NTRU Fatigue: How Stretched is Overstretched ?

Léo Ducas & Wessel van Woerden

Cryptology Group, CWI, Amsterdam, The Netherlands

Abstract. Until recently lattice reduction attacks on NTRU lattices were thought to behave similar as on (ring-)LWE lattices with the same parameters. However several works (Albrecht-Bai-Ducas 2016, Kirchner-Fouque 2017) showed a significant gap for large moduli q, the so-called overstretched regime of NTRU.
With the NTRU scheme being a finalist to the NIST PQC competition it is important to understand —both asymptotically and concretely— where the fatigue point lies exactly, *i.e.* at which q the overstretched regime begins. Unfortunately the analysis by Kirchner and Fouque is based on an impossibility argument, which only results in an asymptotic upper bound on the fatigue point. It also does not really *explain* how lattice reduction actually recovers secret-key information.
We propose a new analysis that asymptotically improves on that of Kirchner and Fouque, narrowing down the fatigue point for ternary NTRU from $q \leqslant n^{2.783+o(1)}$ to $q = n^{2.484+o(1)}$, and finally explaining the mechanism behind this phenomenon. We push this analysis further to a concrete one, settling the fatigue point at $q \approx 0.004 \cdot n^{2.484}$, and allowing precise hardness predictions in the overstretched regime. These predictions are backed by extensive experiments.

1 Introduction

1.1 Context

One should certainly recognize that in the field of lattice-based cryptography the NTRU cryptosystem of Hoffstein, Pipher and Silverman [HPS98, CDH+20] was particularly ahead of its time. After two decades spent basing cryptography [Ajt99, Reg05, SSTX09] on the worst-case hardness of lattice problems and concretising this theory into practical cryptosystems for standardisation [PAA+19, SAB+20, DKR+20], it is quite remarkable to see these constructions landing not so far away from the original design of NTRU (q-ary lattices, module structure over similar polynomial rings). In fact, it was even discovered a posteriori, that, up to the choice of parameters, the NTRU scheme itself can also be supported by worst-case hardness [SS11].

Regarding cryptanalysis, it was only recently discovered that the security of NTRU is in fact more subtle than the problem of finding a single unusually short vector in a lattice. The first dent in this status quo came in 2016, from two concurrent works work of Albrecht et al., and Cheon et al. [ABD16, CJL16], which exploits the specific algebraic structure of the NTRU lattice to improve

upon pure lattice reduction attacks[1]. This approach was shown to be applicable when the modulus q is large enough (say, super-polynomial), a regime coined "overstretched".

Shortly thereafter Kirchner and Fouque [KF17] showed that this improved complexity does *not* require any algebraic structure, and is instead rooted in the purely geometrical fact that the NTRU lattice contains an unusually dense sublattice of large dimension, *i.e.* a sublattice of small determinant.[2] They also go further in their analysis, and conclude that moduli q as small as $n^{2.783+o(1)}$ already belong to the overstretched regime —for random ternary secrets. In particular, for q larger than this bound, the security of NTRU is significantly less than that of Learning With Errors [Reg04] and of its Ring variant [SSTX09, LPR13] using similar parameters.[3]

However, it is not so clear from the analysis of Kirchner and Fouque whether this asymptotic quantity $n^{2.783+o(1)}$ is an estimate or merely an upper bound on the *fatigue point*, that is the value of q separating the standard regime from the overstretched regime. Their analysis is based on a lemma of Pataki and Tural [PT08], that constraints the shape of lattice basis in terms of the volume of their sublattices. While it allows to conclude that the dense sublattice must be discovered after reducing the lattice basis beyond these constraints, it does not really explain *how* lattice reduction ends up discovering the dense sublattice, nor does it exclude that the discovery could happen earlier.

So far, it has been generally considered that only advanced schemes — requiring very large q— such as NTRU-based Homomorphic Encryption [BLLN13] or candidate cryptographic multi-linear maps [GGH13] could be affected by this overstretched regime. Yet, because the analysis of Kirchner and Fouque is only asymptotic, and because it may only provide an upper bound on the fatigue point, there is at the moment little documented evidence that the overstretched regime may not in fact extend further down, maybe down to the NTRU encryption scheme itself [HPS98, CDH+20]! Admittedly, this seems like a far fetched concern: asymptotically this scheme chooses $q = O(n)$, with a hidden constant between 4 and 5 in practice. However, this scheme being now a finalist of the NIST standardisation process for post-quantum cryptography, it appears rather imperious to refine our understanding of the phenomenon, and to finally close this pending question.

We found further motivation to go down this rabbit hole by measuring the concrete value of fatigue point experimentally. Until now, all documented experiments on the overstretched regime [ABD16, KF17, LW20] have focused on rather large values of q, and only used weak lattice reduction (LLL [LLL82], BKZ with blocksize 20): their goal was to demonstrate the claimed general be-

[1] Though the idea had been inconclusively considered already in 2002 by Gentry, Jonsson, Nguyen Stern and Szydlo as reported in [GS02, Sec. 6].

[2] Note that one may associate a short vector to a dense sublattice of dimension 1.

[3] In fact, the presence of n rotations of the secret key already implies a minor security degradation compared to (Ring)-LWE already in the standard regime [MS01, DDGR20].

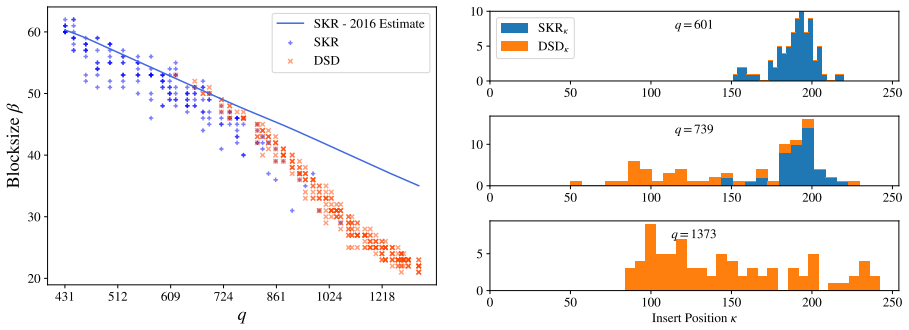

Fig. 1. Progressive BKZ with 8 tours per blocksize on matrix NTRU instances with parameters $n = 127, \sigma^2 = \frac{2}{3}$ for several moduli q. **Left:** the first blocksize β at which Progressive BKZ detects the Secret Key Recovery (SKR_κ) or Dense Sublattice Discovery (DSD_κ) event. We did 10 runs per modulus q. For the 2016-estimates, we use the geometric series assumption (GSA) for the shape of the basis and a probabilistic model for the discovery of the secret vector (see Section 2.4). **Right:** the positions κ at which a secret key or dense sublattice vector are detected over 80 runs per modulus.

haviour when parameters are far in the overstretched regime. On the contrary, we focus our attention to the fatigue point for this preliminary experiment. That is, we ran strong reduction (progressive-BKZ [Sch87, AWHT16] up to blocksize 60) until a vector related to the secret key appeared for a range of moduli q. We distinguished the standard regime from the overstretched regime by classifying according to which event occurs first

- Secret Key Recovery (SKR_κ): a vector as short as a secret key vector is inserted in the basis at any given position κ.
- Dense Sublattice Discovery (DSD_κ): a vector strictly longer than the secret key but belonging to the dense sublattice generated by the secret key is inserted in the basis at any given position κ.

The result (Fig. 1) is rather striking: for $n = 127$, we start seeing a deviation from the standard regime for q as small as 700, while a naive interpretation of the prediction by Kirchner and Fouque [KF17] would suggest a fatigue point at $q \approx n^{2.783} \approx 700\ 000$. We can conclude either that the asymptotic bound is not tight, or that the hidden asymptotic term (the $o(1)$ in $n^{2.783+o(1)}$) is significantly negative in practice. In any case, the bound of Kirchner and Fouque does not seem to provide accurate concrete predictions.

Remark. At this point, we should clarify why the DSD event should essentially be considered a successful attack. First, for q not too much larger than the fatigue point, an SKR event typically quickly follows after the DSD event; what happens is that DSD events cascade, until the full dense sublattice has been extracted: the first half of the reduced basis precisely generates the dense sublattice. Lattice

reduction will happen independently on each half of the basis, meaning that the dimension of the search space for the secret key has effectively been halved, and therefore making the problem much easier.

However, as q increases, DSD becomes easier and easier, to the point that it becomes even easier than secret key recovery within the dense sublattice. In other terms, there is a *superstretched* regime for larger q, where DSD does not directly lead to SKR.

Nevertheless, we argue —essentially rephrasing [ABD16]— that the DSD event is typically sufficient for an attack. First, the dense sublattice vector discovered is of length significantly lower than q; in an FHE scheme such as [BLLN13] it is sufficient to decrypt fresh ciphertexts.[4] Secondly, in the case of cyclotomic or circulant NTRU, it is possible to recover the secret key from the dense sublattice by other means than pure lattice reduction; in particular the recent line of work on the principal ideal-SVP [EHKS14, CDPR16, BEF+17] showed that this can be done classically in sub-exponential time $\exp(\tilde{O}(\sqrt{n}))$ and quantumly in polynomial time.

1.2 Our work

Having identified precisely what event distinguishes the standard regime of NTRU from its overstretched regime, we may now proceed to a refined analysis, and determine precisely both the fatigue point and the precise cost[5] of attacks in the overstretched regime. Our refined analysis diverges from the one of Kirchner and Fouque [KF17] on the following points:

1. we exploit the fact that BKZ runs SVP on large blocks ($\beta \geqslant 2$) not only to deduce the shape of the basis, but also to actually discover dense sublattice vectors,
2. we do not solely focus on the behaviour at position $\kappa = n - \beta + 1$ out of $d = 2n$ dimensions, but instead predict the most relevant position,
3. we propose an average-case analysis of volumes of the relevant lattices and sublattices, leading to a concrete prediction rather than a worst-case bound,
4. we also validate our intermediate and final predictions quantitatively with extensive experiments.

We note that contributions 1 and 2 alone already give us an important asymptotic result: the fatigue point of NTRU is indeed lower than predicted by Kirchner and Fouque, namely, it should happen at $q = n^{2.484+o(1)}$ instead of $n^{2.783+o(1)}$.

Furthermore, our concrete average case analysis also differentiates the *circulant* version of NTRU [HPS98] from its *matrix* version [CG05, GGH+19]. We note minor deviations in the concrete analysis of volumes of relevant sublattices,

[4] The secret key being shorter is only required to deal with ciphertexts obtained by homomorphic computation.

[5] In this work, we only measure cost of lattice reduction in terms of the required BKZ blocksize; the computational cost of BKZ is essentially an orthogonal question.

that on average slightly favours the attacker in the matrix case, but also shows a larger variance in the concrete hardness of the circulant case.

In summary: we achieve an explicative and predictive model for the fatigue of NTRU, with concrete predictions confirmed in practice. In particular, the fatigue point is estimated to be at $q \approx 0.004 \cdot n^{2.484}$ for $n > 100$. All our artefacts for experiments and predictions are open-source and can be accessed at https://github.com/WvanWoerden/NTRUFatigue. These are based on the FPLLL and FPyLLL libraries [dt21a, dt21b].

Impact. We wish to clarify that this work *does not* contradict the concrete security of the NTRU candidate to the NIST competition [CDH+20]; on the contrary, we close a pending question regarding a potential vulnerability.

Limitation: the Lucky-Lifts. During our experiments, we also noted rare occurrence of DSD events that qualitatively differ from what we expected. Namely, the vector from the dense sublattice was found at positions κ quite larger than what was predicted by our model. More remarkable, these vectors were extremely unbalanced: their $2n - \kappa$ last (Gram-Schmidt) coordinates were much smaller than the κ first coordinates We call these DSD events lucky-lifts (DSD-LL), while the one we model and mostly observe are called after the Pataki-Tural Lemma (DSD-PT). Despite those two phenomena being very distinct, they nevertheless occured for the same BKZ blocksizes β, at least in the range of parameters we could experiment with.

It could very well be that these rare DSD-LL events are just artefacts of the modest parameters of our experiments and that these events vanish as the dimension grows. Yet, as they seem of a very different nature, a definitive conclusion would require a dedicated study.

1.3 Organisation

We introduce some preliminaries, the NTRU lattice, and the state-of-the-art estimates in Section 2. In Section 3 we introduce our new DSD-PT estimate and give an asymptotic analysis. In Section 4 we give an average-case analysis to construct a concrete estimator. In the final Section 5 we compare our estimate with experiments.

1.4 Acknowledgements

We would like to thank Martin Albrecht and Paul Kirchner for valuable comments and discussions. The research of L. Ducas was supported by the European Union's H2020 Programme under PROMETHEUS project (grant 780701) and the ERC-StG-ARTICULATE project (no. 947821). W. van Woerden is funded by the ERC-ADG-ALGSTRONGCRYPTO project (no. 740972).

2 Preliminaries

2.1 Notation and distributions

All vectors and matrices are denoted by bold lower and upper case letters respectively. All vectors are column-vectors and we write $\boldsymbol{B} = [\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}]$ for a matrix where the i-th column vector is $\mathbf{b}_i$. If a matrix $\boldsymbol{B} \in \mathbb{R}^{d \times n}$ has full rank n we denote by $\mathcal{L}(\boldsymbol{B}) := \{\boldsymbol{B}\mathbf{x} : \mathbf{x} \in \mathbb{Z}^n\}$ the lattice spanned by the columns of $\boldsymbol{B}$. We call a lattice vector $\mathbf{v} \in \mathcal{L}$ primitive if it is not a strict integer multiple of another lattice vector. For a basis $\boldsymbol{B}$ and $i \in \{0, \ldots, d-1\}$ we define π_i as the orthogonal projection away from $\mathbf{b}_0, \ldots, \mathbf{b}_{i-1}$, and the Gram-Schmidt vectors as $\mathbf{b}_0^*, \ldots, \mathbf{b}_{d-1}^*$ where $\mathbf{b}_i^* := \pi_i(\mathbf{b}_i)$. We write $\boldsymbol{B}_{[l:r)}$ for the matrix $[\pi_l(\mathbf{b}_l), \ldots, \pi_l(\mathbf{b}_{r-1})]$, and denote the projected[6] sublattice $\mathcal{L}(\boldsymbol{B}_{[l:r)})$ as $\mathcal{L}_{[l:r)}$ when the basis is clear from the context. We denote the Euclidean norm of a vector $\mathbf{v}$ by $\|\mathbf{v}\|$ and the volume of a lattice by $\mathrm{vol}(\mathcal{L}(\boldsymbol{B})) := \prod_{i=0}^{n-1} \|\mathbf{b}_i^*\|$. We write $\lambda_1(\mathcal{L}) := \min_{\mathbf{v} \in \mathcal{L} \setminus \{0\}} \|\mathbf{v}\|$ for the first minimum of a lattice $\mathcal{L}$. For a lattice $\mathcal{L}$ we denote the dual lattice as $\mathcal{L}^* := \{\mathbf{w} \in \mathrm{span}(\mathcal{L}) : \langle \mathbf{w}, \mathbf{v} \rangle \in \mathbb{Z} \text{ for all } \mathbf{v} \in \mathcal{L}\}$. We use 'claim' to refer to an informal statement based on heuristics.

We denote the continuous centered Gaussian (normal) distribution with variance σ^2 by χ_{σ^2}. We denote the unit sphere over k coordinates as $\mathcal{S}^{k-1}$ and call the uniform distribution over $\mathcal{S}^{k-1}$ the spherical distribution. We write $\mathcal{B}_1^d$ for the d-dimensional unit ball. We write the chi-square distribution with k degrees of freedom as $\chi^2_{k,\sigma^2} := \sum_{i=1}^{k} X_i^2$, where $X_1, \ldots, X_k$ are independently distributed as χ_{σ^2}. The chi-square distribution has expectation $k\sigma^2$, but for our concrete estimates we consider the log-expectation.

Lemma 2.1. *Let X be distributed as χ^2_{k,σ^2}, then*

$$\mathbb{E}\left[\ln\left(X\right)\right] = \ln(2\sigma^2) + \psi(k/2),$$

where $\psi(x) := \Gamma'(x)/\Gamma(x)$ is the digamma function.

2.2 NTRU and lattice attacks

We start with the historical definition of NTRU.

Definition 2.2 (NTRU). *Let n be prime, q a positive integer and let $\boldsymbol{f}, \boldsymbol{g} \in (\mathbb{Z}/q\mathbb{Z})[X]$ be polynomials of degree n with small coefficients sampled from some distribution χ under the condition that $\boldsymbol{f}$ is invertible in $\mathcal{R}_q := (\mathbb{Z}/q\mathbb{Z})[X]/(X^n - 1)$. The pair $(\boldsymbol{f}, \boldsymbol{g})$ forms the secret key, and the public key is defined as $\boldsymbol{h} := \boldsymbol{g}/\boldsymbol{f} \bmod \mathcal{R}_q$. The* NTRU problem *is to recover any rotation $(X^i\boldsymbol{f}, X^i\boldsymbol{g})$ of the secret key from $\boldsymbol{h}$.*

For *NTRUencrypt* [HPS98, CDH+20] $\boldsymbol{f}$ and $\boldsymbol{g}$ have ternary coefficients, with a fixed number of about $n/3$ of each value in $\{-1, 0, 1\}$. For our analysis we

[6] When $l = 0$, no projection is applied, and $\mathcal{L}_{[0:r)}$ is simply a sublattice of $\mathcal{L}$.

consider the case where each coefficient is sampled from a discrete Gaussian over $\mathbb{Z}$ with some variance $\sigma^2 > 0$. For simplicity the ternary case is treated as a discrete Gaussian with variance $\sigma^2 = \frac{2}{3}$.

More generally we consider a matrix description of NTRU where the polynomials are replaced by matrices $\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{H} \in \mathbb{Z}^{n \times n}$ such that $\boldsymbol{H} := \boldsymbol{G} \cdot \boldsymbol{F}^{-1} \bmod q$ [CG05, GGH+19]. Variants of NTRU, e.g. based on different algebraic rings [BBC+20], can be encoded in the structure of the matrices. For example, the original problem can be encoded by setting $\boldsymbol{F}_{i,j} := f_{(i+j \bmod n)}$ where $\boldsymbol{f} = \sum_{i=0}^{n-1} f_i X^i$, for each polynomial respectively. We call the original variant *circulant NTRU*, based on the resulting shape of the matrices $\boldsymbol{F}, \boldsymbol{G}$, and we treat $\boldsymbol{f}, \boldsymbol{g}$ as n-dimensional vectors. We also consider the variant, called *matrix NTRU*, where the matrices $\boldsymbol{F}, \boldsymbol{G}$ have no extra structure and the coefficients are independently sampled from a discrete Gaussian.

To reduce the NTRU problem to a lattice problem we define the *NTRU lattice*, which contains a particularly *dense* sublattice generated by the secret key.

Definition 2.3. *Let $(n, q, \boldsymbol{F}, \boldsymbol{G}, \boldsymbol{H})$ be an NTRU instance. We define the NTRU lattice as*

$$\mathcal{L}^{\boldsymbol{H},q} := \begin{pmatrix} q\boldsymbol{I}_n & \boldsymbol{H} \\ \boldsymbol{0} & \boldsymbol{I}_n \end{pmatrix} \cdot \mathbb{Z}^{2n},$$

and its (secret) dense sublattice of rank n by:

$$\mathcal{L}^{\boldsymbol{GF}} := \boldsymbol{B}^{\boldsymbol{GF}} \cdot \mathbb{Z}^n \subset \mathcal{L}^{\boldsymbol{H},q}, \text{ where } \boldsymbol{B}^{\boldsymbol{GF}} := \begin{pmatrix} \boldsymbol{G} \\ \boldsymbol{F} \end{pmatrix}.$$

Solving the NTRU problem is equivalent to recovering the dense sublattice basis $\boldsymbol{B}^{\boldsymbol{GF}} = [\boldsymbol{G}; \boldsymbol{F}]$ up to some permutation of the columns. For uniformity of notation we will denote such a column by $(\boldsymbol{g}|\boldsymbol{f})$. These column vectors have a length of about $\|(\boldsymbol{g}|\boldsymbol{f})\| \approx \sqrt{2n\sigma^2}$, which for common parameters is much shorter than the expected minimal length $\lambda_1(\mathcal{L}^{\boldsymbol{H},q}) \approx \sqrt{nq/(\pi e)}$ of the full lattice $\mathcal{L}^{\boldsymbol{H},q}$ for a truly uniform random $\boldsymbol{H} \in (\mathbb{Z}/q\mathbb{Z})^{n \times n}$. To recover the secret key we thus have to find these exceptionally short vectors in the full lattice $\mathcal{L}^{\boldsymbol{H},q}$.

In [CS97] Coppersmith and Shamir showed that we can slightly relax the problem as any small vector from the dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ is enough to decode a message. We therefore focus our analysis on the recovery of elements from $\mathcal{L}^{\boldsymbol{GF}}$, and not (directly) on the full secret basis $\boldsymbol{B}^{\boldsymbol{GF}}$. To recover short vectors we resort to lattice reduction.

2.3 Lattice Reduction

Any lattice $\mathcal{L} = \mathcal{L}(\boldsymbol{B})$ with basis $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ has (for $d > 1$) an infinite number of other bases $\boldsymbol{B} \cdot \boldsymbol{U}$ with $\boldsymbol{U} \in \mathrm{GL}_d(\mathbb{Z})$. The goal of lattice reduction is to find a *good* basis: the basis vectors are preferably short and somewhat orthogonal. Looking at the Gram-Schmidt vectors $\mathbf{b}_0^*, \ldots, \mathbf{b}_{d-1}^*$ we have the invariant $\prod_{i=0}^{d-1} \|\mathbf{b}_i^*\| =$

$\det(\boldsymbol{B}) = \mathrm{vol}(\mathcal{L})$ which is independent of the basis. Therefore decreasing the length of the first basis vector $\mathbf{b}_0 = \mathbf{b}_0^*$ forces some of the other Gram-Schmidt vectors to increase in length. We call these lengths $(\|\mathbf{b}_i^*\|)_{i=0,\ldots,d-1}$ the *profile* of a basis $\boldsymbol{B}$. A good basis has a well balanced profile; in particular one that does not decrease too fast.

The most famous lattice reduction algorithm is the polynomial time LLL algorithm, which gives some guarantees on the slope of an LLL-reduced basis. We consider a generalisation, namely the BKZ algorithm, that gives a flatter slope, but at a higher cost. A basis is BKZ reduced with blocksize β if $\mathbf{b}_\kappa^*$ is a shortest vector of the projected sublattice $\mathcal{L}_{[\kappa:\min(\kappa+\beta,d))}$ at each position κ. LLL-reduction corresponds to the case that $\beta = 2$.

Definition 2.4 (BKZ). *A basis $\boldsymbol{B} = [\mathbf{b}_0, \ldots, \mathbf{b}_{d-1}]$ is called BKZ-β reduced if*

$$\|\mathbf{b}_\kappa^*\| = \lambda_1(\mathcal{L}_{[\kappa:\min(\kappa+\beta,d))}) \text{ for all } \kappa = 0, \ldots, d-1.$$

A BKZ-reduced basis has several provable bounds on the slope of the profile. In the context of cryptanalysis we are more interested in the average-case behaviour and thus we fall back on heuristics to describe the shape of a BKZ-reduced profile. The most commonly used heuristic for lattices is the Gaussian Heuristic, that states that for a measurable volume $\mathcal{V}$ the number of lattice points $|\mathcal{L} \cap \mathcal{V}|$ approximately equals $\mathrm{vol}(\mathcal{V})/\mathrm{vol}(\mathcal{L})$. Applying this to a ball allows to estimate the first minimum of a lattice.

Heuristic 2.5 *Let $\mathcal{L}$ be a d-dimensional lattice with volume $\mathrm{vol}(\mathcal{L})$. The expectation of the first minimum $\lambda_1(\mathcal{L})$ under the Gaussian Heuristic is given by*

$$\mathrm{gh}(\mathcal{L}) := \frac{\mathrm{vol}(\mathcal{L})^{1/d}}{\mathrm{vol}(\mathcal{B}_1)^{1/d}} \approx \sqrt{d/(2\pi e)} \cdot \mathrm{vol}(\mathcal{L})^{1/d}.$$

We also denote $\mathrm{gh}(d) \approx \sqrt{d/(2\pi e)}$ for the expected first minimum of a d-dimensional lattice with volume 1.

Applying the above heuristic to the value of $\|\mathbf{b}_\kappa^*\| = \lambda_1(\mathcal{L}_{[\kappa:\min(\kappa+\beta,d))})$ at each position κ gives us relations between the Gram-Schmidt lengths $\|\mathbf{b}_0^*\|, \ldots, \|\mathbf{b}_{d-1}^*\|$. Solving these relations for $\beta \ll d$ shows that $\|\mathbf{b}_\kappa^*\| / \|\mathbf{b}_{\kappa+1}^*\| \approx \alpha_\beta$ for some constant α_β only depending on β. So heuristically the profile forms a geometric series. This is made more precise by the Geometric Series Assumption.

Heuristic 2.6 (Geometric Series Assumption (GSA)) *Let $\boldsymbol{B}$ be a BKZ-β reduced basis, then the profile satisfies*

$$\ln(\|\mathbf{b}_i^*\|) = \frac{d-1-2i}{2} \cdot \ln(\alpha_\beta) + \frac{\ln(\det(\boldsymbol{B}))}{d},$$

where $\alpha_\beta = \mathrm{gh}(\beta)^{2/(\beta-1)}$.

The GSA is reasonably precise for say $\beta \geqslant 50$ and a not too large blocksize $\beta \ll d$ compared to the lattice dimension.

The BKZ algorithm (see Algorithm 1) computes a BKZ-reduced basis from any other basis. The algorithm greedily attempts to satisfy the BKZ condition at each position by computing a shortest vector in each *block* $\boldsymbol{B}_{[\kappa:\min(\kappa+\beta,d))}$, and replacing the basis vector $\mathbf{b}_\kappa$ accordingly. This makes the basis BKZ-β reduced at position κ, but might invalidate the condition at other positions. Applying this once to all positions $\kappa = 0, \ldots, d-2$ is called a *tour*. The BKZ algorithms repeats such tours until the basis remains unchanged and is thus BKZ-reduced.

Algorithm 1: The BKZ algorithm.

Data: A lattice basis $\boldsymbol{B}$, blocksize β.
while $\boldsymbol{B}$ *is not BKZ-*β *reduced* **do**
  for $\kappa = 0, \ldots, d-2$ **do** `// A single BKZ-β tour`
    $\mathbf{w} \leftarrow$ a shortest vector in $\mathcal{L}\left(\boldsymbol{B}_{[\kappa:\min(\kappa+\beta,d))}\right)$;
    Lift $\mathbf{w}$ to a full vector $\mathbf{v} \in \mathcal{L}\left(\boldsymbol{B}_{[0:\min(\kappa+\beta,d))}\right)$ s.t. $\pi_\kappa(\mathbf{v}) = \mathbf{w}$;
    Insert $\mathbf{v}$ in $\boldsymbol{B}$ at position κ and use LLL to resolve linear dependencies;

The number of tours is polynomially bounded, and in practice not much improvement is attained after say a few dozen tours. The cost of BKZ is thus mainly dominated by the exponential (in β) cost of finding a shortest vector in a β-dimensional lattice. *Progressive* BKZ reduces this cost in practice, where instead of running many tours of BKZ-β, one runs only a few tours for increasing $\beta' = 2, 3, \ldots, \beta$.

For our experiments we also added a hook to BKZ, using secret key information, to detect if a vector $\mathbf{v}$ is part of the dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ and to abort early if this is the case.

While the Geometric Series Assumption gives a good first order estimate of the basis profile after BKZ-reduction, it is known to be inaccurate in small dimensions or when the dimension is only a small multiple of the blocksize. Additionally it does not account for the slower convergence when running progressive BKZ with only a few tours. To resolve this problem [CN11] introduced a BKZ simulator based on the Gaussian Heuristic, that was later refined in [YD17, BSW18]. These allow for accurate and efficient predictions of the profile shape for *random* lattices, even for progressive BKZ with a limited number of tours.

Behaviour on q-ary lattices. While by now the behaviour of BKZ on random lattices is reasonably understood, this is less the case for q-ary lattices (for certain parameters) such as the NTRU lattice $\mathcal{L}^{\boldsymbol{H},q}$.

Definition 2.7 (q-ary lattices). *A lattice $\mathcal{L}$ of dimension d is said to be q-ary if for some $q > 0$ we have*

$$q\mathbb{Z}^d \subset \mathcal{L} \subset \mathbb{Z}^d.$$

Note that the first n basis vectors of $\mathcal{L}^{\boldsymbol{H},q}$ are orthogonal q-vectors $(q, 0, \ldots, 0)$, $(0, q, 0, \ldots, 0), \ldots$, and so the initial basis profile starts with $\|\mathbf{b}_0^*\| = \cdots =$

$\|\mathbf{b}^*_{n-1}\| = q$. Additionally after projecting away from these q-vectors, the remaining basis vectors are again orthogonal with length 1, and thus we have $\|\mathbf{b}^*_n\| = \cdots = \|\mathbf{b}^*_{d-1}\| = 1$. Note that in the BKZ algorithm the length of $\mathbf{b}_0$ can not increase, and is thus always at most q. Also $\mathbf{b}_1$ can not increase in length if $\mathbf{b}_0$ remains unchanged, and so on. For dual-BKZ or the self-dual LLL the profile lengths can not drop below 1 anywhere by the same reasoning. Still LLL and BKZ guarantee that the profile slope in the middle is not too steep. So after LLL reduction the profile must be flat at the start and end, and have a sloped part in the middle, we call this a Z-shape [AD21]. Because BKZ is not self-dual we do not have any guarantee that the last profile elements do not drop below 1, however we could for example run BKZ only on an appropriate middle context $\mathcal{L}_{[n-m:n+m)}$ to force this behaviour. With this description one would expect the middle part to follow the GSA, leading to an alternative heuristic for q-ary lattices.

Heuristic 2.8 (ZGSA) *Let $\boldsymbol{B}$ be a basis of a $2n$-dimensional q-ary lattice $\mathcal{L}$ with n q-vectors. After BKZ-β reduction the profile has the following shape:*

$$\|\mathbf{b}^*_i\| = \begin{cases} q & \text{if } i \leqslant n-m, \\ \sqrt{q} \cdot \alpha_\beta^{\frac{2n-1-2i}{2}}, & \text{if } n-m < i < n+m-1, \\ 1, & \text{if } i \geqslant n+m-1, \end{cases}$$

where $\alpha_\beta = \mathrm{gh}(\beta)^{2/(\beta-1)}$, and $m = \frac{1}{2} + \frac{\ln(q)}{2\ln(\alpha_\beta)}$.

Again this gives us a good first order estimate. Asymptotically setting $\beta = \mathcal{B} \cdot n$ and $q = n^{\mathcal{Q}}$, we obtain $\ln(\alpha_\beta) = \frac{\ln(n)}{\mathcal{B} \cdot n} + O\left(n^{-1}\right)$, and $m = \frac{1}{2}\mathcal{Q}\mathcal{B} \cdot n + O\left(\frac{n}{\ln(n)}\right)$.

2.4 Estimates

The main question of our work is to better understand how BKZ recovers the dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ from an NTRU lattice $\mathcal{L}^{\boldsymbol{H},q}$. Several works exist that give estimates on the blocksize β for which BKZ successfully recovers the secret key $(\boldsymbol{g}, \boldsymbol{f})$, or more generally a vector from the dense sublattice. We discuss the state-of-the-art estimates, one known as the *2016 Estimate* [ADPS16] with further refinements [DDGR20, PV21], and one by Kirchner and Fouque [KF17].

While the 2016 Estimate already gives a clear explanation *how* BKZ recovers a suitable vector, the Kirchner and Fouque estimate is only based on an impossibility result. To be more precise about what we mean with recovery we define the following two events.

Definition 2.9 (BKZ Events). *For a BKZ run on an NTRU lattice $\mathcal{L}$ with dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ we define two events:*

1. ***Secret Key Recovery (SKR)**: The first time one the secret keys $(\boldsymbol{g}|\boldsymbol{f})$ is inserted.*
2. ***Dense Sublattice Discovery (DSD)**: The first time a dense lattice vector $\mathbf{v} \in \mathcal{L}^{\boldsymbol{GF}}$ strictly longer than the secret key(s) is inserted.*

We further specify SKR_κ and DSD_κ when the insertion takes place at position κ in the basis.

2016 Estimate [ADPS16] for SKR. The 2016 Estimate is aimed at the more general problem of detecting an unusually short vector in a lattice. To obtain an estimate for the NTRU problem, and more specifically the SKR event, we apply it to the unusually short vector $(\boldsymbol{g}|\boldsymbol{f}) \in \mathcal{L}^{\boldsymbol{H},q}$.

Claim 2.10 (SKR – 2016 Estimate) *Let $\mathcal{L}$ be a lattice of dimension d and let $\mathbf{v} \in \mathcal{L}$ be a unusually short vector $\|\mathbf{v}\| \ll \mathrm{gh}(\mathcal{L})$. Then under the Geometric Series Assumption BKZ recovers $\mathbf{v}$ if*

$$\sqrt{\beta/d} \cdot \|\mathbf{v}\| < \sqrt{\alpha_\beta}^{2\beta-d-1} \cdot \mathrm{vol}(\mathcal{L})^{1/d},$$

where $\alpha_\beta = \mathrm{gh}(\beta)^{2/(\beta-1)}$.

The left hand side of the inequality is an estimate for $\|\pi_{d-\beta}(\mathbf{v})\|$, while the right hand size is the expected norm of $\mathbf{b}^*_{d-\beta}$ under the GSA. When the inequality is satisfied we expect that the shortest vector in $\mathcal{L}_{[d-\beta:d)}$ is in fact (a projection of) the unusually short vector, and thus it is inserted by BKZ at position $d-\beta$.

For q-ary lattices we can easily change the estimate to make use of the ZGSA instead, although for successful blocksizes $\mathbf{b}^*_{d-\beta}$ will not lie on the flat tail-part, and thus this will not change anything. Additionally for q-ary lattices it can be beneficial to apply the estimate not to the full lattice but on some projected sublattice $\mathcal{L}_{[i:d)}$ for $i \leqslant n$; the left hand side of the equation is expected to remain unchanged, while the right hand side might decrease as $\mathrm{vol}(\mathcal{L})$ loses a factor q^i. Note that we do not necessarily have to explicitly let BKZ act on this projected sublattice, as BKZ already does this naturally.

Asymptotics. Consider the NTRU lattice $\mathcal{L}^{\boldsymbol{H},q}$ and suppose that $q = \Theta(n^{\mathcal{Q}})$, $\|\mathbf{v}\| = \|(\boldsymbol{g},\boldsymbol{f})\| = \Theta(n^{\mathcal{S}})$ and $\beta = (\mathcal{B} + o(1))n$. Applying the 2016 Estimate the right hand side of the inequality is minimised when only keeping $k = \min\left((\sqrt{2\mathcal{B}\mathcal{Q}} - 1)n, n\right)$ of the q-vectors, so by applying the estimate to the projected sublattice $\mathcal{L}^{\boldsymbol{H},q}_{[n-k:2n)}$. For $\mathcal{S} \geqslant 1$ we have $k = n$, and solving the equation gives $\mathcal{B} = \frac{2}{\mathcal{Q}+2-2\mathcal{S}}$. For $\mathcal{S} < 1$ we have $k = (\sqrt{2\mathcal{B}\mathcal{Q}} - 1)n$, and solving gives $\mathcal{B} = \frac{2\mathcal{Q}}{(\mathcal{Q}+1-\mathcal{S})^2}$. Note in particular that in terms of q we require a blocksize of $\beta = \tilde{\Theta}\left(n/\ln(q)\right)$.

Refinements. The 2016 Estimate gives a clear explanation on how and where the secret vector is recovered. This also allows to further refine the estimate and give concrete predictions. For example by using a BKZ-simulator instead of the GSA, and by accounting for the probability that after the projection $\|\pi_{d-\beta}(\mathbf{v})\|$ has been found, it is successfully lifted to the full vector $\mathbf{v}$. Also instead of working with the expected length of the projection, we can directly model the probability distribution under the assumption that $\mathbf{v}$ is distributed as a Gaussian vector. Such refinements were applied in [DDGR20, PV21], and the resulting concrete

predictions match with experiments to recover an unusually short vector. In this work, we use the (Z)GSA for the basis shape, but adjusting the slope to account for the speed of convergence using experimentally determined values. However, we do use the advanced probabilistic model for the detection and lifting of the short vector.

For NTRU there is not just a single unusually short vector, but there are $n = d/2$ of them, which makes it more likely that at least one of them is recovered. Because the refined concrete estimator already works with a probability distribution, we can easily take multiple vectors into account. The resulting predictions for the SKR event match the experiments reasonably well for smallish q as can be seen in Figure 1. For large q, the so-called *overstretched* regime, the estimate is however too pessimistic.

Kirchner–Fouque Estimate [KF17] for DSD. In 2016 Albrecht, Bai and Ducas [ABD16] showed that for very large values of q one can mount an algebraic *subfield attack* on the cyclotomic NTRU problem with sub-exponential or even polynomial complexity. This allowed them to break several homomorphic encryption schemes that relied on NTRU in the overstretched regime.

However soon after, Kirchner–Fouque [KF17] showed that this elaborate algebraic attack was unnecessary: (dual-)BKZ already behaves much better in this regime than the 2016 Estimate predicts, leading to the same asymptotic improvements. The key idea behind their analysis is that in the overstretched regime the NTRU lattice $\mathcal{L}^{\boldsymbol{H},q}$ contains an exceptionally dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ of low volume. This gives a constraint on the basis profile via the following lemma by Pataki and Tural.

Lemma 2.11 (Pataki and Tural [PT08]). *Let $\mathcal{L}$ be a d-dimensional lattice with basis $\mathbf{b}_0, \ldots, \mathbf{b}_{d-1}$. For any k-dimensional sublattice $\mathcal{L}' \subset \mathcal{L}$ we have*

$$\operatorname{vol}(\mathcal{L}') \geqslant \min_J \prod_{j \in J} \left\|\mathbf{b}_j^*\right\|,$$

where J ranges over the k-size subsets of $\{0, \ldots, d-1\}$.

Applying Lemma 2.11 to the n-dimensional sublattice $\mathcal{L}^{\boldsymbol{GF}} \subset \mathcal{L}^{\boldsymbol{H},q}$, and assuming a non-increasing profile, we obtain an upper bound on the volume of $\mathcal{L}^{\boldsymbol{H},q}_{[n:2n)}$. Assuming the ZGSA the latter volume increases when running BKZ-β for increasing blocksizes, eventually contradicting the upper bound. This allows us to detect if a q-ary lattice is in fact an NTRU lattice, but additionally Kirchner–Fouque argue that BKZ must *somehow* have detected the dense sublattice after this point. Based on this impossibility argument they introduced the following estimate.

Claim 2.12 (DSD – Kirchner–Fouque Estimate) *Let $\mathcal{L}^{\boldsymbol{H},q}$ be an NTRU lattice of dimension $2n$, with dense sublattice $\mathcal{L}^{\boldsymbol{GF}} \subset \mathcal{L}^{\boldsymbol{H},q}$. Under the Z-shape Geometric Series Assumption BKZ-β triggers the DSD event if*

$$\operatorname{vol}(\mathcal{L}^{\boldsymbol{GF}}) < q^{\frac{m-1}{2}} \cdot \alpha_\beta^{-\frac{1}{2}(m-1)^2},$$

where $\alpha_\beta = \mathrm{gh}(\beta)^{2/(\beta-1)}$, *and* $m = \frac{1}{2} + \frac{\ln(q)}{2\ln(\alpha_\beta)}$.

To apply this estimate we can bound $\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}})$ using the Hadamard inequality by $\|(\boldsymbol{g}|\boldsymbol{f})\|^n$. As a first approximation this is reasonably tight because the secret basis $\boldsymbol{B}^{\boldsymbol{GF}}$ is close to orthogonal.

Asymptotics. Consider the NTRU lattice $\mathcal{L}^{\boldsymbol{H},q}$ and suppose that $q = \Theta(n^{\mathcal{Q}})$, $\|(\boldsymbol{g},\boldsymbol{f})\| = \Theta(n^{\mathcal{S}})$ and $\beta = (\mathcal{B}+o(1))n$. We apply the Kirchner–Fouque Estimate using that $m \approx \frac{\mathcal{B}\mathcal{Q}}{2}n$ and $\alpha_\beta \approx (\mathcal{B}n)^{1/(\mathcal{B}n)}$.The left hand side of the inequality is bounded by $n^{n\mathcal{S}+o(n)}$ and the right hand side equals $n^{\frac{\mathcal{B}\mathcal{Q}^2}{8}n+o(n)}$; solving gives $\mathcal{B} \geqslant \frac{8\mathcal{S}}{\mathcal{Q}^2}$. Note that in terms of q we require a blocksize of $\beta = \tilde{\Theta}\left(n/\ln^2(q)\right)$, improving upon the 2016 Estimate by a factor $\ln(q)$. So for large enough q the Kirchner–Fouque Estimate predicts a lower successful blocksize than the 2016 Estimate. We call the value of q for which BKZ starts to behave better than predicted by the 2016 Estimate the *fatigue point*. For the common situation that $\mathcal{S} = \frac{1}{2}$, e.g. when each secret coefficient has standard deviation $\sigma = \Theta(1)$, the Kirchner–Fouque Estimate predicts that the fatigue point lies at some $q \leqslant n^{2.783+o(1)}$.

3 A new Estimate

3.1 Preliminary Experiments

Both the 2016 Estimate and the Kirchner–Fouque Estimate analyse an event that leads to successful recovery of a vector of the dense NTRU sublattice. This only gives an upper bound on the hardness; a different event leading to the recovery might happen at a lower blocksize. Additionally the Kirchner–Fouque Estimate is only based on an impossibility result and gives no explanation as to how BKZ actually recovers a vector from the dense sublattice. In order to derive a tight estimate we first run experiments to track down at which point a dense sublattice vector is actually found during the BKZ tours, i.e. when the DSD_κ event is triggered and at what position. Then we model this event in order to hopefully derive a tight estimate.

We run progressive BKZ on NTRU lattices $\mathcal{L}^{\boldsymbol{H},q}$ for fixed parameters $n = 127, \sigma^2 = \frac{2}{3}$, and several moduli q. For each BKZ insertion at position κ we check if the inserted vector belongs to the dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$, and thereby if the SKR_κ or DSD_κ event takes place, after which we stop.

The results are shown in Figure 1. We take a closer look at the observed SKR_κ and DSD_κ events and where they are triggered. We can group our observations in three typical circumstances.

– **SKR-2016**. The SKR_κ event is mostly triggered for small values of q, and this mostly happens at the position $\kappa = 2n-\beta$, so in the last block $[2n-\beta : 2n)$, or slightly earlier. This coincides exactly with the $\mathrm{SKR}_{2n-\beta}$ event as predicted by the 2016 Estimate [ADPS16, AGVW17].

– **DSD-PT**. The DSD_κ event is mostly triggered at positions $\kappa = n + k - \beta$ for $0 < k \ll n$. The inserted dense vector $\mathbf{v}$ is often significantly longer than the secret key but still shorter than the q-vectors. On closer inspection the projected length $\|\pi_{n+k-\beta}(\mathbf{v})\|$ is close to the expected length $\sqrt{\frac{\beta}{n+k}}\,\|\mathbf{v}\|$ for all instances, more specifically the length of $\mathbf{v}$ is well balanced over the Gram-Schmidt directions $\mathbf{b}_0^*, \ldots, \mathbf{b}_{n+k-1}^*$. We name these events after the Pataki–Tural Lemma (DSD-PT).
– **DSD-LL**. For a few instances the DSD_κ event is triggered at large positions κ, up to $2n - \beta$. The inserted dense vector $\mathbf{v}$ is again significantly longer than the secret key, but it has an unexpectedly short projection $\pi_\kappa(\mathbf{v})$ on the BKZ block $[\kappa : \kappa + \beta)$. We call these events lucky-lifts (DSD-LL).

The DSD-LL event could potentially be explained by the relatively large amount of shortish vectors in the close to orthogonal dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ compared to what one would expect based on the Gaussian Heuristic. These many vectors might compensate for the low probability event that: (1) such a long vector has such a short projection, and (2) the projected vector is correctly lifted by Babai's nearest plane algorithm (thus a *lucky lift*). The DSD-LL event remains rare for all parameters we used in our experiments, and the successful blocksizes do not seem to deviate from the DSD-PT events. Although we think this circumstance deserves further analysis we therefore base our estimate on the more common DSD-PT event.

For the DSD-PT event the projected length $\|\pi_{n+k-\beta}(\mathbf{v})\|$ is close to $\sqrt{\frac{\beta}{n+k}}\,\|\mathbf{v}\|$, and thus the inserted dense vector $\mathbf{v}$ must in fact be (close to) a shortest vector of the intersected sublattice $\mathcal{L}^{\boldsymbol{H},q}_{[0:n+k)} \cap \mathcal{L}^{\boldsymbol{GF}}$. If not, the shortest vector would typically have an even smaller projection and would thus be inserted instead. For ease of analysis we therefore assume that $\mathbf{v}$ is a shortest vector of $\mathcal{L}^{\boldsymbol{H},q}_{[0:n+k)} \cap \mathcal{L}^{\boldsymbol{GF}}$. In short our new estimate can be described as follows.

Claim 3.1 (DSD-PT estimate) *A tour of BKZ-β triggers the DSD event if*

$$\pi_{n+k-\beta}(\mathbf{v}) < \left\|\mathbf{b}^*_{n+k-\beta}\right\|,$$

where $\mathbf{v}$ is a shortest vector of $\mathcal{L}^{\boldsymbol{H},q}_{[0:)} \cap \mathcal{L}^{\boldsymbol{GF}}$ for some $0 < k \leqslant n$.

3.2 Asymptotic analysis

We denote the intersected sublattice by $\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:r)} := \mathcal{L}^{\boldsymbol{H},q}_{[0:r)} \cap \mathcal{L}^{\boldsymbol{GF}}$. To directly apply Claim 3.1 we are interested in the length of $\mathbf{v}$, and thus the value of $\lambda_1\big(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\big)$. We break down the analysis into several steps. In order to obtain a bound on the first minimum we first compute a bound on the volume of the intersection $\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}$ in terms of the basis profile and the volume of $\mathcal{L}^{\boldsymbol{GF}}$. Together with the GSA and a simple bound for $\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}})$ we can then apply Minkowski's bound on the first minimum. By optimising $\kappa = n + k - \beta$ we obtain our new asymptotic estimate.

Intersection. To understand the behaviour of the volume of the intersected lattice we first need a small technical Lemma.

Lemma 3.2 ([DDGR20]). *Given a lattice $\mathcal{L}$ with volume $\mathrm{vol}(\mathcal{L})$, and a primitive vector $\mathbf{v}$ with respect to $\mathcal{L}^*$. Let $\mathbf{v}^\perp$ denote the subspace orthogonal to $\mathbf{v}$. Then $\mathcal{L} \cap \mathbf{v}^\perp$ is a lattice with volume $\mathrm{vol}(\mathcal{L} \cap \mathbf{v}^\perp) = \|\mathbf{v}\| \cdot \mathrm{vol}(\mathcal{L})$.*

The following Lemma generalises the Pataki–Tural Lemma on which the estimate of Kirchner–Fouque is based. More specifically the Pataki–Tural Lemma only considers the case where the intersection is always trivial ($s = 0$).

Lemma 3.3 (Generalisation of [PT08]). *Let $\mathcal{L}$ be a d-dimensional lattice with basis $\mathbf{b}_0, \ldots, \mathbf{b}_{d-1}$, and consider the sublattice $\mathcal{L}_{[0:s)}$. For any n-dimensional sublattice $\mathcal{L}' \subset \mathcal{L}$ we have*

$$\mathrm{vol}(\mathcal{L}_{[0:s)} \cap \mathcal{L}') \leqslant \mathrm{vol}(\mathcal{L}') \cdot \left(\min_J \prod_{j \in J} \left\| \mathbf{b}_j^* \right\| \right)^{-1},$$

where $k := \dim(\mathcal{L}_{[0:s)} \cap \mathcal{L}')$ and J ranges over the $(n-k)$-size subsets of $\{s, \ldots, d-1\}$.

Proof. We write $\mathcal{L}'_{\cap[0:r)} := \mathcal{L}_{[0:r)} \cap \mathcal{L}'$. For $j = k, \ldots, n$ we define $s_j \in \{s, \ldots, d\}$ as the maximal index such that $\dim\left(\mathcal{L}'_{\cap[0:s_j)}\right) = j$, i.e. we obtain the following strict chain of sublattices:

$$\mathcal{L}'_{\cap[0:s)} = \mathcal{L}'_{\cap[0:s_k)} \subsetneq \mathcal{L}'_{\cap[0:s_{k+1})} \subsetneq \cdots \subsetneq \mathcal{L}'_{\cap[0:s_n)} = \mathcal{L}'.$$

Fix $j \in \{k, \ldots, n-1\}$. Because the basis vectors $\mathbf{b}_0, \ldots, \mathbf{b}_{d-1}$ are linearly independent we have that $\mathcal{L}'_{\cap[0:s_{(j+1)})} = \mathcal{L}'_{\cap[0:s_j+1)}$. This allows us to focus on the volume decrease from index $s_j + 1$ to s_j, for which we know that

$$\mathcal{L}'_{\cap[0:s_j)} = \mathcal{L}'_{\cap[0:s_j+1)} \cap (\mathbf{b}_{s_j}^*)^\perp,$$

where $(\mathbf{b}_{s_j}^*)^\perp$ denotes the subspace orthogonal to $\mathbf{b}_{s_j}^*$. The corresponding dual basis of $\mathbf{b}_0, \ldots, \mathbf{b}_{s_j}$ contains a dual vector $\mathbf{d} \in \mathcal{L}^*_{[0:s_j+1)}$ of length $\left\|\mathbf{b}_{s_j}^*\right\|^{-1}$ with $\mathrm{span}(\mathbf{d}) = \mathrm{span}(\mathbf{b}_{s_j}^*)$. Let π be the orthogonal projection onto $\mathrm{span}\left(\mathcal{L}'_{\cap[0:s_j+1)}\right)$, then $\pi(\mathbf{d}) \in \left(\mathcal{L}'_{\cap[0:s_j+1)}\right)^*$. Let $m \in \mathbb{Z}_{\geqslant 1}$ be such that $\pi(\mathbf{d})/m$ is primitive w.r.t. $\left(\mathcal{L}'_{\cap[0:s_j+1)}\right)^*$, then by Lemma 3.2 we obtain:

$$\mathrm{vol}\left(\mathcal{L}'_{\cap[0:s_j)}\right) = \mathrm{vol}\left(\mathcal{L}'_{\cap[0:s_j+1)}\right) \cdot \|\pi(\mathbf{d})/m\| \leqslant \mathrm{vol}\left(\mathcal{L}'_{\cap[0:s_j+1)}\right) \cdot \left\|\mathbf{b}_{s_j}^*\right\|^{-1}.$$

We conclude the proof by chaining the above inequality for $j = k, \ldots, n-1$. □

Before recovering a dense lattice vector we heuristically assume that there is no special relation between the current lattice basis and the dense sublattice. More specific we can consider that the span of $\mathbf{b}_0, \ldots, \mathbf{b}_{n-1}$ and that of $\mathcal{L}^{\boldsymbol{GF}}$ behave

like random n-dimensional subspaces, and thus they have a trivial intersection with high probability in the $2n$-dimensional space. As a direct result we have that $\dim\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right) = k$ for $k = 0, \ldots, n$. Applying this to Lemma 3.3 we obtain the following corollary.

Corollary 3.4. *Let $\mathcal{L}^{\boldsymbol{H},q}$ be an NTRU lattice with dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$ of dimension n, if $\dim\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right) = k$ for some $k \geqslant 0$, then*

$$\mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right) \leqslant \mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}\right) \cdot \left(\prod_{j=n+k}^{d-1} \left\|\mathbf{b}^*_j\right\|\right)^{-1}.$$

Note that Corollary 3.4 already shows that the new estimate can not be worse than the Kirchner–Fouque Estimate. Namely if the Kirchner–Fouque Estimate is triggered, then for intersection dimension $k = 1$ the right hand side is smaller than $\|\mathbf{b}^*_n\|$. Assuming a non-decreasing profile we then have $\lambda_1\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+1)}\right) = \mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+1)}\right) \leqslant \|\mathbf{b}^*_n\| \leqslant \left\|\mathbf{b}^*_{n+1-\beta}\right\|$, which implies that BKZ-β would find a dense sublattice vector in this block (or earlier).

Volume dense sublattice. To use Corollary 3.4 we also need to bound the volume of the dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$. Because the secret basis is close to orthogonal the Hadamard Inequality $\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}) \leqslant \|(\boldsymbol{g}|\boldsymbol{f})\|^n$ is sufficient as a first order approximation.

Conclusion. To obtain a heuristic asymptotic estimate we will assume that before finding a dense lattice vector the basis follows the ZGSA shape.

Claim 3.5 *The BKZ algorithm with blocksize $\beta = \mathcal{B}n$ applied to an NTRU instance with parameters $q = \Theta(n^{\mathcal{Q}}), \|(\boldsymbol{g}|\boldsymbol{f})\| = O(n^{\mathcal{S}})$ triggers the DSD event if*

$$\mathcal{B} = \frac{8\mathcal{S}}{\mathcal{Q}^2 + 1} + o(1).$$

Justification. By the Hadamard Inequality we have $\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}})) \leqslant \mathcal{S}n\ln(n) + O(n)$. Let $k := \mathcal{K}n > 0$. Heuristically we expect that $\dim(\mathcal{L}^{\boldsymbol{H},q}_{[0:n+k)} \cap \mathcal{L}^{\boldsymbol{GF}}) = k$, and thus by Corollary 3.4 and by assuming the ZGSA we obtain a bound on the volume of the intersected sublattice:

$$\begin{aligned}\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{H},q}_{[0:n+k)} \cap \mathcal{L}^{\boldsymbol{GF}})) &\leqslant \mathcal{S}n\ln(n) - \frac{1}{2}\sum_{i=n+k}^{n+m-1}\left(\mathcal{Q} + \frac{2n-1-2i}{\mathcal{B}n}\right)\ln(n) + O(n)\\ &= \mathcal{S}n\ln(n) - \frac{(\mathcal{B}\mathcal{Q}-2\mathcal{K})^2}{8\mathcal{B}}n\ln(n) + O(n)\end{aligned}$$

By Minkowski's bound we bound the first minimum using the above volume

$$\begin{aligned}\ln(\lambda_1(\mathcal{L}^{\boldsymbol{H},q}_{[0:n+k)} \cap \mathcal{L}^{\boldsymbol{GF}})) &\leqslant \frac{1}{2}\ln(\mathcal{K}n) + \frac{\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{H},q}_{[0:n+k)} \cap \mathcal{L}^{\boldsymbol{GF}}))}{\mathcal{K}n} + O(1)\\ &\leqslant \left(-\frac{(\mathcal{B}\mathcal{Q}-2\mathcal{K})^2}{8\mathcal{B}\mathcal{K}} + \frac{\mathcal{S}}{\mathcal{K}} + \frac{1}{2}\right)\ln(n) + O(1).\end{aligned}$$

After projecting onto the block $[n+k-\beta : n+k)$ the above short vector does not increase in length.[7] BKZ detects the projected dense lattice vector in this block if the length is less than $\|\mathbf{b}^*_{n+k-\beta}\| = \left(\frac{1}{2}\mathcal{Q} + \frac{\mathcal{B}-\mathcal{K}}{\mathcal{B}}\right)\ln(n) + O(1)$. Solving for $\mathcal{B}$ shows that this is the case when

$$\mathcal{B} \geqslant \frac{2\sqrt{(2\mathcal{S}-\mathcal{K})^2 + \mathcal{K}^2\mathcal{Q}^2} + 2(2\mathcal{S}-\mathcal{K})}{\mathcal{Q}^2}.$$

When $\mathcal{K} = \frac{4\mathcal{S}}{\mathcal{Q}^2+1}$ the right hand side is minimised and we obtain that BKZ detects the projected dense lattice vector when $\mathcal{B} \geqslant \frac{8\mathcal{S}}{\mathcal{Q}^2+1}$, which concludes the claim. This routine computation can be verified symbolically via our sage notebook `claim3_5.ipynb`. △

Our new estimate gives an asymptotic improvement over the Kirchner–Fouque Estimate ($\frac{8\mathcal{S}}{\mathcal{Q}^2}$). Asymptotically the optimal position is at $\kappa = n+k-\beta \approx n - \frac{1}{2}\beta$. Interestingly, if we do not optimize k and only consider $k = O(1)$ we obtain the same asymptotic estimate as Kirchner–Fouque, which again emphasizes that we generalised their analysis.

For the fatigue point we compare the relative blocksize of $\frac{8\mathcal{S}}{\mathcal{Q}^2+1}$ to that of the 2016 Estimate given by $\frac{2\mathcal{Q}}{(\mathcal{Q}+1-\mathcal{S})^2}$ for $\mathcal{S} < 1$ and by $\frac{2}{\mathcal{Q}+2-2\mathcal{S}}$ for $\mathcal{S} \geqslant 1$. For ternary secrets ($\mathcal{S} = \frac{1}{2}$) this narrows down the fatigue point from $q \leqslant n^{2.783+o(1)}$ to $q = n^{2.484+o(1)}$ compared to the Kirchner–Fouque Estimate. This is still far above the (sub)linear parameters used for NTRU encryption schemes, and thus asymptotically we can close the pending question if these parameters fall in the weaker overstretched regime or not. In practice however we do observe fatigue points that are significantly lower than the naive value of $q = n^{2.484}$, which motivates a concrete analysis with concrete predictions.

4 Concrete Analysis

In this section we consider a concrete analysis of our new DSD-PT estimate, based on simple heuristics, to better predict the behaviour in practice, and to show that our analysis matches experiments and is thus likely to be tight. The first order asymptotics shown in Section 3.2 will remain unchanged, but the differences are significant for practical parameters. Again we split the analysis

[7] One may also be concerned that the short vector would collapse to $\vec{0}$ after projection onto the block $[n+k-\beta : n+k)$, but this becomes increasingly unlikely as the dimension β of the block grows.

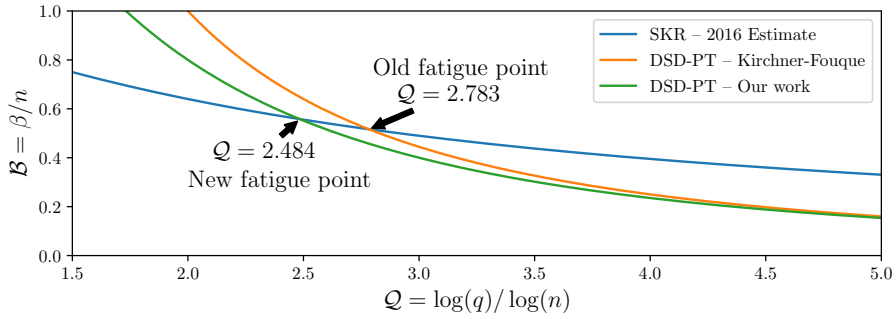

Fig. 2. Comparison of asymptotic estimates and new fatigue point for $n \to \infty$ when the secret key coefficients have standard deviation $\sigma = \Theta(1)$.

into several steps, but now derive heuristic expectations instead of loose upper bounds.

We assume that lattice vectors we encounter follow the Gaussian heuristic, and thus in particular that vectors are spherically distributed after normalisation. When projecting such vectors to a lower dimension they become shorter. The following Lemma shows how much shorter we expect them to become.

Lemma 4.1. *Let $\mathbf{x} \in \mathcal{S}^{d-1}$ follow a spherical distribution, and let $\pi_V : \mathbb{R}^d \to V$ be a projection to some k-dimensional subspace $V \subset \mathbb{R}^d$, then*

$$\mathbb{E}[\ln(\|\pi_V(\mathbf{x})\|)] = \frac{1}{2}(\psi(k/2) - \psi(d/2)).$$

Proof. Let $X_0, \ldots, X_{d-1}$ be standard normal random variables, then the vector $\mathbf{x} = (x_0, \ldots, x_{d-1})$, with $x_j = X_j/\sqrt{\sum_{i=0}^{d-1} X_i^2}$, is spherically distributed. Without loss of generality we can assume that π_V projects onto the first k-coordinates. Then we conclude by Lemma 2.1 that

$$\begin{aligned}
\mathbb{E}[\ln(\|\pi_V(\mathbf{x})\|)] &= \frac{1}{2}\mathbb{E}\left[\ln\left(\frac{\sum_{i=0}^{k-1} X_i^2}{\sum_{i=0}^{d-1} X_i^2}\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\ln\left(\sum_{i=0}^{k-1} X_i^2\right)\right] - \frac{1}{2}\mathbb{E}\left[\ln\left(\sum_{i=0}^{d-1} X_i^2\right)\right] \\
&= \frac{1}{2}(\psi(k/2) - \psi(d/2)).
\end{aligned}$$

□

4.1 Intersection

We start by giving a concrete average-case estimate for the intersection volumes. Assuming that projections behave as random we obtain the following concrete estimate.

Claim 4.2 *Let $\mathcal{L}$ be a $2n$-dimensional NTRU lattice with dense sublattice $\mathcal{L}^{\boldsymbol{GF}}$, before the DSD event is triggered we have for $k = 1, \ldots, n$ that* $\dim\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right) = k$, *and*

$$\mathbb{E}[\ln \operatorname{vol}(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)})] = \ln \operatorname{vol}(\mathcal{L}^{\boldsymbol{GF}}) - \left(\sum_{j=n+k}^{2n-1} \ln \left\|\mathbf{b}_j^*\right\|\right) + \sum_{l=k+1}^{n} \psi\left(\frac{l}{2}\right) - \psi\left(\frac{n+l}{2}\right) + \frac{\zeta'(l)}{\zeta(l)},$$

where $\zeta(l) := \sum_{m=1}^{\infty} \frac{1}{m^l}$ is the Riemann zeta function and $\zeta'(l) := \sum_{m=1}^{\infty} \frac{\ln(m)}{m^l}$ its derivative.

Justification. We follow the proof of Lemma 3.3. It is tight except for the length decrease from $\|\mathbf{d}\|$ to the projected and primitive vector $\|\pi(\mathbf{d})\|/m$. Note that when obtaining $\ln \operatorname{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+l-1)}\right)$ from $\ln \operatorname{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+l)}\right)$ for some $l = k+1, \ldots, n$, the dual vector $\mathbf{d}$ lives in a $(n+l)$-dimensional space and is projected to an l-dimensional space. Heuristically we assume that the normalisation of $\mathbf{d}$ is spherically distributed (or that π projects to a random l-dimensional subspace). By Lemma 4.1 the log-expected decrease in length from this projection then equals

$$\mathbb{E}\left[\ln(\pi(\|\mathbf{d}\|)) - \ln(\|\mathbf{d}\|)\right] = \psi\left(\frac{l}{2}\right) - \psi\left(\frac{n+l}{2}\right).$$

To conclude we also have to include the primitivity of $\pi(\mathbf{d})$ and thus the log-expectation of $m \geqslant 1$ such that $\pi(\mathbf{d})/m$ is primitive. For any basis $\mathbf{d}_0, \ldots, \mathbf{d}_{l-1}$ and $\pi(\mathbf{d}) = \sum_{i=0}^{l-1} x_i \mathbf{d}_i$, we have $m = \gcd(x_0, \ldots, x_{l-1})$. Heuristically we assume that the absolute coefficients $|x_0|, \ldots, |x_{l-1}|$ are random integers in the interval $\{1, \ldots, B\}$ and we let $B \to \infty$. For $l \geqslant 2$ we have (see e.g. [DE+04])

$$\mathbb{P}_{\mathbf{x} \in \{1,\ldots,B\}^l}\left[\gcd(x_0, \ldots, x_{l-1}) = m\right] = \frac{1}{\zeta(l)} \cdot \frac{1}{m^l} + O(\ln(B)/(Bm^{l-1})),$$

where the Riemann zeta function $\zeta(l) = \sum_{m=1}^{\infty} \frac{1}{m^l}$ is just the normalisation factor. From this we conclude that

$$\begin{aligned}\lim_{B\to\infty} \mathbb{E}_{\mathbf{x} \in \{1,\ldots,B\}^l}[\ln \gcd(x_0, \ldots, x_{l-1})] &= \lim_{B\to\infty} \frac{1}{\zeta(l)} \sum_{m=1}^{B} \left[\frac{\ln(m)}{m^l} + O\left(\frac{\ln(m)\ln(B)}{Bm^{l-1}}\right)\right] \\ &= -\frac{\zeta'(l)}{\zeta(l)}\end{aligned}$$

for $l \geqslant k+1 \geqslant 2$. △

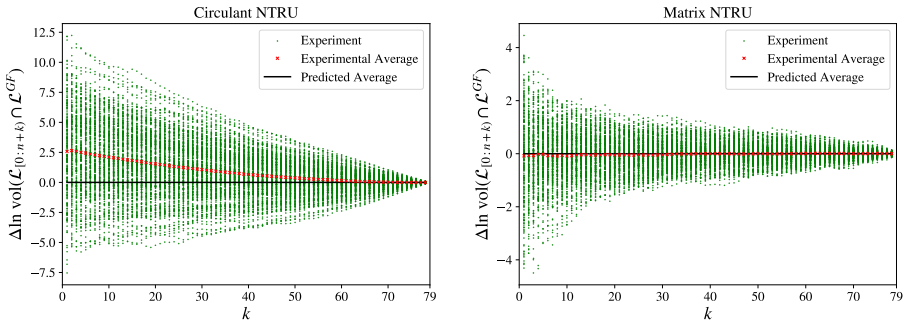

Fig. 3. Experimental values of $\ln \mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right)$ versus Claim 4.2 for circulant and matrix NTRU respectively. For each variant we used 256 LLL reduced NTRU lattices with parameters $q = 257, n = 79, \sigma^2 = \frac{2}{3}$ and computed the intersection for each k.

Validation. To validate Claim 4.2 we computed the actual intersection volumes $\mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right)$ for LLL reduced NTRU instances. We observed here, and also in further experiments, that the dimension assumption $\dim\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right) = k$ holds before we get close to triggering the DSD event. Figure 3 shows that our prediction perfectly matches the experiments for matrix NTRU. For circulant NTRU we see both that the expectation is slightly off and that the variance is much higher. The higher variance can be explained from the fact that the projections are very much dependent due to the circulant structure; in fact a closer inspection shows that for k close to n the differences with the prediction are highly correlated. We were not able to explain the error in the predicted expectation, but it seems to be caused by the circulant structure in combination with the Z-shape: the error decreased and eventually disappeared for large values of q and σ, for which the Z-shape disappeared (and before the DSD event was triggered). A maximal log-error of 2.5 is reached at $k = 1$. Note that a log-error of ϵ on $\mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right)$ translate into a factor of $e^{\epsilon/k}$ on the predicted length for the shortest vector. Except for very small k, this error appears benign.

4.2 Dense sublattice

In this section we give a concrete estimate for the expected volume of the dense NTRU sublatice $\mathcal{L}^{\boldsymbol{GF}}$. Directly from the construction we obtain a basis $[\boldsymbol{G}|\boldsymbol{F}]$ of $\mathcal{L}^{\boldsymbol{GF}}$, with $\boldsymbol{F}$ invertible. We consider two cases, that of regular NTRU where $\boldsymbol{F}$ and $\boldsymbol{G}$ are circulant matrices, and that of matrix NTRU, where all entries are independently sampled. For both constructions the entries are sampled from independent discrete Gaussians over $\mathbb{Z}$, with some standard deviation $\sigma > 0$. As the only heuristic we assume that the individual entries in fact follow a *continuous* Gaussian instead of the discrete one.

Matrix NTRU. We start with matrix NTRU, where we heuristically assume that all $2n \times n$ coefficients of the basis $[\boldsymbol{G}|\boldsymbol{F}]$ are sampled according to independent continuous Gaussians with standard deviation σ. Under this heuristic we can derive an exact expression for the expected log-volume of the dense sublattice.

Lemma 4.3. *Let $[\boldsymbol{G}|\boldsymbol{F}]$ be a basis of the lattice $\mathcal{L}^{\boldsymbol{GF}}$ where all sampled entries are i.i.d. continous Gaussians with standard deviation $\sigma > 0$, then*

$$\mathbb{E}[\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}))] = \frac{1}{2}n\left(\ln(2\sigma^2) + \psi(n)\right) + \sum_{i=0}^{n-1}\left[\psi\left(\frac{2n-i}{2}\right) - \psi(n)\right].$$

Proof. By Lemma 2.1 the log-expectation of the norm of each basis element equals $(\ln(2\sigma^2)+\psi(n))/2$. Note that the i-th Gram-Schmidt vector $\mathbf{b}_i^*$ is obtained after projecting the i-th basis vector orthogonally away from an i-dimensional subspace, and thus onto a $2n-i$ dimensional subspace. However after normalisation the basis vectors follow a spherical distribution and thus by Lemma 4.1 we have

$$\mathbb{E}[\ln \|\mathbf{b}_i^*\|] = (\ln(2\sigma^2) + \psi(n))/2 + \psi\left(\frac{2n-i}{2}\right) - \psi(n).$$

We conclude by noting that $\mathbb{E}[\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}))] = \sum_{i=0}^{n-1}\mathbb{E}[\ln \|\mathbf{b}_i^*\|]$. □

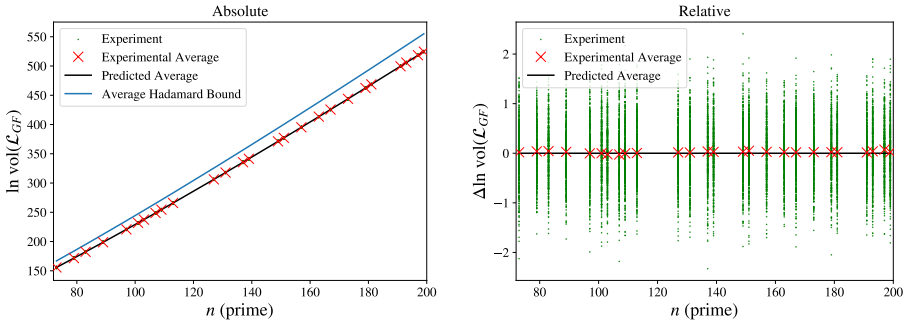

Fig. 4. Experimental values of $\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}))$ versus Lemma 4.3 for matrix NTRU with discrete Gaussians and variance $\sigma^2 = \frac{2}{3}$. For each parameter n we generated 512 instances.

Circulant NTRU. For circulant NTRU both $\boldsymbol{G}$ and $\boldsymbol{F}$ in the basis $[\boldsymbol{G}|\boldsymbol{F}]$ are circulant matrices. Again we replace discrete with continuous Gaussians. The eigenvalues and eigenvectors of a circulant matrix are well known and we use this to obtain an exact expression for the expected volume of the dense sublattice.

Lemma 4.4. *Let* $[\boldsymbol{G}|\boldsymbol{F}]$ *be a basis of the lattice* $\mathcal{L}^{\boldsymbol{GF}}$ *where* $\boldsymbol{G}, \boldsymbol{F}$ *are circulant and all sampled entries are i.i.d. continous Gaussians with with standard deviation* $\sigma > 0$*, then*

$$\mathbb{E}[\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}))] = \frac{1}{2}n\left(\ln(2n\sigma^2) + \psi(1)\right) + \frac{1}{2}(n-1)(1-\ln(2)).$$

Proof. For $n \times n$ circulant matrices $\boldsymbol{G}, \boldsymbol{F}$ the eigenvectors are identical and given by $\mathbf{v}_j := (1, \omega^j, \omega^{2j}, \omega^{(n-1)j})$ for $j = 0, \ldots, n-1$, where $\omega := e^{2\pi i/n} \in \mathbb{C}$ is a primitive n-th root of unity. Suppose that the circulant matrix $\boldsymbol{G}$ is generated by the vector $\mathbf{c} = (c_0, \ldots, c_{n-1})$, then the corresponding eigenvalues are given by the DFT coefficients of $\mathbf{c}$, namely $\lambda_j := c_0 + c_{n-1}\omega^j + \ldots + c_1\omega^{(n-1)j}$. We have that $\lambda_0 = \sum_{j=0}^{n-1} c_j$, and thus λ_0 follows a Gaussian distribution with variance $n\sigma^2$, and in particular $\lambda_0^2 \sim \chi^2_{1,n\sigma^2}$. Additionally for $j = 1, \ldots, n-1$ we can write $\lambda_j = X + i \cdot Y \in \mathbb{C}$ where $X, Y \in \mathbb{R}$ are both linear combinations of the c_i's and thus (X, Y) follows a jointly Gaussion distribution. A simple computation shows that X and Y both have variance $n\sigma^2/2$ and that they are uncorrelated, which for Gaussians implies that they are independent [PS89, p. 212]. So $|\lambda_j|^2 = X^2 + Y^2 \sim \chi^2_{2,n\sigma^2/2}$. Note that all circulant matrices have the same eigenvectors and thus the squared singular values of the concatenation of two circulant matrices are the sum of the squared absolute eigenvalues. So $[\boldsymbol{G}|\boldsymbol{F}]$ has one squared singular value s_0^2 distributed as $\chi^2_{1,n\sigma^2} + \chi^2_{1,n\sigma^2} = \chi^2_{2,n\sigma^2}$, and $n-1$ squared singular values $s_1^2, \ldots, s_{n-1}^2$ distributed as $\chi^2_{2,n\sigma^2/2} + \chi^2_{2,n\sigma^2/2} = \chi^2_{4,n\sigma^2/2}$. By Lemma 2.1 they have a log-expectation of

$$\mathbb{E}[\ln s_0^2] = \ln(2n\sigma^2) + \psi(1), \text{ and } \mathbb{E}[\ln s_j^2] = \ln(n\sigma^2) + \psi(2)$$

for $j = 1, \ldots, n-1$. We conclude by noting that $\mathbb{E}[\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}))] = \frac{1}{2}\sum_{i=0}^{n-1} \ln(s_i^2)$. □

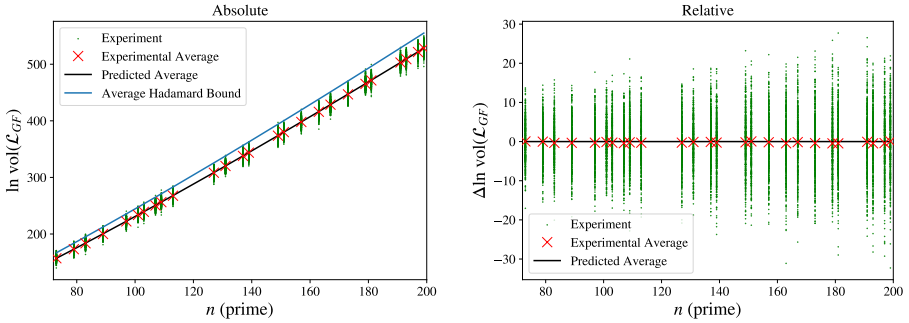

Fig. 5. Experimental values of $\ln(\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}}))$ versus Lemma 4.4 for circulant NTRU with discrete Gaussians and variance $\sigma^2 = \frac{2}{3}$. For each parameter n we generated 512 instances.

Validation. To validate our concrete estimate for $\mathrm{vol}(\mathcal{L}^{\boldsymbol{GF}})$ we generated the NTRU sublattice for several dimensions and computed its volume. We sample the secret coefficients following a discrete Gaussian with variance $\sigma^2 = \frac{2}{3}$ and ran experiments for both matrix NTRU and circulant NTRU. In Figures 4 and 5 we see that the predictions from Lemmas 4.3 and 4.4 perfectly fit the observed volumes in all dimensions. We do note that the variance is quite significant for the circulant case, but it can be fully explained by the computed eigenvalue distributions in the proof of Lemma 4.4.

4.3 Further Refinements

We discuss some further refinements, some of which were already successfully applied to the 2016 Estimate [AGVW17, DDGR20, PV21].

Gaussian Heuristic. For our asymptotic analysis we used Minkowski's bound to estimate the length $\lambda_1\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right)$ in terms of the volume $\mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right)$. A natural way to obtain a concrete estimate for the expected minimal length is by assuming that the intersection $\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}$ follows the Gaussian Heuristic and thus for our prediction we assume that

$$\lambda_1\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right) = \mathrm{gh}(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}) \approx \sqrt{k/(2\pi e)} \cdot \mathrm{vol}\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:n+k)}\right)^{1/k}.$$

We should however be careful with this assumption, as in fact it is false for $k = n$. E.g. the above predicts that $\lambda_1(\mathcal{L}^{\boldsymbol{GF}}) \approx \sqrt{n/(2\pi e)} \cdot \sqrt{2n\sigma^2}$, while we know that $\lambda_1(\mathcal{L}^{\boldsymbol{GF}}) = \|(\boldsymbol{g}, \boldsymbol{f})\| \approx \sqrt{2n\sigma^2}$, a factor $\Theta(\sqrt{n})$ shorter than predicted. The reason for this is that the dense sublattice is up to rotation and scaling very similar to the orthogonal lattice $\mathbb{Z}^n$, precisely the lattice for which it is well known that the Gaussian Heuristic is false. For small $k \ll n$ we do observe that the intersected lattice $\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:s)}$ follows the Gaussian Heuristic; the orthogonal structure seems to be broken by the intersection. However we do not have a clear idea how large k can become before the orthogonal structure returns and the minimal length stops following the prediction from the Gaussian Heuristic. We think this behaviour deserves some further investigation, e.g. if the transition is very sudden or not, and we leave it as an open problem. This near-orthogonality of $\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:s)}$ may be critical to model the DSD-LL events.

Probabilities. So far we have only considered expectations of volumes and projections. While this is enough to give a rough concrete estimate we want to be more precise. Success probabilities can accumulate up over multiple BKZ blocks and (progressive) tours, possibly leading to success at much lower blocksizes than the rough estimate. We continue using the expected values for the volume of the dense sublattice and the intersection volumes to obtain the expected length $\lambda_1\left(\mathcal{L}^{\boldsymbol{GF}}_{\cap[0:s)}\right)$ of the dense sublattice vector via the Gaussian Heuristic. However we then model the short dense sublattice vector $\mathbf{v} \in \mathcal{L}^{\boldsymbol{GF}}_{\cap[0:s)}$ as an s-dimensional Gaussian vector with the same expected length; allowing us to compute the

exact probability that $\|\pi_{s-\beta}(\mathbf{v})\| \leqslant \|\mathbf{b}^*_{s-\beta}\|$ using the CDF of the chi-square distribution with β degrees of freedom.

Up to now we have ignored the probability that after $\pi_{s-\beta}(\mathbf{v})$ is inserted, it is also correctly lifted to the full vector $\mathbf{v}$ by later BKZ tours. While this almost always happens for higher blocksizes, it is not so likely for lower blocksizes, and ignoring this leads to overly optimistic predictions. For BKZ-β to successfully lift or *eventually* pull the vector $\mathbf{v}$ to the front it should also satisfy $\|\pi_i(\mathbf{v})\| \leqslant \|\mathbf{b}^*_i\|$ for all $i = s - 2\beta + 1, s - 3\beta + 2, \ldots$. These conditions are not independent which makes them hard to compute exactly. We simplify the computation by only considering the dependence for consecutive positions $i, i - \beta + 1$ as done in [DDGR20]. We iteratively run our estimator for progressive $\beta = 2, 3, \ldots$ and take account of all probabilities assuming that all tours behave completely independently. Our new concrete estimate will be the expected successful blocksize. Additionally this allows us to combine both the (probabilistic) SKR 2016 Estimate and our new DSD-PT estimate in a single estimator. With some more administration we can also predict the distribution of the successful location κ, and predict the probability that the SKR event happens before the DSD event.

BKZ shape for low blocksizes. While the formulas for the (Z)GSA slope α_β and the expected first minimum $\mathrm{gh}(\beta)$ convert to the experimental values for large blocksizes of say $\beta \geqslant 50$, they are not as accurate for small β. As expected the convergence is worse for progressive BKZ when we only use a few tours of each blocksize. We ran some experiment on random low dimensional q-ary lattices to obtain practical estimates for $\mathrm{gh}(\beta)$ with $\beta \leqslant 50$. Earlier works about the 2016 Estimate resorted to BKZ simulators to predict the BKZ shape, which account for the number of tours and also the special shape of the head and tail that do not perfectly follow the GSA shape. Together with the earlier mentioned refinements this resulted in very precise predictions [DDGR20, PV21]. However how BKZ acts on a Z-shaped basis is much less understood [AD21] and as of yet there are no accurate BKZ simulators. Understanding the behaviour and creating an accurate simulator would be very interesting, but is out of the scope of this work. We continue using the ZGSA, but we resort to experimental values for α_β obtained by running BKZ on random q-ary lattices for large q. To remain consistent we also do not use a simulator for the GSA shape, and accept the small discrepancy between the predictions and practical experiments.

5 Experimental Verification

In this section we experamentally confirm our predicitions. Further detailed experimental data and discussion is given in the eprint version [8].

5.1 Successful blocksize

We start with comparing our concrete predictions to the preliminary experiment from Section 3.1. We ran progressive BKZ with 8 tours on matrix NTRU

[8] Section 5.2 in https://eprint.iacr.org/2021/999.

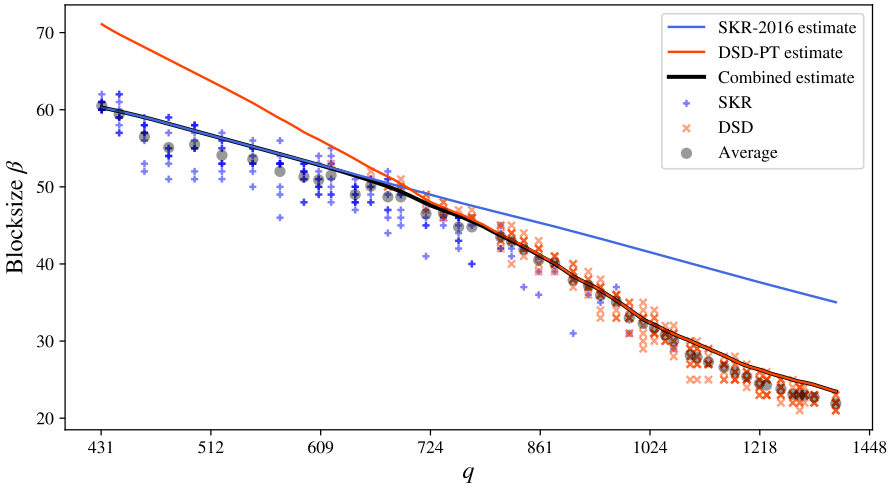

Fig. 6. Experiment versus prediction for progressive BKZ with 8 tours on matrix NTRU instances with parameters $n = 127, \sigma^2 = \frac{2}{3}$ for several moduli q. We did 10 runs per modulus q.

instances with parameters $n = 127, \sigma^2 = \frac{2}{3}$ for several moduli q. In Figure 6 we show the blocksizes at which the SKR or DSD event is first detected, and compare them to our concrete estimator. We ran the estimator three times for each modulus q: only accounting for SKR, only accounting for DSD-PT, and accounting for both. Note that the combined estimate can be strictly lower than both the first two because the probabilities to succeed accumulate over both events. We calibrated the values of α_β by running the same BKZ routine on $(2 \cdot 127)$-dimensional q-ary lattices with $q \approx 2^{20}$.

We observe that the experiments match the estimates reasonably well, with an average blocksize error of less than 2 for the DSD events and less than 3 for the SKR events. We shortly discuss potential sources of the small errors error.

- We do not actually run the classical BKZ algorithm, but the BKZ 2.0 algorithm as it is more feasible to run for large blocksizes. One part of the latter algorithm is that in each BKZ block $[\kappa : \kappa + \beta)$ the last $\beta - 1$ vectors are randomised before finding a short projected vector. This temporarily breaks the GSA shape and results a small 'bump' in the profile that is pushed to the right during a tour. On average we measured at the SKR events a log-increase of 0.048 on the value of $\|\mathbf{b}_\kappa^*\|$ compared to the GSA (while the rest of the basis matches very closely). Although anecdotal, adjusting our estimator with this offset of 0.048 resulted in very close predictions for the SKR events.
- For small blocksizes $\beta \leqslant 30$ we see that our DSD-PT estimate is slightly pessimistic compared to the experiments. However the successful profile slope α_β (computed from the profile at the moment of detection) does closely match the predicted slope $\alpha_{\beta_{pred}}$, pointing to a wrong calibration of the

slope parameter for very low blocksizes. Note that the non-flat part of the Z-shape in our experiments has size less than the $2 \cdot 127$ dimensional lattice used for calibration, which plausibly explain why the slope converges more quickly than expected.

5.2 Fatigue Point

Our concrete estimator follows the experiments reasonably well and thus we can use it to estimate the concrete fatigue point for dimensions that are not feasible in practice. To verify our estimate of the fatigue point we also did some experiments in dimensions that are still feasible. For this we ran a *soft* binary search, only decreasing the interval length by 3/4 so as not view a probabilistic result as a definitive answer. More specifically, starting with a range of $[q_{\min}, q_{\max}]$ we ran an experiment for a prime $q \approx (q_{\min} + q_{\max})/2$. If it succeeds with an SKR event we update $q_{\min}$ to $(q_{\min} + q)/2 + 1$, if it succeeds with a DSD event we update $q_{\max}$ to $(q_{\max} + q)/2 - 1$. We repeat this until the interval does not contain any prime and we return $(q_{\min} + q_{\max})/2$ as a rough estimate of the fatigue point. We averaged this over 20 experiments for each parameter n. We chose for matrix NTRU because of the lower variance in the hardness of these instances.

We compared this to our prediction. Because the estimator accounts for probabilities of events, we can predict for which value of q about 50% of the instances succeeds with a DSD event. Because it would be unreasonable to calibrate the low blocksize slope values α_β for each dimension we reused those of the $2 \cdot 127$ dimensional q-ary lattice from an earlier experiment. This might make the esti-

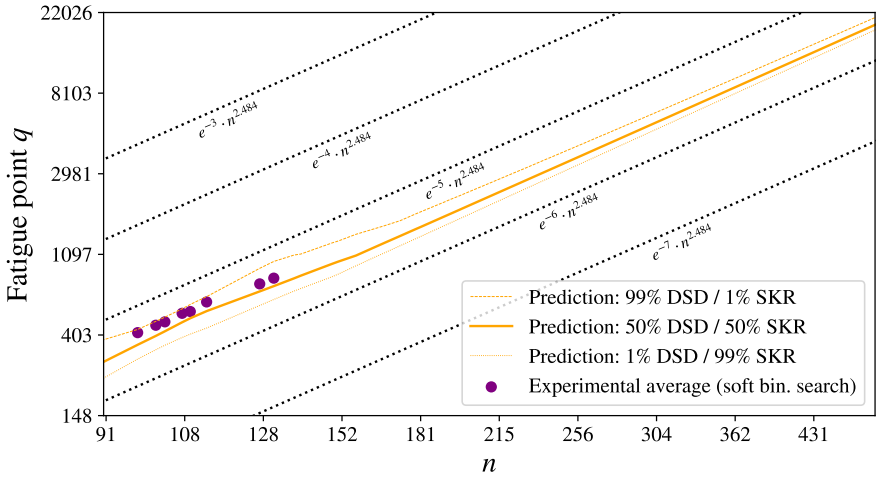

Fig. 7. Concrete fatigue point versus asymptotics using progressive BKZ with 8 tours on matrix NTRU instances with variance $\sigma^2 = \frac{2}{3}$. The 0.5 percentile line shows for which q we estimate that the DSD event is triggered before the SKR event for about 50% of the instances.

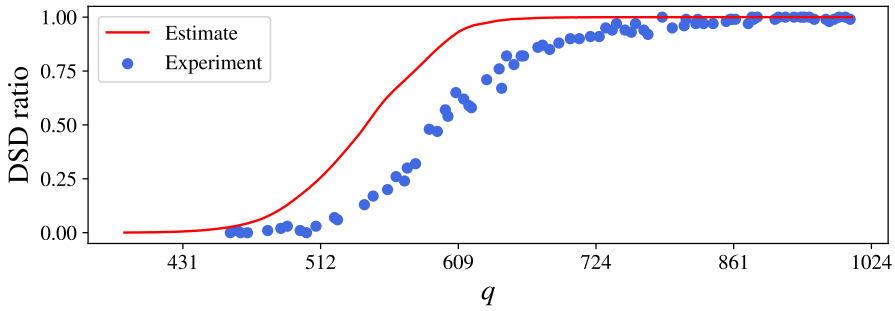

Fig. 8. Experiment versus prediction for progressive BKZ with 8 tours on matrix NTRU instances with parameters $n = 113, \sigma^2 = \frac{2}{3}$ for several moduli q. We did 100 runs per modulus q and the plot shows the ratio of these runs succeeding with a DSD event (before an SKR event).

mates a bit less precise for $n \ll 127$, and $n \gg 127$ if the successful blocksize is small around the fatigue point.

The results are shown in Figure 7 and plotted against $Cn^{2.484}$ for several constants C. Remarkably the experiments and concrete predictions closely follow the asymptotics already for reasonably small values of n. A loglog-linear regression of the 50% DSD-PT estimate over all primes $199, \ldots, 499$ gives $0.0034 \cdot n^{2.506}$. Restricting the exponent to 2.484 gives $0.0038 \cdot n^{2.484}$ with a log-standard deviation of only 0.006.

The experimental average appears slightly higher than the estimator prediction for 50% DSD - 50% SKR. The main reason for this seems to be that the estimator is slightly pessimistic for detecting the SKR event, as already observed and explained in Section 5.1. Another small detail is that the binary search is slightly biased to higher values of q because at each iteration we pick the *next* prime after $(q_{\min} + q_{\max})/2$.

5.3 Zoom on the Fatigue point: a smooth probabilistic transition

We take a closer look at the transition from the non-overstretched to the overstretched regime. For this we ran several experiments on matrix NTRU instances with parameters $n = 113, \sigma^2 = \frac{2}{3}$ for several moduli q, with 100 runs each. We compare the DSD success ratio with our probabilistic concrete estimate. The results are shown in Figure 8. Just as in Figure 7 we see a shift between the experiment and prediction, which can again be explained by our SKR estimator being too pessimistic. Note however that while the discrepancy looks significant in this zoomed plot, it only emphasises a small error of about 2 block sizes between the experiments and our predictions. Ignoring this shift the shape of the predicted transition matches the experiments very well.

References

[ABD16] Martin R. Albrecht, Shi Bai, and Léo Ducas. A subfield lattice attack on overstretched NTRU assumptions - cryptanalysis of some FHE and graded encoding schemes. In *CRYPTO 2016, Part I*, volume 9814 of *LNCS*, pages 153–178. Springer, Heidelberg 2016. 1, 2, 4, 12

[AD21] Martin Albrecht and Léo Ducas. Lattice attacks on NTRU and LWE: A history of refinements. Cryptology ePrint Archive, Report 2021/799, 2021. https://eprint.iacr.org/2021/799. 10, 24

[ADPS16] Erdem Alkim, Léo Ducas, Thomas Pöppelmann, and Peter Schwabe. Post-quantum key exchange - A new hope. In *USENIX Security 2016*, pages 327–343. USENIX Association 2016. 10, 11, 13

[AGVW17] Martin R. Albrecht, Florian Göpfert, Fernando Virdia, and Thomas Wunderer. Revisiting the expected cost of solving uSVP and applications to LWE. In *ASIACRYPT 2017, Part I*, volume 10624 of *LNCS*, pages 297–322. Springer, Heidelberg 2017. 13, 23

[Ajt99] Miklós Ajtai. Generating hard instances of the short basis problem. In *ICALP*, pages 1–9, 1999. 1

[AWHT16] Yoshinori Aono, Yuntao Wang, Takuya Hayashi, and Tsuyoshi Takagi. Improved progressive BKZ algorithms and their precise cost estimation by sharp simulator. In *EUROCRYPT 2016, Part I*, volume 9665 of *LNCS*, pages 789–819. Springer, Heidelberg 2016. 3

[BBC+20] Daniel J. Bernstein, Billy Bob Brumley, Ming-Shing Chen, Chitchanok Chuengsatiansup, Tanja Lange, Adrian Marotzke, Bo-Yuan Peng, Nicola Tuveri, Christine van Vredendaal, and Bo-Yin Yang. NTRU Prime. Technical report, National Institute of Standards and Technology, 2020. 7

[BEF+17] Jean-François Biasse, Thomas Espitau, Pierre-Alain Fouque, Alexandre Gélin, and Paul Kirchner. Computing generator in cyclotomic integer rings - A subfield algorithm for the principal ideal problem in $\mathrm{L}_{|\Delta_{\mathbb{K}}|}(\frac{1}{2})$ and application to the cryptanalysis of a FHE scheme. In *EUROCRYPT 2017, Part I*, volume 10210 of *LNCS*, pages 60–88. Springer, Heidelberg 2017. 4

[BLLN13] Joppe W. Bos, Kristin Lauter, Jake Loftus, and Michael Naehrig. Improved security for a ring-based fully homomorphic encryption scheme. In *14th IMA International Conference on Cryptography and Coding*, volume 8308 of *LNCS*, pages 45–64. Springer, Heidelberg 2013. 2, 4

[BSW18] Shi Bai, Damien Stehlé, and Weiqiang Wen. Measuring, simulating and exploiting the head concavity phenomenon in BKZ. In *ASIACRYPT 2018, Part I*, volume 11272 of *LNCS*, pages 369–404. Springer, Heidelberg 2018. 9

[CDH+20] Cong Chen, Oussama Danba, Jeffrey Hoffstein, Andreas Hulsing, Joost Rijneveld, John M. Schanck, Peter Schwabe, William Whyte, Zhenfei Zhang, Tsunekazu Saito, Takashi Yamakawa, and Keita Xagawa. NTRU. Technical report, National Institute of Standards and Technology, 2020. 1, 2, 5, 6

[CDPR16] Ronald Cramer, Léo Ducas, Chris Peikert, and Oded Regev. Recovering short generators of principal ideals in cyclotomic rings. In *EUROCRYPT 2016, Part II*, volume 9666 of *LNCS*, pages 559–585. Springer, Heidelberg 2016. 4

[CG05] Michael Coglianese and Bok-Min Goi. Matru: A new ntru-based cryptosystem. In *International Conference on Cryptology in India*, pages 232–243. Springer, 2005. 4, 7

[CJL16] Jung Hee Cheon, Jinhyuck Jeong, and Changmin Lee. An algorithm for ntru problems and cryptanalysis of the ggh multilinear map without a low-level encoding of zero. *LMS Journal of Computation and Mathematics*, 19(A):255–266, 2016. 1

[CN11] Yuanmi Chen and Phong Q. Nguyen. BKZ 2.0: Better lattice security estimates. In *ASIACRYPT 2011*, volume 7073 of *LNCS*, pages 1–20. Springer, Heidelberg 2011. 9

[CS97] Don Coppersmith and Adi Shamir. Lattice attacks on NTRU. In *EUROCRYPT'97*, volume 1233 of *LNCS*, pages 52–61. Springer, Heidelberg 1997. 7

[DDGR20] Dana Dachman-Soled, Léo Ducas, Huijing Gong, and Mélissa Rossi. LWE with side information: Attacks and concrete security estimation. In *CRYPTO 2020, Part II*, volume 12171 of *LNCS*, pages 329–358. Springer, Heidelberg 2020. 2, 10, 11, 15, 23, 24

[DE+04] Persi Diaconis, Paul Erdös, et al. On the distribution of the greatest common divisor. In *A festschrift for Herman Rubin*, pages 56–61. Institute of Mathematical Statistics, 2004. 19

[DKR+20] Jan-Pieter D'Anvers, Angshuman Karmakar, Sujoy Sinha Roy, Frederik Vercauteren, Jose Maria Bermudo Mera, Michiel Van Beirendonck, and Andrea Basso. SABER. Technical report, National Institute of Standards and Technology, 2020. 1

[dt21a] The FPLLL development team. FPLLL, a lattice reduction library, Version: 5.4.1. Available at https://github.com/fplll/fplll, 2021. 5

[dt21b] The FPLLL development team. fpylll, a Python wraper for the fplll lattice reduction library, Version: 0.5.6. Available at https://github.com/fplll/fpylll, 2021. 5

[EHKS14] Kirsten Eisenträger, Sean Hallgren, Alexei Kitaev, and Fang Song. A quantum algorithm for computing the unit group of an arbitrary degree number field. In *46th ACM STOC*, pages 293–302. ACM Press 2014. 4

[GGH13] Sanjam Garg, Craig Gentry, and Shai Halevi. Candidate multilinear maps from ideal lattices. In *EUROCRYPT 2013*, volume 7881 of *LNCS*, pages 1–17. Springer, Heidelberg 2013. 2

[GGH+19] Nicholas Genise, Craig Gentry, Shai Halevi, Baiyu Li, and Daniele Micciancio. Homomorphic encryption for finite automata. In *ASIACRYPT 2019, Part II*, volume 11922 of *LNCS*, pages 473–502. Springer, Heidelberg 2019. 4, 7

[GS02] Craig Gentry and Michael Szydlo. Cryptanalysis of the revised NTRU signature scheme. In *EUROCRYPT 2002*, volume 2332 of *LNCS*, pages 299–320. Springer, Heidelberg 2002. 2

[HPS98] Jeffrey Hoffstein, Jill Pipher, and Joseph H. Silverman. NTRU: A ring-based public key cryptosystem. In *ANTS*, pages 267–288, 1998. 1, 2, 4, 6

[KF17] Paul Kirchner and Pierre-Alain Fouque. Revisiting lattice attacks on overstretched NTRU parameters. In *EUROCRYPT 2017, Part I*, volume 10210 of *LNCS*, pages 3–26. Springer, Heidelberg 2017. 2, 3, 4, 10, 12

[LLL82] Arjen K. Lenstra, Hendrik W. Lenstra, Jr., and László Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261(4):515–534 1982. 2

[LPR13] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. *Journal of the ACM*, 60(6):43:1–43:35 2013. Preliminary version in Eurocrypt 2010. 2

[LW20] Changmin Lee and Alexandre Wallet. Lattice analysis on MiNTRU problem. Cryptology ePrint Archive, Report 2020/230, 2020. https://eprint.iacr.org/2020/230. 2

[MS01] Alexander May and Joseph H Silverman. Dimension reduction methods for convolution modular lattices. In *International Cryptography and Lattices Conference*, pages 110–125. Springer, 2001. 2

[PAA+19] Thomas Poppelmann, Erdem Alkim, Roberto Avanzi, Joppe Bos, Léo Ducas, Antonio de la Piedra, Peter Schwabe, Douglas Stebila, Martin R. Albrecht, Emmanuela Orsini, Valery Osheter, Kenneth G. Paterson, Guy Peer, and Nigel P. Smart. NewHope. Technical report, National Institute of Standards and Technology, 2019. 1

[PS89] Athanasios Papoulis and H Saunders. Probability, random variables and stochastic processes. 1989. 22

[PT08] Gábor Pataki and Mustafa Tural. On sublattice determinants in reduced bases. *arXiv preprint arXiv:0804.4014*, 2008. 2, 12, 15

[PV21] Eamonn W. Postlethwaite and Fernando Virdia. On the success probability of solving unique SVP via BKZ. In *PKC 2021, Part I*, volume 12710 of *LNCS*, pages 68–98. Springer, Heidelberg 2021. 10, 11, 23, 24

[Reg04] Oded Regev. Quantum computation and lattice problems. *SIAM J. Comput.*, 33(3):738–760, 2004. Preliminary version in FOCS 2002. 2

[Reg05] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *37th ACM STOC*, pages 84–93. ACM Press 2005. 1

[SAB+20] Peter Schwabe, Roberto Avanzi, Joppe Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, John M. Schanck, Gregor Seiler, and Damien Stehlé. CRYSTALS-KYBER. Technical report, National Institute of Standards and Technology, 2020. 1

[Sch87] Claus-Peter Schnorr. A hierarchy of polynomial time lattice basis reduction algorithms. *Theor. Comput. Sci.*, 53:201–224, 1987. 3

[SS11] Damien Stehlé and Ron Steinfeld. Making NTRU as secure as worst-case problems over ideal lattices. In *EUROCRYPT 2011*, volume 6632 of *LNCS*, pages 27–47. Springer, Heidelberg 2011. 1

[SSTX09] Damien Stehlé, Ron Steinfeld, Keisuke Tanaka, and Keita Xagawa. Efficient public key encryption based on ideal lattices. In *ASIACRYPT 2009*, volume 5912 of *LNCS*, pages 617–635. Springer, Heidelberg 2009. 1, 2

[YD17] Yang Yu and Léo Ducas. Second order statistical behavior of LLL and BKZ. In *SAC 2017*, volume 10719 of *LNCS*, pages 3–22. Springer, Heidelberg 2017. 9