Sharper and Closed-Form Attacks on SIS When Modulus Is Small

Navid Abapour

Surrey Centre for Cyber Security, Guildford, United Kingdom

Abstract. The Large Norm attacks of Ducas-Espitau-Postlethwaite (CRYPTO 2023) on the ISIS problem have shown that small values of q can be used to recover short solutions. This was applied to Falcon and Mitaka. Two issues are left, however. First, the cost model of the attack oversimplifies the BDGL sieve; it does not account for how long vectors will be distributed and treats two dependent probabilistic events as if they were independent, which results in an overestimation of the attack cost. Secondly, the analysis only deals with the ℓ_2 norm and has not yet been extended to ISIS$^\infty$, which underlies Dilithium-type systems.

This work addresses both issues. First, we extend the cost model of the Large Norm attack by including the principal sieve length distribution into our estimation of success probabilities. In addition, we use a joint probability rather than an approximate factor where possible. Since all other parts of the original θ convolution framework are reused, the extension is relatively minor. We reduce the attack cost of Large Norm on Falcon-256 by a $\approx 11\times$ cheaper model, and successfully forge a Mitaka-512 signature in ≈ 4.5 seconds at a higher success rate. We additionally introduce a closed-form ℓ_∞ variant as a Z-shape attack against Dilithium-type ISIS$^\infty$ at small-to-moderate modulus, which succeeds in ≤ 1.6 seconds across three presets.

Keywords: Lattice Cryptanalysis · Large Norm Attack · Lattice Sieves

1 Introduction

A *digital signature scheme* is a public-key primitive in which a signer publishes a verification key vk and keeps a secret signing key sk; a signature on a message msg is a string σ that an honest signer computes from (sk,msg) and that anyone holding vk can check [CMHST25, ACC+21]. The scheme is secure if no efficient adversary can produce, without knowing sk, a pair (msg*,σ^*) that the verifier accepts [GJM+24, KMSS25]. In *lattice-based* signatures the verification key encodes a high-dimensional integer linear system $\mathbf{Ax} \equiv \mathbf{u} \pmod q$, and forging a signature reduces to producing a nonzero integer vector $\mathbf{x}$ that solves this system and is *short* according to a chosen norm [SY25]. The set $\Lambda_q^\perp(\mathbf{A}) = \{\mathbf{x} \in \mathbb{Z}^m : \mathbf{Ax} \equiv \mathbf{0} \pmod q\}$ of solutions to the homogeneous system is a *lattice*: an additively closed grid of integer points sitting inside $\mathbb{Z}^m$. The forger's task is therefore to find a lattice point that lies inside a small geometric region (a Euclidean ball of radius ν in the ℓ_2 case, a coordinate-aligned box of half-side β_∞ in the ℓ_∞ case); the harder this geometric search is believed to be, the larger the security level claimed by the scheme [LPMSW19]. The modulus q controls the "coarseness" of the lattice: smaller q produces fewer short vectors per unit volume, and is therefore proposed as a bandwidth-saving optimization for compact signatures [DEP23], but it also exposes the lattice to specialized attacks that this work studies and improves.

The Inhomogeneous Short Integer Solution problem is the central computational assumption underlying lattice-based digital signatures [ETWY22]. An instance of ISIS is

E-mail: n.abapour@surrey.ac.uk (Navid Abapour)

This work is licensed under a "CC BY 4.0" license.
Date of this document: 2026-06-30.

a uniformly random matrix $\mathbf{A}\in\mathbb{Z}_q^{n\times m}$ together with a uniformly random syndrome $\mathbf{u}\in\mathbb{Z}_q^n$, and the task is to find a vector $\mathbf{x}\in\mathbb{Z}^m$ satisfying $\mathbf{A}\mathbf{x}\equiv\mathbf{u}\pmod q$ and $\|\mathbf{x}\|\le\nu$ for a public norm bound $\nu>0$. Replacing the Euclidean norm bound by the entry-wise bound $\|\mathbf{x}\|_\infty\le\beta_\infty$ defines the variant ISIS^∞, which underlies signature schemes whose forgery condition is geometrically expressed as a coordinate box rather than a ball. After the 2022 standardization round of the NIST [Nat22], the deployed lattice signatures Falcon [FHK+20, AAB+24, HPT25, YZZ+24] and Dilithium [LDK+22, BBD+23, ABC+23], together with their successors Mitaka [EFG+22, LZS+24, Pre23] and Hawk [DPPvW22, APMvW25, FH23, Mur25, CME+25], all reduce existential forgery to one of these two problems. The concrete security claims of these schemes are based on lattice-reduction attacks against the kernel basis $\mathbf{B}_\mathbf{A}\in\mathbb{Z}^{m\times m}$ of the q-ary lattice $\Lambda_q^\perp(\mathbf{A})=\{\mathbf{x}\in\mathbb{Z}^m:\mathbf{A}\mathbf{x}\equiv\mathbf{0}\pmod q\}$, run with the BKZ-$\beta$ algorithm of Schnorr–Euchner [SE94b]. For the typical regime $m=2n$ and prime q, the BKZ-reduced basis is known to exhibit a piecewise-linear "Z-shape" Gram–Schmidt profile: a flat plateau of n_q vectors at log-norm $\log q$ (the unmoved q-vectors of Zone I), a sloped Geometric Series Assumption region of length n_{GSA} (Zone II), and a flat tail of n_1 unit vectors at log-norm 0 (Zone III) [HG07]. The Large Norm [1] attack of Ducas, Espitau and Postlethwaite [DEP23], by addressing the work of [ETWY22], exploits this Z-shape: it sieves the projected sublattice $\Lambda_{[\ell:\ell+\beta]}$ immediately past Zone I to obtain a database of short Euclidean vectors, then *lifts* each candidate over the n_q q-vectors by reducing the missing coordinates modulo q around zero, accepting the lift when the full vector lies in the ball $\mathcal{B}_m(\nu)$. Two open lines of work follow from this attack: refining its cost model in the ℓ_2 regime that is already supported, and extending it to the ℓ_∞ regime that it does not currently target.

The Large Norm attack [DEP23] proceeds in three stages whose intuition is independent of the formal definitions to follow. The first stage is *lattice reduction*: an algorithm called BKZ-β [SE94b] processes the kernel basis $\mathbf{B}_\mathbf{A}$ to produce an equivalent basis whose vectors are shorter and closer to mutually orthogonal, with a tunable quality parameter β (called the block-size) controlling the trade-off between running time and shortness. The peculiar geometry of a q-ary lattice causes the output of BKZ-β to split into three contiguous regions, the so-called Z-shape: the first region consists of basis vectors that BKZ has been unable to shorten and that retain length exactly q, the middle region consists of properly reduced vectors of intermediate length, and the third region consists of trivial unit vectors [HG07]. The second stage is a *lattice sieve* [NV08, BDGL16]: an algorithm that takes the middle region of the reduced basis and outputs a large database of short vectors lying in the smaller-dimensional sub-lattice spanned by it. The third stage is *lifting*: each short vector returned by the sieve, although short with respect to the middle region only, can be extended into a candidate full vector of the original lattice by deterministically filling in the missing coordinates of the first region, using the modular constraint $\mathbf{A}\mathbf{x}\equiv\mathbf{0}\pmod q$ to fix them uniquely modulo q and centering each one in the interval $(-q/2,q/2]$. Whether the resulting full vector is short enough to constitute a valid forgery depends on the random outcome of these centered residues; computing the probability of this random event accurately is the only place where the forger's cost estimate has any flexibility that we improve it in this work.

Problem 1. The cost estimate of Large Norm attack [DEP23] factors as $|\mathcal{P}|\cdot p^\dagger$, where $|\mathcal{P}|$ is the size of the sieve database and $p^\dagger$ is the per-vector probability that a uniform-mod-q residue lifts into the target ball. Three pessimistic choices sit inside this product. First, the BDGL on-the-fly sieve [BDGL16] populates a spherical shell $[0,\sqrt{3/2}\mathrm{gh}]$, not a single sphere; evaluating $p^\dagger$ at the worst-case radius $L_{\max}=\sqrt{4/3}\mathrm{gh}$ underestimates the average per-vector success, because the cube–ball intersection probability $\pi(R;q,n_q):=|\mathrm{Cube}_{n_q}(q)\cap\mathcal{B}_{n_q}(R)|/q^{n_q}$ is monotonically decreasing in L. Second, the inhomogeneous reduction of ISIS to a punctured SIS* instance is charged the detached factor $2/q$, treating the coordinate event $\{x'_0\in\{\pm1\}\}$ as independent of the length event $\{\|\mathbf{x}'\|\le\nu\}$,

[1] In this work, we also refer the Large Norm attack as *Small-q* attack.

even though they are positively correlated. Third, the BDGL database size $(3/2)^{\beta/2}$ is downgraded to $(4/3)^{\beta/2}$ for the cost evaluation. The [DEP23] acknowledge these gaps but do not address them; the combined slack is of the order of several bits and propagates into a consequential under-attack of Falcon, Mitaka and the families that sit on top of ISIS.

Problem 2. To our knowledge, there is no practical ℓ_∞ Z-shape attack on ISIS, and the attack of [DEP23] is ℓ_2-specific. Its lift-success probability is computed by truncated theta convolution, which counts integer points in a Euclidean ball; no closed-form analogue is given for the box $\mathcal{C}_{n_q}(\beta_\infty) := \{\mathbf{x} \in \mathbb{R}^{n_q} : \|\mathbf{x}\|_\infty \le \beta_\infty\}$, and no practical instantiation of the attack against ISIS^∞ is reported. In [DEP23] it has been mentioned that the Z-shape "did not lead to the best attacks" on Dilithium's ℓ_∞-ISIS, an observation that holds for the standardised parameter set Dilithium-2 ($q = 8380417$, $\beta_\infty \approx 2^{17}$), where the box is so loose that Zone I lifting is essentially free and generic BKZ on the kernel basis dominates the attack. The argument is, however, silent on the parameter regime of small-to-moderate modulus ($q \sim 2^4 - 2^5$) that arises in Dilithium-adjacent constructions: lightweight instantiations, threshold signatures [CAL21], and MPC-in-the-head follow-ups [FJR22]. In that regime the ratio $\beta_\infty/q \in [0.15, 0.35]$ is sufficiently tight that Zone I lifting is non-trivial, and whether the Z-shape attack succeeds, with what cost, and at what concrete blocksize, is an open problem.

Contributions

To address both problems above, we present two attacks in this work: First, we instantiate an efficient variant of Large Norm attack [DEP23] by sharpening it. Second, we design an ℓ_∞ variant of the Z-shape attack, which targets Dilithium-family signatures with small/moderate modulus.

Tightening the ℓ_2 cost model (Sharp Large Norm Attack). We resolve Problem 1 by replacing the worst-case point-mass evaluation of $\pi(R; q, n_q)$ with an integral against the BDGL spherical-shell density $f_L(L) = \beta_S L^{\beta_S - 1} / L_{\max}^{\beta_S}$ supported on $[0, \sqrt{3/2}\,\mathrm{gh}]$, and by replacing the detached $2/q$ factor with the exact joint probability $p_{\mathsf{ISIS}}^{\mathrm{sharp}}(L) = (2/q)\pi(\sqrt{\nu^2 - L^2 - 1}; q, n_q - 1)$. The latter follows from conditioning on $x_0'^2 = 1$, which reduces the residual radius budget by 1 and the residual dimension by 1, and is computed by extending the truncated-theta engine of [DEP23, §2.4] with a single additional convolution. Our refactored cost engine predicts a CoreSVP cost of 36.85 bits on Falcon-256 versus the 40.36 bits of the Large Norm attack [DEP23], a reduction of $\Delta = 3.51$ bits, equivalent to a factor of $2^{3.51} \approx 11\times$ in attack work. Feeding the same expected-success count into the practical attack pipeline of [DEP23] permits the terminal sieve dimension on Mitaka-512 ($n = 512$, $q = 257$, $\nu = 1470.71$) giving a $1.36\times$ wall-clock speed-up on seeded trials $\beta_{\mathrm{sieve}} = 58$, for the Large Norm [DEP23] configuration $\beta_{\mathrm{sieve}} = 60$. To our knowledge, this is the first cost engine for the Large Norm attack that simultaneously accounts for sieve-shell length dispersion and positively-correlated ISIS inhomogeneity in a single closed-form expression.

Practical ℓ_∞ Z-shape attack (Small-q^∞ Attack). We address Problem 2 by porting the Large Norm attack to ISIS^∞. The truncated-theta cube–ball intersection $\pi(R; q, n_q)$ is replaced by the closed-form cube–box intersection $\pi^\infty(\beta_\infty; q, n_q) = ((2\beta_\infty + 1)/q)^{n_q}$, and the Euclidean acceptance check $\|\mathbf{x}\| \le \nu$ is replaced by the box check $\|\mathbf{x}\|_\infty \le \beta_\infty$. Because $\{\pm 1\} \subseteq [-\beta_\infty, +\beta_\infty]$ for every $\beta_\infty \ge 1$, the inhomogeneity event $\{x_0' \in \{\pm 1\}\}$ is strictly contained in the box event, and the joint ℓ_∞-ISIS probability factorises exactly as $(2/q) \cdot ((2\beta_\infty + 1)/q)^{n_q - 1}$ with no correction term. The attack therefore dispenses with the truncated-theta convolution engine of [DEP23, §2.4] entirely. We instantiate the construction on three Dilithium-type presets at $(n, q, \beta_\infty) \in \{(56, 23, 10), (64, 31, 14), (48, 13, 5)\}$ with BKZ-18 on the full kernel basis followed by a terminal sieve of dimension 30; every

preset completes each forgery in wall-clock time ≤ 1.6 seconds, and recovers solutions with $\|\mathbf{x}\|_\infty/\beta_\infty \leq 0.86$. To our knowledge, these are the first practical ℓ_∞ Z-shape attacks on ISIS^∞ in the small-modulus regime, which [DEP23] explicitly flagged as unaddressed.

2 Preliminaries

Notation. Vectors are column vectors written in lowercase bold, $\mathbf{x}=(x_1,\ldots,x_m)^\top$. The standard ℓ_2 norm is $\|\mathbf{x}\|$ and the ℓ_∞ norm is $\|\mathbf{x}\|_\infty=\max_i|x_i|$. For a positive integer q we write $\mathbb{Z}_q$ for the centred residue system $\{-\lfloor(q-1)/2\rfloor,\ldots,\lfloor q/2\rfloor\}$ and $\mathrm{Cube}_n(q)=\mathbb{Z}_q^n$. The Euclidean ball of radius R in $\mathbb{R}^n$ is $\mathcal{B}_n(R)$ and the ℓ_∞ box of half-side β_∞ is $\mathcal{C}_n(\beta_\infty)=\{\mathbf{x}\in\mathbb{R}^n:\|\mathbf{x}\|_\infty\leq\beta_\infty\}$. We write $\mathrm{U}(\cdot)$ for the uniform distribution.

2.1 Lattices

Before recalling the formal definition we summarize the geometric picture relied upon throughout our work. A *lattice* is a regular grid of points in $\mathbb{R}^d$, generalizing the integer grid $\mathbb{Z}^d$: it is the set of all integer-coefficient combinations of a fixed list of m linearly independent vectors $\mathbf{b}_1,\ldots,\mathbf{b}_m$ called a basis [AD21]. Different bases can describe the same lattice (any uni-modular change of basis preserves it), but a generic basis is far from orthogonal and contains very long vectors compared with the actual shortest non-zero point of the lattice [CCSL24]. *Lattice reduction algorithms* [LLL82, SE94b] replace such a basis with an equivalent one in which the basis vectors are shorter and closer to orthogonal; their output is the standard input for any quantitative cryptanalytic computation [GJ21]. Two derived quantities recur in our work: the *volume* $\mathrm{vol}(\Lambda)$, which is invariant of the choice of basis and quantifies how widely spaced the lattice points are [BCD+16]; and the *Gaussian heuristic* $\mathrm{gh}(\Lambda)$, an estimate for the length of the shortest non-zero lattice point [AGPS20], whose accuracy is useful for the lattices that arise in our work. The lattices we consider are all q-ary, that is, they contain the scaled integer grid $q\mathbb{Z}^m$ [PS24]; the modulus q drives the coarseness of the lattice and ultimately the difficulty of the attacks.

A *lattice* $\Lambda\subset\mathbb{R}^d$ of rank m is the integer span of the columns of a matrix $\mathbf{B}\in\mathbb{R}^{d\times m}$ of full column rank, called a *basis*; Λ is *full-rank* if $d=m$. Bases $\mathbf{B},\mathbf{C}$ generate the same lattice iff $\mathbf{B}=\mathbf{C}\mathbf{U}$ for some $\mathbf{U}\in\mathrm{GL}_m(\mathbb{Z})$, so $\mathrm{vol}(\Lambda):=\sqrt{\det(\mathbf{B}^\top\mathbf{B})}$ is well defined [DEP23]. The first minimum $\lambda_1(\Lambda)=\min_{\mathbf{x}\in\Lambda\setminus\{0\}}\|\mathbf{x}\|$ is heuristically estimated by the *Gaussian heuristic* $\mathrm{gh}(\Lambda)=v_m^{-1/m}\mathrm{vol}(\Lambda)^{1/m}\approx\sqrt{m/(2\pi e)}\mathrm{vol}(\Lambda)^{1/m}$, with $v_m=\pi^{m/2}/\Gamma(1+m/2)$ the unit-ball volume.

The *Gram–Schmidt* orthogonalization $\mathbf{B}^*=(\mathbf{b}_1^*\,|\,\cdots\,|\,\mathbf{b}_m^*)$ satisfies $\mathbf{B}=\mathbf{B}^*\mathbf{M}$ for an upper-triangular unipotent $\mathbf{M}$ and induces orthogonal projections $\pi_{\mathbf{B},i}:\mathbb{R}^d\to\mathrm{span}(\mathbf{b}_1,\ldots,\mathbf{b}_{i-1})^\perp$. The *profile* of $\mathbf{B}$ is the tuple $(\log\|\mathbf{b}_i^*\|)_{i=1}^m$; note $\mathrm{vol}(\Lambda)=\prod_i\|\mathbf{b}_i^*\|$. For $1\leq\ell<r\leq m+1$ the projected sublattice $\Lambda_{[\ell:r]}$ is generated by $\mathbf{B}_{[\ell:r]}=(\mathbf{b}_\ell^*\,|\,\pi_\ell(\mathbf{b}_{\ell+1})\,|\,\cdots\,|\,\pi_\ell(\mathbf{b}_{r-1}))$. A lattice is *$q$-ary* [MR09] if $q\mathbb{Z}^m\subseteq\Lambda\subseteq\mathbb{Z}^m$.

2.2 Computational Problems

Let $\mathbf{A}\leftarrow\mathrm{U}(\mathbb{Z}_q^{n\times m})$, the *Short Integer Solution* problem $\mathrm{SIS}_{n,m,q,\nu}$ is to find $\mathbf{x}\in\mathbb{Z}^m\setminus\{0\}$ with $\mathbf{A}\mathbf{x}\equiv 0\pmod q$ and $\|\mathbf{x}\|\leq\nu$. The *inhomogeneous* variant $\mathsf{ISIS}_{n,m,q,\nu}$ adds a syndrome $\mathbf{u}\leftarrow\mathrm{U}(\mathbb{Z}_q^n)$ and asks for $\mathbf{x}$ with $\mathbf{A}\mathbf{x}\equiv\mathbf{u}\pmod q$ and $\|\mathbf{x}\|\leq\nu$ [Ajt96]. The *punctured* variant $\mathrm{SIS}^*_{n,m,q,\nu}$ disallows trivial solutions of the form $q\mathbf{e}_i$ [MP12, LS15]. For prime q and $m=2n$, $\mathbf{A}$ admits with overwhelming probability a column permutation $\mathbf{A}=(\mathbf{A}_1\,|\,\mathbf{A}_2)$ with $\mathbf{A}_1\in\mathrm{GL}_n(\mathbb{Z}_q)$, giving the *kernel basis* $\mathbf{B}_\mathbf{A}=\left(\begin{smallmatrix}q\mathbf{I}_n & -\mathbf{A}_1^{-1}\mathbf{A}_2\\ 0 & \mathbf{I}_{m-n}\end{smallmatrix}\right)$ of $\Lambda_q^\perp(\mathbf{A})=\{\mathbf{x}\in\mathbb{Z}^m:\mathbf{A}\mathbf{x}\equiv 0\pmod q\}$, which has rank m and volume q^n.

The ℓ_∞ *inhomogeneous SIS* problem $\mathsf{ISIS}^{\infty}_{n,m,q,\beta_\infty}$ replaces the bound $\|\mathbf{x}\| \le \nu$ with $\|\mathbf{x}\|_\infty \le \beta_\infty$. This is the form used in the Dilithium signature scheme [LDK+22], where forging a signature reduces to recovering an $\mathbf{x}$ inside the ℓ_∞ box $\mathcal{C}_m(\beta_\infty)$.

2.3 Reduction Algorithms and Z-Shape

Lenstra–Lenstra–Lovász (LLL) [LLL82] is a polynomial-time lattice reduction algorithm that finds a short, nearly orthogonal basis for an integer lattice; BKZ-β [SE94b, SE94a] is a block-based generalization of LLL that produces shorter vectors by trading higher computation time for better quality; i.e., it iteratively calls an SVP oracle on rank-β projected sub-blocks [ADH+19]. Its is captured by the *root Hermite factor* $\delta_\beta = \left(\frac{\beta}{2\pi e}(\pi\beta)^{1/\beta}\right)^{1/(2(\beta-1))}$ [GN08], and by the *Geometric Series Assumption* (GSA) [Sch03]: the reduced profile decreases as a geometric series with ratio $\gamma(\beta) = \delta_\beta^{-2}$, hence $\|\mathbf{b}_i^*\| \approx \delta_\beta^{m-2i+1}\mathrm{vol}(\Lambda)^{1/m}$.

Specializing to $\Lambda_q^\perp(\mathbf{A})$ with the kernel basis above, the first n Gram–Schmidt vectors are $q\mathbf{e}_i$ as long as $q < \delta_\beta^{m-2i+1}\mathrm{vol}(\Lambda)^{1/m}$ [MSD20]. This condition is satisfied only past some index ℓ, producing the *Z-shape* profile [HG07, DSDGR20, EK20]: a flat plateau of n_q many q-vectors (Zone I), a sloped GSA region of length n_{GSA} (Zone II) with slope $-2\log\delta_\beta$, and a flat tail of n_1 many unit vectors at log-norm 0 (Zone III), with $n_q + n_{\mathrm{GSA}} + n_1 = m$.

2.4 Lattice Sieves

A *lattice sieve* is an algorithmic procedure that, given a basis of a lattice of moderate rank, returns a large set of short non-zero lattice vectors [DP23]. The procedure starts from a pool of long lattice vectors and repeatedly replaces pairs of vectors $(\mathbf{u},\mathbf{v})$ in the pool by their sum or difference whenever the result is shorter than at least one of the operands [PV21]; iterating this rule converges to a pool of vectors whose lengths concentrate near the Gaussian heuristic $\mathrm{gh}(\Lambda)$ of the lattice. The number of vectors in the converged pool is exponential in the rank, and is referred to throughout this paper as the *sieve database* [NV08, BDGL16]. Two regimes of the same algorithm are distinguished in the literature: a base regime in which the database has size $(4/3)^{m/2}$ and all output vectors have length close to $\sqrt{4/3}\mathrm{gh}(\Lambda)$, and the *on-the-fly* (OTF) regime of [BDGL16] which produces a strictly larger database of size $(3/2)^{m/2}$ whose vector lengths are spread over $[0, \sqrt{3/2}\mathrm{gh}(\Lambda)]$. The cost of running a sieve on a rank-β lattice is conventionally modeled as $\kappa\beta$ bits in the *CoreSVP* cost model [BDGL16], with $\kappa = \log_2\sqrt{3/2} \approx 0.292$, and dominates every other operation in the Large Norm attack pipeline. Throughout this paper, an instance of the Large Norm attack runs a single sieve invocation, on the projected sub-block immediately past the Zone I plateau of the BKZ-reduced kernel basis; the sieve database it returns is the input to the lifting step described above.

A *lattice sieve* [NV08, BDGL16] on a basis of rank m outputs in time $2^{\Theta(m)}$ a constant fraction of $\{\mathbf{x} \in \Lambda \setminus \{0\} : \|\mathbf{x}\| \le \sqrt{4/3}\mathrm{gh}(\Lambda)\}$. Its *database* has expected size $(4/3)^{m/2}$, and lengths concentrate around $\sqrt{4/3}\mathrm{gh}(\Lambda)$. The BDGL on-the-fly (OTF) variant [BDGL16] additionally enumerates pairs within a wider radius and admits a database of size $(3/2)^{m/2}$ whose lengths spread over $[0, \sqrt{3/2}\mathrm{gh}(\Lambda)]$. Throughout, the cost of a single rank-β sieve is taken as the CoreSVP estimate $\kappa\beta$ with $\kappa = \log_2\sqrt{3/2} \approx 0.292$.

2.5 Theta Functions and the Cube–Ball Intersection

For a lattice Λ the *theta function* is $\Theta_\Lambda(\tau) = \sum_{\mathbf{x}\in\Lambda} e^{\pi i \tau \|\mathbf{x}\|^2}$ for $\Im(\tau) > 0$ [CS99]. Setting $X = e^{\pi i \tau}$, the coefficient of X^j in $\Theta_{\mathbb{Z}^m}$ is the number of $\mathbf{x} \in \mathbb{Z}^m$ with $\|\mathbf{x}\|^2 = j$, and $\Theta_{\mathbb{Z}^m} = (\Theta_{\mathbb{Z}})^m$. Truncating $\Theta_{\mathbb{Z}}$ to coefficients of X^j with $j \le R^2$ and projecting each

coordinate to its centred residue mod q yields the *cube–ball intersection probability*

$$\pi(R;q,n)=\frac{|\mathrm{Cube}_n(q)\cap\mathcal{B}_n(R)|}{q^n},\qquad \pi^{\infty}(\beta_\infty;q,n)=\frac{|\mathrm{Cube}_n(q)\cap\mathcal{C}_n(\beta_\infty)|}{q^n}=\Big(\tfrac{2\beta_\infty+1}{q}\Big)^n,$$

the second of which is the *closed-form* ℓ_∞ *analogue* we use throughout the Dilithium variant of our attack. The function π is computed by truncated theta convolution in time $O(nM\sqrt{N})$ [HG07] with $N=\lfloor R^2\rfloor+1$.

3 The Large Norm attack on Short Integer Solution

Here, we recall the attack of [DEP23]. The Large Norm attack of [DEP23] solves $\mathsf{ISIS}_{n,m,q,\nu}$ by combining four components in sequence: (**i**) construction of the *kernel basis* $\mathbf{B}_\mathbf{A}$ of the q-ary lattice $\Lambda_q^\perp(\mathbf{A})$ from a uniform public matrix $\mathbf{A}$ (the lattice on which short solutions live); (**ii**) BKZ-β reduction of $\mathbf{B}_\mathbf{A}$, producing a basis whose Gram–Schmidt log-norm sequence forms the *Z-shape* of Fig. 1; (**iii**) a single call to a lattice sieve on a rank-β projected sub-lattice positioned immediately after the Zone I plateau, returning a database $\mathcal{P}$ of short vectors; and (**iv**) a *lifting* step that extends each sieve vector into a candidate full solution by reducing the missing Zone I coordinates modulo q around zero, accepting the candidate when the resulting full vector has ℓ_2 length at most ν. The cost of the attack is the cost of one BKZ-β reduction (which dominates the sieve in the CoreSVP model), divided by the probability that one lift produces a valid solution; the optimal block-size β^* is the integer that minimizes this ratio. The remainder of this section formalizes components (**i**)–(**iv**) and culminates in the explicit cost procedure $\mathsf{Smallq}_{\mathrm{cost}}$ of Fig. 5, which putes them into a callable routine.

As Figure 1 shows, given $\mathbf{A}\leftarrow\mathrm{U}(\mathbb{Z}_q^{n\times m})$, the attacker forms the kernel basis $\mathbf{B}_\mathbf{A}\in\mathbb{Z}^{m\times m}$ of $\Lambda_q^\perp(\mathbf{A})$, of volume q^n, and applies BKZ-β reduction to obtain a basis $\mathbf{B}$ whose profile decomposes into three zones: **Zone I**, indices $1\le i\le n_q$, in which $\mathbf{b}_i^*=q\mathbf{e}_i$ (the unmoved q-vectors, $\log\|\mathbf{b}_i^*\|=\log q$); **Zone II**, indices $\ell:=n_q+1\le i\le n_q+n_{\mathrm{GSA}}$, the GSA slope of length n_{GSA} that drops from $\log q$ at slope $-2\log\delta_\beta$; and **Zone III**, indices $i>n_q+n_{\mathrm{GSA}}$, of n_1 unit vectors with $\log\|\mathbf{b}_i^*\|=0$. The three counts sum to m; the volume identity $\prod_{i=1}^m\|\mathbf{b}_i^*\|=q^n$ fixes them once β is chosen, via the simulator of [DEP23].

The three-zone decomposition of Figure 1 can be read as a partition of the m Gram–Schmidt vectors into "rigid", "reduced", and "trivial" parts. The Zone I prefix of length n_q consists of the standard q-vectors $q\mathbf{e}_1,\ldots,q\mathbf{e}_{n_q}$ that BKZ has not been able to shorten because no shorter lattice vector lies in their orthogonal complement at this block-size; the Zone II middle of length n_{GSA} is the genuine output of BKZ-β on the remaining sub-block, with log-norms decreasing arithmetically at slope $-2\log\delta_\beta$ as predicted by the geometric series assumption [Sch03, GN08]; and the Zone III suffix of length n_1 consists of unit vectors $\mathbf{e}_{m-n_1+1},\ldots,\mathbf{e}_m$ at log-norm 0, whose presence is forced by the volume identity $\prod_{i=1}^m\|\mathbf{b}_i^*\|=q^n$. The integer n_q is the only zone count that matters for the attack: it counts *exactly* the number of q-coordinates the lift step has to randomize, and therefore it sets both the dimension of the cube–ball intersection event and the block-size beyond which the Zone I plateau collapses (n_q is decreasing in β).

As we demonstrate in Figure 2, the attack focuses on the projected sublattice $\Lambda_{[\ell:r]}$ with $r=\min(\ell+\beta,m+1)$; this is the rank-β sub-block immediately past the Zone I boundary. Its first coordinate axis is spanned by the q-vector $q\mathbf{e}_1$, drawn green in Fig. 2; orthogonally to it sits the $(\ell-1)$-dimensional slab $\mathrm{Cube}_{\ell-1}(q)\times\mathbb{R}$ (blue) inside which short lattice points may live. A vector $\mathbf{x}\in\Lambda_q^\perp(\mathbf{A})$ of ℓ_2-length at most ν lies in the red ball $\mathcal{B}_n(\nu)$; its projection $\pi_\ell(\mathbf{x})$ onto the Zone II directions is the quantity directly searched by the sieve. *Lifting* a sieve output $\mathbf{v}\in\Lambda_{[\ell:r]}$ over the n_q q-vectors of Zone I amounts to pulling back $\mathbf{v}$ to a full $\Lambda_q^\perp(\mathbf{A})$-vector $\mathbf{x}$ by reducing each Zone I coordinate of $\mathbf{A}\mathbf{v}$ mod q to its centered representative in $\mathrm{Cube}_{n_q}(q)$. The lift succeeds if $\|\mathbf{x}\|\le\nu$, equivalently if the random Zone I residue lands inside the residual ball of radius $R=\sqrt{\nu^2-\|\pi_\ell(\mathbf{x})\|^2}$.

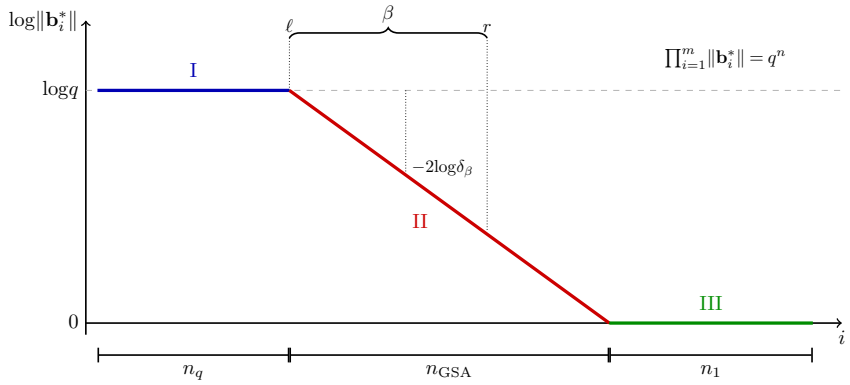


Figure 1: The Large Norm attack's Z-shape Gram–Schmidt profile of the kernel basis $\mathbf{B_A}$ of $\Lambda_q^\perp(\mathbf{A})$ after BKZ-β reduction: a flat plateau **I** of n_q unmoved q-vectors at log-norm $\log q$, a sloped **II** region of n_{GSA} properly reduced vectors decreasing at slope $-2\log\delta_\beta$ as predicted by the geometric series assumption, and a flat tail **III** of n_1 unit vectors at log-norm 0, with $n_q+n_{\mathrm{GSA}}+n_1=m$ fixed by the volume identity $\prod_{i=1}^m\|\mathbf{b}_i^*\|=q^n$; the rank-$\beta$ projected sub-block $[\ell:r]$ on which the sieve is invoked sits immediately past the Zone I boundary at index $\ell=n_q+1$.

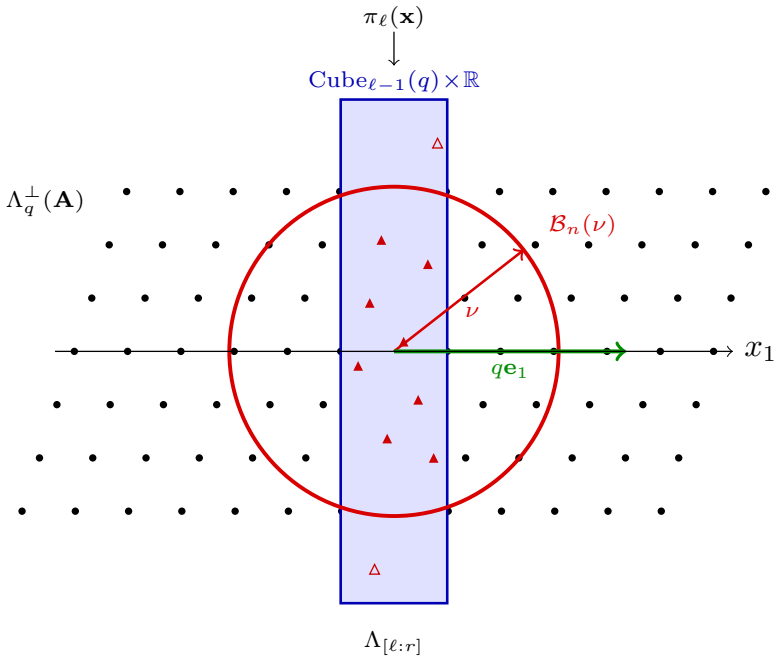


Figure 2: Lifting with two-dimensional slice in Large Norm attack

In fact, the lifting step takes as input a vector $\mathbf{v}\in\Lambda_{[\ell:r]}$ produced by the sieve and outputs either an admissible full vector $\mathbf{x}\in\Lambda_q^\perp(\mathbf{A})$ with $\|\mathbf{x}\|\le\nu$ or the symbol $\perp$. It proceeds in three operations: first, embed $\mathbf{v}$ into $\mathbb{R}^m$ by zero-padding its Zone I coordinates, obtaining a vector $\tilde{\mathbf{v}}$ of dimension m; second, compute the syndrome $\mathbf{s}=\mathbf{A}\tilde{\mathbf{v}}$ mod q and center it in $\mathrm{Cube}_{n_q}(q)$ by mapping every coordinate of $\mathbf{s}$ to the unique representative in $\{-(q-1)/2,\ldots,(q-1)/2\}$; third, set the Zone I block of $\mathbf{x}$ equal to $-\mathbf{s}$ and the Zone II/III block equal to $\mathbf{v}$. The resulting $\mathbf{x}$ satisfies $\mathbf{A}\mathbf{x}\equiv\mathbf{0}\pmod q$ by construction; the lift is

accepted if and only if $\|\mathbf{x}\| \le \nu$. Because the centred Zone I residue $-\mathbf{s}$ behaves heuristically as a uniform draw from $\mathrm{Cube}_{n_q}(q)$ [DEP23], the per-vector acceptance probability equals the cube–ball intersection probability $\pi(R;q,n_q)$ at the residual radius $R = \sqrt{\nu^2 - \|\mathbf{v}\|^2}$, which justifies the formula used in the cost procedure.

Treating the Zone I residues as uniform in $\mathrm{Cube}_{n_q}(q)$, the per-vector SIS^* lift-success probability is $\pi(R;q,n_q) = |\mathrm{Cube}_{n_q}(q) \cap \mathcal{B}_{n_q}(R)|/q^{n_q}$, computed by the truncated theta convolution [DEP23, §2.4]. For ISIS, an instance $(\mathbf{A},\mathbf{u})$ is reduced to a SIS^* instance on the augmented matrix $(\mathbf{u} \,|\, \mathbf{A})$ [DEP23, §3.5]: a solution $\mathbf{x}' = (x'_0,x'_1,\ldots,x'_m)$ with $x'_0 \in \{\pm 1\}$ inverts to an ISIS solution by setting $f = -x'_0$ and outputting $f \cdot (x'_1,\ldots,x'_m)$. The original cost model penalizes this by a detached factor of $2/q$, treating the event $\{x'_0 \in \{\pm 1\}\}$ as independent of $\{\|\mathbf{x}'\| \le \nu\}$, giving $p_{\mathsf{ISIS}}(R;q,n_q) = (2/q)\pi(R;q,n_q)$. So, this is cost of one lift, and ISIS reduction.

The cost formula (left column of Figure 5). Procedure $\mathsf{SmallQ}_{\mathrm{cost}}(q,w,h,\nu)$ constructs $\mathbf{B}_{\mathbf{A}}$ over a SIS lattice of rank w and log-volume exponent h, optimizes a single block-size $\beta^* = \mathrm{argmin}_\beta \mathrm{Cost}(\beta)$, runs BKZ-$\beta^*$, extracts the profile (returning the Zone I count n_q and the Gaussian heuristic gh of the Zone II block), sieves $\Lambda_{[\ell:r]}$ to obtain a database $\mathcal{P}$ of size $|\mathcal{P}| = (4/3)^{\beta^*/2}$ whose vectors all have length exactly $L_{\max} = \sqrt{4/3}\mathrm{gh}$ in the model, computes a single residual radius $R = \sqrt{\nu^2 - L_{\max}^2}$, evaluates $\log p_{\mathrm{lift}} = \log \pi(R;q,n_q)$, applies the detached $\log p_{\mathsf{ISIS}} = \log(2/q)$ factor, and returns $\mathrm{Cost} = \kappa\beta^* - (\log p_{\mathrm{lift}} + \log|\mathcal{P}| + \log p_{\mathsf{ISIS}})/\log 2$ [DEP23, §3]. The model is intentionally conservative on three counts: (i) sieve outputs are placed at the maximum radius rather than distributed across the support; (ii) the OTF database size $(3/2)^{\beta^*/2}$ is downgraded to $(4/3)^{\beta^*/2}$; (iii) the ISIS event is detached from the length event.

4 Our Sharp Large Norm attack on SIS

In Section 3, we described the Large Norm attack as the composition of a Z-shape reduction step, a sieve step, and a lifting step, parameterized by a single block-size β and a single sieve-output radius $L_{\max}$. The cost procedure $\mathsf{SmallQ}_{\mathrm{cost}}$ of Figure 5 (left) collapses two random quantities into deterministic worst cases: the actual length of a sieve output is replaced by $L_{\max}$, and the actual joint event $\{x'_0 \in \{\pm 1\}\} \cap \{\|\mathbf{x}'\| \le \nu\}$ is replaced by an independence approximation. Both replacements are conservative in the sense that they *decrease* the predicted attack-success probability, hence *increase* the predicted attack cost; tightening either of them lowers the attack-cost estimate without changing the attack itself. This section quantifies both tightening and combines them into a single procedure $\mathsf{SharpSmallQ}_{\mathrm{cost}}$ (Figure 5, right), whose evaluation reuses the truncated theta convolution of [DEP23] with one additional invocation per block-size.

The original cost model separates the sieve database to a single radius $L_{\max} = \sqrt{4/3}\mathrm{gh}$ (the blue Dirac $\delta_{L_{\max}}$ in Figure 3). Under the BDGL on-the-fly sieve [BDGL16], however, the actual database has size $(3/2)^{\beta_S/2}$ and its lengths populate the spherical-shell density $f_L(L) = \beta_S L^{\beta_S - 1}/L_{\max}^{\beta_S}$ on $[0, \sqrt{3/2}\mathrm{gh}]$ (the red curve in Figure 3). Because the per-vector lift-success probability $\pi(\sqrt{\nu^2 - L^2};q,n_q)$ is monotone *decreasing* in L, integrating against f_L rather than evaluating at $L_{\max}$ is strictly favorable to the attacker: $\int \pi(L) f_L(L) \mathrm{d}L \ge \pi(L_{\max})$, with equality only in the degenerate saturated case. We bucketize f_L and reuse the existing fast boxed theta engine of [DEP23] unchanged. So, in Figure 3, the model from [DEP23] is at $L_{\max} = \sqrt{(4/3)}gh$, and our model in Figure 3 is the shell density L^{beta-1} truncated at $\sqrt{(3/2)}gh$.

The detached $2/q$ factor of [DEP23] considers the event $\{x'_0 \in \{\pm 1\}\}$ as independent of $\{\|\mathbf{x}'\| \le \nu\}$, which yields the flat blue marginal of Figure 4 at height $\frac{1}{q}\pi(R;q,n_q - 1,\nu)$. The truth is the joint event, whose marginal is the red bell $\Pr[x'_0 = c \wedge \|\mathbf{x}'\| \le \nu] = \frac{1}{q}\pi(\sqrt{R^2 - c^2};q,n_q - 1)$ peaked at $c = 0$. Restricting to the green strips at $c = \pm 1$ yields the

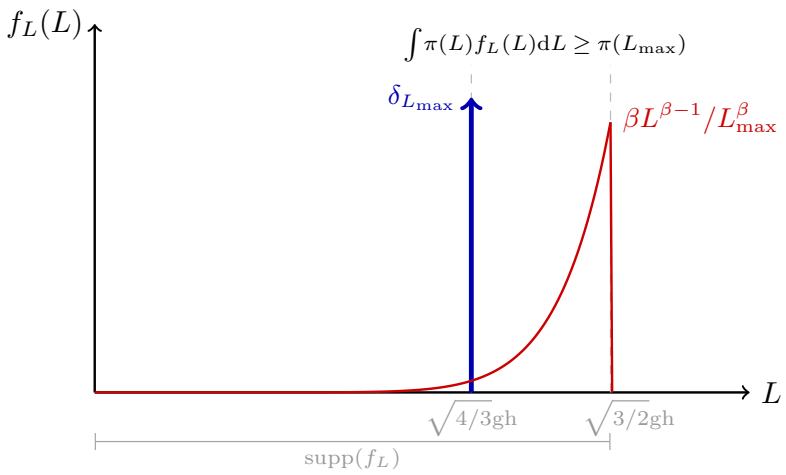


Figure 3: Length distribution of vectors in the projected sieve database $\mathcal{P} = \mathsf{Sieve}(\Lambda_{[\ell:r]})$ used by the Large Norm attack: the original cost model of [DEP23] (blue) collapses all $|\mathcal{P}|$ vectors to a single radius $L_{\max} = \sqrt{4/3}\mathrm{gh}$, whereas under the BDGL on-the-fly sieve (red) the database actually populates the spherical-shell density $f_L(L) = \beta_S L^{\beta_S - 1}/L_{\max}^{\beta_S}$ on $[0, \sqrt{3/2}\mathrm{gh}]$, captured by our $\mathsf{SharpSmallq}$ attack.

sharp ISIS per-lift probability

$$p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L;q,n_q,\nu) = \tfrac{2}{q}\,\pi\big(\sqrt{\nu^2 - L^2 - 1};q,n_q - 1\big),$$

which reuses the same theta engine with dimension reduced by 1 and budget reduced by 1 (accounting for $x_0'^2 = 1$). One verifies that $p_{\mathrm{ISIS}}^{\mathrm{sharp}}/(2/q\cdot\pi)$ exceeds 1 whenever typical squared coordinates dominate 1, i.e. throughout the attack-tight regime $R \approx q\sqrt{n_q/12}$.

So, in Figure 4, the marginal condition of [DEP23] is the uniform line at $(2/q)\cdot\pi(R)$, here the $2/q$ factor is uniform over the cube, so the marginal is a flat line. Our sharp condition in Figure 4 is biased toward c; curve $= P_{n_q - 1}(R^2 - c^2)/q$.

In Figure 5, the procedure $\mathsf{SharpSmallq}_{\mathrm{cost}}(q,w,h,\nu)$ shares lines 1 and 4 with $\mathsf{Smallq}_{\mathrm{cost}}$ (basis construction; profile extraction) and changes the remaining ones as follows. The block-size search becomes a one-dimensional minimization of the sharp cost; under the OTF model the database size becomes $|\mathcal{P}| = (3/2)^{\beta_S/2}$ at the same sieve dimension; the detached $\log(2/q)$ factor disappears; and $\log p_\star$ is now $\log \int_0^{L_{\max}} p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L;q,n_q,\nu) f_L(L)\mathrm{d}L$, returning $\mathrm{Cost}_{\mathrm{sharp}} = \kappa\beta^* - (\log p_\star + \log|\mathcal{P}|)/\log 2$.

Compared with $\mathsf{Smallq}_{\mathrm{cost}}$, our $\mathsf{SharpSmallq}_{\mathrm{cost}}$ differs in phases in which each replacing a conservative choice with the more precise one and each strictly favorable to the attacker: **(i)** the sieve length is treated as a distribution f_L rather than the point mass $\delta_{L_{\max}}$, capturing the $\beta_S - 1$ extra degrees of freedom of the BDGL spherical shell; **(ii)** the OTF database is taken at its precise size $(3/2)^{\beta_S/2}$ instead of the conservative $(4/3)^{\beta_S/2}$; **(iii)** the detached $2/q$ factor is replaced by the joint sharp probability $p_{\mathrm{ISIS}}^{\mathrm{sharp}} = (2/q)\pi(\sqrt{R^2 - 1};q,n_q - 1)$, which charges the engine with the correct event and not with a strictly loose bound; **(iv)** the cost formula collapses $\log p_{\mathrm{lift}} + \log p_{\mathrm{ISIS}}$ into a single integrated term $\log p_\star$, so that the BDGL shell density and the sharp ISIS event are evaluated jointly rather than as independent multiplicative corrections. The Lemma 1, investigates further superiority of our integrated cost engine.

Lemma 1. *Fix q,n_q,ν and a BDGL-OTF sieve of dimension $\beta_S \geq 2$ whose outputs have ℓ_2*

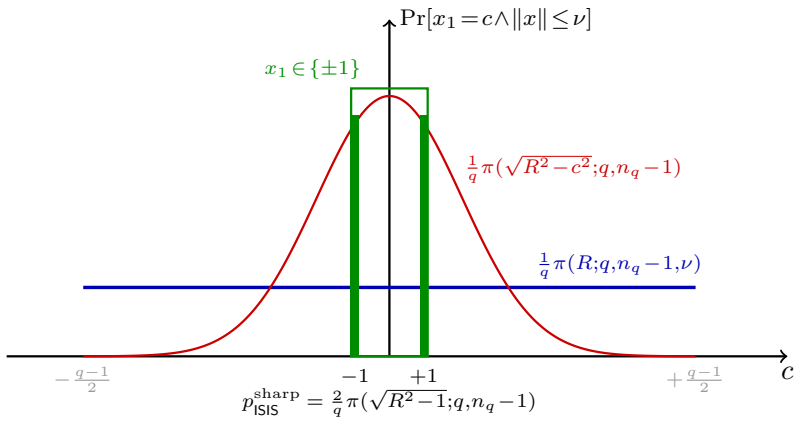


Figure 4: Marginal distribution of the inhomogeneity coordinate x_1 jointly with the length event $\{\|\mathbf{x}\| \le \nu\}$ over $\mathbf{x} \sim \mathrm{U}(\mathrm{Cube}_{n_q}(q))$, which contrasts the detached approximation of [DEP23] (blue) with our sharp closed form (red): the original cost model treats the events $\{x_1 \in \{\pm 1\}\}$ and $\{\|\mathbf{x}\| \le \nu\}$ as independent, which produces a flat marginal at constant height $\frac{1}{q}\pi(R;q,n_q-1,\nu)$, whereas the joint marginal is the centered bell $\frac{1}{q}\pi(\sqrt{R^2-c^2};q,n_q-1)$ peaked at $c=0$, whose mass restricted to the green strips $c=\pm 1$ defines the sharp ISIS probability $p_{\mathrm{ISIS}}^{\mathrm{sharp}} = \frac{2}{q}\pi(\sqrt{R^2-1};q,n_q-1)$ used by our cost engine, with $R=\sqrt{\nu^2-L^2}$ the residual radius left after the Zone II/III sieve output of length L.

$\mathsf{Smallq}_{\mathrm{cost}}(q,w,h,\nu)$

$\mathbf{B_A} \leftarrow \mathsf{Basis}(\Lambda_q^{\perp}(\mathbf{A}))$
$\beta^* \leftarrow \operatorname{argmin}_{\beta} \mathrm{Cost}(\beta)$
$\mathbf{B} \leftarrow \mathsf{BKZ}_{\beta^*}(\mathbf{B_A})$
$(n_q, \mathrm{gh}) \leftarrow \mathsf{Profile}(\mathbf{B}, \beta^*)$
$r \leftarrow \min(\ell+\beta^*, m+1)$
$\mathcal{P} \leftarrow \mathsf{Sieve}(\Lambda_{[\ell:r]})$
$L_{\max} \leftarrow \sqrt{4/3}\cdot\mathrm{gh}$
$R \leftarrow \sqrt{\nu^2 - L_{\max}^2}$
$\log p_{\mathrm{lift}} \leftarrow \log \pi(R;q,n_q)$
$\log|\mathcal{P}| \leftarrow (\beta^*/2)\log(4/3)$
$\log p_{\mathsf{ISIS}} \leftarrow \log(2/q)$
$\mathrm{Cost} \leftarrow \kappa\beta^* - (\log p_{\mathrm{lift}} + \log|\mathcal{P}| + \log p_{\mathsf{ISIS}})/\log 2$
return Cost

$\mathsf{SharpSmallq}_{\mathrm{cost}}(q,w,h,\nu)$

$\mathbf{B_A} \leftarrow \mathsf{Basis}(\Lambda_q^{\perp}(\mathbf{A}))$
$(\beta_R^*, \beta_S^*) \leftarrow \operatorname{argmin}_{\beta_R \le \beta_S} \mathrm{Cost}(\cdot)$
$\mathbf{B} \leftarrow \mathsf{BKZ}_{\beta_R^*}(\mathbf{B_A})$
$(n_q, \mathrm{gh}) \leftarrow \mathsf{Profile}(\mathbf{B}, \beta_R^*)$
$r \leftarrow \min(\ell+\beta_S^*, m+1)$
$\mathcal{P} \leftarrow \mathsf{Sieve}_{\beta_S^*}^{\mathsf{OTF}}(\Lambda_{[\ell:r]})$
$f_L \leftarrow \mathsf{ShellDensity}(\beta_S^*, \mathrm{gh})$
$\log p_\star \leftarrow \log \int p_{\mathsf{ISIS}}^{\mathrm{sharp}}(L) f_L(L)\mathrm{d}L$
$\log|\mathcal{P}| \leftarrow (\beta_S^*/2)\log(3/2)$
$\mathrm{Cost} \leftarrow \kappa\max(\beta_R^*, \beta_S^*) - (\log p_\star + \log|\mathcal{P}|)/\log 2$
return Cost

Figure 5: Left (gray): $\mathsf{Smallq}_{\mathrm{cost}}$ of [DEP23]. Right: our $\mathsf{SharpSmallq}_{\mathrm{cost}}$; lines identical to the left are gray, the differences in black. Here $\pi(R;q,n_q) = |\mathrm{Cube}_{n_q}(q) \cap \mathcal{B}_{n_q}(R)|/q^{n_q}$, $p_{\mathsf{ISIS}}^{\mathrm{sharp}}(L) = (2/q)\pi(\sqrt{\nu^2-L^2-1};q,n_q-1)$, $\mathsf{ShellDensity}(\beta,\mathrm{gh}) = \beta L^{\beta-1}/L_{\max}^{\beta}$ on $[0,L_{\max}]$ with $L_{\max} = \sqrt{3/2}\mathrm{gh}$.

length density $f_L(L) = \beta_S L^{\beta_S-1}/L_{\max}^{\beta_S}$ *supported on* $[0,L_{\max}]$ *with* $L_{\max} = \sqrt{3/2}\mathrm{gh}$. *Then*

$$\int_0^{L_{\max}} p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L;q,n_q,\nu) f_L(L)\mathrm{d}L \ge \frac{2}{q}\pi(R_{\max};q,n_q)$$

with $R_{\max}=\sqrt{\nu^2-L_{\max}^2}$ *(the [DEP23] point-mass evaluation), and the inequality is strict whenever* $\Pr[|x_0'|>1\wedge\|\mathbf{x}'\|\le\nu]>0$.

Proof. *The map* $L\mapsto\pi(\sqrt{\nu^2-L^2-1};q,n_q-1)$ *is monotonically non-increasing on* $[0,L_{\max}]$ *(the integrand of the truncated theta function* $\Theta_{\mathbb{Z}^{n_q-1}}$ *is term-wise monotone in the radius), hence so is* $p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L)=(2/q)\pi(\sqrt{\nu^2-L^2-1};q,n_q-1)$. *By the first mean-value theorem for integrals [Tao16],* $\mathbb{E}_{f_L}[p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L)]\ge p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L_{\max})$. *Second, shifting the radius budget by* 1 *to condition on* $x_0'^2=1$ *strictly enlarges the event counted by* $\pi(\cdot;q,n_q-1)$ *over the generic* $(2/q)\pi(\cdot;q,n_q)$ *whenever non-unit* x_0' *contribute to the length event. So, combining the two monotone enlargements yields the inequality.*

In Figure, 5, lines 1 and 3–4 of $\mathsf{SharpSmallq}_{\mathrm{cost}}$ are inherited from $\mathsf{Smallq}_{\mathrm{cost}}$: the kernel basis $\mathbf{B_A}$ is built from $\mathbf{A}$, BKZ-β_R^* is run on it to obtain a Z-shape, and the simulator of [DEP23] returns the Zone I count n_q together with the Gaussian heuristic gh of the rank-β_S^* projected sub-block. The remaining lines diverge from [DEP23] on four points. Line 2 selects the block-size as the minimizer of the *integrated* cost rather than the worst-case cost (and we collapse the search to a single block-size $\beta^*=\beta_R^*=\beta_S^*$ since the joint optimum lies on this diagonal in the parameter regimes considered, see Sec. 4.1). Line 6 invokes the BDGL on-the-fly sieve, whose database has size $|\mathcal{P}|=(3/2)^{\beta_S^*/2}$ rather than the conservative $(4/3)^{\beta_S^*/2}$. Line 7 instantiates the spherical-shell density $f_L(L)=\beta_S^*L^{\beta_S^*-1}/L_{\max}^{\beta_S^*}$ on $[0,\sqrt{3/2}\,\mathrm{gh}]$, supplied by the simulator, that describes the actual length distribution of the sieve database (Figure 3). Line 8 evaluates the integral $\log p_\star=\log\int_0^{L_{\max}}p_{\mathrm{ISIS}}^{\mathrm{sharp}}(L;q,n_q,\nu)f_L(L)\mathrm{d}L$, where $p_{\mathrm{ISIS}}^{\mathrm{sharp}}$ is the joint probability obtained by Lemma 1; this is the only line that calls the truncated theta convolution, and it does so a number of times equal to the number of buckets in the discretizations of f_L. Lines 9–12 assemble the cost as $\mathrm{Cost}=\kappa\beta^*-(\log p_\star+\log|\mathcal{P}|)/\log 2$, with $\kappa=\log_2\sqrt{3/2}$ the CoreSVP slope, and return it.

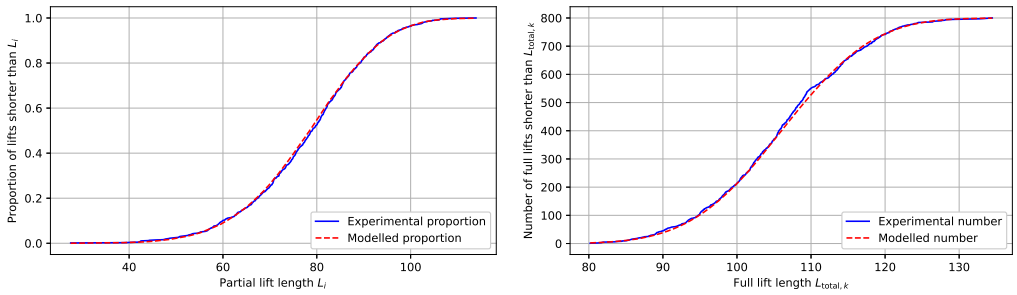
4.1 Experiments on Sharp Large Norm Attack

We instantiate both the cost model of [DEP23] and our sharp cost model in OTF-lift, ISIS "specific" mode, and predict the CoreSVP cost on Falcon-256 ($n=256,q=257,\nu=\tau\sigma\sqrt{2nq}\approx 560.98$). Our model lowers the predicted cost by $\Delta=3.51$ bits, a multiplicative factor of $2^{3.51}\approx 11.4\times$ in the predicted attack work; the optimal block-size also drops from $\beta^*=138$ to 126. To corroborate this concretely, we re-implement the sharp cost engine inside the pipeline of https://github.com/verdiverdiverdi/ISIS-LargeNorm and target the Mitaka-512 instance $(n,q,\nu)=(512,257,1470.71)$. Sweeping the terminal sieve dimension $\beta_{\mathrm{sieve}}\in\{54,56,58,60\}$ over five seeded trials each, the empirical knee is at $\beta_{\mathrm{sieve}}=58$ (mean 4.56 s), versus the safety choice $\beta_{\mathrm{sieve}}=60$ of [DEP23] (mean 4.95 s, expected 6.19 s). "Expected (s)" is mean wall-time divided by success rate, the cost per successful forgery. On identical seeded instances our variant yields a 1.36× speed-up, with every recovered solution satisfying the bound $\|\mathbf{x}\|<\nu=1470.71$. Our code is available at https://github.com/navid-divan/sharper_small_q_attacks. Our experiments are conducted on an Apple MacBook Air (2022) with the M2 chip, A15 Bionic (64-bit ARM-based), an 8-core CPU featuring a base clock speed of 3.49 GHz, and 16GB of memory, Tahoe OS 26.2 build 25C56, SageMath v.10.4, and Python v.3.12.3.

Figure 6(a) plots the empirical Cumulative Distribution Function (CDF) of partial Zone I lift lengths against the cube–ball intersection probability $\pi(L_i;q,n_q)$ that drives our cost engine of Sec. 4. The two curves coincide across the support, the maximum vertical deviation falling within Monte-Carlo noise at $N=800$ samples, which confirms that the truncated theta-convolution computation of π is statistically concrete at light parameters: the cube–ball geometry that $\mathsf{SharpSmallq}_{\mathrm{cost}}$ relies on for its lift-success probability matches the actual statistics of uniform residues modulo q to within sampling error. Figure 6(b)

Table 1: Comparison of the Large Norm vs. our sharp variant. Cost-model row: predicted CoreSVP cost on Falcon-256 (Sec. 4). Practical row: Mitaka forgery on $(n,q,\nu)=(512,257,1470.71)$ averaged over five seeded trials, random instances across both variants.

	Large Norm Attack [DEP23]		**Sharp Large Norm [Ours]**	
	block-size	cost / time	block-size	cost / time
Cost model, Falcon-256	$\beta^*=138$	40.36 bits	$\beta^*=126$	36.85 bits
Practical, Mitaka-512	$\beta_{\text{sieve}}=60$	6.19 seconds	$\beta_{\text{sieve}}=58$	4.56 seconds



(a) CDF of partial-lift lengths L_i.

(b) Counts of full-lift lengths $L_{\text{total},k}$.

Figure 6: Empirical validation of the SharpSmallq cost model at light parameters $(n,m,q,\beta_S,n_q)=(80,160,97,30,8)$. (a) Proportion of partial Zone I lifts of length at most L_i; the experimental curve (blue) is computed from $N=800$ uniform-mod-q residues drawn from $\text{Cube}_{n_q}(q)$, while the modeled curve (red dashed) is the cube–ball intersection probability $\pi(L_i;q,n_q)=|\text{Cube}_{n_q}(q)\cap\mathcal{B}_{n_q}(L_i)|/q^{n_q}$ used by our cost engine. (b) Number of full lifts whose total ℓ_2 length is at most $L_{\text{total},k}$; experimental (blue) over the same N samples, modeled (red dashed) by the integrated count $E_k=\sum_i\pi(\sqrt{L^2_{\text{total},k}-L^2_{\text{proj},i}};q,n_q)$ over BDGL-OTF sieve outputs $\{L_{\text{proj},i}\}_{i=1}^N$.

extends the validation from partial to full lifts, plotting the cumulative count of admissible full vectors as a function of the total ℓ_2 length $L_{\text{total},k}$ that the integrated cost engine is supposed to predict. The integrated count E_k, formed by summing π at the residual radius $\sqrt{L^2_{\text{total},k}-L^2_{\text{proj},i}}$ over the BDGL-OTF length distribution $\{L_{\text{proj},i}\}$, again tracks the experimental number across the entire support of $L_{\text{total},k}$, confirming that the integration of π against the spherical-shell density f_L in line 8 of $\text{SharpSmallq}_{\text{cost}}$ (Figure 5, right column) reproduces the joint behavior of the full attack and not merely a worst-case bound.

5 Our Small-q^∞ Attack: From ℓ_2 to ℓ_∞

The Dilithium-style ℓ_∞ analogue $\text{ISIS}^\infty_{n,m,q,\beta_\infty}$ replaces the Euclidean acceptance condition $\|\mathbf{x}\|\le\nu$ of the original problem by the coordinate-wise condition $\|\mathbf{x}\|_\infty\le\beta_\infty$. Geometrically, this swaps the Euclidean ball $\mathcal{B}_m(\nu)$ used by [DEP23] for the ℓ_∞ box $\mathcal{C}_m(\beta_\infty)$ and forces every coordinate of the recovered vector, including the Zone II/III coordinates that are returned directly by the sieve, to lie inside the box. We show that this regime is solvable by the Z-shape mechanism of Sec. 3 when the modulus q is small to moderate, and that the resulting cost procedure is strictly simpler than $\text{SharpSmallq}_{\text{cost}}$: the truncated theta convolution disappears, replaced by a closed-form expression in q,n_q and β_∞ (Lemma 2); only the Zone II/III tail-survival event of the sieve output requires a quantitative model.

The Z-shape mechanism only contributes to the attack when the Zone I prefix is long enough to make the cube–box intersection probability $\pi^\infty(\beta_\infty;q,n_q)=((2\beta_\infty+1)/q)^{n_q}$ a meaningful constraint on the lift step. When $\beta_\infty\ge q/2$, the box already covers the entire

centered cube and $\pi^\infty = 1$, so Zone I lifting succeeds unconditionally and the attack reduces to generic BKZ on the kernel basis; this is the regime of standardized Dilithium-2 with $q = 8380417$ and $\beta_\infty \approx 2^{17}$, in which [DEP23] observed that the Z-shape did not lead to the best attacks. In the opposite regime $\beta_\infty < q/2$, the ratio $(2\beta_\infty+1)/q$ is strictly less than 1 and π^∞ decays exponentially in n_q, which makes the lift step the dominant contribution to the success probability; the attack then succeeds when BKZ is strong enough to drive n_q down to single-digit values. This is the regime of Dilithium-adjacent constructions such as lightweight instantiations, threshold signatures [CAL21], and MPC-in-the-head follow-ups [FJR22], where $\beta_\infty/q \in [0.15, 0.35]$, and it is the regime our experiments target.

The Dilithium signature scheme [LDK+22] is secure against an adversary exactly to the extent that $\mathrm{ISIS}^\infty_{n,m,q,\beta_\infty}$ is hard: a forgery must exhibit a vector $\mathbf{x} \in \mathbb{Z}^m$ with $\mathbf{A}\mathbf{x} \equiv \mathbf{u} \pmod q$ and $\|\mathbf{x}\|_\infty \le \beta_\infty$. Running the [DEP23] pipeline on this problem preserves every step of Sections 3–4 up to one component: the cube–ball intersection probability $\pi(R; q, n_q)$ of [DEP23, §2.4], which was computed by truncated theta convolution, collapses in the ℓ_∞ setting to the closed-form *cube–box* intersection probability

$$\pi^\infty(\beta_\infty; q, n_q) = \frac{|\mathrm{Cube}_{n_q}(q) \cap \mathcal{C}_{n_q}(\beta_\infty)|}{q^{n_q}} = \left(\tfrac{2\beta_\infty+1}{q}\right)^{n_q}, \qquad p^\infty_{\mathrm{ISIS}}(\beta_\infty; q, n_q) = \tfrac{2}{q}\pi^\infty(\beta_\infty; q, n_q - 1)$$

the latter being *exact* in the ℓ_∞ case (unlike the ℓ_2 sharpening of Section 4): the event $\{x'_0 \in \{\pm 1\}\}$ is strictly inside the box for every $\beta_\infty \ge 1$, so the inhomogeneity costs only the marginal factor $2/q$ with no correction term. This observation dispenses with the convolution engine entirely in the ℓ_∞ regime (cf. Figure 7, top panel).

A sieve output $\mathbf{v} \in \Lambda_{[\ell:r]}$ is a vector with ℓ_2-length concentrated on the BDGL spherical shell $[0, \sqrt{3/2}\,\mathrm{gh}(\Lambda_{[\ell:r]})]$ [BDGL16]; it is usable for the box-lift only if its Zone II/III coordinates *individually* satisfy $|v_i| \le \beta_\infty$. Under the standard isotropy heuristic $v_i \sim \mathcal{N}(0, L^2/\beta_S)$ (cf. Fig. 7, bottom panel, red contour), the joint survival probability of one sieve vector of length L is $s_\infty(L) = \mathrm{erf}(\beta_\infty/(L\sqrt{2/\beta_S}))^{\beta_S}$, and the expected count of admissible lifts factorises as $\mathbb{E}[\#\mathrm{succ}] = |\mathcal{P}| \cdot \mathbb{E}_L[s_\infty(L)] \cdot p^\infty_{\mathrm{ISIS}}(\beta_\infty; q, n_q)$. The cost formula of Section 4 rewrites accordingly, with the theta-convolution integrand replaced by the closed-form π^∞; the attack's block-size optimum is the unique β where the CoreSVP cost $\kappa\beta$ balances the log-success count $-\log_2 \mathbb{E}[\#\mathrm{succ}]$.

The attack is applicable in a regime that [DEP23] addressed as open: small to moderate modulus $q \sim 2^4 - 2^5$ with the Dilithium-style structure $m = 2n$ and $\beta_\infty \le q/2$. In that regime BKZ-18 on the kernel basis $\mathbf{B}_\mathbf{A}$ already drives n_q to single-digit values, so the lift probability π^∞ is $\Theta(1)$ and the attack succeeds with overwhelming probability per trial; full Dilithium-2 [LDK+22] with $q = 8380417$ and $\beta_\infty \approx 2^{17}$ remains out of reach because the box is so loose that Zone I lifting imposes no constraint at all (Zone II dominates, which reduces the attack to generic BKZ on the full kernel). Our attack's success criterion changes from $\|\mathbf{x}\| < \nu$ of [DEP23] to $\|\mathbf{x}\|_\infty \le \beta_\infty$ and the theta convolution is replaced by the closed form. The Lemma 2 addresses the closed-form exact ISIS probability in ℓ_∞.

Lemma 2. *For $q, n_q \ge 1$, $\beta_\infty \ge 1$ with $2\beta_\infty + 1 \le q$, and $\mathbf{x}' = (x'_0, \ldots, x'_{n_q-1})$ drawn uniformly from $\mathrm{Cube}_{n_q}(q)$,*

$$\Pr\left[x'_0 \in \{\pm 1\} \wedge \|\mathbf{x}'\|_\infty \le \beta_\infty\right] = \frac{2}{q}\left(\frac{2\beta_\infty+1}{q}\right)^{n_q-1} = p^\infty_{\mathrm{ISIS}}(\beta_\infty; q, n_q).$$

The ℓ_∞ ISIS probability is exact*, with no correction term beyond the detached $2/q$ factor.*

Proof. *Since $\beta_\infty \ge 1$, the event $\{x'_0 \in \{\pm 1\}\}$ is strictly contained in $\{|x'_0| \le \beta_\infty\}$ and the conjunction factorises as the marginal $\Pr[x'_0 \in \{\pm 1\}]$ times the conditional event $\{|x'_i| \le \beta_\infty\}_{i \ge 1}$. Under $\mathbf{x}' \sim U(\mathrm{Cube}_{n_q}(q))$ the coordinates are independent and identically uniform over $\mathbb{Z}_q$, giving the marginal $2/q$ and the conditional $((2\beta_\infty+1)/q)^{n_q-1}$ respectively; the assumption $2\beta_\infty + 1 \le q$ guarantees no coordinate overflows the centered residue window.*

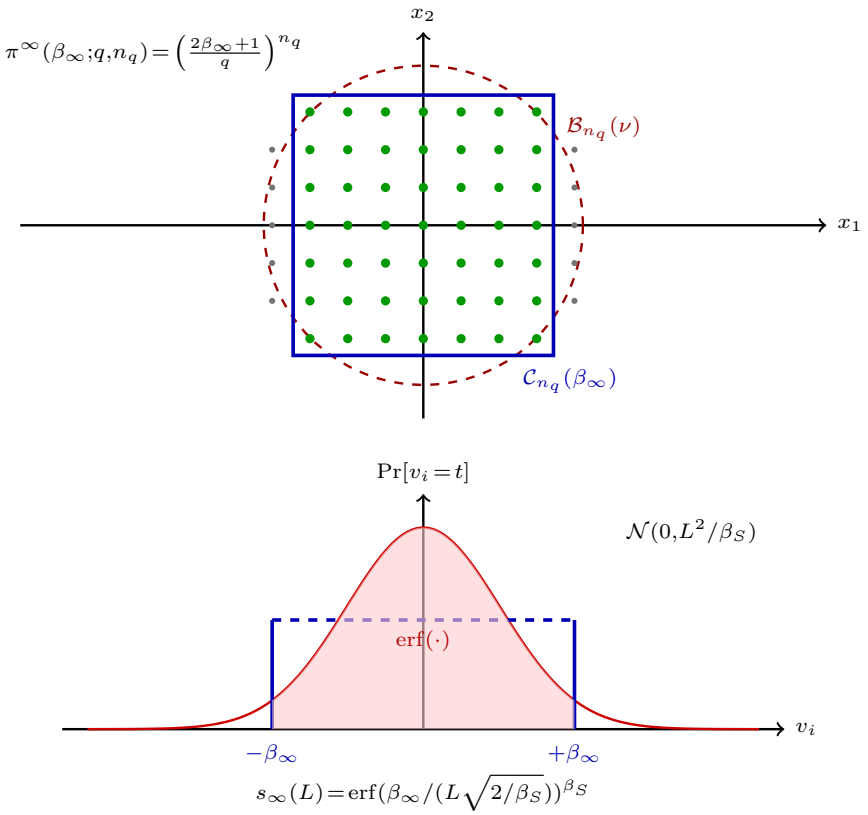


Figure 7: Small-q^∞ attack. *Top:* the Zone I lift event is driven by the cube–box intersection $\mathcal{C}_{n_q}(\beta_\infty)\cap\mathrm{Cube}_{n_q}(q)$ (blue box, green lattice points), giving the closed form $\pi^\infty=((2\beta_\infty+1)/q)^{n_q}$; the red dashed ball $\mathcal{B}_{n_q}(\nu)$ is the [DEP23] ℓ_2 event. *Bottom:* the Zone II/III tail-survival event on a single sieve coordinate v_i, Gaussian with variance L^2/β_S (red), clipped to the box of half-side β_∞ (blue); the shaded mass equals $\mathrm{erf}(\beta_\infty/(L\sqrt{2/\beta_S}))$.

Fig. 8 adapts $\mathsf{SharpSmallq}_{\mathrm{cost}}$ to the ℓ_∞ regime in exactly two places. Lines 1–7 are almost unchanged: the kernel basis is built, a single block-size β^* is selected, BKZ-β^* is run, the Zone I count n_q and the Gaussian heuristic gh are extracted, the rank-β^* projected sub-block is set up, the on-the-fly sieve is invoked, and the spherical-shell density f_L is supplied by the simulator. Line 8 introduces the Zone II/III tail-survival function $s_\infty(L)=\mathrm{erf}(\beta_\infty/(L\sqrt{2/\beta^*}))^{\beta^*}$, which estimates the probability that all β^* coordinates of a sieve output of length L individually lie in $[-\beta_\infty,\beta_\infty]$, under the standard isotropy heuristic that those coordinates are jointly distributed as $\mathcal{N}(0,L^2/\beta^*)$. Line 9 evaluates the closed-form Zone I lift probability $\log p^\infty_{\mathrm{lift}}=(n_q-1)\log((2\beta_\infty+1)/q)+\log(2/q)$ from Lemma 2; no theta convolution is invoked. Line 10 multiplies the two contributions across the length distribution to obtain $\log p^\infty_\star=\log\int s_\infty(L)f_L(L)\mathrm{d}L+\log p^\infty_{\mathrm{lift}}$. Lines 11–12 close the procedure with the same CoreSVP balance $\mathrm{Cost}=\kappa\beta^*-(\log p^\infty_\star+\log|\mathcal{P}|)/\log 2$ that $\mathsf{SharpSmallq}_{\mathrm{cost}}$ uses.

5.1 Experiments on Small-q^∞ Attack

We evaluate the Small-q^∞ attack on three preset $\mathsf{ISIS}^\infty_{n,m,q,\beta_\infty}$ instances chosen to exercise the small-modulus regime of the previous paragraph. Each preset fixes $m=2n$, picks a prime $q\in\{13,23,31\}$ corresponding to the modulus used in lightweight Dilithium-style constructions, and sets $\beta_\infty\approx q/3$ so that the cube–box ratio $(2\beta_\infty+1)/q$ stays in the non-trivial

$\mathsf{SmallQ}^{\infty}_{\mathrm{cost}}(q,w,h,\beta_\infty)$

$\mathbf{B_A} \leftarrow \mathsf{Basis}(\Lambda_q^{\perp}(\mathbf{A}))$
$\beta^* \leftarrow \operatorname{argmin}_{\beta} \mathrm{Cost}(\beta)$
$\mathbf{B} \leftarrow \mathsf{BKZ}_{\beta^*}(\mathbf{B_A})$
$(n_q,\mathrm{gh}) \leftarrow \mathsf{Profile}(\mathbf{B},\beta^*)$
$r \leftarrow \min(\ell+\beta^*, m+1)$
$\mathcal{P} \leftarrow \mathsf{Sieve}^{\mathsf{OTF}}_{\beta^*}(\Lambda_{[\ell:r]})$
$f_L \leftarrow \mathsf{ShellDensity}(\beta^*,\mathrm{gh})$
$s_\infty(L) \leftarrow \mathrm{erf}\left(\beta_\infty/(L\sqrt{2/\beta^*})\right)^{\beta^*}$
$\log p^{\infty}_{\mathrm{lift}} \leftarrow (n_q-1)\log\big((2\beta_\infty+1)/q\big)+\log(2/q)$
$\log p^{\infty}_{\star} \leftarrow \log\int s_\infty(L) f_L(L)\mathrm{d}L + \log p^{\infty}_{\mathrm{lift}}$
$\log|\mathcal{P}| \leftarrow (\beta^*/2)\log(3/2)$
$\mathrm{Cost} \leftarrow \kappa\beta^* - (\log p^{\infty}_{\star} + \log|\mathcal{P}|)/\log 2$
return Cost

Figure 8: Cost procedure of the Small-q^∞ attack. Compared with $\mathsf{SharpSmallq}_{\mathrm{cost}}$ of Fig. 5, the truncated-theta term $\pi(R;q,n_q)$ is replaced by the closed-form $\pi^\infty=((2\beta_\infty+1)/q)^{n_q}$, and the inner-product-to- Euclidean length bridge becomes the coordinate-wise erf-survival $s_\infty(L)$, so no theta-convolution engine is invoked here.

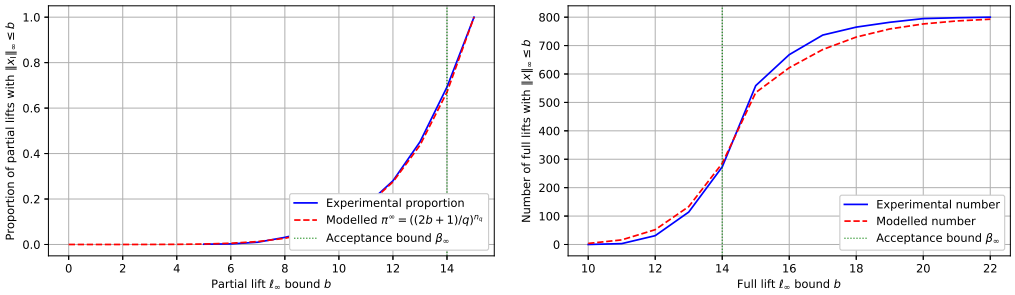
range [0.39,0.85]. For every preset we run BKZ at a fixed block-size $\beta^*=18$ on the full kernel basis $\mathbf{B_A}$ and follow it with a single on-the-fly sieve at the same block-size on the rank-β^* projected sub-block past Zone I; for each sieve output we apply the lifting procedure of Sec. 3 and accept the result whenever it satisfies $\|\mathbf{x}\|_\infty \le \beta_\infty$. The column "pred. $\log_2\mathbb{E}[\#\mathrm{succ}]$" of Table 2 records the cost-engine prediction $\log_2(|\mathcal{P}|\cdot\mathbb{E}_L[s_\infty(L)]\cdot p^{\infty}_{\mathrm{ISIS}}(\beta_\infty;q,n_q))$ at these block-size settings; values strictly above zero indicate that more than one admissible lift is expected per sieve database, and accordingly that the attack succeeds with overwhelming probability per trial. The remaining columns report the empirical evaluation: the number of seeded trials, the mean wall-clock time of one attack execution, the mean number tr. of sieve candidates examined before an admissible lift was found, and the mean ℓ_∞ norm of the recovered solution; comparing the latter to the bound β_∞ is the operational success indicator for an ISIS^∞ forgery.

We instantiate the Small-q^∞ attack against uniformly random $\mathrm{ISIS}^{\infty}_{n,m,q,\beta_\infty}$ instances for three Dilithium-style preset parameter regimes with $m=2n$ and $\beta_\infty \approx q/3$. On every preset the measured success rate equals 1 over the seeded trials, the measured wall-clock time per successful forgery stays below 1.6 seconds, and the number of sieve candidates examined is ≤ 15 in all trials. The predicted $\log_2\mathbb{E}[\#\mathrm{succ}] \in [+3.65,+4.23]$ translates to 12–18 expected admissible lifts per sieve pass, within a factor of 2 of the measured candidate counts, and the measured ℓ_∞ norms of recovered solutions sit strictly below β_∞ ($\|\mathbf{x}\|_\infty/\beta_\infty \in [0.76,0.86]$). To our knowledge these are the first practical ℓ_∞ Z-shape attacks on Dilithium-type ISIS^∞ in the small-modulus regime; both the cost-engine predictions and the runtime demonstrator confirm the regime is solvable with a cheap $\beta=18$ BKZ reduction and a single OTF sieve.

In Figure 9(a), we substitute the Euclidean cube–ball intersection of Figure 6(a) by the cube–box intersection $\pi^\infty(b;q,n_q)=((2b+1)/q)^{n_q}$ of Lemma 2, evaluated against the empirical CDF of $\|\mathbf{x}_{\mathrm{I}}\|_\infty$ for $\mathbf{x}_{\mathrm{I}} \sim \mathrm{U}(\mathrm{Cube}_{n_q}(q))$. The two curves agree closely across the entire range, which demonstrate that the closed form of Lemma 2 is exact and that no truncated theta-convolution machinery is required for the ℓ_∞ regime: at $b=\beta_\infty=14$ (the green vertical bound) the modeled and experimental proportions coincide at approximately

Table 2: Verification of the Small-q^∞ attack on three Dilithium-type $\mathrm{ISIS}^\infty_{n,m,q,\beta_\infty}$ preset instances. Every preset uses BKZ-18 and a terminal sieve of dimension 30. "Predicted $\log_2\mathbb{E}[\#\mathrm{succ}]$" is output of our closed-form cost engine; remaining columns are measured over seeded trials. "tr." is number of sieve candidates examined before a valid lift was found.

preset	n	q	β_∞	pred. $\log_2\mathbb{E}[\#\mathrm{succ}]$	#tri.	mean (s)	mean tr.	mean $\|\mathbf{x}\|_\infty$
fast	56	23	10	+4.23	5	1.04	9	8.6
medium	64	31	14	+3.76	3	1.56	12	10.7
small	48	13	5	+3.65	3	1.30	$11^\dagger$	$4.2^\dagger$



(a) CDF of partial-lift ℓ_∞ norm.

(b) Counts of full-lift ℓ_∞ norm.

Figure 9: Empirical validation of the Small-q^∞ cost model at the medium preset $(n,m,q,\beta_S,n_q,\beta_\infty)=(64,128,31,30,6,14)$. (a) Proportion of partial Zone I lifts with $\|\mathbf{x}_\mathrm{I}\|_\infty \le b$; experimental (blue) over $N=800$ uniform-mod-q residues, modeled (red dashed) by the closed form $\pi^\infty(b;q,n_q)=((2b+1)/q)^{n_q}$ of Lemma 2. (b) Number of full lifts with $\|\mathbf{x}\|_\infty \le b$; experimental (blue) and modeled (red dashed) by $\sum_i s_\infty(b;L_{\mathrm{proj},i})\cdot\pi^\infty(b;q,n_q)$ over the same N sieve outputs. The vertical green dotted line marks the acceptance bound β_∞, the only operationally relevant ordinate/abscissa for forgery.

0.65, the precise value used by the cost engine to predict per-lift success. In Figure 9(b), we then validate the joint count of admissible full lifts under the box constraint, where the modeled count combines $\pi^\infty(b;q,n_q)$ with the Zone II/III box-survival fraction $s_\infty(b;L_{\mathrm{proj},i})$ across the BDGL-OTF sieve outputs. The two curves agree to within sampling noise on the operationally relevant region $b\le\beta_\infty$; for $b>\beta_\infty$ the experimental number rises slightly above the modeled number, which reflects a mild conservatism in the Gaussian-coordinate isotropy heuristic underlying s_∞. Since $b\le\beta_\infty$ is the only regime in which a recovered solution constitutes a valid forgery, this conservatism preserves the soundness of every cost prediction reported in Table 2.

The four panels of Figure 6 and Figure 9 demonstrate that the two cost engines proposed in Section 4 and Section 5 predict the joint statistics of the attack, and not merely point-mass averages or independent marginals as in [DEP23]. In Figure 6 the agreement between the integrated count E_k and the experimental count substantiates Lemma 1: the integrated quantity $\int p^{\mathrm{sharp}}_{\mathrm{ISIS}}(L)f_L(L)\mathrm{d}L$ that drives the 3.51-bit reduction reported in Table 1 is supported by direct measurement of the full lift-length distribution. In Figure 9 the agreement between the closed form π^∞ and the empirical ℓ_∞ proportion substantiates Lemma 2: the elimination of the truncated theta-convolution engine is shown to incur no statistical penalty at the parameter regime in which Small-q^∞ operates. Both figures use $N=800$ Monte-Carlo samples drawn from the same probabilistic model (uniform-mod-q residues for partial lifts, BDGL-OTF spherical-shell density for sieve outputs) on which $\mathsf{SharpSmallq}_{\mathrm{cost}}$ and $\mathsf{Smallq}^\infty_{\mathrm{cost}}$ are themselves built, which provides internally consistent verification of probabilistic components that are invoked by our two attacks.

6 Future Directions

Here, we bring up technical open problems that deserve further investigations, each refers either to an extension of the Z-shape mechanism to a setting it does not yet cover, or to a structural strengthening of the cost engines of Section 4 and Section 5.

Babai-lifting extension to Hawk. The Hawk signature scheme [DPPvW22] reduces unforgeability to module-LIP rather than to ISIS, and an attack on Hawk requires recovering a short basis of the secret module rather than a single short syndrome solution. When the public-key Gaussian width σ_{pk} is small, Babai-style nearest-plane lifting [Bab86] over the Zone I plateau may lower the per-lift acceptance probability enough to bring the Z-shape attack into range; quantifying this threshold and combining it with the sharp cost engine of Section 4 is open. The work of [DEP23] is also marked this question as open.

Decoupled blocksize search beyond the diagonal. Our experiments show that the joint (β_R,β_S) optimum of $\mathsf{SharpSmallq}_{\mathrm{cost}}$ lies on the diagonal $\beta_R=\beta_S$ for Falcon [FHK+20] and Mitaka [EFG+22] parameters, and we therefore collapsed the search to one dimension. Whether this remains true on rebalanced parameter sets such as Falcon-1024 or on rank-2 module variants in the sense of [LPMSW19, MSD20] is open; a counter-example would expose an additional bit-margin our attack does not currently exploit.

Practical knee of β_{sieve} for ISIS-ℓ_2. On Mitaka-512 we identified the empirical knee at $\beta_{\mathrm{sieve}}=58$, two units below the safety choice of [DEP23]. Predicting this knee analytically as a function of (n,q,ν) rather than locating it by sweep would let us compute the BKZ-vs-sieve trade-off in closed form and would integrate naturally with progressive-sieving domain in the sense of G6K [ADH+19] and recent dual-sieve refinements [DP23, GJ21, CMHST25].

Lattice-aware refinement of the Gaussian-coordinate isotropy heuristic. The Zone II/III tail-survival function $s_\infty(L)$ that drives the cost prediction of $\mathsf{Smallq}^{\infty}_{\mathrm{cost}}$ assumes the coordinates of a sieve output are jointly $\mathcal{N}(0,L^2/\beta_S)$. Our Monte-Carlo validation indicates this is mildly conservative; replacing it with a measured-from-BKZ-profile model [AGVW17, PV21] or with the worst-case sphere-on-cube bounds of [CCSL24] would tighten Table 2 predictions.

References

[AAB+24] Marius A. Aardal, Diego F. Aranha, Katharina Boudgoust, Sebastian Kolby, and Akira Takahashi. Aggregating falcon signatures with labrador. In *Advances in Cryptology – CRYPTO 2024: 44th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2024, Proceedings, Part I*, page 71–106, Berlin, Heidelberg, 2024. Springer-Verlag. doi:10.1007/978-3-031-68376-3_3.

[ABC+23] Melissa Azouaoui, Olivier Bronchain, Gaëtan Cassiers, Clément Hoffmann, Yulia Kuzovkova, Joost Renes, Tobias Schneider, Markus Schönauer, François-Xavier Standaert, and Christine Van Vredendaal. Protecting dilithium against leakage: Revisited sensitivity analysis and improved implementations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, page 58–79, August 2023. URL: http://dx.doi.org/10.46586/tches.v2023.i4.58-79, doi:10.46586/tches.v2023.i4.58-79.

[ACC+21] Martin Albrecht, Melissa Chase, Hao Chen, Jintai Ding, Shafi Goldwasser, Sergey Gorbunov, Shai Halevi, Jeffrey Hoffstein, Kim Laine, Kristin Lauter, Satya Lokam, Daniele Micciancio, Dustin Moody, Travis Morrison, Amit Sahai, and Vinod Vaikuntanathan. *Homomorphic Encryption Standard*, page 31–62. Springer International Publishing, 2021. URL: http://dx.doi.org/10.1007/978-3-030-77287-1_2, doi:10.1007/978-3-030-77287-1_2.

[AD21] Martin R. Albrecht and Léo Ducas. *Lattice Attacks on NTRU and LWE: A History of Refinements*, pages 15–40. London Mathematical Society Lecture Note Series. Cambridge University Press, 2021.

[ADH+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*, page 717–746. Springer International Publishing, 2019. URL: http://dx.doi.org/10.1007/978-3-030-17656-3_25, doi:10.1007/978-3-030-17656-3_25.

[AGPS20] Martin R. Albrecht, Vlad Gheorghiu, Eamonn W. Postlethwaite, and John M. Schanck. Estimating quantum speedups for lattice sieves. In *ASIACRYPT 2020*, volume 12492 of *LNCS*, pages 583–613. Springer, 2020.

[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*, volume 10624 of *LNCS*, pages 297–322. Springer, 2017.

[Ajt96] M. Ajtai. Generating hard instances of lattice problems (extended abstract). In *ACM STOC*, pages 99–108, 1996.

[APMvW25] Bill Allombert, Alice Pellet-Mary, and Wessel van Woerden. *Cryptanalysis of Rank-2 Module-LIP: A Single Real Embedding Is All It Takes*, page 184–212. Springer Nature Switzerland, 2025. URL: http://dx.doi.org/10.1007/978-3-031-91124-8_7, doi:10.1007/978-3-031-91124-8_7.

[Bab86] László Babai. On Lovász' lattice reduction and the nearest lattice point problem. *Combinatorica*, 6(1):1–13, 1986. doi:10.1007/BF02579403.

[BBD+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. In *Advances in Cryptology – CRYPTO 2023: 43rd Annual International Cryptology Conference, CRYPTO 2023, Santa Barbara, CA, USA, August 20–24, 2023, Proceedings, Part V*, page 358–389, Berlin, Heidelberg, 2023. Springer-Verlag. doi:10.1007/978-3-031-38554-4_12.

[BCD+16] Joppe W. Bos, Craig Costello, Léo Ducas, Ilya Mironov, Michael Naehrig, Valeria Nikolaenko, Ananth Raghunathan, and Douglas Stebila. Frodo: Take off the ring! practical, quantum-secure key exchange from LWE. In *ACM CCS 2016*, pages 1006–1018. ACM, 2016.

[BDGL16] Anja Becker, Léo Ducas, Nicolas Gama, and Thijs Laarhoven. New directions in nearest neighbor searching with applications to lattice sieving. In Robert Krauthgamer, editor, *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2016)*, pages 10–24. ACM-SIAM, January 2016. doi:10.1137/1.9781611974331.ch2.

[CAL21] Rutchathon Chairattana-Apirom and Anna Lysyanskaya. Compact threshold signatures from Lattice-based zero-knowledge. In *ACM CCS 2021*, 2021.

[CCSL24] Lynn Chua, Hao Chen, Yongsoo Song, and Kristin Lauter. On the concrete security of LWE with small secret. 2024. Springer, open access.

[CME+25] Clémence Chevignard, Guilhem Mureau, Thomas Espitau, Alice Pellet-Mary, Heorhii Pliatsok, and Alexandre Wallet. *A Reduction from Hawk to the Principal Ideal Problem in a Quaternion Algebra*, page 154–183. Springer Nature Switzerland, 2025. URL: http://dx.doi.org/10.1007/978-3-031-91124-8_6, doi:10.1007/978-3-031-91124-8_6.

[CMHST25] Kévin Carrier, Charles Meyer-Hilfiger, Yixin Shen, and Jean-Pierre Tillich. Assessing the impact of a variant of MATZOV's dual attack on Kyber. In *CRYPTO 2025*, volume 16000 of *LNCS*. Springer, 2025.

[CS99] John H. Conway and Neil J. A. Sloane. *Sphere Packings, Lattices and Groups*, volume 290 of *Grundlehren der mathematischen Wissenschaften*. Springer, New York, 3rd edition, 1999. doi:10.1007/978-1-4757-6568-7.

[DEP23] Léo Ducas, Thomas Espitau, and Eamonn W. Postlethwaite. *Finding Short Integer Solutions When the Modulus Is Small*, page 150–176. Springer Nature Switzerland, 2023. URL: http://dx.doi.org/10.1007/978-3-031-38548-3_6, doi:10.1007/978-3-031-38548-3_6.

[DP23] Léo Ducas and Ludo N. Pulles. Does the dual-sieve attack on learning with errors even work? In *CRYPTO 2023*, volume 14083 of *LNCS*, pages 37–69. Springer, 2023.

[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 *ASIACRYPT 2022*, volume 13794 of *LNCS*, pages 65–94. Springer, 2022.

[DSDGR20] Dana Dachman-Soled, Léo Ducas, Huijing Gong, and Mike Rossi. LWE with side information: Attacks and concrete security estimation. In *CRYPTO 2020*, LNCS, 2020.

[EFG+22] Thomas Espitau, Pierre-Alain Fouque, François Gérard, Mehdi Tibouchi, and Alexandre Wallet. Mitaka: A simpler, parallelizable, maskable variant of falcon. In *EUROCRYPT 2022*, volume 13277 of *LNCS*, pages 222–253. Springer, 2022.

[EK20] Thomas Espitau and Paul Kirchner. The nearest-colattice algorithm: Time-approximation tradeoff for Approx-CVP. In *ANTS XIV*, 2020.

[ETWY22] Thomas Espitau, Mehdi Tibouchi, Alexandre Wallet, and Yang Yu. *Shorter Hash-and-Sign Lattice-Based Signatures*, page 245–275. Springer Nature Switzerland, 2022. URL: http://dx.doi.org/10.1007/978-3-031-15979-4_9, doi:10.1007/978-3-031-15979-4_9.

[FH23] Serge Fehr and Yu-Hsuan Huang. *On the Quantum Security of HAWK*, page 405–416. Springer Nature Switzerland, 2023. URL: http://dx.doi.org/10.1007/978-3-031-40003-2_15, doi:10.1007/978-3-031-40003-2_15.

[FHK+20] 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. In *NIST PQC Round 3 Submission*, 2020.

[FJR22] Thibauld Feneuil, Antoine Joux, and Matthieu Rivain. Syndrome decoding in the head: Shorter signatures from zero-knowledge proofs. In *CRYPTO 2022*, LNCS. Springer, 2022.

[GJ21] Qian Guo and Thomas Johansson. Faster dual lattice attacks for solving LWE with applications to CRYSTALS. In *ASIACRYPT 2021*, volume 13093 of *LNCS*, pages 33–62. Springer, 2021.

[GJM+24] Sanjam Garg, Abhishek Jain, Pratyay Mukherjee, Rohit Sinha, Mingyuan Wang, and Yinuo Zhang. hints: Threshold signatures with silent setup. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3034–3052, 2024. doi:10.1109/SP54263.2024.00057.

[GN08] Nicolas Gama and Phong Q. Nguyen. Predicting lattice reduction. In *EUROCRYPT 2008*, volume 4965 of *LNCS*, pages 31–51. Springer, 2008.

[HG07] Nick Howgrave-Graham. A hybrid lattice-reduction and meet-in-the-middle attack against NTRU. In Alfred Menezes, editor, *Advances in Cryptology – CRYPTO 2007*, volume 4622 of *Lecture Notes in Computer Science*, pages 150–169. Springer, Berlin, Heidelberg, 2007. doi:10.1007/978-3-540-74143-5_9.

[HPT25] Calvin Abou Haidar, Quentin Payet, and Mehdi Tibouchi. *Crowhammer: Full Key Recovery Attack on Falcon with a Single Rowhammer Bit Flip*, page 103–135. Springer Nature Switzerland, 2025. URL: http://dx.doi.org/10.1007/978-3-032-01901-1_4, doi:10.1007/978-3-032-01901-1_4.

[KMSS25] Aniket Kate, Pratyay Mukherjee, Samipa Samanta, and Pratik Sarkar. Dyna-hints: Silent threshold signatures for dynamic committees. In *2025 Crypto Valley Conference (CVC)*, pages 106–126, 2025. doi:10.1109/CVC65719.2025.00019.

[LDK+22] Jacob Lichtinger, Quynh Dang, John Kelsey, Yi-Kai Liu, Daniel Smith-Tone, Angela Robinson, Daniel Apon, Gorjan Alagic, David Cooper, Thinh Dang, Carl Miller, Dustin Moody, Rene Peralta, and Ray Perlner. Status report on the third round of the NIST post-quantum cryptography standardization process. Technical Report NISTIR 8413-upd1, National Institute of Standards and Technology, September 2022. doi:10.6028/NIST.IR.8413-upd1.

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

[LPMSW19] Changmin Lee, Alice Pellet-Mary, Damien Stehlé, and Alexandre Wallet. An LLL algorithm for module lattices. In *ASIACRYPT 2019*, volume 11922 of *LNCS*, pages 59–90. Springer, 2019.

[LS15] A. Langlois and D. Stehle. Worst-case to average-case reductions for module lattices. *Des. Codes Cryptogr.*, 75(3):565–599, 2015.

[LZS+24] Wai-Kong Lee, Raymond K. Zhao, Ron Steinfeld, Amin Sakzad, and Seong Oun Hwang. High throughput lattice-based signatures on gpus: Comparing falcon and mitaka. *IEEE Transactions on Parallel and Distributed Systems*, 35(4):675–692, 2024. doi:10.1109/TPDS.2024.3367319.

[MP12] Daniele Micciancio and Chris Peikert. Trapdoors for lattices: Simpler, tighter, faster, smaller. In *Advances in Cryptology – EUROCRYPT*, pages 700–718, 2012.

[MR09] Daniele Micciancio and Oded Regev. Lattice-based cryptography. In Daniel J. Bernstein, Johannes Buchmann, and Erik Dahmen, editors, *Post-Quantum Cryptography*, pages 147–191. Springer, Berlin, Heidelberg, 2009. doi:10.1007/978-3-540-88702-7_5.

[MSD20] Tamalika Mukherjee and Noah Stephens-Davidowitz. Lattice reduction for modules, or how to reduce ModuleSVP to ModuleSVP. In *CRYPTO 2020*, volume 12171 of *LNCS*, pages 213–242. Springer, 2020.

[Mur25] Guilhem Mureau. *Special Genera of Hermitian Lattices and Applications to HAWK*, page 289–317. Springer Nature Switzerland, December 2025. URL: http://dx.doi.org/10.1007/978-3-032-12290-2_10, doi:10.1007/978-3-032-12290-2_10.

[Nat22] National Institute of Standards and Technology. Request for additional digital signature schemes for the post-quantum cryptography standardization process. NIST Post-Quantum Cryptography Project, September 2022. Published September 6, 2022. URL: https://csrc.nist.gov/news/2022/request-additional-pqc-digital-signature-schemes.

[NV08] Phong Q. Nguyen and Thomas Vidick. Sieve algorithms for the shortest vector problem are practical. In *Journal of Mathematical Cryptology*, 2008.

[Pre23] Thomas Prest. *A Key-Recovery Attack Against Mitaka in the t-Probing Model*, page 205–220. Springer Nature Switzerland, 2023. URL: http://dx.doi.org/10.1007/978-3-031-31368-4_8, doi:10.1007/978-3-031-31368-4_8.

[PS24] Amaury Pouly and Yixin Shen. Provable dual attacks on learning with errors. In *EUROCRYPT 2024*, volume 14657 of *LNCS*, pages 256–285. Springer, 2024.

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

[Sch03] Claus-Peter Schnorr. Lattice reduction by random sampling and birthday methods. In *STACS 2003*, pages 145–156, 2003.

[SE94a] Claus-Peter Schnorr and M. Euchner. Lattice basis reduction: improved practical algorithms and solving subset sum problems. In *Mathematical Programming*, 1994.

[SE94b] Claus-Peter Schnorr and Martin Euchner. Lattice basis reduction: Improved practical algorithms and solving subset sum problems. *Mathematical Programming*, 66(1–3):181–199, 1994. doi:10.1007/BF01581144.

[SY25] Kiarash Sedghighadikolaei and Attila Altay Yavuz. A survey of threshold signatures: Nist standards, post-quantum cryptography, exotic techniques, and real-world applications. *ACM Comput. Surv.*, October 2025.

[Tao16] Terence Tao. *Analysis I*, volume 37 of *Texts and Readings in Mathematics*. Springer, Singapore / Hindustan Book Agency, 3rd edition, 2016. doi:10.1007/978-981-10-1789-6.

[YZZ+24] Ouyang Yi, Yihong Zhu, Wenping Zhu, Bohan Yang, Zirui Zhang, Hanning Wang, Qichao Tao, Min Zhu, Shaojun Wei, and Leibo Liu. Falconsign: An efficient and high-throughput hardware architecture for falcon signature generation. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2025:203–226, 2024. URL: https://api.semanticscholar.org/CorpusID:274649233.