06/10/2022) (cit. on pp. 17, 38).

[BLNS21] Jonathan Bootle, Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. “More Efficient Amortization of Exact Zero-Knowledge Proofs for LWE”. In: *Computer Security – ESORICS 2021*. Ed. by Elisa Bertino, Haya Shulman, and Michael Waidner. Cham: Springer International Publishing, 2021, pp. 608–627. ISBN: 978-3-030-88428-4 (cit. on p. 116).

[BLNS23b] Jonathan Bootle, Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Alessandro Sorniotti. "A Framework for Practical Anonymous Credentials from Lattices". In: *CRYPTO 2023, Part II*. Ed. by Helena Handschuh and Anna Lysyanskaya. Vol. 14082. LNCS. Springer, Cham, Aug. 2023, pp. 384–417. DOI: 10.1007/978-3-031-38545-2_13 (cit. on p. 116).

[BGMWZ13] J. M. Borwein, M. L. Glasser, R. C. McPhedran, J. G. Wan, and I. J. Zucker. *Lattice Sums Then and Now*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2013 (cit. on p. 150).

[Bos+23] Joppe W. Bos, Olivier Bronchain, Léo Ducas, Serge Fehr, Yu-Hsuan Huang, Thomas Pornin, Eamonn W. Postlethwaite, Thomas Prest, Ludo N. Pulles, and Wessel van Woerden. *HAWK Signature Specification Document*. June 2023 (cit. on p. 66).

[Brø83] Arne Brøndsted. *An Introduction to Convex Polytopes*. Springer New York, NY, 1983. DOI: https://doi.org/10.1007/978-1-4612-1148-8 (cit. on p. 46).

[BHLY16] Leon Groot Bruinderink, Andreas Hülsing, Tanja Lange, and Yuval Yarom. "Flush, Gauss, and Reload - A Cache Attack on the BLISS Lattice-Based Signature Scheme". In: *CHES 2016*. Ed. by Benedikt Gierlichs and Axel Y. Poschmann. Vol. 9813. LNCS. Springer, Berlin, Heidelberg, Aug. 2016, pp. 323–345. DOI: 10.1007/978-3-662-53140-2_16 (cit. on p. 117).

[CGS14] Peter Campbell, Michael Groves, and Dan Shepherd. *Soliloquy: A cautionary tale*. ETSI 2nd Quantum-Safe Crypto Workshop. Available at http://docbox.etsi.org/Workshop/2014/201410_CRYPTO/S07_Systems_and_Attacks/S07_Groves_Annex.pdf. 2014 (cit. on pp. 17, 38).

[CJMS23] Marcelo Campos, Matthew Jenssen, Marcus Michelen, and Julian Sahasrabudhe. *A new lower bound for sphere packing*. 2023. arXiv: 2312.10026 [math.MG]. URL: https://arxiv.org/abs/2312.10026 (cit. on p. 146).

[CMST22] Kevin Carrier, Charles Meyer-Hilfiger, Yixin Shen, and Jean-Pierre Tillich. *Assessing the Impact of a Variant of MATZOV's Dual Attack on Kyber*. Cryptology ePrint Archive, Paper 2022/1750. 2022. URL: https://eprint.iacr.org/2022/1750 (cit. on p. 51).

[Cas18] Castryck, Wouter and Lange, Tanja and Martindale, Chloe and Panny, Lorenz and Renes, Joost. "CSIDH : an efficient post-quantum commutative group action". eng. In: *Advances in cryptology, ASIACRYPT 2018 : proceedings*. Ed. by Peyrin, Thomas and Galbraith, Steven D. Vol. 11274. Brisbane, QLD, Australia: Springer, 2018, 395–427. ISBN: 9783030033316. URL: %7Bhttp://dx.doi.org/10.1007/978-3-030-03332-3_15%7D (cit. on pp. 23, 43).

[CGL09] Denis X. Charles, Eyal Z. Goren, and Kristin E. Lauter. "Cryptographic hash functions from expander graphs". In: *J. Cryptology* 22.1 (2009), pp. 93–113. ISSN: 0933-2790. DOI: 10.1007/s00145-007-9002-x. URL: https://doi.org/10.1007/s00145-007-9002-x (cit. on p. 168).

[Che+20] Cong Chen, Oussama Danba, Jeffrey Hoffstein, Andreas Hulsing, Joost Rijneveld, John M. Schanck, Tsunekazu Saito, Peter Schwabe, William Whyte, Keita Xagawa, Takashi Yamakawa, and Zhenfei Zhang. *NTRU Algorithm Specifications And Supporting Documentation*. Sept. 2020 (cit. on pp. 51, 82, 85–87, 109).

[Che13] Yuanmi Chen. "Réduction de réseau et sécurité concrète du chiffrement complètement homomorphe". PhD thesis. Univ. Paris 7, 2013 (cit. on p. 50).

[CN11] Yuanmi Chen and Phong Q. Nguyen. "BKZ 2.0: Better Lattice Security Estimates". In: *Advances in Cryptology – ASIACRYPT 2011*. Ed. by Dong Hoon Lee and Xiaoyun Wang. Vol. 7073. Lecture Notes in Computer Science. Seoul, South Korea: Springer Berlin Heidelberg, Germany, Dec. 2011, pp. 1–20. DOI: 10.1007/978-3-642-25385-0_1 (cit. on pp. 50, 51, 74, 75).

[Che+23] Jung Hee Cheon, Hyeongmin Choe, Julien Devevey, Tim Güneysu, Dongyeon Hong, Markus Krausz, Georg Land, Junbum Shin, Damien Stehlé, and MinJune Yi. *HAETAE Algorithm Specifications and Supporting Documentation*. Submission to the NIST's post-quantum cryptography standardization process. 2023 (cit. on pp. 116, 120, 121, 132, 134–136, 138, 140).

[CCF22] Augustin Chevallier, Frédéric Cazals, and Paul Fearnhead. *Efficient computation of the volume of a polytope in high-dimensions using Piecewise Deterministic Markov Processes*. 2022. arXiv: 2202.09129 [stat.CO] (cit. on p. 123).

[Clo24] Cloudflare. *A look at the latest post-quantum signature standardization candidates*. https://blog.cloudflare.com/another-look-at-pq-signatures/. Accessed: 17/09/2025. Nov. 2024 (cit. on p. 96).

[Coh93] Henri Cohen. *A Course in Computational Algebraic Number Theory*. Vol. 138. Graduate Texts in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993. ISBN: 9783642081422 9783662029459. DOI: 10.1007/978-3-662-02945-9. URL: http://link.springer.com/10.1007/978-3-662-02945-9 (visited on 07/08/2025) (cit. on pp. 54, 60).

[Cou06] Jean-Marc Couveignes. *Hard Homogeneous Spaces*. Cryptology ePrint Archive, Paper 2006/291. 2006. URL: https://eprint.iacr.org/2006/291 (cit. on pp. 22, 43).

[Cox13] David A. Cox. *Primes of the form $x^2 + ny^2$*. Second. Pure and Applied Mathematics (Hoboken). John Wiley & Sons, Inc., Hoboken, NJ, 2013, pp. xviii+356. ISBN: 978-1-118-39018-4. DOI: 10.1002/9781118400722. URL: https://doi.org/10.1002/9781118400722 (cit. on pp. 174, 177, 178, 182, 183, 193).

[CDPR16] Ronald Cramer, Léo Ducas, Chris Peikert, and Oded Regev. "Recovering Short Generators of Principal Ideals in Cyclotomic Rings". In: *Advances in Cryptology – EUROCRYPT 2016*. Ed. by Marc Fischlin and Jean-Sébastien Coron. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 559–585. ISBN: 978-3-662-49896-5 (cit. on pp. 17, 38).

[CDW17] Ronald Cramer, Léo Ducas, and Benjamin Wesolowski. "Short Stickelberger Class Relations and Application to Ideal-SVP". In: *Advances in Cryptology – EUROCRYPT 2017*. Ed. by Jean-Sébastien Coron and Jesper B. Nielsen. Cham: Springer International Publishing, 2017, pp. 324–348 (cit. on pp. 17, 38, 64).

[CDW21] Ronald Cramer, Léo Ducas, and Benjamin Wesolowski. "Mildly Short Vectors in Cyclotomic Ideal Lattices in Quantum Polynomial Time". In: *J. ACM* 68.2 (Jan. 2021). ISSN: 0004-5411. DOI: 10.1145/3431725. URL: https://doi.org/10.1145/3431725 (cit. on pp. 17, 38).

[DDGR20] Dana Dachman-Soled, Léo Ducas, Huijing Gong, and Mélissa Rossi. "LWE with Side Information: Attacks and Concrete Security Estimation". In: *Advances in Cryptology – Proc. CRYPTO 2020*. Berlin, Heidelberg: Springer-Verlag, 2020, pp. 329–358 (cit. on pp. 51, 64, 66, 68, 70, 73, 76).

[DM13] Daniel Dadush and Daniele Micciancio. "Algorithms for the Densest Sub-Lattice Problem". In: *24th Annual ACM-SIAM Symposium on Discrete Algorithms*. Ed. by Sanjeev Khanna. New Orleans, LA, USA: ACM-SIAM, Jan. 2013, pp. 1103–1122. DOI: 10.1137/1.9781611973105.79 (cit. on p. 110).

[DJP14] Luca De Feo, David Jao, and Jérôme Plût. "Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies". In: *Journal of Mathematical Cryptology* 8.3 (Jan. 2014). ISSN: 1862-2976, 1862-2984. DOI: 10.1515/jmc-2012-0015 (cit. on p. 168).

[dBoe22] Koen de Boer. "Random Walks on Arakelov Class Groups". Ph.D. Thesis. Universiteit Leiden, Sept. 2022 (cit. on p. 58).

[dBDPW20] Koen de Boer, Léo Ducas, Alice Pellet-Mary, and Benjamin Wesolowski. "Random Self-reducibility of Ideal-SVP via Arakelov Random Walks". In: *Advances in Cryptology – CRYPTO 2020: 40th Annual International Cryptology Conference, CRYPTO 2020, Santa Barbara, CA, USA, August 17–21, 2020, Proceedings, Part II*. Santa Barbara, CA, USA: Springer-Verlag, 2020, pp. 243–273. ISBN: 978-3-030-56879-5. DOI: 10.1007/978-3-030-56880-1_9. URL: https://doi.org/10.1007/978-3-030-56880-1_9 (cit. on pp. 56, 58, 150, 153, 154).

[dPLS18] Rafael del Pino, Vadim Lyubashevsky, and Gregor Seiler. "Lattice-Based Group Signatures and Zero-Knowledge Proofs of Automorphism Stability". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. CCS '18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 574–591. ISBN: 9781450356930. DOI: 10.1145/3243734.3243852. URL: https://doi.org/10.1145/3243734.3243852 (cit. on p. 116).

[DFPS23] Julien Devevey, Pouria Fallahpour, Alain Passelègue, and Damien Stehlé. "A Detailed Analysis of Fiat-Shamir with Aborts". In: *Advances in Cryptology – CRYPTO 2023*. Ed. by Helena Handschuh and Anna Lysyanskaya. Cham: Springer Nature Switzerland, 2023, pp. 327–357. ISBN: 978-3-031-38554-4 (cit. on pp. 53, 121).

[DFPS22] Julien Devevey, Omar Fawzi, Alain Passelègue, and Damien Stehlé. "On Rejection Sampling in Lyubashevsky's Signature Scheme". In: *ASIACRYPT 2022, Part IV*. Ed. by Shweta Agrawal and Dongdai Lin. Vol. 13794. LNCS. Springer, Cham, Dec. 2022, pp. 34–64. DOI: 10.1007/978-3-031-22972-5_2 (cit. on pp. 21, 42, 118–120).

[DH76] Whitfield Diffie and Martin Hellman. "New directions in cryptography". In: *IEEE Transactions on Information Theory* 22.6 (1976), pp. 644–654. DOI: 10.1109/TIT.1976.1055638 (cit. on pp. 4, 26).

[Duc18] Léo Ducas. "Shortest Vector from Lattice Sieving: A Few Dimensions for Free". In: *Advances in Cryptology – EUROCRYPT 2018*. Ed. by Jesper Buus Nielsen and Vincent Rijmen. Cham: Springer International Publishing, 2018, pp. 125–145. ISBN: 978-3-319-78381-9 (cit. on p. 90).

[Duc23] Léo Ducas. "Provable lattice reduction of $\mathbb{Z}^n$ with blocksize $n/2$". In: *Designs, Codes and Cryptography* (Nov. 2023) (cit. on pp. 19, 40, 50, 81–84, 87–89).

[DDLL13] Léo Ducas, Alain Durmus, Tancrède Lepoint, and Vadim Lyubashevsky. "Lattice Signatures and Bimodal Gaussians". In: *CRYPTO 2013, Part I*. Ed. by Ran Canetti and Juan A. Garay. Vol. 8042. LNCS. Springer, Berlin, Heidelberg, Aug. 2013, pp. 40–56. DOI: 10.1007/978-3-642-40041-4_3 (cit. on p. 141).

[DEL25] Léo Ducas, Lynn Engelberts, and Johanna Loyer. *Wagner's Algorithm Provably Runs in Subexponential Time for SIS$^\infty$*. Cryptology ePrint Archive, Paper 2025/575. 2025. URL: https://eprint.iacr.org/2025/575 (cit. on p. 140).

[DEP23] Léo Ducas, Thomas Espitau, and Eamonn W. Postlethwaite. "Finding Short Integer Solutions When the Modulus Is Small". In: *Advances in Cryptology – CRYPTO 2023, Part III*. Ed. by Yevgeniy Dodis and Thomas Shrimpton. Vol. 14083. Lecture Notes in Computer Science. Santa Barbara, CA, USA: Springer, Cham, Switzerland, Aug. 2023, pp. 150–176. DOI: 10.1007/978-3-031-38548-3_6 (cit. on pp. 107, 129).

[Duc+21] Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, Peter Schwabe, Gregor Seiler, and Damien Stehlé. *CRYSTALS–Dilithium: A Lattice-Based Digital Signature Scheme*. Submission to the NIST's post-quantum cryptography standardization process (update from February 2021). 2021 (cit. on pp. 64, 95, 116, 117, 120, 121, 134–136, 139, 140).

[DPPvW22] Léo Ducas, Eamonn W. Postlethwaite, Ludo N. Pulles, and Wessel P. J. van Woerden. "Hawk: Module LIP Makes Lattice Signatures Fast, Compact and Simple". In: *Advances in Cryptology - Proc. ASIACRYPT 2022*. Vol. 13794. Lecture Notes in Computer Science. Springer, 2022, pp. 65–94 (cit. on pp. 64, 66, 73, 80, 99, 113).

[DP23] Léo Ducas and Ludo N. Pulles. "Does the Dual-Sieve Attack on Learning with Errors Even Work?" In: *Advances in Cryptology – CRYPTO 2023*. Ed. by Helena Handschuh and Anna Lysyanskaya. Cham: Springer Nature Switzerland, 2023, pp. 37–69 (cit. on p. 70).

[DS21] Léo Ducas and John Schanck. *pq-crystals/security-estimates*. https://github.com/pq-crystals/security-estimates. 2021 (cit. on p. 140).

[DvW21] Léo Ducas and Wessel van Woerden. "NTRU Fatigue: How Stretched is Overstretched?" In: *Advances in Cryptology – ASIACRYPT 2021*. Ed. by Mehdi Tibouchi and Huaxiong Wang. Cham: Springer International Publishing, 2021, pp. 3–32 (cit. on pp. 73, 98, 107, 110).

[DvW22] Léo Ducas and Wessel P. J. van Woerden. "On the Lattice Isomorphism Problem, Quadratic Forms, Remarkable Lattices, and Cryptography". In: *Advances in Cryptology – EUROCRYPT 2022, Part III*. Ed. by Orr Dunkelman and Stefan Dziembowski. Vol. 13277. Lecture Notes in Computer Science. Trondheim, Norway: Springer, Cham, Switzerland, June 2022, pp. 643–673. DOI: 10.1007/978-3-031-07082-2_23 (cit. on pp. 64, 113).

[Dud09] Jarek Duda. "Asymmetric numeral systems". In: *CoRR* abs/0902.0271 (2009). arXiv: 0902.0271. URL: http://arxiv.org/abs/0902.0271 (cit. on pp. 135, 140).

[DK22] Samed Düzlü and Juliane Krämer. "Application of automorphic forms to lattice problems". In: *Journal of Mathematical Cryptology* 16 (2022), pp. 156–197 (cit. on p. 56).

[DF88] Martin E. Dyer and Alan M. Frieze. "On the Complexity of Computing the Volume of a Polyhedron." In: *SIAM J. Comput.* 17.5 (1988), pp. 967–974. URL: http://dblp.uni-trier.de/db/journals/siamcomp/siamcomp17.html#DyerF88 (cit. on p. 123).

[ELMV11] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel, and Akshay Venkatesh. "Distribution of periodic torus orbits and Duke's theorem for cubic fields". en. In: *Annals of Mathematics* 173.2 (Mar. 2011). ISSN: 0003-486X. DOI: 10.4007/annals.2011.173.2.5. URL: http://annals.math.princeton.edu/2011/173-2/p05 (visited on 07/11/2025) (cit. on pp. 153, 164).

[ELMV12] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel, and Akshay Venkatesh. "The distribution of closed geodesics on the modular surface, and Duke's theorem". In: *L'Enseignement Mathématique* 58.3 (Dec. 2012), pp. 249–313. ISSN: 0013-8584, 2309-4672. DOI: 10.4171/lem/58-3-2. URL: https://ems.press/doi/10.4171/lem/58-3-2 (visited on 07/09/2025) (cit. on p. 153).

[EV22] Friedrich Eisenbrand and Moritz Venzin. "Approximate CVP_p in time $2^{0.802n}$". In: *J. Comput. Syst. Sci.* 124 (2022), pp. 129–139 (cit. on pp. 82, 90).

[EO06] Alex Eskin and Hee Oh. "Ergodic theoretic proof of equidistribution of Hecke points". In: *Ergodic Theory and Dynamical Systems* 26.1 (2006), pp. 163–167. DOI: 10.1017/S0143385705000428 (cit. on pp. 16, 37).

[EFGT17] Thomas Espitau, Pierre-Alain Fouque, Benoît Gérard, and Mehdi Tibouchi. "Side-Channel Attacks on BLISS Lattice-Based Signatures: Exploiting Branch Tracing against strongSwan and Electromagnetic Emanations in Microcontrollers". In: *ACM CCS 2017*. Ed. by Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu. ACM Press, Nov. 2017, pp. 1857–1874. DOI: 10.1145/3133956.3134028 (cit. on p. 117).

[FPS22] Joël Felderhoff, Alice Pellet-Mary, and Damien Stehlé. "On Module Unique-SVP and NTRU". In: *Advances in Cryptology – ASIACRYPT 2022*. Ed. by Shweta Agrawal and Dongdai Lin. Cham: Springer Nature Switzerland, 2022, pp. 709–740. ISBN: 978-3-031-22969-5 (cit. on p. 58).

[Fel71] William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. New York: John Wiley & Sons Inc., 1971, pp. xxiv+669 (cit. on p. 128).

[Feu24] Martin Feussner. *DEFI Reference implementation*. June 2024. URL: %5Curl%7Bhttps://github.com/martinfeussner/DEFI/tree/dd038b3/DEFI128%7D (cit. on pp. 100, 111).

[FS23] Martin Feussner and Igor Semaev. *DIGITAL SIGNATURE ALGORITHMS EHTv3 AND EHTv4 – Submission to round 1 of the NIST additional call for post-quantum signatures*. 2023 (cit. on p. 96).

[FS24a] Martin Feussner and Igor Semaev. *Isotropic Quadratic Forms, Diophantine Equations and Digital Signatures*. Cryptology ePrint Archive, Report 2024/679, version 1. Announced on May 6th, 2024 on the NIST-pqc mailing list. 2024. URL: https://eprint.iacr.org/archive/2024/679/20240503:175841 (cit. on pp. 96–100, 102, 111, 112).

[FS24b] Martin Feussner and Igor Semaev. *Isotropic Quadratic Forms, Diophantine Equations and Digital Signatures, DEFIv2*. Cryptology ePrint Archive, Report 2024/679, version 2. Last updated November 2024. 2024. URL: https://eprint.iacr.org/archive/2024/679/20241105:105112 (cit. on pp. 97, 113).

[Fou+19] Pierre-Alain Fouque, Jeffrey Hoffstein, Paul Kirchner, Vadim Lyubashevsky, Thomas Pornin, Thomas Prest, Thomas Ricosset, Gregor Seiler, William Whyte, and Zhenfei Zhang. *Falcon: Fast-Fourier Lattice-based Compact Signatures over NTRU*. Mar. 2019 (cit. on pp. 52, 64, 82, 85–87, 95, 117).

[FM02] Mireille Fouquet and Franois Morain. “Isogeny Volcanoes and the SEA Algorithm”. In: *Algorithmic Number Theory*. Ed. by Claus Fieker and David Kohel. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 276–291. ISBN: 978-3-540-45455-7 (cit. on pp. 22, 43, 168, 171).

[fSic25] Bundesamt für Sicherheit in der Informationstechnik (BSI). *Entwicklungsstand Quantencomputer. Version 2.1*. Studie / Technical Report. Accessed: 17/09/2025. BSI – Bundesamt für Sicherheit in der Informationstechnik, Jan. 2025. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Studien/Quantencomputer/Entwicklungstand%5C_QC%5C_V%5C_2%5C_1.html?nn=916616 (cit. on pp. 6, 27).

[GHN06] Nicolas Gama, Nick Howgrave-Graham, and Phong Q. Nguyen. “Symplectic Lattice Reduction and NTRU”. In: *Advances in Cryptology - Proc. EUROCRYPT 2006*. Ed. by Serge Vaudenay. Vol. 4004. Lecture Notes in Computer Science. Springer, 2006, pp. 233–253 (cit. on pp. 19, 40, 81, 82, 85, 87).

[GINX16] Nicolas Gama, Malika Izabachène, Phong Q. Nguyen, and Xiang Xie. “Structural Lattice Reduction: Generalized Worst-Case to Average-Case Reductions and Homomorphic Cryptosystems”. In: *Advances in Cryptology – EUROCRYPT 2016*. Ed. by Marc Fischlin and Jean-Sébastien Coron. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 528–558. ISBN: 978-3-662-49896-5 (cit. on pp. 16, 37).

[GN08a] Nicolas Gama and P. Q. Nguyen. “Finding short lattice vectors within Mordell’s inequality”. In: *Proc. 40th ACM Symp. on Theory of Computing (STOC)*. 2008 (cit. on pp. 19, 50, 82–84).

[GN08b] Nicolas Gama and Phong Q. Nguyen. “Predicting Lattice Reduction”. In: *Advances in Cryptology – EUROCRYPT 2008*. Ed. by Nigel P. Smart. Vol. 4965. Lecture Notes in Computer Science. Istanbul, Turkey: Springer Berlin Heidelberg, Germany, Apr. 2008, pp. 31–51. DOI: 10.1007/978-3-540-78967-3_3 (cit. on pp. 50, 51, 107).

[GNR10] Nicolas Gama, Phong Q. Nguyen, and Oded Regev. “Lattice enumeration using extreme pruning”. In: *Advances in Cryptology - Proc. EUROCRYPT 2010*. Vol. 6110. LNCS. Springer, 2010 (cit. on p. 82).

[GSV24] Nihar Gargava, Vlad Serban, and Maryna Viazovska. *Moments of the number of points in a bounded set for number field lattices*. 2024. arXiv: 2308.15275 [math.NT]. URL: https://arxiv.org/abs/2308.15275 (cit. on p. 146).

[GSVV24] Nihar Gargava, Vlad Serban, Maryna Viazovska, and Ilaria Viglino. *Effective module lattices and their shortest vectors*. 2024. arXiv: 2402.10305 [math.NT]. URL: https://arxiv.org/abs/2402.10305 (cit. on p. 146).

[GV24] Nihar Gargava and Maryna Viazovska. *Mean Value for Random Ideal Lattices*. 2024. arXiv: 2411.14973 [math.NT]. URL: https://arxiv.org/abs/2411.14973 (cit. on pp. 20, 41, 145, 146, 163–165, 199).

[GS02] Craig Gentry and Mike Szydlo. “Cryptanalysis of the Revised NTRU Signature Scheme”. In: *Advances in Cryptology — EUROCRYPT 2002*. Ed. by Lars R. Knudsen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 299–320. ISBN: 978-3-540-46035-0 (cit. on p. 99).

[GMRR22] Morgane Guerreau, Ange Martinelli, Thomas Ricosset, and Mélissa Rossi. “The Hidden Parallelepiped Is Back Again: Power Analysis Attacks on Falcon”. In: *IACR TCHES* 2022.3 (2022), pp. 141–164. DOI: 10.46586/tches.v2022.i3.141-164 (cit. on p. 117).

[HHHW09] Philip S. Hirschhorn, Jeffrey Hoffstein, Nick Howgrave-Graham, and William Whyte. "Choosing NTRUEncrypt Parameters in Light of Combined Lattice Reduction and MITM Approaches". In: *Proc. ACNS 2009*. Vol. 5536. Lecture Notes in Computer Science. 2009, pp. 437–455 (cit. on p. 51).

[Hof+17] Jeff Hoffstein, Jill Pipher, John M. Schanck, Joseph H. Silverman, William Whyte, and Zhenfei Zhang. "Choosing Parameters for NTRUEncrypt". In: *Topics in Cryptology – CT-RSA 2017*. Ed. by Helena Handschuh. Cham: Springer International Publishing, 2017, pp. 3–18 (cit. on p. 51).

[HPS98] Jeffrey Hoffstein, Jill Pipher, and Joseph H. Silverman. "NTRU: A Ring Based Public Key Cryptosystem". In: *Proc. of ANTS III*. Vol. 1423. LNCS. Springer-Verlag, 1998, pp. 267–288 (cit. on pp. 51, 52, 82, 85–87).

[HPRR20] James Howe, Thomas Prest, Thomas Ricosset, and Mélissa Rossi. "Isochronous Gaussian Sampling: From Inception to Implementation". In: *Post-Quantum Cryptography - 11th International Conference, PQCrypto 2020*. Ed. by Jintai Ding and Jean-Pierre Tillich. Springer, Cham, 2020, pp. 53–71. DOI: 10.1007/978-3-030-44223-1_5 (cit. on pp. 120, 129).

[How07] Nick Howgrave-Graham. "A Hybrid Lattice-Reduction and Meet-in-the-Middle Attack Against NTRU". In: *Proc. CRYPTO 2007*. Vol. 4622. Lecture Notes in Computer Science. Springer, 2007, pp. 150–169 (cit. on p. 51).

[Hus04] Dale Husemöller, ed. *Elliptic Curves*. eng. Second Edition. Graduate Texts in Mathematics 111. New York, NY: Springer-Verlag New York, Inc, 2004. ISBN: 9780387215778 (cit. on p. 54).

[IJ10] Sorina Ionica and Antoine Joux. "Pairing the Volcano". In: *Algorithmic Number Theory*. Ed. by Guillaume Hanrot, François Morain, and Emmanuel Thomé. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 201–218. ISBN: 978-3-642-14518-6 (cit. on pp. 23, 43, 168).

[JT15] Christian U. Jensen and Anders Thorup. "Gorenstein orders". In: *Journal of Pure and Applied Algebra* 219.3 (2015), pp. 551–562. DOI: doi.org/10.1016/j.jpaa.2014.05.013. URL: https://doi.org/10.1016/j.jpaa.2014.05.013 (cit. on p. 174).

[Jia+23] Kaijie Jiang, Anyu Wang, Hengyi Luo, Guoxiao Liu, Yang Yu, and Xiaoyun Wang. "Exploiting the Symmetry of Zn: Randomization and the Automorphism Problem". In: *Advances in Cryptology – ASIACRYPT 2023: 29th International Conference on the Theory and Application of Cryptology and Information Security, Guangzhou, China, December 4–8, 2023, Proceedings, Part IV*. Guangzhou, China: Springer-Verlag, 2023, pp. 167–200. ISBN: 978-981-99-8729-0. DOI: 10.1007/978-981-99-8730-6_6. URL: https://doi.org/10.1007/978-981-99-8730-6_6 (cit. on pp. 20, 40, 64, 74, 77, 78, 80).

[KTX08] Akinori Kawachi, Keisuke Tanaka, and Keita Xagawa. "Concurrently Secure Identification Schemes Based on the Worst-Case Hardness of Lattice Problems". In: *ASIACRYPT 2008*. Ed. by Josef Pieprzyk. Vol. 5350. LNCS. Springer, Berlin, Heidelberg, Dec. 2008, pp. 372–389. DOI: 10.1007/978-3-540-89255-7_23 (cit. on p. 116).

[KLS18] Eike Kiltz, Vadim Lyubashevsky, and Christian Schaffner. "A Concrete Treatment of Fiat-Shamir Signatures in the Quantum Random-Oracle Model". In: *EUROCRYPT 2018, Part III*. Ed. by Jesper Buus Nielsen and Vincent Rijmen. Vol. 10822. LNCS. Springer, Cham, Apr. 2018, pp. 552–586. DOI: 10.1007/978-3-319-78372-7_18 (cit. on pp. 53, 121, 135).

[KP23] Jonghyun Kim and Jong Hwan Park. "NTRU+: Compact Construction of NTRU Using Simple Encoding Method". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 4760–4774 (cit. on p. 52).

[KF17] Paul Kirchner and Pierre-Alain Fouque. "Revisiting Lattice Attacks on Overstretched NTRU Parameters". In: *Advances in Cryptology – EUROCRYPT 2017, Part I*. Ed. by Jean-Sébastien Coron and Jesper Buus Nielsen. Vol. 10210. Lecture Notes in Computer Science. Paris, France: Springer, Cham, Switzerland, May 2017, pp. 3–26. DOI: 10.1007/978-3-319-56620-7_1 (cit. on pp. 73, 107).

[Kla25] Boaz Klartag. *Lattice packing of spheres in high dimensions using a stochastically evolving ellipsoid*. 2025. arXiv: 2504.05042 [math.MG]. URL: https://arxiv.org/abs/2504.05042 (cit. on p. 146).

[Koh96] David Kohel. "Endomorphism rings of elliptic curves over finite fields". PhD thesis. University of California at Berkeley, 1996 (cit. on pp. 23, 43, 59, 168).

[Lan87] Serge Lang. *Elliptic Functions*. Vol. 112. Graduate Texts in Mathematics. New York, NY: Springer New York, 1987. ISBN: 9781461291428. DOI: 10.1007/978-1-4612-4752-4. URL: http://link.springer.com/10.1007/978-1-4612-4752-4 (visited on 08/06/2021) (cit. on pp. 59, 60).

[LS15] Adeline Langlois and Damien Stehlé. "Worst-case to average-case reductions for module lattices". In: *Designs, Codes and Cryptography* 75.3 (June 2015), pp. 565–599. ISSN: 1573-7586. DOI: 10.1007/s10623-014-9938-4. URL: https://doi.org/10.1007/s10623-014-9938-4 (cit. on pp. 17, 38).

[LLL82] Arjen K. Lenstra, Hendrik W. Lenstra Jr., and László. Lovász. "Factoring polynomials with rational coefficients". In: *Mathematische Ann.* 261 (1982), pp. 513–534 (cit. on pp. 11, 33, 50).

[LN24] Jianwei Li and Phong Q. Nguyen. "A Complete Analysis of the BKZ Lattice Reduction Algorithm". In: *Journal of Cryptology* 38 (Dec. 2024). DOI: 10.1007/s00145-024-09527-0 (cit. on pp. 49, 50, 80, 109).

[LNSW13] San Ling, Khoa Nguyen, Damien Stehlé, and Huaxiong Wang. "Improved Zero-Knowledge Proofs of Knowledge for the ISIS Problem, and Applications". In: *PKC 2013*. Ed. by Kaoru Kurosawa and Goichiro Hanaoka. Vol. 7778. LNCS. Springer, Berlin, Heidelberg, Mar. 2013, pp. 107–124. DOI: 10.1007/978-3-642-36362-7_8 (cit. on p. 116).

[Lyu09] Vadim Lyubashevsky. "Fiat-Shamir with Aborts: Applications to Lattice and Factoring-Based Signatures". In: *ASIACRYPT 2009*. Ed. by Mitsuru Matsui. Vol. 5912. LNCS. Springer, Berlin, Heidelberg, Dec. 2009, pp. 598–616. DOI: 10.1007/978-3-642-10366-7_35 (cit. on p. 116).

[Lyu12] Vadim Lyubashevsky. "Lattice Signatures without Trapdoors". In: *EUROCRYPT 2012*. Ed. by David Pointcheval and Thomas Johansson. Vol. 7237. LNCS. Springer, Berlin, Heidelberg, Apr. 2012, pp. 738–755. DOI: 10.1007/978-3-642-29011-4_43 (cit. on pp. 47, 120).

[LNP22] Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Maxime Plançon. "Lattice-Based Zero-Knowledge Proofs and Applications: Shorter, Simpler, and More General". In: *CRYPTO 2022, Part II*. Ed. by Yevgeniy Dodis and Thomas Shrimpton. Vol. 13508. LNCS. Springer, Cham, Aug. 2022, pp. 71–101. DOI: 10.1007/978-3-031-15979-4_3 (cit. on p. 116).

[LNS20] Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. "Practical Lattice-Based Zero-Knowledge Proofs for Integer Relations". In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. CCS '20. Virtual Event, USA: Association for Computing Machinery, 2020, pp. 1051–1070. ISBN: 9781450370899. DOI: 10.1145/3372297.3417894. URL: https://doi.org/10.1145/3372297.3417894 (cit. on p. 116).

[LNS21a] Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. "Shorter Lattice-Based Zero-Knowledge Proofs via One-Time Commitments". In: *PKC 2021, Part I*. Ed. by Juan Garay. Vol. 12710. LNCS. Springer, Cham, May 2021, pp. 215–241. DOI: 10.1007/978-3-030-75245-3_9 (cit. on p. 116).

[LNS21b] Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. "SMILE: Set Membership from Ideal Lattices with Applications to Ring Signatures and Confidential Transactions". In: *Advances in Cryptology – CRYPTO 2021*. Ed. by Tal Malkin and Chris Peikert. Cham: Springer International Publishing, 2021, pp. 611–640. ISBN: 978-3-030-84245-1 (cit. on p. 116).

[LPR10] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. "On Ideal Lattices and Learning with Errors over Rings". In: *Advances in Cryptology – EUROCRYPT 2010*. Ed. by Henri Gilbert. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 1–23. ISBN: 978-3-642-13190-5 (cit. on pp. 17, 38).

[MM92] Marvin Marcus and Henryk Minc. *A survey of matrix theory and matrix inequalities*. New York: Dover, 1992. ISBN: 9780486671024 (cit. on p. 108).

[Mar03] Jacques Martinet. *Perfect Lattices in Euclidean Spaces*. Springer Berlin, Heidelberg, 2003 (cit. on pp. 49, 76).

[MAT22] MATZOV. *Report on the Security of LWE: Improved Dual Lattice Attack*. Apr. 2022 (cit. on p. 51).

[MS01] Alexander May and Joseph H. Silverman. "Dimension Reduction Methods for Convolution Modular Lattices". In: *Cryptography and Lattices*. Ed. by Joseph H. Silverman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 110–125. ISBN: 978-3-540-44670-5 (cit. on pp. 63, 68, 76, 80).

[MR07] Daniele Micciancio and Oded Regev. "Worst-Case to Average-Case Reductions Based on Gaussian Measures". In: *SIAM Journal on Computing* 37.1 (2007), pp. 267–302. DOI: 10.1137/S0097539705447360. eprint: https://doi.org/10.1137/S0097539705447360. URL: https://doi.org/10.1137/S0097539705447360 (cit. on p. 147).

[MW16] Daniele Micciancio and Michael Walter. "Practical, Predictable Lattice Basis Reduction". In: *Advances in Cryptology - Proc. EUROCRYPT 2016, Part I*. Vol. 9665. Lecture Notes in Computer Science. Springer, 2016, pp. 820–849 (cit. on p. 50).

[MSTTV07] J. Miret, D. Sadornil, J. Tena, R. Tomàs, and M. Valls. "Isogeny cordillera algorithm to obtain cryptographically good elliptic curves". In: *ACSW*. 2007 (cit. on pp. 168, 171).

[Mol10] Pascal Molin. "Intégration numérique et calculs de fonctions L". Theses. Université Sciences et Technologies - Bordeaux I, Oct. 2010. URL: https://theses.hal.science/tel-00537489 (cit. on p. 165).

[MPPW24] Guilhem Mureau, Alice Pellet-Mary, Georgii Pliatsok, and Alexandre Wallet. "Cryptanalysis of Rank-2 Module-LIP in Totally Real Number Fields". In: *Advances in Cryptology – EUROCRYPT 2024*. Ed. by Marc Joye and Gregor Leander. Cham: Springer Nature Switzerland, 2024, pp. 226–255. ISBN: 978-3-031-58754-2 (cit. on p. 99).

[Nag55] Trygve Nagell. "Contributions to the theory of a category of Diophantine equations of the second degree with two unknowns". In: *Nova Acta Soc. Sci. Upsaliensis (4)* 16.2 (1955), p. 38. ISSN: 0029-5000 (cit. on pp. 183, 187).

[Neu99a] Jürgen Neukirch. *Algebraic Number Theory*. eng. Grundlehren der Mathematischen Wissenschaften Ser v.322. Berlin, Heidelberg: Springer Berlin / Heidelberg, 1999. ISBN: 9783662039830 (cit. on p. 54).

[Neu99b] Jürgen Neukirch. *Algebraic number theory*. Vol. 322. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Translated from the 1992 German original and with a note by Norbert Schappacher, With a foreword by G. Harder. Springer-Verlag, Berlin, 1999, pp. xviii+571. ISBN: 3-540-65399-6. DOI: 10.1007/978-3-662-03983-0. URL: https://doi.org/10.1007/978-3-662-03983-0 (cit. on p. 193).

[NS97] Phong Nguyen and Jacques Stern. "Merkle-Hellman revisited: A cryptanalysis of the Qu-Vanstone cryptosystem based on group factorizations". In: *Advances in Cryptology — CRYPTO '97*. Ed. by Burton S. Kaliski. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 198–212. ISBN: 978-3-540-69528-8 (cit. on p. 76).

[NS06] Phong Q. Nguyen and Damien Stehlé. "LLL on the Average". In: *Proc. ANTS*. 2006, pp. 238–256 (cit. on p. 50).

[NV08] Phong Q. Nguyen and Thomas Vidick. In: *Journal of Mathematical Cryptology* 2.2 (2008), pp. 181–207. DOI: doi:10.1515/JMC.2008.009. URL: https://doi.org/10.1515/JMC.2008.009 (cit. on p. 90).

[NIS22a] NIST. *Call for Additional Digital Signature Schemes for the Post-Quantum Cryptography Standardization Process*. 2022. URL: https://csrc.nist.gov/csrc/media/Projects/pqc-dig-sig/documents/call-for-proposals-dig-sig-sept-2022.pdf (cit. on p. 96).

[NIS22b] NIST. *Post-Quantum Cryptography Standardization*. 2022. URL: https://csrc.nist.gov/projects/post-quantum-cryptography/post-quantum-cryptography-standardization (cit. on p. 95).

[PHS19] Alice Pellet–Mary, Guillaume Hanrot, and Damien Stehlé. *Approx-SVP in Ideal Lattices with Pre-processing*. Cryptology ePrint Archive, Report 2019/215. https://ia.cr/2019/215. 2019 (cit. on pp. 17, 38, 64).

[PBY17] Peter Pessl, Leon Groot Bruinderink, and Yuval Yarom. "To BLISS-B or not to be: Attacking strongSwan's Implementation of Post-Quantum Signatures". In: *ACM CCS 2017*. Ed. by Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu. ACM Press, Nov. 2017, pp. 1843–1855. DOI: 10.1145/3133956.3134023 (cit. on p. 117).

[Phi60] J. R. Philip. "The Function inverfc θ". In: *Australian Journal of Physics* 13 (Mar. 1960) (cit. on p. 72).

[Piz90] Arnold K. Pizer. "Ramanujan graphs and Hecke operators". In: *Bull. Amer. Math. Soc. (N.S.)* 23.1 (1990), pp. 127–137. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-1990-15918-X. URL: https://doi.org/10.1090/S0273-0979-1990-15918-X (cit. on p. 168).

[PS24] Amaury Pouly and Yixin Shen. *Solving the Shortest Vector Problem in* $2^{0.63269n+o(n)}$ *time on Random Lattices*. Cryptology ePrint Archive, Paper 2024/1805. 2024. URL: https://eprint.iacr.org/2024/1805 (cit. on pp. 147, 148, 150).

[PQS24] PQShield. *Post-Quantum signatures zoo*. https://pqshield.github.io/nist-sigs-zoo/. Oct. 2024 (cit. on p. 96).

[Pre23] Thomas Prest. "A Key-Recovery Attack Against Mitaka in the t-Probing Model". In: *PKC 2023, Part I*. Ed. by Alexandra Boldyreva and Vladimir Kolesnikov. Vol. 13940. LNCS. Springer, Cham, May 2023, pp. 205–220. DOI: 10.1007/978-3-031-31368-4_8 (cit. on p. 117).

[PTT18] Joscha Prochno, Christoph Thäle, and Nicola Turchi. *Geometry of ℓ_p^n-balls: Classical results and recent developments*. 2018. arXiv: 1808.10435 [math.PR] (cit. on p. 133).

[Reg05] Oded Regev. "On lattices, learning with errors, random linear codes, and cryptography". en. In: *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing - STOC '05*. Accessed: 15/08/2022. Baltimore, MD, USA: ACM Press, 2005, p. 84. ISBN: 9781581139600. DOI: 10.1145/1060590.1060603. URL: http://portal.acm.org/citation.cfm?doid=1060590.1060603 (cit. on pp. 17, 38).

[Rog55] Claude A. Rogers. "Mean values over the space of lattices". In: *Acta Mathematica* 94.0 (1955), pp. 249–287 (cit. on p. 146).

[RS06] Alexander Rostovtsev and Anton Stolbunov. *PUBLIC-KEY CRYPTOSYSTEM BASED ON ISOGENIES*. Cryptology ePrint Archive, Paper 2006/145. 2006. URL: https://eprint.iacr.org/2006/145 (cit. on pp. 22, 43).

[RH23] Keegan Ryan and Nadia Heninger. "Fast Practical Lattice Reduction Through Iterated Compression". In: *Advances in Cryptology – CRYPTO 2023, Part III*. Ed. by Helena Handschuh and Anna Lysyanskaya. Lecture Notes in Computer Science. Santa Barbara, CA, USA: Springer, Cham, Switzerland, Aug. 2023, pp. 3–36. DOI: 10.1007/978-3-031-38548-3_1 (cit. on pp. 111, 112).

[Sar80] Peter Sarnak. "Prime geodesic theorems". PhD thesis. Stanford, 1980 (cit. on p. 153).

[SZ90] Gideon Schechtman and Joel Zinn. "On the Volume of the Intersection of Two Ln p Balls". In: *Proceedings of the American Mathematical Society* 110.1 (1990), pp. 217–224. ISSN: 00029939, 10886826. URL: http://www.jstor.org/stable/2048262 (visited on 02/13/2024) (cit. on p. 133).

[Sch60] Wolfgang M. Schmidt. "A metrical theorem in geometry of numbers". en. In: *Transactions of the American Mathematical Society* 95.3 (1960), pp. 516–529 (cit. on p. 147).

[SG] Michael Schneider and Nicolas Gama. "SVP Challenge". URL: http://www.latticechallenge.org/svp-challenge/ (cit. on pp. 74, 82).

[Sch03] Claus Peter Schnorr. "Lattice reduction by random sampling and birthday methods". In: *Proc. STACS 2003*. Vol. 2607. LNCS. Springer, 2003, pp. 145–156 (cit. on p. 50).

[SE94] Claus-Peter Schnorr and M. Euchner. "Lattice basis reduction: improved practical algorithms and solving subset sum problems". In: *Math. Programming* 66 (1994), pp. 181–199 (cit. on pp. 50, 51).

[Sch95] René Schoof. “Counting points on elliptic curves over finite fields”. en. In: *Journal de Théorie des Nombres de Bordeaux* 7.1 (1995), pp. 219–254. URL: http://www.numdam.org/item/JTNB_1995___7_1_219_0/ (cit. on p. 60).

[Sch08] René Schoof. “Computing Arakelov Class Groups”. In: *Algorithmic Number Theory: Lattices, Number Fields, Curves and Cryptography* (2008), pp. 447–495 (cit. on p. 58).

[Sho94] Peter W. Shor. “Algorithms for quantum computation: discrete logarithms and factoring”. In: *Proceedings 35th Annual Symposium on Foundations of Computer Science*. 1994, pp. 124–134. DOI: 10.1109/SFCS.1994.365700 (cit. on pp. 5, 27).

[Sie45] Carl Ludwig Siegel. “A Mean Value Theorem in Geometry of Numbers”. In: *The Annals of Mathematics* 46.2 (Apr. 1945), p. 340 (cit. on pp. 16, 36, 146).

[Sil94] Joseph H. Silverman. *Advanced topics in the arithmetic of elliptic curves*. Vol. 151. Graduate Texts in Mathematics. Springer-Verlag, New York, 1994, pp. xiv+525. ISBN: 0-387-94328-5. DOI: 10.1007/978-1-4612-0851-8. URL: https://doi.org/10.1007/978-1-4612-0851-8 (cit. on pp. 59, 174).

[Sil09] Joseph H. Silverman. *The Arithmetic of Elliptic Curves*. Graduate Texts in Mathematics. Springer New York, 2009. ISBN: 9780387094946. URL: https://books.google.co.uk/books?id=Z90CA%5C_EUCCkC (cit. on pp. 54, 58, 171).

[SSTX09] Damien Stehlé, Ron Steinfeld, Keisuke Tanaka, and Keita Xagawa. “Efficient Public-Key Encryption Based on Ideal Lattices (Extended Abstract)”. In: *Asiacrypt 2009*. Japan, 2009, pp. 617–635. URL: https://hal.archives-ouvertes.fr/hal-00550978 (cit. on pp. 17, 38).

[Ste17] Noah Stephens-Davidowitz. “On the Gaussian measure over lattices”. PhD Thesis. New York University, 2017 (cit. on p. 46).

[ST16] Ian Stewart and David Orme Tall. *Algebraic number theory and Fermat’s last theorem*. Fourth edition. Boca Raton: CRC Press, Taylor & Francis Group, 2016. ISBN: 9781498738392 (cit. on p. 54).

[Sut12] Andrew Sutherland. “Identifying supersingular elliptic curves”. en. In: *LMS Journal of Computation and Mathematics* 15 (Sept. 2012), pp. 317–325. ISSN: 1461-1570. DOI: 10.1112/S1461157012001106. URL: https://www.cambridge.org/core/product/identifier/S1461157012001106/type/journal_article (visited on 08/26/2021) (cit. on p. 168).

[Sut13] Andrew Sutherland. “Isogeny volcanoes”. en. In: *The Open Book Series* 1.1 (Nov. 2013), pp. 507–530. ISSN: 2329-907X, 2329-9061. DOI: 10.2140/obs.2013.1.507. URL: http://msp.org/obs/2013/1-1/p25.xhtml (visited on 05/18/2021) (cit. on pp. 23, 43, 59, 168, 176).

[Szy03] Michael Szydlo. “Hypercubic Lattice Reduction and Analysis of GGH and NTRU Signatures”. In: *Advances in Cryptology - Proc. EUROCRYPT 2003*. Ed. by Eli Biham. Vol. 2656. Lecture Notes in Computer Science. Springer, 2003, pp. 433–448 (cit. on pp. 64, 81).

[TM73] Hidetosi Takahasi and Masatake Mori. “Double Exponential Formulas for Numerical Integration”. In: *Publications of the Research Institute for Mathematical Sciences* 9.3 (Dec. 1973), pp. 721–741. ISSN: 0034-5318, 1663-4926. DOI: 10.2977/prims/1195192451. URL: https://ems.press/doi/10.2977/prims/1195192451 (visited on 02/25/2025) (cit. on p. 165).

[Tat66] John Tate. “Endomorphisms of abelian varieties over finite fields”. In: *Invent. Math.* 2 (1966), pp. 134–144. ISSN: 0020-9910. DOI: 10.1007/BF01404549. URL: https://doi.org/10.1007/BF01404549 (cit. on p. 172).

[Tem92] Nico M. Temme. “Asymptotic inversion of the incomplete beta function”. In: *Journal of Computational and Applied Mathematics* 41.1 (1992), pp. 145–157 (cit. on pp. 71, 92).

[The24] The FPLLL development team. “fpylll, a Python wrapper for the fplll lattice reduction library, Version: 0.6.1”. Available at https://github.com/fplll/fpylll. 2024. URL: https://github.com/fplll/fpylll (cit. on p. 111).

[Thu98] Jeffrey Lin Thunder. “Higher-dimensional analogs of Hermite’s constant.” In: *Michigan Mathematical Journal* 45 (1998), pp. 301–314 (cit. on p. 110).

[Ven13] Akshay Venkatesh. “A Note on Sphere Packings in High Dimension”. In: *International Mathematics Research Notices* 2013.7 (Mar. 2013), pp. 1628–1642. ISSN: 1073-7928. DOI: 10.1093/imrn/rns096. eprint: https://academic.oup.com/imrn/article-pdf/2013/7/1628/19078935/rns096.pdf. URL: https://doi.org/10.1093/imrn/rns096 (cit. on pp. 20, 41, 146).

[W] Weisstein Eric W. “Double Series.” In: *MathWorld–A Wolfram Web Resource.* (). URL: https://mathworld.wolfram.com/DoubleSeries.html (cit. on p. 150).

[Was97] L. C. Washington. *Introduction to Cyclotomic Fields*. Vol. 83. Graduate Texts in Mathematics. New York, NY: Springer New York, 1997. ISBN: 9781461273462. DOI: 10.1007/978-1-4612-1934-7. URL: http://link.springer.com/10.1007/978-1-4612-1934-7 (visited on 07/29/2022) (cit. on p. 55).

[Wat69] William C. Waterhouse. “Abelian varieties over finite fields”. In: *Annales scientifiques de l’École normale supérieure* 2.4 (1969), pp. 521–560. ISSN: 0012-9593, 1873-2151. DOI: 10.24033/asens.1183. URL: http://www.numdam.org/item?id=ASENS_1969_4_2_4_521_0 (visited on 05/17/2021) (cit. on pp. 59, 170, 172, 174, 179).

[WW21] Thom Wiggers and Bas Westerbaan. *Sizing up post-quantum signatures*. Cloudflare Blog. Accessed: 2025-07-11. Nov. 2021. URL: https://blog.cloudflare.com/sizing-up-post-quantum-signatures/ (cit. on p. 117).

[Yam70] Yoshihiko Yamamoto. “On unramified Galois extensions of quadratic number fields”. In: *Osaka Math. J.* 7 (1970), pp. 57–76. ISSN: 0388-0699. URL: http://projecteuclid.org/euclid.ojm/1200692686 (cit. on pp. 170, 183).

RÉSUMÉ

La menace que représente l'arrivée prochaine d'un ordinateur quantique assez puissant pour briser la cryptographie du début du XXIe siècle a contraint les chercheurs à développer de nouveaux algorithmes pour le chiffrement et la signature numérique. La sécurité de ces protocoles repose sur la difficulté présumée de problèmes mathématiques autres que la factorisation ou du logarithme discret, dont le principal est le problème du plus court vecteur dans les réseaux euclidiens. Pour des raisons d'efficacité, les réseaux utilisés dans les premières constructions élevées au rang de norme sont dotés d'une structure particulière, souvent issue de la géométrie des corps de nombres. Il est donc crucial de comprendre les attaques qui exploitent cette structure afin de choisir des paramètres assurant réellement leur sécurité.

Le premier grand axe de cette thèse est l'étude de différentes familles de réseaux utilisés en cryptologie, sous divers aspects :

- Sur le plan algorithmique, nous étudions des attaques prouvées et heuristiques sur le problème du plus court vecteur dans les réseaux NTRU et les réseaux hypercubiques.
- Sur le plan mathématique, nous étudions le comportement moyen des vecteurs les plus courts dans les réseaux issus d'idéaux de corps de nombres.

Le second grand axe est l'amélioration de potentielles normes post-quantiques pour les signatures numériques :

- Sur le plan offensif, nous élaborons une attaque de récupération de clé par réduction de réseaux sur le schéma de signature DEFI, prétendu post-quantique par ses auteurs.
- Sur le plan défensif, nous proposons un nouveau type de distribution pour cacher l'information secrète dans les schémas de signature de type Fiat-Shamir avec Rejets à base de réseaux. Cela nous permet d'élaborer sous les mêmes hypothèses de sécurité un schéma de signature plus compact que la norme Dilithium, mais qui contrairement à la norme Haetae n'utilise pas de distributions gaussiennes, difficiles à implémenter de façon sécurisée.

MOTS CLÉS

Cryptographie Post-Quantique ⋆ Réseaux Euclidiens ⋆ Cryptanalyse ⋆ Signatures ⋆ Théorie des Nombres

ABSTRACT

Early 21st-century cryptography is threatened by the fact that a robust quantum computer could be built in the near future. This threat has forced researchers into designing new algorithms for encryption and digital signatures, whose security relies on the presumed computational hardness of solving mathematical problems different from factoring or discrete logarithm, such as the shortest vector problem in Euclidean lattices.
Due to efficiency concerns, the first standardised schemes rely on lattices with extra structure, often derived from the geometry of well-chosen number fields. Understanding the attacks that exploit this particular structure is crucial in order to select parameters that truly ensure security.

The first main topic of this thesis is the study of different families of lattices used in cryptology:

- On the algorithmic level, we study provable and heuristic attacks on the shortest vector problem in NTRU and hypercubic lattices.
- On a more mathematical level, we study the average behaviour of the shortest vectors in ideal lattices: lattices that inherit structure from ideals of number fields.

The second main topic is the improvement of potential post-quantum standards for digital signatures:

- From an attacker's perspective, we develop a key recovery attack using lattice reduction on the DEFI signature scheme, which its authors claimed to be post-quantum.
- From a defender's perspective, we propose a new type of distribution to hide the secret information in Fiat-Shamir with Aborts lattice-based signature schemes. This allows us to construct, under the same security assumptions, a more compact signature scheme than the Dilithium standard, but which, unlike the Haetae standard, does not use Gaussian distributions, known to be difficult to implement securely.

KEYWORDS

Post-Quantum Cryptography ⋆ Lattices ⋆ Cryptanalysis ⋆ Signatures ⋆ Number Theory