Learning With Quantization: A Ciphertext Efficient Lattice Problem with Tight Security Reduction from LWE

Shanxiang Lyu[1], Ling Liu[2], and Cong Ling[3]

[1] Jinan University, Guangzhou, China
lsx07@jnu.edu.cn
[2] Xidian University, Xi'an, China
liuling@xidian.edu.cn
[3] Imperial College London, London, UK
c.ling@imperial.ac.uk

Abstract. In this paper, we introduce Learning With Quantization (LWQ), a new problem related to the Learning With Errors (LWE) and Learning With Rounding (LWR) problems. LWQ provides a tight security reduction from LWE while enabling efficient ciphertext compression comparable to that of LWR. We adopt polar lattices to instantiate the quantizer of LWQ. Polar lattices are a specific instance of the classical Construction D, which utilizes a set of nested polar codes as component codes. Due to the polarization phenomenon of polar codes, the distribution of the quantization error converges to a discrete Gaussian. Moreover, the quantization algorithm can be executed in polynomial time. Our main result establishes a security reduction from LWE to LWQ, ensuring that the LWQ distribution remains computationally indistinguishable from the uniform distribution. The technical novelty lies in bypassing the noise merging principle often seen in the security reduction of LWR, instead employing a more efficient noise matching principle. We show that the compression rate is ultimately determined by the capacity of the "LWE channel," which can be adjusted flexibly. Additionally, we propose a full information-rate encryption framework based on LWQ, demonstrating its advantage over constructions based on LWE and quantized LWE. Our result answers affirmatively a question left open by Micciancio and Schultz (CRYPTO 2023).

Keywords: Lattice-Based Cryptography · Learning With Quantization · Polar Lattice · Ciphertext Compression · Source Coding.

1 Introduction

Regev's Learning with Errors (LWE) problem [43] is fundamental to modern cryptography, offering both versatility and robust security guarantees. The LWE assumption states that the decision LWE problem is hard to solve: With proper parameters $n, m, q \in \mathbb{N}$ and a small error distribution χ_e over $\mathbb{Z}^m$, for uniformly

random matrices $\mathbf{A} \leftarrow \mathbb{Z}_q^{m \times n}$, vectors $\mathbf{s} \leftarrow \mathbb{Z}_q^n$, $\mathbf{u} \leftarrow \mathbb{Z}_q^m$, and an error vector $\mathbf{e} \leftarrow \chi_e$, the pair $(\mathbf{A}, \mathbf{As}+\mathbf{e})$ is computationally indistinguishable from $(\mathbf{A}, \mathbf{u})$. It is known that when the modulus q is sufficiently large compared to n, certain error distributions make solving LWE as hard as tackling worst-case computational problems on lattices [43, 40, 13]. These problems are conjectured to remain difficult even for quantum computers. Beyond its strong hardness guarantees, LWE has proven extremely useful in cryptographic applications. Since its introduction, a significant amount of research has focused on LWE-based constructions for a wide array of known cryptographic primitives (e.g., [23, 33, 24, 14, 42], among many others).

However, the inherent randomness in the LWE problem—specifically, the randomness involved in generating the error vector $\mathbf{e}$—prevents straightforward constructions of certain cryptographic primitives that require determinism. To address the issue, the Learning With Rounding (LWR) problem was introduced by Banerjee, Peikert, and Rosen [8] as a derandomized version of the LWE problem. Instead of adding an error vector $\mathbf{e}$ to $\mathbf{As}$ to hide its exact values, LWR releases a deterministically rounded version of $\mathbf{As}$. In particular, for some $p < q$, an element-wise rounding function $\lfloor \cdot \rceil_p : \mathbb{Z}_q^m \to \mathbb{Z}_p^m$ is applied. The LWR assumption is expressed as follows: $(\mathbf{A}, \lfloor \mathbf{As} \rceil_p)$ is computationally indistinguishable from $(\mathbf{A}, \lfloor \mathbf{u} \rceil_p)$. We can also write $\lfloor \mathbf{As} \rceil_p = \mathbf{As} + \mathbf{e}_Q$ with the rounding error $\mathbf{e}_Q$, but the storage size for the term $\lfloor \mathbf{As} \rceil_p$ is smaller than that of LWE. The applications of LWR span various areas, including pseudorandom functions [8], reusable randomness extractors [2], and public key encryption schemes such as Saber [20] and Lizard [15].

The hardness of LWR is mostly established through a reduction from the quantized LWE problem. The reduction of Banerjee, Peikert, and Rosen requires the the modulus q has to be super-polynomial, which makes all of the computations less efficient. Moreover, the ratio of the input-to-output modulus q/p is super-polynomial, meaning that we must throw away a lot of information when rounding and therefore get fewer bits of output per LWR sample. In practical applications, it is advantageous to use a smaller modulus q to enable more efficient implementations. However, establishing the hardness of LWR with polynomial modulus q is a significant open question as noted in [8].

Subsequent research [2, 10, 37] has further examined this area. The size of q was reduced to a polynomial by assuming it is a prime in [2, 37], but their results do not address cases where q is a power of two, where the rounding function becomes particularly straightforward. Restricting on the number of query samples, [10] also showed that q can be polynomial. From the terminology of this paper, a principle called *noise merging* is frequently involved in the hardness proof of LWR. For instance, the security reduction of [8] is

$$(\mathbf{A}, \lfloor \mathbf{As} + \mathbf{e} \rceil_p) \approx_c (\mathbf{A}, \lfloor \mathbf{u} \rceil_p) \to (\mathbf{A}, \lfloor \mathbf{As} \rceil_p) \approx_c (\mathbf{A}, \lfloor \mathbf{u} \rceil_p). \tag{1}$$

Its noise merging principle is that a large uniform noise $\mathbf{e}_Q$ can merge a small noise $\mathbf{e}$ to itself:

$$\mathbf{e}_Q + \mathbf{e} \approx_s \mathbf{e}_Q. \tag{2}$$

Additionally, noise merging is utilized in the lossy code-based security reduction in [2] and is applied using the Rényi divergence metric in [10].

Informally, we can think of the width of a Gaussian $\mathbf{e}$ as σ, and the element-wise width of $\mathbf{e}_Q$ as $\sigma_Q = q/(2p)$. We have to choose $\sigma_Q \gg \sigma$ to enable the noise merging technique. This situation is not ideal as a smaller modulo-to-noise ratio q/σ implies more secure LWE [43]. Moreover, the noise merging technique does not ensure that a larger compression ratio for quantized LWE samples (and therefore a large σ_Q) corresponds to a more difficult LWE problem; it only says that a large σ_Q makes LWR as hard as LWE of noise width σ. The above analysis leads us to the following question: *can we design a variant with tighter security reduction from LWE?*

1.1 Our Contribution

Our primary contribution is the introduction of a variant termed Learning With Quantization (LWQ), along with a reduction from LWE. This approach utilizes a lattice to quantize the vector $\mathbf{As}$ in a randomized and vector-wise manner, resulting in an observation term represented as $\mathbf{As} + \mathbf{e}_Q$, where $\mathbf{e}_Q$ denotes the error introduced by quantization. This method offers several advantages: first, it eliminates the error vector $\mathbf{e}$ of LWE and reduces the size of $\mathbf{As}$ similar to LWR; second, it achieves greater quantization efficiency compared to LWR due to the more flexible choice of Λ, where LWR is only a degenerate case of LWQ with deterministic quantization and $\Lambda = \frac{q}{p}\mathbb{Z}^m$; and third, it provides a tight security reduction from LWE, where a higher degree of quantization corresponding to an increased level of security. The main result of this paper is the following theorem:

Theorem 1 (Informal, see Theorem 5 for formal statement). *If there exists an oracle that can distinguish the LWQ distribution* $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ *from the uniform distribution with non-negligible advantage, then it can also distinguish the LWE distribution from uniform with non-negligible advantage.*

In a sense, LWQ can be seen as an advanced method for approximating the LWE distribution while simultaneously compressing ciphertexts. The proof techniques and results are new and flexible. Specifically, we build reduction from LWE to LWQ, rather than from quantized LWE.

An observant reader may realize that it is impossible to prove indistinguishability between *naive* LWQ and LWE, as the support set is different: the quantization $\in \mathbb{Z}_q^m \cap \Lambda$, while $\mathbf{As} + \mathbf{e} \in \mathbb{Z}_q^m$. To do so, we resort to an adapted form of dithering, which is the process of adding a small amount of artificial noise to the data/signal to align the support set of quantization errors. In general, dithering leads to $Q_\Lambda(\mathbf{As} - \mathbf{d})$ using a uniform vector $\mathbf{d}$ which is public. Our adapted form is to transmit $Q_\Lambda(\mathbf{As} - \mathbf{d}) + \mathbf{d} = \mathbf{As} + \mathbf{e}_Q \in \mathbb{Z}_q^m$, where the quantization error $\mathbf{e}_Q$ becomes independent of the input and is uniformly distributed over the Voronoi region of the quantization lattice. Then we can focus on proving

$$(\mathbf{A}, \mathbf{As} + \mathbf{e}_Q) \approx_s (\mathbf{A}, \mathbf{As} + \mathbf{e}). \tag{3}$$

The consequence of this technique is that the parameters can be chosen based on *noise matching*: $\mathbf{e}_Q \approx_s \mathbf{e}$, rather than noise merging. This allows for more flexible parameter choices for LWQ, including polynomial and power-of-2 moduli, among others. Up to this point, $(\mathbf{A}, \mathbf{As}+\mathbf{e}_Q)$ appears to be merely an alternative implementation of LWE. We further produce the dither from a random oracle, thus compressing the ciphertext while still maintaining computational security.

Our second contribution, as detailed in Section 4, is the introduction of polar-lattice-aided quantization to prove $\mathbf{e}_Q \approx_s \mathbf{e}$. We can consider $\mathbf{e}_Q$ as being uniformly distributed over the Voronoi region of Λ: $\mathbf{e}_Q \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$. A powerful theory of lattices states that if the normalized second moment (NSM) of a lattice $G(\Lambda)$ converges to $\frac{1}{2\pi e} \approx 0.0585$, then the Voronoi region takes the shape of a sphere, and the uniform distribution over this sphere is equivalent to a Gaussian distribution. For example, the NSMs for the integer lattice $\mathbb{Z}$, checker-board lattice D_4, and Gosset lattice E_8 are: $0.08333, 0.07660, 0.07168$ [18]. We identify two technical hurdles in adopting this theory. First, a randomized construction of Λ becomes hard to decode as the problem dimension increases. Second, the convergence speed of $G(\Lambda) \to \frac{1}{2\pi e}$ is important. Fortunately, polar lattices have efficient polynomial-time decoding, and the distribution of its quantization error $\mathbf{e}_Q$ can be analyzed via either the statistical distance or the Kullback-Leibler divergence. The takeaway of this contribution is that, the quantization error $\mathbf{e}_Q$ of polar-lattice-aided LWQ is close to a discrete Gaussian distribution, while that of LWR is close to a uniform distribution over a hypercube. Since LWQ permits a tight security reduction based on noise matching rather than noise merging, it offers stronger hardness guarantees than LWR.

Lastly, we highlight the advantages of LWQ by illustrating its benefits via an encryption framework based on it. There is growing interest in enhancing the information rate—the size ratio of plaintext to ciphertext—in lattice-based homomorphic encryption schemes, which has led to the development of constructions achieving rates asymptotically close to 1 [12]. Recently, Micciancio and Schultz [35] introduced a quantized LWE-based encryption framework to analyze the information rate of lattice-based encryption, where the scaled $\mathbb{Z}^m$ lattice and dual of Davenport's lattice (a generalization of D_m^*) were used. In particular, their work [35, Bound 2] demonstrates that, under a heuristic assumption, if σ_Q (the width of the quantization noise) and σ (the width of the LWE noise) satisfy $\sigma_Q \le O(\sigma)$, it becomes impossible for a perfectly-correct quantized LWE-based framework to achieve an asymptotic rate of $1 - o\left(\frac{1}{\log_2 q}\right)$. This scenario can be interpreted as the failure of noise merging, where $\mathbf{e}_Q + \mathbf{e} \approx_s \mathbf{e}_Q$. LWQ offers a straightforward solution to overcome this limitation by excluding the error term $\mathbf{e}$, enabling it to achieve a rate of 1 with polynomial modulus. To the best of our knowledge, the proposed LWQ-based encryption scheme is the first to achieve a full information rate.

To summarize, LWQ serves as a "one code for all" solution: it unifies noise addition, compression, and error correction into a single step, where the noise naturally arises from quantization. In contrast, the standard approach typically requires separate codes for error correction and compression [35]. In LWQ, a

single lattice code not only achieves capacity for the "LWE channel," but also provides optimal ciphertext compression under the given noise constraint.

Remark 1 (Role of the public dither). While $\mathbf{d}$ is indeed public, this does not compromise the security reduction. We can define a slight variation of the LWQ distribution $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ in the following triplet form [4]:

$$\mathsf{LWQ}'_{\Lambda,\mathbf{d}} = (\mathbf{A}, \mathbf{b} = Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}), \mathbf{d}).$$

Here, the effective noise can be written as $\mathbf{e}' = Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}) - \mathbf{A}\mathbf{s}$, which is a function of $\mathbf{d}$ and hence depends on it. However, this dependency does not affect the reduction from LWQ to LWE as we argue below.

The security argument proceeds in three steps:

1. The two distributions $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ and $\mathsf{LWQ}'_{\Lambda,\mathbf{d}}$ are equivalent. From the distribution $\mathsf{LWQ}'_{\Lambda,\mathbf{d}}$, we can recover the LWQ distribution by simply adding the second and third elements:

   $$\mathsf{LWQ}_{\Lambda,\mathbf{d}} = (\mathbf{A}, \mathbf{b} = Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}) + \mathbf{d}).$$

   Conversely, since $\mathbf{A}\mathbf{s} + \mathbf{e}_Q = Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}) + \mathbf{d}$, one can recover $Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d})$ by subtracting $\mathbf{d}$.
2. We show that $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ is statistically close to the standard LWE distribution $(\mathbf{A}, \mathbf{A}\mathbf{s} + \mathbf{e})$, where $\mathbf{e}$ is a discrete Gaussian. This follows from the fact that polar lattice aided dithered quantization noise $\mathbf{e}_Q$ converges to a discrete Gaussian under appropriate conditions.
3. By the hardness of the LWE assumption, distinguishing $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ from uniform is computationally hard. Therefore, distinguishing $\mathsf{LWQ}'_{\Lambda,\mathbf{d}}$ from uniform must also be hard, since the transformation from $\mathsf{LWQ}'_{\Lambda,\mathbf{d}}$ to $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ is efficiently computable.

Thus, although $\mathbf{d}$ is public and appears in the distribution, it does not invalidate the reduction. Our security reduction remains sound: any adversary capable of distinguishing $\mathsf{LWQ}'_{\Lambda,\mathbf{d}}$ samples $(\mathbf{A}, Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}), \mathbf{d})$ from uniform can be used to construct an adversary that breaks the LWE problem.

There is, however, a trade-off associated with LWQ. Even if

$$(\mathbf{A}, Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}) + \mathbf{d}) \approx_s (\mathbf{A}, \mathbf{A}\mathbf{s} + \mathbf{e}),$$

the public dither $\mathbf{d}$ cannot be reused as a source of pseudorandomness. Consequently, only the quantized component $Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d})$ can contribute to pseudorandom output. This means that LWQ inherently produces fewer pseudorandom bits than LWE, which in turn explains its ability to compress ciphertexts.

[4] Quantization for LWE with a public dither in a similar form has been used in [35, Fig. 3]. It is also stated in [35] that "security of our construction easily follows".

1.2 Technical Overview

We show that the LWQ and LWE distributions are statistically indistinguishable:

Theorem 2 (Informal, see Theorem 4 for formal statement). *There exist a sequence of efficient lattice quantizers $Q_{\Lambda+\mathbf{d}}$ with random dither $\mathbf{d}$ such that the LWQ distribution $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ is statistically indistinguishable from the LWE distribution.*

In this work, we will adopt polar lattices to instantiate LWQ, whose running time is quasilinear with binary component codes. The technical novelty is to prove the quantization noise $\mathbf{e}_Q$ of dithered quantization converges to a discrete Gaussian distribution. This is key to prove the closeness of the LWQ and LWE distributions, therefore justifying the hardness of LWQ. Readers unfamiliar with coding theory may treat polar lattices as a black box. Next we briefly explain how our method works (see Section 4 for technical details).

The central idea is to use polar lattice quantization to simulate the "LWE channel." Recall that the LWE problem involves an *additive noise channel* model, represented by $\mathbf{b} = \mathbf{A}\mathbf{s} + \mathbf{e}$, where the received signal $\mathbf{b}$ is the sum of the transmitted data $\mathbf{A}\mathbf{s}$ and a noise component $\mathbf{e}$ added during transmission. In lattice quantization-based data compression, a *test channel* serves as a hypothetical model of the quantization process, analogous to the additive noise channel, aiming to describe the statistical relationship between the input and output for a target distortion; see [19, Chapter 10].

Definition 1 (Test channel). *The statistic of the test channel for polar quantization is described by the relationship*

$$Y = X + E \mod q, \tag{4}$$

where E is an additive discrete Gaussian noise.

In Section 4, we construct a polar lattice over this test channel. Due to the polarization phenomenon, we obtain two types of bits: "frozen bits," which are nearly independent of the input, and "information bits," which can be determined from other bits. In a polar code, frozen bits are replaced with random bits, which essentially form the random dither of a polar lattice. We prove that the polar lattice approximates the test channel very well:

Theorem 3 (Informal, see Theorem 6 for formal statement). *The statistical distance between the joint distribution $\mathsf{Q}_{X^{[m]},Y^{[m]}}$ induced by the polar lattice and $\mathsf{P}_{X^{[m]},Y^{[m]}}$ induced by the above test channel is negligible.*

Remark 2. LWQ is dithered, meaning the so-called *frozen bits* in polar codes are assumed to be uniformly random. This is not merely a technical aspect of the proof but is also crucial for achieving statistical indistinguishability between the LWE and LWQ distributions.

Remark 3. For simplicity, we will assume the modulus size q is a prime power in the proofs of the above theorems. However, it is possible to remove this small restriction by using polar codes of arbitrary alphabet size [46, 44].

Compression rate vs capacity We will show in Section 3.2 that, asymptotically, the minimum compression rate R_c of LWQ is characterized by the capacity of the "LWE channel" (4). A noisier channel has smaller capacity, thus enabling more compression. Since the "LWE channel" is virtual here, one can tune the noise variance to obtain any rate $0 < R_c < \log_2 q$, offering great flexibility.

Thus, we uncover a novel connection between information theory and the LWE problem. By linking the compression rate to the capacity of the "LWE channel," we unite the concepts of information-theoretic compression and computational security, which may have far-reaching implications to lattice-based cryptography.

Running time Assuming $q = p^r$ where p is a prime and $r \in \mathbb{N}$. The running time of both encoding and decoding for a p-ary polar code of blocklength m using an $\ell \times \ell$ kernel is $O(p^\ell m \log_2 m)$ [45]. Since for a polar lattice, the number of levels $r = \log_p q$, the overall running time of a polar lattice quantizer is given by $O(p^\ell m \log_2 m \cdot \log_p q)$. In practical implementation, we often use binary polar codes where $p = \ell = 2$ such that the running time is reduced to $O(m \log_2 m \cdot \log_2 q)$. However, there is a trade-off between the running time and convergence speed $2^{-\omega(\lambda^\beta)}$ of the statistical distance where $0 < \beta < 1$ is a parameter depending on p. A large p and ℓ enables the parameter $\beta \to 1$, while for $p = \ell = 2$, we only obtain $\beta < 1/2$. To sum up, the running time of a polar lattice using binary codes is quasilinear $\tilde{O}(m)$. See Remark 7 for more details.

1.3 Related Work

Quantization in lattice-based cryptography Nowadays, lattice-based cryptography can operate as quickly as conventional public-key cryptosystems such as RSA. However, their ciphertexts are significantly larger, necessitating the use of compression algorithms to save bandwidth. A prevalent compression technique is scalar quantization, also known as modulus switching/modulus reduction. For instance, CKKS homomorphic encryption [14] employs simple modulus reduction to a smaller modulus before computation on ciphertexts at different levels.

Another variant of LWE, known as LWER, was introduced in CRYSTALS-Kyber [47]. This variant essentially combines LWE and LWR. In LWER, LWE ciphertexts are compressed using scalar quantization, resulting in two main advantages: bandwidth savings due to compression and an increased noise level resulting from quantization error.

Ciphertext compression in lattice based cryptography is closely tied to lattice-aided quantization. Unlike computationally-hard random lattices for security, here the quantization lattice should be fast-decodable. By increasing the dimension of quantization, vector quantization can be expected to outperform scalar quantization [51]. Certain performance benefits of vector quantization have been justified in the secret-key encryption framework [35], and to reduce the ciphertext rate of CRYSTALS-Kyber [31].

The inquiry into optimal lattices for quantization, aiming for the smallest average distortion, is different from sphere packing [49, 17]. The theoretical proof of optimal lattice quantizers has been limited to dimensions up to 3 (*i.e.*, $\mathbb{Z}$, A_2, A_3^*) [9], although efforts to identify good lattice quantizers have resulted in periodic updates of tables for small-dimensional lattices ≤ 24 [1]. Closely related research focuses on the pursuit of optimal quantization lattices in the information theory community. In this context, dithered quantization has been under development for decades [53], where a (pseudo-)random signal, known as a dither, is introduced to the input signal before quantization. This regulated perturbation has the potential to enhance the statistical characteristics of the quantization error. While obtaining the rate-distortion bound with random lattices seems feasible [51], decoding a high-dimensional random lattice poses challenges. For a continuous Gaussian source, an explicit construction of polar lattices to achieve the rate-distortion bound has been presented in [29], where the computational complexity of the quantizer is $O(m \log_2 m)$.

Polar codes and polar lattices The polar lattices investigated in this work originate from polar codes [3]. Polar codes represent a significant breakthrough in coding theory, as they are the first class of codes that are efficiently encodable and decodable while achieving both channel capacity and Shannon's data compression limit [3]. The effectiveness of polar codes lies in the polarization phenomenon: through Arıkan's polar transform, the information measures of synthesized sources or channels converge to either 0 or 1, simplifying the coding process. Additionally, the state-of-the-art decoding algorithm operates with $O(m \log_2 \log_2 m)$ complexity for blocklength m [50]. Due to their outstanding performance, polar codes have been widely adopted in various practical applications, including fifth-generation (5G) wireless communication networks [21]. To help readers understand polar quantizers, an overview of polar codes is provided in Appendix A.

Polar lattices are an instance of the well-known Construction D [18, p.232] which uses a set of nested polar codes as component codes. Thanks to the nice structure of Construction D, both the encoding and decoding complexity of polar lattices are quasilinear in the block length (*i.e.*, dimension of the lattice). A construction of polar lattices achieving the Shannon capacity of the Gaussian noise channel was presented in [30]. A follow-up work [29] gave a construction of polar lattices to achieve the rate-distortion bound of source coding for Gaussian sources. Note that the two types of polar lattices constructed in [30, 29] are related but not the same (*i.e.*, one for channel coding and the other for source coding). The multilevel structure of polar lattices enables not only efficient encoding and decoding algorithms, but also a layer-by-layer implementation.

2 Preliminaries

Table 1 summarizes a few important notations in this paper for easy reference. We follow the standard asymptotic notations $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot), \Theta(\cdot)$ etc. We let

λ denote the security parameter throughout the paper. A function $\text{negl} : \mathbb{N} \to \mathbb{R}^+$ is negligible if for every positive polynomial $p(\lambda)$, there exists $\lambda_0 \in \mathbb{N}$ such that $\text{negl}(\lambda) < \frac{1}{p(\lambda)}$ for all $\lambda > \lambda_0$. The notation $X \approx_s Y$ (resp. $X \approx_c Y$) means that the random variables X and Y are statistically indistinguishable (resp. computationally indistinguishable) throughout the paper.

Symbol	Definition
$\mathbf{x}$	a boldface lower case for vectors
$\mathbf{X}$	a boldface capital for matrices
$x \sim U$	(random variable) x admits a uniform distribution on $\mathcal{U}$
$x \leftarrow \chi$	(sample) x is drawn according to distribution χ
$\mathbb{Z}_q$	set $\{0, 1, ..., q-1\}$
$\mathbb{Z}_q^{n*}$	set of integer vectors $(s_1, ..., s_n) \in \mathbb{Z}_q^n$ with $\gcd(s_1, ..., s_n, q) = 1$
X_ℓ	binary representation random variable of X at level ℓ
x_ℓ^i	i-th realization of X_ℓ
$x_\ell^{i:j}$	shorthand for $(x_\ell^i, ..., x_\ell^j)$
$x_{\ell:\jmath}^i$	realization of i-th random variable from level ℓ to level $\jmath$
$[m]$	set of all integers from 1 to m
$X^{\mathcal{I}}$	subvector of $X^{[m]}$ with indices limited in $\mathcal{I} \subseteq [m]$

Table 1: Notations

2.1 Lattices and Quantization

A lattice Λ is a discrete additive subgroup of $\mathbb{R}^n$. The rank of a lattice is the dimension of the subspace of $\mathbb{R}^n$ that it spans. A lattice is called full-rank if its rank equals its dimension. A basis $\mathbf{B}$ of a full-rank lattice $\Lambda \subset \mathbb{R}^n$ is a set of linearly independent vectors $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ in $\mathbb{R}^n$ such that every vector in the lattice Λ can be written as an integer linear combination of the basis vectors. The dual of a lattice Λ in $\mathbb{R}^n$, denoted $\hat{\Lambda}$, is the lattice given by the set of all vectors $\mathbf{y} \in \mathbb{R}^n$ such that $\langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{Z}$ for all vectors $\mathbf{x} \in \Lambda$.

For $\mathbf{v} \in \mathbb{R}^n$ and $\Lambda \subset \mathbb{R}^n$, a lattice coset $\mathbf{v} + \Lambda$ is defined as:

$$\mathbf{v} + \Lambda = \{\mathbf{v} + \mathbf{w} \mid \mathbf{w} \in \Lambda\}.$$

A coset representative is a specific vector chosen from each coset to uniquely represent that coset. The notation Λ/Λ' denotes the set of all distinct cosets of Λ' in Λ. The coset representatives of Λ/Λ' can be described as a set of vectors $\mathbf{v}_i \in \Lambda$ such that:

$$\Lambda = \bigcup_i (\mathbf{v}_i + \Lambda').$$

Definition 2 (Fundamental Region). *A fundamental region of the lattice Λ is a bounded set $\mathcal{P}_\Lambda$ that satisfies the following properties:*

1. *Covering Property: The union of translates of $\mathcal{P}_\Lambda$ by lattice points covers the entire space $\mathbb{R}^n$, i.e., $\cup_{\mathbf{v}\in\Lambda}(\mathbf{v}+\mathcal{P}_\Lambda)=\mathbb{R}^n$.*
2. *Partitioning Property: For any pair of distinct lattice points $\mathbf{v}$ and $\mathbf{w}$ in Λ, if their corresponding translated fundamental regions intersect, then $\mathbf{v}$ must equal $\mathbf{w}$.*

The half-open Voronoi region $\mathcal{V}_\Lambda$ is a fundamental region which encompasses the set of points in $\mathbb{R}^n$ that are closer to the origin than to any other lattice point.

A nearest neighbor quantizer refers to a function that maps a vector $\mathbf{y}\in\mathbb{R}^n$ to the closest lattice point in Λ via the following rule:

$$Q_\Lambda(\mathbf{y})=\arg\min_{\lambda\in\Lambda}\|\mathbf{y}-\lambda\| \tag{5}$$

where ties are broken in a systematic manner (such that $\mathbf{y}-Q_\Lambda(\mathbf{y})\in\mathcal{V}_\Lambda$). The quantization can be implemented in polynomial time by choosing fast-decodable lattices for Λ, such as $\mathbb{Z}^n$, the tensor product $\mathbb{Z}^{n/8}\otimes E_8$ (based on the Gosset lattice), or polar lattices.

In lossy source coding, quantization is often combined with a dithering technique. Our approach employs a compensated dithered quantizer where the dither is added back prior to transmission rather than in the reconstruction step, slightly differing from conventional subtractive dithering (cf. [51, 35]):

Definition 3 (Compensated Dithered Quantizer). *A compensated dithered quantizer over lattice Λ samples $\mathbf{d}\leftarrow\mathcal{P}_\Lambda$ and outputs*

$$Q_\Lambda(\mathbf{y}-\mathbf{d})+\mathbf{d}. \tag{6}$$

Equivalently, this process corresponds to quantizing $\mathbf{y}$ to a coset $\Lambda+\mathbf{d}$ of Λ:

$$Q_{\Lambda+\mathbf{d}}(\mathbf{y})=\arg\min_{\lambda\in\Lambda}\|(\mathbf{y}-\mathbf{d})-\lambda\|+\mathbf{d} \tag{7}$$

$$=Q_\Lambda(\mathbf{y}-\mathbf{d})+\mathbf{d}. \tag{8}$$

The quantization error $\mathbf{e}_Q=Q_{\Lambda+\mathbf{d}}(\mathbf{y})-\mathbf{y}\sim\mathcal{U}(\mathcal{V}_\Lambda)$, i.e., $\mathbf{e}_Q$ is uniformly distributed over the Voronoi region $\mathcal{V}_\Lambda$ and independent of $\mathbf{y}$.

Definition 4 (Second moment [51]). *The second moment of a lattice is defined as the second moment per dimension of a random variable $\mathbf{u}$ which is uniformly distributed over the Voronoi region $\mathcal{V}_\Lambda$:*

$$\gamma^2(\Lambda)=\frac{1}{n}\mathbb{E}\|\mathbf{u}\|^2=\frac{1}{n}\frac{1}{\det(\Lambda)}\int_{\mathcal{V}_\Lambda}\|\mathbf{x}\|^2\,d\mathbf{x}$$

where $\mathbb{E}$ denotes expectation, and $\det(\Lambda)$ is the volume of the Voronoi region.

The averaged quantization error of the (compensated) dithered quantizer can be quantified by $\gamma^2(\Lambda)$: for any distribution of $\mathbf{y}$, with $\mathbf{d}\sim\mathcal{U}(\mathcal{P}_\Lambda)$, then

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{e}_Q\right\|^2=\gamma^2(\Lambda). \tag{9}$$

The figure of merit for a lattice quantizer is the normalized second moment (NSM), i.e., the second-moment to volume ratio, defined as

$$G(\Lambda) = \frac{\gamma^2(\Lambda)}{\det^{2/n}(\Lambda)}. \tag{10}$$

Given a fixed dimension, a lattice with a smaller NSM is considered better. The minimum possible value of $G(\Lambda)$ over all lattices in $\mathbb{R}^n$ is denoted by G_n.

Definition 5 (Quantization-good [51]). *A sequence of lattices $\Lambda^{(n)}$ with growing dimension is called good for mean squared error quantization if*

$$\lim_{n\to\infty} G(\Lambda^{(n)}) = \frac{1}{2\pi e}. \tag{11}$$

For any $r > 0$, define the Gaussian function on $\mathbb{R}^n$ with width parameter r:

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \rho_r(\mathbf{x}) = e^{-\pi \|\mathbf{x}\|^2 / r^2}.$$

Note that although we refer to r as the width of ρ_r, the actual standard deviation of ρ_r is $\frac{r}{\sqrt{2\pi}}$. A discrete Gaussian distribution is defined as follows: For any $\mathbf{c} \in \mathbb{R}^n$, $r > 0$,

$$\mathcal{D}_{\Lambda,r,\mathbf{c}}(\mathbf{x}) = \frac{\rho_r(\mathbf{x}-\mathbf{c})}{\rho_r(\Lambda-\mathbf{c})}, \quad \forall \mathbf{x} \in \Lambda \tag{12}$$

Sampling from $\mathcal{D}_{\Lambda,r,\mathbf{c}}$ yields a distribution centered at $\mathbf{c}$. We abbreviate $\mathcal{D}_{\Lambda,r,\mathbf{0}}$ as $\mathcal{D}_{\Lambda,r}$.

2.2 Statistics

To demonstrate that the distribution of the quantization errors closely resembles discrete Gaussians, we introduce the following statistical measures.

Definition 6 (Statistical Distance). *The statistical distance between two probability distributions P and Q over the same sample space $\mathcal{X}$ is defined as:*

$$\Delta(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|.$$

Definition 7 (KL Divergence). *The Kullback-Leibler (KL) divergence between two probability distributions P and Q over the same sample space $\mathcal{X}$ is defined as:*

$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log_2\frac{P(x)}{Q(x)}.$$

Definition 8 (Computational Indistinguishability). *Two probability ensembles $\mathcal{P} = \{P_n\}_{n\in\mathbb{N}}$ and $\mathcal{Q} = \{Q_n\}_{n\in\mathbb{N}}$ are* computationally indistinguishable *(denoted $\mathcal{P} \approx_c \mathcal{Q}$) if for every probabilistic polynomial-time (PPT) distinguisher $\mathcal{D}$, there exists a negligible function* $\mathrm{negl}(\cdot)$ *such that:*

$$\left| \Pr_{x\leftarrow P_n}[\mathcal{D}(x) = 1] - \Pr_{x\leftarrow Q_n}[\mathcal{D}(x) = 1] \right| = \mathrm{negl}(n).$$

Definition 9 (Statistical Indistinguishability). *Two probability ensembles* $\mathcal{P} = \{P_n\}_{n\in\mathbb{N}}$ *and* $\mathcal{Q} = \{Q_n\}_{n\in\mathbb{N}}$ *are* statistically indistinguishable *(denoted* $\mathcal{P} \approx_s \mathcal{Q}$*) if their statistical distance is negligible in* n*:*

$$\Delta(P_n, Q_n) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P_n(x) - Q_n(x)| = \mathrm{negl}(n).$$

By Pinsker's inequality $\Delta(P_n, Q_n) \le \sqrt{\frac{\ln 2}{2} D_{\mathrm{KL}}(P_n \| Q_n)}$, negligible KL divergence implies statistical indistinguishability: $D_{\mathrm{KL}}(P_n\|Q_n) = \mathrm{negl}(n) \implies \Delta(P_n, Q_n) = \mathrm{negl}(n)$.

3 Hardness Results of LWQ

3.1 Definition

In the following, we review the definitions of LWE and LWR, and present our generalization called LWQ.

Definition 10 (LWE/LWR/LWQ distributions). *Let* n, m, p, q *be positive integers with* $q > p > 1$*, and* Λ *be an* m*-dimensional integer lattice satisfying* $q^m > \det(\Lambda) > 1$*. For a "secret"* $\mathbf{s} \in \mathbb{Z}_q^n$*, and an error distribution* χ_e^m *over* $\mathbb{Z}^m$*, samples for the LWE/LWR/LWQ distributions are respectively generated by*

- $\mathsf{LWE}_{\chi_e^m}$*:* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$*,* $\mathbf{e} \leftarrow \chi_e^m$*, and output* $(\mathbf{A}, \mathbf{b} = \mathbf{A}\mathbf{s} + \mathbf{e}) \in \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$*.*
- $\mathsf{LWR}_{\frac{q}{p}\mathbb{Z}^m}$*:* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$*, and output* $(\mathbf{A}, \mathbf{b} = \lfloor \mathbf{A}\mathbf{s} \rceil_p) \in \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_p^m$*.*
- $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$*:* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$*,* $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$ *and output* $(\mathbf{A}, \mathbf{b} = Q_{\Lambda+\mathbf{d}}(\mathbf{A}\mathbf{s})) \in \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$*.*

Note that our definition of the LWQ distribution always includes a dither term $\mathbf{d}$. In essence, we quantize to a lattice coset $\Lambda + \mathbf{d}$ for a random dither $\mathbf{d}$.

Definition 11 (LWE/LWR/LWQ problems). ***Decision problem:*** *It challenges an adversary to distinguish between LWE/LWR/LWQ distributions and the respective uniform distributions.* ***Search problem:*** *Given* m *samples from the LWE/LWR/LWQ distribution, where* $\mathbf{s}$ *is sampled from some distribution* χ_s^n *(fixed for all samples), the search problem asks to recover* $\mathbf{s}$*.*

For convenience, this paper considers q such that $q\mathbb{Z}^m \subset \Lambda$. LWQ generalizes LWR in two ways: i) it employs vector quantization instead of scalar quantization, thereby the quantization error admits a distribution that more closely resembles a Gaussian, and ii) it introduces dithering, ensuring that the quantization error is independent of the input. This approach enables LWQ to benefit from a tight security reduction from LWE. When instantiated with $\Lambda = \frac{q}{p}\mathbb{Z}^m$ where p divides q, this scalar-LWQ amounts to dithered LWR, which enjoys better security guarantees than LWR.

Remark 4. The primary advantage of LWQ over LWE is the reduced size of the samples, as the dithering vector $\mathbf{d}$ is public. Given an LWQ sample $(\mathbf{A}, \mathbf{b} = Q_{\Lambda+\mathbf{d}}(\mathbf{As}) = Q_\Lambda(\mathbf{As} - \mathbf{d}) + \mathbf{d})$, the coset representative $\mathbf{d} = \mathbf{b} \bmod \Lambda$ can be efficiently identified in polynomial time [51]. Here the $\bmod \Lambda$ operation maps the input to its coset representative in a fundamental parallelepiped $\mathcal{P}_\Lambda$, which can be computed using standard techniques such as Babai's nearest plane algorithm [5] (or the nearest neighbor quantizer (5) if it runs in polynomial time). Then we may rewrite the LWQ sample as

$$(\mathbf{A}, Q_\Lambda(\mathbf{As} - \mathbf{d}), \mathbf{d}).$$

Note that a uniform distribution over $(\mathbb{Z}_q^m \cap \Lambda) \times (\mathbb{Z}^m/\Lambda)$ is the same as that over $\mathbb{Z}_q^m$. Thus, $Q_\Lambda(\mathbf{As} - \mathbf{d})$ is pseudorandom, while the matrix $\mathbf{A}$ and dither $\mathbf{d}$ can be transmitted as seeds of an extendable-output function (XOF). These seeds, which do not need to remain secret, can effectively serve as the public information in practice.

Let $\mathcal{H} : \{0,1\}^k \to \mathbb{Z}_q^{m\times n} \times (\mathbb{Z}^m/\Lambda)$ be an XOF, and $(\mathbf{A}, \mathbf{d}) = \mathcal{H}(seed)$ (with proper domain separation). Then storing an LWQ sample in the form of $(\mathbf{A}, Q_\Lambda(\mathbf{As}-\mathbf{d}), \mathbf{d})$ requires $k+\log_2\left(\frac{q^m}{\det(\Lambda)}\right)$ bits. On the contrary, LWE requires $k + \log_2(q^m)$ bits.

From the results of Regev [43] and Peikert [40], for any $m = n^{O(1)}$, any modulus $q \le 2^{n^{O(1)}}$, and for a (discrete) Gaussian distribution χ_e with parameter $\sigma \ge 2\sqrt{n}$, solving decision LWE is at least as hard as solving GapSVP_γ and SIVP_γ on arbitrary n-dimensional lattices, where $\gamma = \tilde{O}\left(\frac{nq}{\sigma}\right)$. Moreover, for moduli q of a certain form, the (average-case decision) LWE problem is equivalent to the (worst-case search) LWE problem, up to a $\mathrm{poly}(n)$ factor in the number of samples used. Although the primary hardness justification of LWE [43] is based on continuous Gaussian errors, it also holds when the error distribution is a discrete Gaussian, $\chi_e^m = \mathcal{D}_{\mathbb{Z}^m,\sigma}$. This reduction can be proved by applying a randomized rounding algorithm to the $\mathbf{b}$ samples of $\mathsf{LWE}_{\chi_e^m=\rho_\sigma/\sigma^m}$ (cf. [41, Theorem 3.1]). Unless otherwise specified, we define the hardness of the LWE assumption as the computational indistinguishability between LWE samples and uniform random samples:

$$\mathsf{LWE}_{\chi_e^m=\mathcal{D}_{\mathbb{Z}^m,\sigma}} \approx_c \mathcal{U}(\mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m). \tag{13}$$

3.2 Hardness of LWQ with polar lattices

We prove the asymptotic hardness of LWQ by showing the distributions of LWQ and LWE are statistically indistinguishable, for carefully designed polar lattice quantizers. The polar lattice presented in Section 4 is inherently dithered (cf. Section 1.2), and the quantizer can be described by $Q_{\Lambda+\mathbf{d}}$ for a random dither $\mathbf{d}$. Nevertheless, we will show later in this subsection that dithering can be generated by an extendable-output function as far as the computational indistinguishability of LWQ is concerned.

We will establish the following bound on the statistical distance between the LWQ and LWE distributions. The proof is essentially a translation of Theorem 6 in Section 4.2 from the language of coding theory into that of cryptography. We assume $\mathbf{s} \in \mathbb{Z}_q^{n*}$ such that $\mathbf{As}$ admits a uniform distribution on $\mathbb{Z}_q^m$. This is a minor condition as the probability $\mathbf{s} \in \mathbb{Z}_q^{n*}$ is at least $1 - O(1/2^n)$ for $\mathbf{s} \in \mathbb{Z}_q^n$.

We rewrite the following distributions given earlier to serve our purpose.

- Consider the LWE distribution $\mathsf{LWE}_{\chi_e^m = \mathcal{D}_{\mathbb{Z}^m,\sigma}}$: $\mathsf{P}_{\mathbf{A},\mathbf{b}}$ where $\mathbf{b} = X^{[m]} = Y^{[m]} + \mathbf{e} \mod q\mathbb{Z}$ where $Y^{[m]} = \mathbf{As}$ and $e_i \sim \mathcal{D}_{\mathbb{Z},\sigma}$.
- Consider the LWQ distribution $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$: $\mathsf{Q}_{\mathbf{A},\mathbf{b}}$ where $\mathbf{b} = X^{[m]} = Q_{\Lambda+\mathbf{d}}(Y^{[m]})$ where $Y^{[m]} = \mathbf{As}$ and $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$.

Theorem 4 (LWQ $\approx_s$ LWE). *Let $m = m(\lambda)$, $n = n(\lambda)$, $q = p(\lambda)^{r(\lambda)}$ where λ is the security parameter, $p(\lambda)$ is a prime number and $r(\lambda) \in \mathbb{N}$. Let $\mathbf{s} \in \mathbb{Z}_q^{n*}$. There exist a sequence of efficient lattice quantizers $Q_{\Lambda+\mathbf{d}}$, indexed by dimension m, such that the statistical distance between the LWE distribution $\mathsf{P}_{\mathbf{A},\mathbf{b}}$ and the LWQ distribution $\mathsf{Q}_{\mathbf{A},\mathbf{b}}$ satisfy:*

$$\Delta\left(\mathsf{P}_{\mathbf{A},\mathbf{b}}, \mathsf{Q}_{\mathbf{A},\mathbf{b}}\right) = 2^{-\omega(\lambda^\beta)}, \quad \forall \;\; 0 < \beta < 1. \tag{14}$$

Proof. Given the secret $\mathbf{s}$, the LWE distribution satisfies

$$\mathsf{P}_{\mathbf{A},\mathbf{b}} = \sum_{\mathbf{As}} \mathsf{P}_{\mathbf{A},\mathbf{As},\mathbf{b}} = \sum_{\mathbf{As}} \mathsf{P}_{\mathbf{A}} \cdot \mathsf{P}_{\mathbf{As}|\mathbf{A}} \cdot \mathsf{P}_{\mathbf{b}|\mathbf{As}}, \tag{15}$$

which is due to the Markov chain[5] $\mathbf{A} \to \mathbf{As} \to \mathbf{b}$. Notice that for given $\mathbf{s}$ and samples $Y^{[m]}$, $\mathsf{P}_{\mathbf{As}|\mathbf{A}}$ is indeed an indicator function $\mathbb{1}\{\mathbf{As} = Y^{[m]}\}$. Therefore, recalling that $\mathbf{b} = X^{[m]}$, we have

$$\mathsf{P}_{\mathbf{A},\mathbf{b}} = \mathsf{P}_{\mathbf{A}} \mathsf{P}_{X^{[m]}|Y^{[m]}}. \tag{16}$$

Analogously, the LWQ distribution satisfies

$$\mathsf{Q}_{\mathbf{A},\mathbf{b}} = \mathsf{P}_{\mathbf{A}} \mathsf{Q}_{X^{[m]}|Y^{[m]}} \tag{17}$$

because $\mathbf{A}$ and $Y^{[m]}$ are the same as those in the LWE distribution.

[5] In information theory, random variables X, Y, Z are said to form a Markov chain $X \to Y \to Z$ if their joint probability distribution function satisfy $P(x, y, z) = P(x)P(y|x)P(z|y)$ [19].

Now we have

$$
\begin{aligned}
&\Delta\left(\mathsf{P}_{\mathbf{A},\mathbf{b}}, \mathsf{Q}_{\mathbf{A},\mathbf{b}}\right) \\
&= \frac{1}{2} \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \frac{1}{2} \sum_{\mathbf{As}} \mathsf{P}_{\mathbf{As}|\mathbf{A}}(\cdot) \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \frac{1}{2} \sum_{\mathbf{As}} \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{As},\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \qquad (18) \\
&= \frac{1}{2} \sum_{Y^{[m]}} \mathsf{P}_{Y^{[m]}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \Delta\left(\mathsf{P}_{X^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right) \\
&\leq r \cdot m \sqrt{\ln 2 \cdot 2^{-m^{\beta'}}}, \quad \beta < \beta' < 1
\end{aligned}
$$

where the second equality of (18) holds since $\mathsf{P}_{\mathbf{As}|\mathbf{A}}$ is an indicator function when $\mathbf{s}$ is given, and the last inequality is obtained by instantiating with a polar lattice and applying Theorem 6 and Remark 7 in Section 4. Note that Remark 7 allows to choose a parameter $\beta' \in (\beta, 1)$ for any given $0 < \beta < 1$.

Since $r = \log_p q$ is fixed by LWE, and since the term $2^{-m^{\beta'}/2}$ dictates the bound $r \cdot m\sqrt{\ln 2 \cdot 2^{-m^{\beta'}}}$, we may set $m = \Theta(\lambda)$ such that $r \cdot m\sqrt{\ln 2 \cdot 2^{-m^{\beta'}}} = 2^{-\omega(\lambda^{\beta})}$. □

Informally, this theorem shows that the LWE noise can be substituted with the quantization noise of LWQ while preserving security.

Remark 5. Theorem 4 holds under KL divergence too, by applying Lemma 5 in Appendix D.

Theorem 5 (LWE reduced to LWQ). *Let $m = m(\lambda)$, $n = n(\lambda)$, $q = p(\lambda)^{r(\lambda)}$ where λ is the security parameter, $p(\lambda)$ is a prime number and $r(\lambda) \in \mathbb{N}$. Let $\mathbf{s} \in \mathbb{Z}_q^{n*}$. Let LWQ be instantiated with the quantization lattice Λ in Theorem 4, of dimension m and modulus q. If there exists an oracle that can distinguish the LWQ distribution from the uniform distribution $\mathcal{U}(\mathbb{Z}_q^{m \times n} \times \mathbb{Z}_q^m)$ with non-negligible advantage, then it can also distinguish the LWE distribution from uniform with non-negligible advantage.*

Proof. We consider an oracle interacting as part of probabilistic experiments called games in the following.

- $\mathcal{G}_0$: The oracle is given LWQ samples: $c \leftarrow \mathsf{LWQ}_{\Lambda,\mathbf{d}}$.
- $\mathcal{G}_1$: In this game, we give the oracle samples from LWE: $c \leftarrow \mathsf{LWE}_{\chi_e^m = \mathcal{D}_{\mathbb{Z}^m,\sigma}}$.
- $\mathcal{G}_2$: In this game, uniform samples are given: $c \leftarrow \mathbb{Z}_q^{m \times n} \times \mathbb{Z}_q^m$.

$\mathcal{G}_0$ and $\mathcal{G}_1$ are statistically indistinguishable as $\mathsf{LWQ}_{\Lambda,\mathbf{d}}$ and $\mathsf{LWE}_{\chi_e^m=\mathcal{D}_{\mathbb{Z}^m,\sigma}}$ are statistically indistinguishable, thus

$$\mathsf{Adv}_{\mathcal{G}_0,\mathcal{G}_1}(\mathcal{A}) = \left|\Pr(\mathcal{A}(\mathsf{LWQ}_{\Lambda,\mathbf{d}}) = 1) - \Pr(\mathcal{A}(\mathsf{LWE}_{\chi_e^m=\mathcal{D}_{\mathbb{Z}^m,\sigma}}) = 1)\right| \tag{19}$$

$$\leq \Delta(\mathsf{LWQ}_{\Lambda,\mathbf{d}}, \mathsf{LWE}_{\chi_e^m=\mathcal{D}_{\mathbb{Z}^m,\sigma}}) \tag{20}$$

$$= 2^{-\omega(\lambda^\beta)}, \quad \forall \ \ 0 < \beta < 1 \tag{21}$$

where the inequality is due to the data processing inequality of distributions, and the last equality is due to Theorem 4.

Now we use proof by contradiction. Assuming the oracle can only distinguish the LWE distribution with negligible advantage, then we have

$$\mathsf{Adv}_{\mathcal{G}_0,\mathcal{G}_2}(\mathcal{A}) \leq \mathsf{Adv}_{\mathcal{G}_0,\mathcal{G}_1}(\mathcal{A}) + \mathsf{Adv}_{\mathcal{G}_1,\mathcal{G}_2}(\mathcal{A}) = \mathrm{negl}(\lambda) \tag{22}$$

which contradicts with our assumption of LWQ. □

A direct consequence of Equation (22) is the following result.

Corollary 1 (LWQ $\approx_c$ Uniform). *Under the LWE assumption and the settings of Theorem 4, the LWQ distribution is computationally indistinguishable from the uniform distribution* $\mathcal{U}\left(\mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m\right)$.

3.3 Compression rate

In essence, we simulate the LWE channel using a polar lattice (which may also be viewed as a q-ary polar code) so that $\mathsf{LWE}_{\chi_e^m=\mathcal{D}_{\mathbb{Z}^m,\sigma}} \approx_s \mathsf{LWQ}_{\Lambda,\mathbf{d}}$. This is illustrated in Fig. 1 [6]. The bits of $U^{\mathcal{I}}$ are determined by the LWE channel, while those of $U^{\mathcal{F}}$ serve as the random dither. Basically, the bits of $U^{\mathcal{I}}$ are compressed LWE samples, and the LWE assumption implies that they are pseudorandom.

The LWE channel (4) is a so-called $\mathbb{Z}/q\mathbb{Z}$ channel with well-defined capacity $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$ [22] (see Appendix A.4 for details). Define the rate of the compressed ciphertext $R_c \triangleq \frac{1}{m}\log_2\left(\frac{q^m}{\det(\Lambda)}\right)$ bits per sample. The theory of polar lattices shows that any rate R_c above channel capacity $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$ is achievable for source coding [29]. In the language of Shannon's lossy compression theory, for a given test $\mathbb{Z}/q\mathbb{Z}$ channel between the source random variable and its reconstruction, the lowest achievable compression rate is the capacity $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$. Note that the $\mathbb{Z}/q\mathbb{Z}$ channel inherits the symmetric nature of the lattice partition channels [22], and the standard channel polarization technique for the symmetric rate-distortion function [26] works well in this context. In fact, by channel polarization, as $m \to \infty$,

$$R_c \to C(\mathbb{Z}/q\mathbb{Z}, \sigma^2).$$

[6] Note that we write $\mathbf{As} = \mathbf{b} + \mathbf{e} \mod q\mathbb{Z}$ in the figure, which is statistically equivalent to $\mathbf{b} = \mathbf{As} + \mathbf{e} \mod q\mathbb{Z}$ due to the symmetry of $\chi_e^m = \mathcal{D}_{\mathbb{Z}^m,\sigma}$. Reversing the input/output is a common practice for the test channel in source coding theory [19].

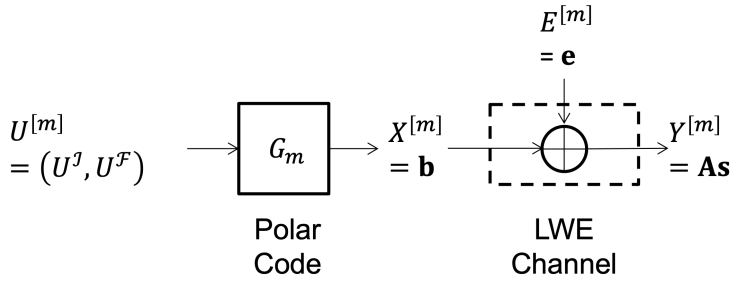

Fig. 1: Simulating the LWE channel using a polar lattice (which may be viewed as a q-ary polar code due to the mod q operation).

Thus, the compression rate R_c is ultimately determined by the capacity of the LWE channel. For the parameters $q = \text{poly}(n)$ and $\sigma = \Omega(\sqrt{n})$ in LWE, the capacity can be made explicit: it is possible to show $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2) \approx \log_2\left(\frac{q}{\sqrt{2\pi e}\cdot\sigma}\right)$ [22]. Intuitively, $\log_2\left(\sqrt{2\pi e}\cdot\sigma\right)$ of the $\log_2(q)$ bits are buried under noise.

If the noise variance σ^2 increases, the channel capacity $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$ decreases, and does the rate R. Conversely, if σ^2 decreases, channel capacity $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$ increases, as does R. However, it is important to note that the "LWE channel" is virtual, meaning σ^2 is a free parameter that can be adjusted. Consequently, any rate $0 < R_c < \log_2(q)$ can be achieved by appropriately tuning σ^2. Remarkably, efficiency and security (R_c vs. σ^2) align: higher compression is accompanied by increased security. The trade-off, however, is that greater compression results in fewer pseudorandom numbers being generated.

4 Polar Lattice for Quantization

The idea of polar quantizer is to use multilevel error correction codes to decode (quantize) inputs at each level. Throughout this section, we assume $\mathbf{s} \in \mathbb{Z}_q^{n*}$ such that $\mathbf{As}$ admits a uniform distribution on $\mathbb{Z}_q^m$. We often assume $q = 2^r$ for $r \in \mathbb{N}$ and it is straightforward to extend to the case $q = p^r$ for prime p.

4.1 Polar Quantizer: Construction

In this subsection, we present an explicit construction of polar lattices [30, 29] for the dithered quantization of random integers, which produces Gaussian-like quantization errors. In a nutshell, the quantizer consists of a series of decoders for polar codes according to the multilevel structure of "Construction D" [22]. For convenience, we present Construction D using binary polar codes, whereas the extension to nonbinary codes is straightforward [18].

For those unfamiliar with polar codes or polar lattices, it could be useful to treat the polar lattice quantizer as a black box, as shown in the dashed box

in Fig. 2, whose task is to mimic the test channel between X and Y in Fig. 1. From the perspective of lossy compression, the test channel for the source $Y \sim \mathsf{P}_Y$ is defined by the transition probability $\mathsf{P}_{Y|X}$, where X is referred to as the reconstruction of the source. As can be seen in Fig. 1, the statistic of the test channel is described by the relationship $Y = X + E \mod q\mathbb{Z}$, where E is an additive discrete Gaussian noise. Note that for this test channel, defined from the information theory, is purely based on the statistic of E, which is not necessarily generated by the lattice quantization operation. However, Theorem 6 illustrates that the difference between these two test channels can be negligible, which confirms the motivation of introducing lattice quantization in our LWQ scheme. Moreover, the relationship between the lattice quantization from $Y^{[m]}$ to $X^{[m]}$ and the lattice construction based on the test channel from $X^{[m]}$ to $Y^{[m]}$ will be explained in Remark 8.

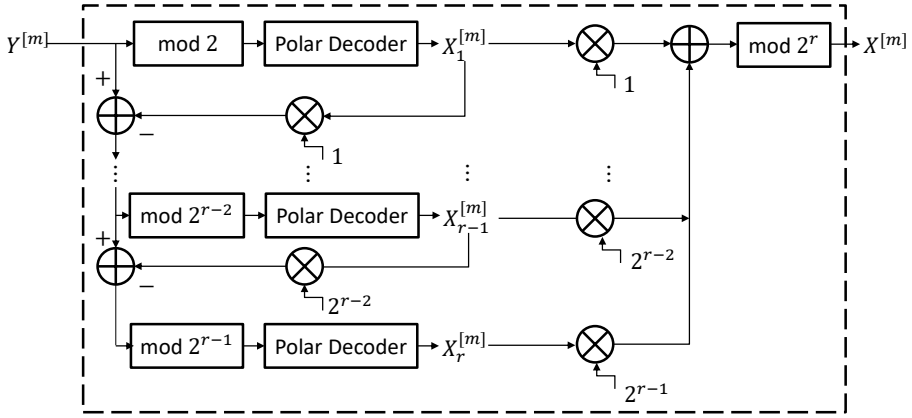

Fig. 2: The internal structure of a polar lattice quantizer.

Definition 12 (Partition Chain). *A sublattice $\Lambda' \subset \Lambda$ induces a partition (denoted by Λ/Λ') of Λ into equivalence groups modulo Λ'. The order of the partition is denoted by $|\Lambda/\Lambda'|$, which is equal to the number of cosets. If $|\Lambda/\Lambda'| = 2$, this is called a binary partition. A lattice partition chain, which is denoted by $\Lambda(\Lambda_0)/\Lambda_1/\cdots/\Lambda_{r-1}/\Lambda'(\Lambda_r)$ for $r \geq 1$, is a sequence of nested lattices.*

If only one level is used ($r = 1$), the construction is called Construction A. If multiple levels are used, it is called Construction D. For each partition $\Lambda_{\ell-1}/\Lambda_\ell$ ($1 \leq \ell \leq r$), a code C_ℓ over $\Lambda_{\ell-1}/\Lambda_\ell$ selects a sequence of coset representatives a_ℓ in a set A_ℓ of representatives for the cosets of Λ_ℓ. This construction requires a set of nested linear binary codes C_ℓ with block length m and dimension k_ℓ, represented as $[m, k_\ell]$ codes for $1 \leq \ell \leq r$, with $C_1 \subseteq C_2 \subseteq \cdots \subseteq C_r$. We use the partition chain $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}/\cdots$ in the following.

Algorithm 1 Polar Lattice Quantization Algorithm

Require: Source $Y^{[m]} \in [-2^{r-1}, 2^{r-1})^m$.
Ensure: Quantized output $X^{[m]}$.
1: Initialize $R^{[m]} = Y^{[m]}$;
2: **for** $\ell = 1$ to r **do**
3: Execute SC decoding to obtain $X_\ell^{[m]}$ from $R^{[m]} \mod 2^\ell$. ▷ Decodes binary polar codes
4: $R^{[m]} = R^{[m]} - 2^{\ell-1} X_\ell^{[m]}$. ▷ Interference cancellation
5: **end for**
6: Return $X^{[m]} = X_1^{[m]} + 2X_2^{[m]} + \cdots + 2^{r-1} X_r^{[m]} \mod 2^r \mathbb{Z}$.

Definition 13 (Construction D). *Let ψ be the natural embedding of $\mathbb{F}_2^m$ into $\mathbb{Z}^m$, where $\mathbb{F}_2$ is the binary field. Consider a basis $\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m$ of $\mathbb{F}_2^m$ such that $\mathbf{d}_1, \cdots, \mathbf{d}_{k_\ell}$ span C_ℓ. The Construction D lattice consists of all vectors of the form*

$$\sum_{\ell=1}^{r} 2^{\ell-1} \sum_{j=1}^{k_\ell} u_\ell^j \psi(\mathbf{d}_j) + 2^r z, \tag{23}$$

where $u_\ell^j \in \{0, 1\}$, $z \in \mathbb{Z}^m$, and ψ denotes the embedding into $\mathbb{R}^m$.

The quality of a subchannel is generally identified based on its associated Bhattacharyya parameter.

Definition 14. *Given a binary-input memoryless symmetric channel (BMSC) W with transition probability $\mathsf{P}_{Y|X}$, the Bhattacharyya parameter $Z \in [0, 1]$ is defined as*

$$Z(W) = Z(X|Y) \triangleq \sum_y \sqrt{\mathsf{P}_{Y|X}(y|0)\mathsf{P}_{Y|X}(y|1)}. \tag{24}$$

E.g., in [4], the rate of channel polarization is characterized in terms of the Bhattacharyya parameter as

$$\lim_{m\to\infty} \mathsf{Pr}\left(Z(W_m^{(i)}) < 2^{-m^\beta}\right) = C, \quad \text{for any } 0 < \beta < \frac{1}{2}.$$

This means that as the block length m becomes very large, the probability that the Bhattacharyya parameter $Z(W_m^{(i)})$ of a subchannel $W_m^{(i)}$ is less than 2^{-m^β} approaches the channel capacity C. For efficient construction of polar codes, $Z(W_m^{(i)})$ can be evaluated using the methods introduced in [48, 39].

In the context of lossy compression, polar codes can achieve the rate-distortion bound for binary symmetric sources [27]. To achieve a target distortion:

– A test channel $W : X \to Y$ is constructed for the source Y and the reconstruction X.

– Polar codes for compression are constructed according to the test channel W, with the information set defined as $\mathcal{I} \triangleq \{i \in [m] : Z(W_m^{(i)}) < 1 - 2^{-m^\beta}\}$.

By the duality between channel coding and source coding, the Successive Cancellation (SC) decoding algorithm for polar channel coding transforms into the SC encoding algorithm for polar source coding. Given m i.i.d. sources $Y^{[m]}$:

– The polarized bits $U^{\mathcal{F}}$ are almost independent of $Y^{[m]}$ since $Z(W_m^{(i)}) \geq 1 - 2^{-m^\beta}$ by definition.
– Compression of $Y^{[m]}$ is achieved by replacing $U^{\mathcal{F}}$ with random bits and saving the relevant bits $U^{\mathcal{I}}$, which are determined from $Y^{[m]}$ and $U^{\mathcal{F}}$ using the SC encoder.

The channel splitting process also leads to a simple SC decoding algorithm to achieve the so-called symmetric capacity [3], which executes maximum a posteriori (MAP) decoding for each subchannel sequentially from $i = 1$ to m. By the union bound, the block error probability of SC decoding can be upper-bounded by $\sum_{i \in \mathcal{I}} Z(W_m^{(i)})$. See Appendix A.2 or [3, Section VIII] for more details on the SC decoding algorithm.

Pseudo-codes of the polar lattice quantization algorithm are given in Algorithm 1 where q is a power of 2. For the samples $Y^{[m]}$, the decoder at each level tries to find the best binary representative of the lattice point $X^{[m]}$ close to $Y^{[m]}$, using the results of all previous levels. The multilevel structure of polar lattices not only provides us a feasible complexity of the quantization operation for high dimensional lattices, but also paves for us a path to the rich theory of binary polar codes.

The next subsection will show that the distribution of $Y^{[m]} - X^{[m]}$ is close to that of m i.i.d. discrete Gaussian random variables. Fig. 3 shows a comparison between the distribution of quantization noise $Y - X$ achieved by the polar lattice quantizer and the genuine discrete Gaussian distribution $\mathcal{D}_{\mathbb{Z},\sigma}$ with parameters $\sigma = 3$, $r = 8$ and $m = 2^{20}$.

Dithered quantization with polar lattices In the literature on traditional lattice quantization [52], the source vector is shifted by dithering $\mathbf{d}$ while the quantization lattice remains fixed (the output is $Q_\Lambda(\mathbf{As} - \mathbf{d})$). In contrast, our dithered quantization compensates the dither vector and output: $Q_\Lambda(\mathbf{As} - \mathbf{d}) + \mathbf{d}$. This type of quantization can be easily implemented via a polar lattice. Specifically, when the frozen bits are chosen randomly, the output of a polar lattice quantizer $Q_{\Lambda + \mathbf{d}}$ belongs to a random coset $\Lambda + \mathbf{d}$, where the randomness $\mathbf{d}$ is determined by the frozen bits. This can be understood as follows. Let $U^{\mathcal{F}_\Lambda} = \{U_1^{\mathcal{F}_1}, \ldots, U_r^{\mathcal{F}_r}\}$ denote the collection of all frozen bits across the r levels. For a specific choice $u^{\mathcal{F}_\Lambda} = \{u_1^{\mathcal{F}_1}, \ldots, u_r^{\mathcal{F}_r}\}$, the resulting offset from Λ can be expressed as

$$\mathbf{d} = \sum_{\ell=1}^{r} 2^{\ell-1} \sum_{j=k_\ell+1}^{N} u_\ell^j \psi(\mathbf{g}_j), \tag{25}$$

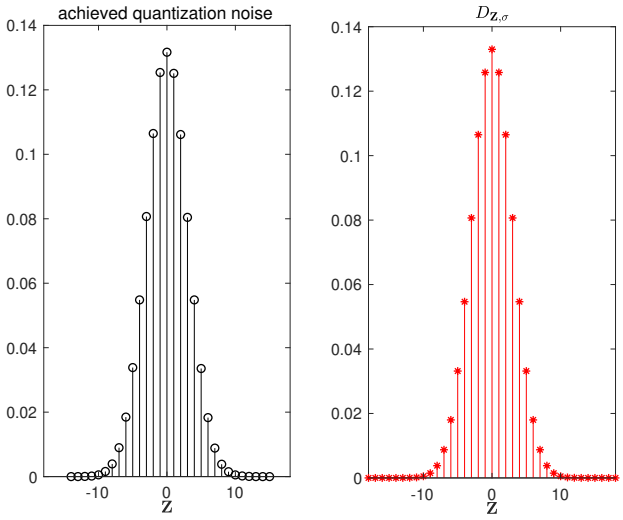

Fig. 3: A comparison between the distribution of quantization noise $Y - X$ and $D_{\mathbb{Z},\sigma=3}$.

where $u_\ell^j \in \{0,1\}$ and $\mathbf{g}_{k_\ell+1}, \ldots, \mathbf{g}_N$ are the remaining base vectors in the vector space spanned by G_N after selecting $\mathbf{g}_1, \ldots, \mathbf{g}_{k_\ell}$ for the binary code at level ℓ. Clearly, Λ corresponds to the all-zero configuration of $U^{\mathcal{F}_\Lambda}$, and $\Lambda + \mathbf{d}$ forms a valid partition of $\mathbb{Z}^m$ as $U^{\mathcal{F}_\Lambda}$ traverses all possible choices.

Remark 6. The dither $\mathbf{d}$ of LWQ is public, thus an extendable-output function can be used to produce the dither, with only the generator's seed needing to be shared as part of the public key. This approach allows LWQ to achieve computational indistinguishability from LWE while also reducing bandwidth.

4.2 Polar Quantizer: Performance Analysis

We now analyze the distribution of quantization noise. Let $Y^{[m]}$ denote m samples drawn from **As**. The quantization result or the so-called reconstruction of $Y^{[m]}$ is denoted by $X^{[m]}$, which is also in $\mathbb{Z}_q^m$.

- Consider the first case in which the correlation between $Y^{[m]}$ and $X^{[m]}$ is due to an i.i.d. discrete Gaussian random vector $E^{[m]}$, i.e., $Y^i = X^i + E^i \mod q\mathbb{Z}$ for each $i \in [m]$, and $E^i \sim \mathcal{D}_{\mathbb{Z},\sigma}$. The joint distribution between $X^{[m]}$ and $Y^{[m]}$ in this case is denoted by $\mathsf{P}_{X^{[m]},Y^{[m]}}$.
- Consider the second case in which the correlation between $Y^{[m]}$ and $X^{[m]}$ is generated by the polar lattice quantizer, i.e., $X^{[m]} = Q_\Lambda(Y^{[m]})$. The joint distribution between $X^{[m]}$ and $Y^{[m]}$ in this case is denoted by $\mathsf{Q}_{X^{[m]},Y^{[m]}}$.

We will show the statistical distance $\Delta(\mathsf{P}_{X^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}})$ vanishes sub-exponentially in m in a layer-by-layer manner, corresponding to the multi-level quantization process of polar lattices. Notice that each $X^i \in \mathbb{Z}_q, i \in [m]$ can be uniquely represented by a binary sequence $X_1^i, ..., X_\ell^i, ..., X_r^i$, and X_ℓ^i determines the coset of the binary partition $2^{\ell-1}\mathbb{Z}/2^\ell\mathbb{Z}$ for $1 \le \ell \le r$. Given a source vector $Y^{[m]}$, the (m-dimensional) polar lattice quantizer tries to find the coset leader $X_1^{[m]}$ at the first level; then it decides the coset leader $X_2^{[m]}$ at the second level using both $X_1^{[m]}$ and $Y^{[m]}$; the process keeps going at level ℓ, where $X_\ell^{[m]}$ is decoded from $Y^{[m]}$ and $X_{1:\ell-1}^{[m]}$; the process ends at the final r-th level, where $X_r^{[m]}$ is decoded from $Y^{[m]}$ and $X_{1:r-1}^{[m]}$.

From the perspective of lossy compression in information theory, $\mathsf{P}_{Y|X}$ is called the test channel with input (reconstruction) X and output (source) Y. As can be seen in Fig. 1, since $Y = X + E \mod q\mathbb{Z}$, the test channel is a discrete additive white Gaussian noise channel with a modulo $q\mathbb{Z}$ operation at the end. Following the step of Forney et al. [22], the test channel can be partitioned into r $2^{\ell-1}\mathbb{Z}/2^\ell\mathbb{Z}$ binary-input channels with $1 \le \ell \le r$, which are called binary partition channels.

In fact, the polar lattice consists of the component polar codes designed for these r partition channels. More explicitly, the first level $\mathbb{Z}/2\mathbb{Z}$ partition channel completely determines the joint distribution $\mathsf{P}_{X_1,Y}$ of X_1 and Y, and $Y \mod 2\mathbb{Z}$ is a sufficient statistic of Y with respect to X_1. The polar code C_1 at the first level is constructed according to the $\mathbb{Z}/2\mathbb{Z}$ channel, which is equivalently described by $W_1 : X_1 \underset{\mathsf{P}_{Y|X_1}}{\longrightarrow} Y$. Let $U_1^{[m]} = X_1^{[m]} G_m$ be the bits after channel polarization at level 1. The information set of C_1 is defined as $\mathcal{I}_1 \triangleq \{i \in [m] : Z(U_1^i | U_1^{1:i-1}, Y^{[m]}) \le 1 - 2^{-m^\beta}\}$ for any $0 < \beta < 0.5$, and the frozen set of C_1 is the complement set $\mathcal{F}_1 \triangleq \mathcal{I}_1^c$. By this definition, the correlation between $U_1^{\mathcal{F}_1}$ and $Y^{[m]}$ is negligible. The polar quantizer assigns uniformly random bits that are independent of $Y^{[m]}$ to $U_1^{\mathcal{F}_1}$, and then determines $U_1^{\mathcal{I}_1}$ from $Y^{[m]}$ and $U_1^{\mathcal{F}_1}$ using the SC encoding algorithm. The reconstruction at level 1 is obtained from the inverse polarization transform $X_1^{[m]} = U_1^{[m]} G_m^{-1} = U_1^{[m]} G_m$.

Lemma 1. *Let $\mathsf{Q}_{U_1^{[m]},Y^{[m]}}$ denote the resulted joint distribution of $U_1^{[m]}$ and $Y^{[m]}$ according to the encoding rules* (26) *and* (27) *at the first partition level.*

$$U_1^i = \begin{cases} 0 \text{ w. p. } P_{U_1^i|U_1^{1:i-1},Y^{[m]}}\left(0|u_1^{1:i-1}, y^{[m]}\right) \\ 1 \text{ w. p. } P_{U_1^i|U_1^{1:i-1},Y^{[m]}}\left(1|u_1^{1:i-1}, y^{[m]}\right) \end{cases} \text{if } i \in \mathcal{I}_1 \tag{26}$$

$$U_1^i = \begin{cases} 0 \text{ w. p. } \dfrac{1}{2} \\ 1 \text{ w. p. } \dfrac{1}{2}. \end{cases} \text{if } i \in \mathcal{F}_1 \tag{27}$$

Let $\mathsf{P}_{U_1^{[m]},Y^{[m]}}$ *denote the joint distribution directly generated from* $\mathsf{P}_{X_1,Y}$*, i.e.,* U_1^i *is generated according to the encoding rule* (26) *for all* $i \in [m]$*. The statistical distance between* $\mathsf{P}_{U_1^{[m]},Y^{[m]}}$ *and* $\mathsf{Q}_{U_1^{[m]},Y^{[m]}}$ *is upper-bounded as follows:*

$$\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right) \leq m\sqrt{\ln 2 \cdot 2^{-m^\beta}}, \quad 0 < \beta < \frac{1}{2}. \tag{28}$$

Proof. See Appendix C.

After finishing the encoding at level 1, the polar lattice quantizer proceeds to level 2 in a similar manner. The $2\mathbb{Z}/4\mathbb{Z}$ partition channel completely determines the joint distribution $\mathsf{P}_{X_2,Y|X_1}$ of X_2 and Y given the previous quantization result X_1, and $Y - X_1 \mod 4\mathbb{Z}$ is a sufficient statistic of Y with respect to X_2. The polar code C_2 at the second level is constructed according to the $2\mathbb{Z}/4\mathbb{Z}$ channel, which is equivalently described by $W_2 : X_2 \underset{\mathsf{P}_{Y,X_1|X_2}}{\longrightarrow} (Y, X_1)$. Let $U_2^{[m]} = X_2^{[m]} G_m$ be the bits after channel polarization at level 2. The information set of C_2 is defined as $\mathcal{I}_2 \triangleq \{i \in [m] : Z(U_2^i | U_2^{1:i-1}, X_1^{[m]}, Y^{[m]}) \leq 1 - 2^{-m^\beta}\}$ for $0 < \beta < 1/2$, and the frozen set is defined as $\mathcal{F}_2 \triangleq \mathcal{I}_2^c$.

Lemma 2. *Let* $\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ *denote the resulted joint distribution of* $U_1^{[m]}$*,* $U_2^{[m]}$ *and* $Y^{[m]}$ *according to the encoding rules* (26) *and* (27) *at the first partition level, and then rules* (29) *and* (30) *at the second partition level.*

$$U_1^i = \begin{cases} 0 \text{ w. p. } P_{U_2^i|U_2^{1:i-1},X_1^{[m]},Y^{[m]}}\left(0|u_2^{1:i-1}, x_1^{[m]}, y^{[m]}\right) \\ 1 \text{ w. p. } P_{U_2^i|U_2^{1:i-1},X_1^{[m]},Y^{[m]}}\left(1|u_2^{1:i-1}, x_1^{[m]}, y^{[m]}\right) \end{cases} \text{if } i \in \mathcal{I}_2 \tag{29}$$

$$U_2^i = \begin{cases} 0 \text{ w. p. } \dfrac{1}{2} \\ 1 \text{ w. p. } \dfrac{1}{2}. \end{cases} \text{if } i \in \mathcal{F}_2 \tag{30}$$

Let $\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ *denote the joint distribution directly generated from* $\mathsf{P}_{X_1,X_2,Y}$*, i.e.,* U_1^i *and* U_2^i *are generated according to the encoding rule* (26) *and rule* (29) *for all* $i \in [m]$*, respectively. The statistical distance between* $\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ *and* $\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ *is upper-bounded as follows:*

$$\Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) \leq 2m\sqrt{\ln 2 \cdot 2^{-m^\beta}}, \quad 0 < \beta < \frac{1}{2}. \tag{31}$$

Proof. Assume an auxiliary joint distribution $\mathsf{Q}'_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ resulted from using the encoding rule (26) for all U_1^i with $i \in [m]$ at the first partition level, and rules (29) and (30) at the second partition. Clearly, $\mathsf{Q}'_{U_1^{[m]},Y^{[m]}} = \mathsf{P}_{U_1^{[m]},Y^{[m]}}$ and $\mathsf{Q}'_{U_2^{[m]}|U_1^{[m]},Y^{[m]}} = \mathsf{Q}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}$. By the triangle inequality,

$$\begin{aligned}&\varDelta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right)\\&\le\varDelta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q}'_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right)+\varDelta\left(\mathsf{Q}'_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right),\end{aligned}\tag{32}$$

where the first term on the right hand side can be upper bounded by $m\sqrt{\ln 2\cdot 2^{-m^\beta}}$ using the same method as in the proof of Lemma 1, and the second term is equal to $\varDelta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)$. □

After the lattice quantization process with r sequential levels, the joint distribution produced by the lattice quantizer is denoted by $\mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}}$, and the joint distribution directly generated from m i.i.d. test channels is denoted by $\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}}$. By induction, we obtain $\varDelta\left(\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}},\mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}}\right)\le rm\sqrt{\ln 2\cdot 2^{-m^\beta}}$. We arrive at the following theorem on the distribution of quantization noise, which shows the quantization noise closely resemble an i.i.d. discrete Gaussian distribution.

Theorem 6. *The statistical distance between the joint distribution induced by the polar lattice and that by an i.i.d. discrete Gaussian distribution is bounded by*

$$\varDelta\left(\mathsf{P}_{X^{[m]},Y^{[m]}},\mathsf{Q}_{X^{[m]},Y^{[m]}}\right)\le r\cdot m\sqrt{\ln 2\cdot 2^{-m^\beta}},\quad 0<\beta<\frac{1}{2}.\tag{33}$$

Proof. By the inverse polarization transform $X_\ell^{[m]}=U_\ell^{[m]}G_m$ from $\ell=1$ to r, we immediately have $\varDelta\left(\mathsf{P}_{X^{[m]},Y^{[m]}},\mathsf{Q}_{X^{[m]},Y^{[m]}}\right)\le r\cdot m\sqrt{\ln 2\cdot 2^{-m^\beta}}$, by induction. □

Remark 7. The restriction $0<\beta<\frac{1}{2}$ in Theorem 6 is due to the standard 2×2 kernel $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ of binary polar codes, which results in sub-exponential decay of the statistical distance. Nevertheless, it is possible to obtain any value $0<\beta<1$ by using nonbinary polar codes with prime alphabet size p and carefully designed kernels [36]; thus we can obtain almost exponential decay of the statistical distance. More precisely, [36] showed that using a p-ary $\ell\times\ell$ kernel ($1<\ell\in\mathbb{N}$), it is possible to obtain $\beta=\log_2(\ell!)/(\ell\log_2\ell)$. But there is a price to pay: the decoding complexity will become $O(p^\ell m\log_2 m)$ [45]. Although using nonbinary polar codes in Construction D will increase the computational complexity, it is still $O(m\log_2 m)$ for fixed p and ℓ.

Remark 8. Observant readers may wonder why our polar lattice quantizer is constructed based on the forward test channel $X\underset{\mathsf{P}_{Y|X}}{\longrightarrow}Y$, with additive noise E mod $q\mathbb{Z}$, whereas the quantization performance shown above is analyzed from the reversed direction $Y\underset{\mathsf{P}_{X|Y}}{\longrightarrow}X$. The reason is that when X and Y are both uniform in $\mathbb{Z}_q$, we have $\mathsf{P}_{X|Y}=\mathsf{P}_{Y|X}$, and the additive noise E is pairwisely

independent of both X and Y. To see this, letting $\mathsf{P}_X(x) = 1/q$, we have $\mathsf{P}_Y(y) = \sum_x \mathsf{P}_{X,Y}(x,y) = \frac{1}{q}\sum_x \mathsf{P}_E(y-x) = 1/q$. Therefore, $\mathsf{P}_X = \mathsf{P}_Y = 1/q$, and hence $\mathsf{P}_{Y|X} = \mathsf{P}_{X|Y}$. The symmetry of the test channel, which is termed as the mod Λ/Λ' channel, is discussed in more detail by Forney et al. in [22].

Remark 9. We note that the validity of polar lattice structure can be easily guaranteed. Taking the above simulation as an example, when constructing multilevel polar codes along the binary partition chain $\mathbb{Z}/2\mathbb{Z}/\cdots/2^r\mathbb{Z}$ for the additive discrete Gaussian test channel ($\sigma = 3$), the capacities of the partition channels from $\ell = 1$ to r are given by 0, 3.2732×10^{-10}, 0.0056, 0.3933, 0.9690, 1.0000 and 1.0000, respectively. The size of the information set is chosen as $|\mathcal{I}_\ell| = \lceil m \cdot C(W_\ell)\rceil$, where $C(W_\ell)$ denotes the capacity of the ℓ-th partition channel. As a result, the component polar codes are consecutively nested by ensuring $\mathcal{I}_\ell \subseteq \mathcal{I}_{\ell+1}$ for $1 \le \ell \le r-1$, and we have an ascertained polar lattice quantizer. Moreover, the constructed polar lattice is roughly sphere-bound achieving, by the capacity-achieving property of polar codes for all partition levels.

5 Improving Lattice-Based Secret-Key Encryption

This section introduces a secret-key LWQ-based encryption framework, denoted as $\mathsf{LWQ}_{E,\Lambda,\mathbf{d}}$, and contrasts it with LWE and quantized LWE (LWEQ) based frameworks from [35]. Specifically, LWE_{E,χ_e} represents the encryption framework without quantization, while $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$ corresponds to $\mathsf{LWE}[E,\Lambda]$ as described in [35]. It is important to note that the LWER problem in CRYSTALS-Kyber [47] represents a special case of LWEQ where the quantization is rounding. In comparison to quantized LWE [35], LWQ streamlines the processes of noise addition and quantization into a single step, where only quantization noise is present while ensuring security. This efficiency suggests that LWQ could serve as a potential alternative to LWE, LWR, or LWER in a range of cryptographic scenarios, offering a favorable balance between efficiency and security. E.g., in addition to the presented secret-key encryption hereby, we show in Appendix E that LWQ can be employed to reduced the size of public key of plain-LWE based public key encryption (PKE).

The presented LWE-, LWEQ-, and LWQ-based secret-key encryption schemes all use a triplet $(\mathsf{KGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$. For fair comparison, they share a common KGen and nested lattice error-correction structures.

The key generation function $\mathsf{KGen}(1^\lambda)$ along with the standard choice of parameters from LWE are defined as follows:

- Select $m = n^{O(1)}$, and $q \in [n^{O(1)}, 2^{O(n)}]$. Let χ_e be a discrete Gaussian error distribution of parameter $\sigma \ge 2\sqrt{n}$, and a private key distribution χ_s^n over $\mathbb{Z}_q^{n*}$ with respect to the security parameter λ. Sample $\mathbf{s} \leftarrow \chi_s^n$ until $\mathbf{s} \in \mathbb{Z}_q^{n*}$ (e.g., $\mathbf{s} \leftarrow \mathbb{Z}_q^n$, which satisfies $\mathbf{s} \in \mathbb{Z}_q^{n*}$ with overwhelming probability).
- Targeting specific error correction capacity and quantization noise level, choose the error correction lattice E and quantization lattice Λ from the

partition chain of polar lattices:

$$q\mathbb{Z}^m \subset E \subseteq \Lambda \subset \mathbb{Z}^m.$$

- Specify the lattice encoding function ec_E that maps a message $\mu \in \{0,1\}^{|\mathsf{pt}|}$, where $|\mathsf{pt}| = \log_2\left(\frac{q^m}{\det(E)}\right)$, to a lattice point within the error correction lattice E, and the lattice decoding function dc_E that recovers the original message by decoding a noisy lattice point back into the message space.
- RETURN the private key $\mathbf{s}$.

Scheme	$\mathsf{Encrypt}_{\mathbf{s}}(\mu)$	$\mathsf{Decrypt}_{\mathbf{s}}(\mathbf{A},\mathbf{b})$	**Ciphertext Error** $\tilde{\mathbf{e}}$	**Ciphertext Size** $\lvert\mathsf{ct}\rvert$
LWE_{E,χ_e} [35]	$\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{e} \leftarrow \chi_e$ $\mathbf{b} = \mathbf{As} + \mathbf{e} + \mathsf{ec}_E(\mu)$ RETURN $(\mathbf{A},\mathbf{b})$	RETURN $\mathsf{dc}_E(\mathbf{b}-\mathbf{As})$	$\mathbf{e}$	$\lvert\mathsf{seed}(\mathbf{A})\rvert + \log_2(q^m)$
$\mathsf{LWEQ}_{E,\chi_e,\Lambda}$ [35]	$\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{e} \leftarrow \chi_e$ $\mathbf{b} = Q_\Lambda(\mathbf{As} + \mathbf{e}) + \mathsf{ec}_E(\mu)$ RETURN $(\mathbf{A},\mathbf{b})$	RETURN $\mathsf{dc}_E(\mathbf{b}-\mathbf{As})$	$\mathbf{e} + \mathbf{e}'_Q$	$\lvert\mathsf{seed}(\mathbf{A})\rvert + \log_2(\frac{q^m}{\det(\Lambda)})$
$\mathsf{LWQ}_{E,\Lambda,\mathbf{d}}$ (proposed)	$\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$ $\mathbf{b} = Q_{\Lambda+\mathbf{d}}(\mathbf{As}) + \mathsf{ec}_E(\mu)$ RETURN $(\mathbf{A},\mathbf{b})$	RETURN $\mathsf{dc}_E(\mathbf{b}-\mathbf{As})$	$\mathbf{e}_Q$	$\lvert\mathsf{seed}(\mathbf{A},\mathbf{d})\rvert + \log_2(\frac{q^m}{\det(\Lambda)})$

Table 2: Comparison of encryption frameworks based on LWE, LWEQ, and LWQ.

The $(\mathsf{Encrypt}, \mathsf{Decrypt})$ algorithms for LWE_{E,χ_e}, $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$, and the proposed $\mathsf{LWQ}_{E,\Lambda,\mathbf{d}}$ are summarized in Table 2. These schemes differ only in the encryption process. Since $E \subseteq \Lambda$, we have $Q_\Lambda(\mathbf{As}+\mathbf{e}+\mathsf{ec}_E(\mu)) = Q_\Lambda(\mathbf{As}+\mathbf{e})+\mathsf{ec}_E(\mu)$ and $Q_{\Lambda+\mathbf{d}}(\mathbf{As}+\mathsf{ec}_E(\mu)) = Q_{\Lambda+\mathbf{d}}(\mathbf{As}) + \mathsf{ec}_E(\mu)$. Thus, in all these schemes, the message $\mathsf{ec}_E(\mu)$ is encrypted by masking it with a pseudorandom vector. Define the effective ciphertext error as

$$\tilde{\mathbf{e}} = \mathbf{b} - \mathbf{As} - \mathsf{ec}_{\mathbf{E}}(\mu).$$

Table 2 shows that the effective ciphertext errors for these schemes are $\mathbf{e}$, $\mathbf{e}+\mathbf{e}'_Q$, and $\mathbf{e}_Q$, respectively, where $\mathbf{e}'_Q$ and $\mathbf{e}_Q$ represent the quantization errors introduced by Q_Λ and $Q_{\Lambda+\mathbf{d}}$, respectively. The storage sizes for $\mathbf{A}$ and $(\mathbf{A},\mathbf{d})$ are denoted as $|\mathsf{seed}(\mathbf{A})|$ and $|\mathsf{seed}(\mathbf{A},\mathbf{d})|$, respectively, leveraging XOF-based compact representations [35]. The performance differences between these schemes are reflected in the ciphertext error $\tilde{\mathbf{e}}$ and the ciphertext size $|\mathsf{ct}|$.

5.1 Security

In these encryption schemes, the pseudorandomness of the ciphertext, ensuring RND-CPA security [35], is derived from the hardness of the decision LWE and LWQ assumptions.

Definition 15 (RND-CPA). *An encryption scheme* $(\mathsf{KGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ *is said to be pseudorandom under chosen plaintext attack if any efficient (probabilistic polynomial-time) adversary* $\mathcal{A}$ *can only achieve at most negligible advantage in the following game, parameterized by a bit* $b \in \{0, 1\}$*:*

1. $\mathbf{s} \leftarrow \mathsf{KGen}(1^\lambda)$,
2. $b' \leftarrow \mathcal{A}^{O_b(\cdot)}$ *where* $O_b(\mu)$ *returns either an encryption* $\mathsf{Encrypt}_{\mathbf{s}}(\mu)$ *of the message* μ *under the key* $\mathbf{s}$ *if* $b = 0$*, or a sample from a uniform distribution that has support* $\left\{\mathsf{Encrypt}_{\mathbf{s}}(\mu) \mid \mathbf{s} \in \mathrm{supp}(\mathsf{KGen}(1^\lambda)), \forall \mu\right\}$ *if* $b = 1$*.*

The adversary's advantage is defined as $\mathrm{Adv}(\mathcal{A}) = |\Pr(b' = 1|b = 0) - \Pr(b' = 1|b = 1)|$.

Theorem 7. *Under the LWE and LWQ indistinguishability assumptions, the schemes* LWE_{E,χ_e}*,* $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$*, and* $\mathsf{LWQ}_{E,\Lambda,\mathbf{d}}$ *are RND-CPA secure.*

Proof. We demonstrate that if an adversary can break the RND-CPA security of LWE_{E,χ_e}, $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$, or $\mathsf{LWQ}_{E,\Lambda,\mathbf{d}}$, it implies the ability to distinguish the LWE/LWEQ/LWQ distributions from uniform distributions. We will focus on the reduction for LWEQ, as the arguments for the other two cases are analogous.

We construct an oracle O'_b for $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$:

- Request the pair $(\mathbf{A}, \mathbf{b})$ from the LWE oracle O_b.
- Compute $Q_\Lambda(\mathbf{b})$.
- Return the output $(\mathbf{A}, Q_\Lambda(\mathbf{b}) + \mathsf{ec}_E(\mu))$.

Since O'_b incorporates O_b, breaking $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$ would consequently imply breaking the LWE assumption, establishing the RND-CPA security of the encryption scheme.

5.2 Efficiency

The information rate R of an encryption scheme is defined as the ratio of plaintext size to ciphertext size:

$$R = \frac{|\mathsf{pt}|}{|\mathsf{ct}|}. \tag{34}$$

Notably, the contribution of XOF seeds will be excluded from $|\mathsf{ct}|$ in the subsequent discussion, as their size is generally negligible compared to the overall ciphertext (cf. [35]). A scheme is said to achieve perfect rate when $R = 1$.

The correctness of the schemes are defined as:

Definition 16. ***(DFR).*** *The decryption failure rate (DFR) of an encryption scheme* $(\mathsf{KGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ *is defined as*

$$\delta = \mathbb{E}_{\mathbf{s}} \max_{\mu} \Pr(\mathsf{Decrypt}_{\mathbf{s}}(\mathsf{Encrypt}_{\mathbf{s}}(\mu)) \neq \mu).$$

The scheme is said to be δ*-correct for a neglegible* δ*, and perfectly correct if* $\delta = 0$*.*

Theorem 8. *There exists a family of polar lattices $\{\Lambda_m\}$ (indexed by dimension m) such that $\mathsf{LWQ}_{E=\Lambda,\Lambda,\mathbf{d}}$ achieves: perfect correctness ($\delta = 0$), perfect rate, polynomial modulus $q = n^{O(1)}$, and security equivalent to the LWE assumption.*

Proof. Instantiate LWQ with the quantization lattice Λ from Theorem 4. By construction, LWQ is as secure as LWE with modulus $q = n^{O(1)}$. Regarding perfect correctness, the decryption noise $\tilde{\mathbf{e}} = \mathbf{b} - \mathbf{A}\mathbf{s} - \mathsf{ec}_E(\mu)$ reduces to $\tilde{\mathbf{e}}_{\mathsf{LWQ}} = \mathbf{e}_Q$, where $\mathbf{e}_Q \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$. Since $\mathbf{s} \in \mathbb{Z}_q^n \setminus \{\mathbf{0}\}$, $\mathbf{A}\mathbf{s}$ is uniform over $\mathbb{Z}_q^m$, and thus $\Pr(\tilde{\mathbf{e}}_{\mathsf{LWQ}} \notin \mathcal{V}_E) = 0$. Perfect rate follows directly from $|\mathsf{pt}| = |\mathsf{ct}| = \log_2\left(\frac{q^m}{\det(\Lambda)}\right)$. □

It has been shown in [35] that $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$ improves the information rate of LWE_{E,χ_e} from $R = 1 - f(m)$ to $1 - \frac{f(m)}{\log_2 m}$, achieving a logarithmic dimensional advantage. Nevertheless, $R = 1 - o\left(\frac{1}{\log_2(q)}\right)$ is impossible within the quantized LWE framework [35, Bound 2].

Theorem 8 shows that LWQ-based encryption can eliminate the $o(1)$ term and achieve a perfect rate of $R = 1$. This improvement is attributed to a key distinction: in LWE-based or LWEQ-based schemes, the $o(1)$ term stems from the additive noise $\mathbf{e}$, which is introduced solely for security purposes. In contrast, LWQ fully utilizes the entire ciphertext error for compression, thereby eliminating this overhead.

6 Conclusions and Open Questions

This paper introduces a new hardness assumption termed LWQ for lattice-based cryptography. By combining the security guarantees of LWE with the efficiency of public dithering via XOFs, LWQ enables primitives that achieve both provable security and practical efficiency. Our results validate two concrete instantiations: Scalar-LWQ enhances LWR's security without sacrificing its storage efficiency, while polar-LWQ optimizes LWE's storage efficiency while retaining its security guarantees.

To reduce the bandwidth of LWE-based applications, algebraic variants of LWE has been developed, including Ring-LWE [32], Module-LWE [28], Middle-Product-LWE [6], and Cyclic-LWE [25]. These variants offer more compact representations and faster arithmetic operations, making them more suitable for practical implementations. Future research could explore the extension of LWQ to its algebraic counterparts.

Although our sub-exponential bound for LWQ in Theorem 4 is significantly tighter than the polynomial bound for LWR (with a polynomial modulus q), we have not achieved an exponential bound, which would be ideal for practical cryptographic applications. This appears to be an inherent limitation of polar codes when analyzed under statistical distance or Kullback-Leibler (KL) divergence. One potential approach to overcome this limitation is to use the Rényi divergence, as a small bound on Rényi divergence is sufficient in many

cases [7]. However, constructing polar codes under Rényi divergence remains an open problem in coding theory, to the best of our knowledge. We encourage further research efforts to address this challenge.

Our analysis of LWQ reveals an interesting phenomenon: in the security analysis of lattice-based cryptosystems involving LWR or LWER, quantization noise is often approximated as a discrete Gaussian with the same variance. Consequently, a quantization lattice with a larger normalized second moment would imply higher security. This, however, contradicts our proposal of using lattices with a small normalized second moment. This observation suggests that the existing security analysis, which models quantization noise as Gaussian in LWR and LWER, may not be tight. We hope our work on LWQ improves the understanding of LWR and LWER and stimulates interest in a tighter analysis.

References

1. Agrell, E., Allen, B.: On the best lattice quantizers. IEEE Transactions on Information Theory **69**(12), 7650–7658 (2023). https://doi.org/10.1109/TIT.2023.3291313
2. Alwen, J., Krenn, S., Pietrzak, K., Wichs, D.: Learning with rounding, revisited - new reduction, properties and applications. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013, Part I. LNCS, vol. 8042, pp. 57–74. Springer, Berlin, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 18–22, 2013). https://doi.org/10.1007/978-3-642-40041-4_4
3. Arıkan, E.: Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Transactions on Information Theory **55**(7), 3051–3073 (July 2009). https://doi.org/10.1109/TIT.2009.2021379
4. Arıkan, E., Telatar, I.: On the rate of channel polarization. In: Proc. 2009 IEEE Int. Symp. Inform. Theory. pp. 1493–1495. Seoul, South Korea (June 2009)
5. Babai, L.: On Lovász' lattice reduction and the nearest lattice point problem. Combinatorica **6**(1), 1–13 (1986). https://doi.org/10.1007/BF02579403
6. Bai, S., Boudgoust, K., Das, D., Roux-Langlois, A., Wen, W., Zhang, Z.: Middle-product learning with rounding problem and its applications. In: Galbraith, S.D., Moriai, S. (eds.) ASIACRYPT 2019, Part I. LNCS, vol. 11921, pp. 55–81. Springer, Cham, Switzerland, Kobe, Japan (Dec 8–12, 2019). https://doi.org/10.1007/978-3-030-34578-5_3
7. Bai, S., Langlois, A., Lepoint, T., Stehlé, D., Steinfeld, R.: Improved security proofs in lattice-based cryptography: Using the rényi divergence rather than the statistical distance. In: Iwata, T., Cheon, J.H. (eds.) Advances in Cryptology – ASIACRYPT 2015. pp. 3–24. Springer Berlin Heidelberg, Berlin, Heidelberg (2015)
8. Banerjee, A., Peikert, C., Rosen, A.: Pseudorandom functions and lattices. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 719–737. Springer, Berlin, Heidelberg, Germany, Cambridge, UK (Apr 15–19, 2012). https://doi.org/10.1007/978-3-642-29011-4_42
9. Barnes, E., Sloane, N.: The optimal lattice quantizer in three dimensions. SIAM Journal on Algebraic Discrete Methods **4**(1), 30–41 (1983)
10. Bogdanov, A., Guo, S., Masny, D., Richelson, S., Rosen, A.: On the hardness of learning with rounding over small modulus. In: Kushilevitz, E., Malkin, T. (eds.) TCC 2016-A, Part I. LNCS, vol. 9562, pp. 209–224. Springer, Berlin, Heidelberg,

Germany, Tel Aviv, Israel (Jan 10–13, 2016). https://doi.org/10.1007/978-3-662-49096-9_9
11. Bos, J., Costello, C., Ducas, L., Mironov, I., Naehrig, M., Nikolaenko, V., Raghunathan, A., Stebila, D.: Annex on FrodoKEM updates, April 18, 2023 version (PDF) (2023), https://frodokem.org/files/FrodoKEM-annex-20230418.pdf
12. Brakerski, Z., Döttling, N., Garg, S., Malavolta, G.: Leveraging linear decryption: Rate-1 fully-homomorphic encryption and time-lock puzzles. In: Hofheinz, D., Rosen, A. (eds.) TCC 2019, Part II. LNCS, vol. 11892, pp. 407–437. Springer, Cham, Switzerland, Nuremberg, Germany (Dec 1–5, 2019). https://doi.org/10.1007/978-3-030-36033-7_16
13. Brakerski, Z., Langlois, A., Peikert, C., Regev, O., Stehlé, D.: Classical hardness of learning with errors. In: Boneh, D., Roughgarden, T., Feigenbaum, J. (eds.) 45th ACM STOC. pp. 575–584. ACM Press, Palo Alto, CA, USA (Jun 1–4, 2013). https://doi.org/10.1145/2488608.2488680
14. Cheon, J.H., Kim, A., Kim, M., Song, Y.S.: Homomorphic encryption for arithmetic of approximate numbers. In: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017, Part I. LNCS, vol. 10624, pp. 409–437. Springer, Cham, Switzerland, Hong Kong, China (Dec 3–7, 2017). https://doi.org/10.1007/978-3-319-70694-8_15
15. Cheon, J.H., Kim, D., Lee, J., Song, Y.: Lizard: Cut off the tail! A practical post-quantum public-key encryption from LWE and LWR. In: Catalano, D., De Prisco, R. (eds.) SCN 18. LNCS, vol. 11035, pp. 160–177. Springer, Cham, Switzerland, Amalfi, Italy (Sep 5–7, 2018). https://doi.org/10.1007/978-3-319-98113-0_9
16. Cheon, J.H., Kim, D., Lee, J., Song, Y.: Lizard: Cut off the tail! A practical post-quantum public-key encryption from LWE and LWR. In: Catalano, D., Prisco, R.D. (eds.) Security and Cryptography for Networks - 11th International Conference, SCN 2018, Amalfi, Italy, September 5-7, 2018, Proceedings. Lecture Notes in Computer Science, vol. 11035, pp. 160–177. Springer (2018). https://doi.org/10.1007/978-3-319-98113-0_9, https://doi.org/10.1007/978-3-319-98113-0_9
17. Cohn, H., Kumar, A., Miller, S., Radchenko, D., Viazovska, M.: The sphere packing problem in dimension 24. Annals of Mathematics **185**(3), 1017–1033 (2017). https://doi.org/10.4007/annals.2017.185.3.8
18. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups. Springer, New York, 3 edn. (1999). https://doi.org/10.1007/978-1-4757-6568-7
19. Cover, T.M.: Elements of Information Theory. John Wiley & Sons, Hoboken, New Jersey (1999)
20. D'Anvers, J.P., Karmakar, A., Roy, S.S., Vercauteren, F.: Saber: Module-LWR based key exchange, CPA-secure encryption and CCA-secure KEM. In: Joux, A., Nitaj, A., Rachidi, T. (eds.) AFRICACRYPT 18. LNCS, vol. 10831, pp. 282–305. Springer, Cham, Switzerland, Marrakesh, Morocco (May 7–9, 2018). https://doi.org/10.1007/978-3-319-89339-6_16
21. Egilmez, Z.B.K., Xiang, L., Maunder, R.G., Hanzo, L.: The development, operation and performance of the 5G polar codes. IEEE Communications Surveys & Tutorials **22**(1), 96–122 (2019)
22. Forney, G., Trott, M., Chung, S.Y.: Sphere-bound-achieving coset codes and multilevel coset codes. IEEE Transactions on Information Theory **46**(3), 820–850 (May 2000). https://doi.org/10.1109/18.841165
23. Gentry, C., Peikert, C., Vaikuntanathan, V.: Trapdoors for hard lattices and new cryptographic constructions. In: Ladner, R.E., Dwork, C. (eds.) 40th ACM STOC. pp. 197–206. ACM Press, Victoria, BC, Canada (May 17–20, 2008). https://doi.org/10.1145/1374376.1374407

24. Gentry, C., Sahai, A., Waters, B.: Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013, Part I. LNCS, vol. 8042, pp. 75–92. Springer, Berlin, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 18–22, 2013). https://doi.org/10.1007/978-3-642-40041-4_5
25. Grover, C., Mendelsohn, A., Ling, C., Vehkalahti, R.: Non-commutative ring learning with errors from cyclic algebras. Journal of Cryptology **35**(3), 22 (Jul 2022). https://doi.org/10.1007/s00145-022-09430-6
26. Korada, S.B.: Polar codes for channel and source coding. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (2009)
27. Korada, S., Urbanke, R.: Polar codes are optimal for lossy source coding. IEEE Transactions on Information Theory **56**(4), 1751–1768 (April 2010). https://doi.org/10.1109/TIT.2010.2040961
28. Langlois, A., Stehlé, D.: Worst-case to average-case reductions for module lattices. Designs, Codes and Cryptography **75**(3), 565–599 (2015)
29. Liu, L., Shi, J., Ling, C.: Polar lattices for lossy compression. IEEE Transactions on Information Theory **67**(9), 6140–6163 (2021), https://doi.org/10.1109/TIT.2021.3097965
30. Liu, L., Yan, Y., Ling, C., Wu, X.: Construction of capacity-achieving lattice codes: Polar lattices. IEEE Trans. Commun. **67**(2), 915–928 (Feb 2019)
31. Liu, S., Sakzad, A.: Crystals-kyber with lattice quantizer. In: 2024 IEEE International Symposium on Information Theory (ISIT). pp. 2886–2891 (2024). https://doi.org/10.1109/ISIT57864.2024.10619497
32. Lyubashevsky, V., Peikert, C., Regev, O.: On ideal lattices and learning with errors over rings. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 1–23. Springer, Berlin, Heidelberg, Germany, French Riviera (May 30 – Jun 3, 2010). https://doi.org/10.1007/978-3-642-13190-5_1
33. Micciancio, D., Peikert, C.: Trapdoors for lattices: Simpler, tighter, faster, smaller. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 700–718. Springer, Berlin, Heidelberg, Germany, Cambridge, UK (Apr 15–19, 2012). https://doi.org/10.1007/978-3-642-29011-4_41
34. Micciancio, D., Regev, O.: Worst-case to average-case reductions based on Gaussian measures. In: 45th FOCS. pp. 372–381. IEEE Computer Society Press, Rome, Italy (Oct 17–19, 2004). https://doi.org/10.1109/FOCS.2004.72
35. Micciancio, D., Schultz, M.: Error correction and ciphertext quantization in lattice cryptography. In: Handschuh, H., Lysyanskaya, A. (eds.) CRYPTO 2023, Part V. LNCS, vol. 14085, pp. 648–681. Springer, Cham, Switzerland, Santa Barbara, CA, USA (Aug 20–24, 2023). https://doi.org/10.1007/978-3-031-38554-4_21
36. Mori, R., Tanaka, T.: Source and channel polarization over finite fields and reed–solomon matrices. IEEE Transactions on Information Theory **60**(5), 2720–2736 (2014). https://doi.org/10.1109/TIT.2014.2312181
37. Newton, P., Richelson, S.: A lower bound for proving hardness of learning with rounding with polynomial modulus. In: Handschuh, H., Lysyanskaya, A. (eds.) CRYPTO 2023, Part V. LNCS, vol. 14085, pp. 805–835. Springer, Cham, Switzerland, Santa Barbara, CA, USA (Aug 20–24, 2023). https://doi.org/10.1007/978-3-031-38554-4_26
38. Park, W., Barg, A.: Polar codes for q-ary channels, $q = 2^r$. IEEE Transactions on Information Theory **59**(2), 955–969 (2013). https://doi.org/10.1109/TIT.2012.2219035

39. Pedarsani, R., Hassani, S., Tal, I., Telatar, I.: On the construction of polar codes. In: Proc. 2011 IEEE Int. Symp. Inform. Theory. pp. 11–15. St. Petersburg, Russia (July 2011). https://doi.org/10.1109/ISIT.2011.6033724
40. Peikert, C.: Public-key cryptosystems from the worst-case shortest vector problem: extended abstract. In: Mitzenmacher, M. (ed.) 41st ACM STOC. pp. 333–342. ACM Press, Bethesda, MD, USA (May 31 – Jun 2, 2009). https://doi.org/10.1145/1536414.1536461
41. Peikert, C.: An efficient and parallel Gaussian sampler for lattices. In: Rabin, T. (ed.) CRYPTO 2010. LNCS, vol. 6223, pp. 80–97. Springer, Berlin, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 15–19, 2010). https://doi.org/10.1007/978-3-642-14623-7_5
42. Peikert, C., Shiehian, S.: Noninteractive zero knowledge for NP from (plain) learning with errors. In: Boldyreva, A., Micciancio, D. (eds.) CRYPTO 2019, Part I. LNCS, vol. 11692, pp. 89–114. Springer, Cham, Switzerland, Santa Barbara, CA, USA (Aug 18–22, 2019). https://doi.org/10.1007/978-3-030-26948-7_4
43. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. In: Gabow, H.N., Fagin, R. (eds.) 37th ACM STOC. pp. 84–93. ACM Press, Baltimore, MA, USA (May 22–24, 2005). https://doi.org/10.1145/1060590.1060603
44. Sasoglu, E.: Polar codes for discrete alphabets. In: 2012 IEEE International Symposium on Information Theory Proceedings. pp. 2137–2141 (2012). https://doi.org/10.1109/ISIT.2012.6283740
45. Sasoglu, E.: Polarization and polar codes. Foundations and Trends in Communications and Information Theory **8**(4), 259–381 (2012). https://doi.org/10.1561/0100000041
46. Sasoglu, E., Telatar, E., Arikan, E.: Polarization for arbitrary discrete memoryless channels. In: 2009 IEEE Information Theory Workshop. pp. 144–148 (2009). https://doi.org/10.1109/ITW.2009.5351487
47. Schwabe, P., Avanzi, R., Bos, J., Ducas, L., Kiltz, E., Lepoint, T., Lyubashevsky, V., Schanck, J.M., Seiler, G., Stehlé, D., Ding, J.: CRYSTALS-KYBER. Tech. rep., National Institute of Standards and Technology (2022), available at https://csrc.nist.gov/Projects/post-quantum-cryptography/selected-algorithms-2022
48. Tal, I., Vardy, A.: How to construct polar codes. IEEE Transactions on Information Theory **59**(10), 6562–6582 (Oct 2013). https://doi.org/10.1109/TIT.2013.2272694
49. Viazovska, M.S.: The sphere packing problem in dimension 8. Annals of Mathematics pp. 991–1015 (2017). https://doi.org/10.4007/annals.2017.185.3.7
50. Wang, H.P., Duursma, I.M.: Log-logarithmic time pruned polar coding. IEEE Transactions on Information Theory **67**(3), 1509–1521 (2021). https://doi.org/10.1109/TIT.2020.3041523
51. Zamir, R.: Lattice Coding for Signals and Networks. Cambridge University Press, Cambridge, UK (2014)
52. Zamir, R., Feder, M.: Information rates of pre/post-filtered dithered quantizers. IEEE Transactions on Information Theory **42**(5), 1340–1353 (1996). https://doi.org/10.1109/18.532876
53. Zamir, R., Feder, M.: On lattice quantization noise. IEEE Transactions on Information Theory **42**(4), 1152–1159 (1996). https://doi.org/10.1109/18.508838
54. Zamir, R., Feder, M.: On lattice quantization noise. IEEE Transactions on Information Theory **42**(4), 1152–1159 (1996)

A Background of Polar Codes/Lattices

A.1 Polar Codes

Polar coding [3] presents arguably the first explicit construction of codes that are capacity-achieving for any binary-input memoryless symmetric channels (BMSCs). Let us break down the concept:

- **BMSC and Polar Code**: A BMSC is a type of communication channel characterized by binary input and output without memory of previous inputs. A polar code is designed specifically for such channels and achieves their capacity.
- **Block Length and Generator Matrix**: For a given BMSC, we construct a polar code with block length $m = 2^t$, where t is a non-negative integer. The polar code employs a generator matrix G_m, derived by iteratively applying the Kronecker product to the base matrix $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$.
- **Information Set and Frozen Set**: Among the rows of the generator matrix G_m, we select K specific rows to form the information set $\mathcal{I}$. The remaining rows constitute the frozen set $\mathcal{F}$. The information set comprises positions used for encoding actual data, whereas the frozen set includes positions predetermined to facilitate decoding.
- **Channel Combination and Polarization Transform**: We consider N identical copies of the BMSC, denoted W_m, which process input vectors $X^{[m]}$ to yield output vectors $Y^{[m]}$. By applying the generator matrix G_m to the input, we obtain $U^{[m]} = X^{[m]}G_m$. This transformation decomposes the channel into m simpler subchannels.
- **Subchannels and Polarization**: Each subchannel $W_m^{(i)}$ processes part of the transformed input U^i and produces output based on the entire output vector $Y^{[m]}$ and previous parts of the transformed input $U^{1:i-1}$. As m (the block length) increases indefinitely, these subchannels polarize into either very reliable (almost error-free) or very unreliable (ineffective for communication).
- **Good Subchannels and Capacity**: Through channel polarization, we can identify the good subchannels. The proportion of good subchannels approaches the channel's capacity C as the block length m becomes large. Hence, to achieve capacity, the K rows selected for encoding should correspond to these good subchannels.

Example 1. When $m = 2$, the generator matrix for binary polar codes is given by $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. One may use one ($r = 1$) partition level $\mathbb{Z}/2\mathbb{Z}$ and choose $[1, 1]$ as the basis for C_1. Therefore, the polar lattice is made by $[1, 1] \cdot U_1 + 2\mathbb{Z}^2$, where U_1 is the information bit of C_1. The generator matrix of the 2-dimensional polar lattice is given by $\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$, which is indeed the famous checkerboard lattice D_2.

Example 2. When $m = 4$, the generator matrix for binary polar codes is given by

$$\begin{bmatrix} 1\,0\,0\,0 \\ 1\,1\,0\,0 \\ 1\,0\,1\,0 \\ 1\,1\,1\,1 \end{bmatrix}.$$

One may use two partition levels $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}$ and construct two binary polar codes according to the Construction D method. For the first level, one may choose $[1,1,1,1]$ as the basis for C_1. For the second level, C_2 can have bases $[1,1,0,0]$, $[1,0,1,0]$ and $[1,1,1,1]$. Clearly, $C_1 \subset C_2$. Therefore, the polar lattice is made by $[1,1,1,1]\cdot U_1 + 2\cdot[1,1,0,0]\cdot U_2 + 2\cdot[1,1,0,0]\cdot U_3 + 2\cdot[1,1,0,0]\cdot U_4 + 4\mathbb{Z}^4$, where U_1 is the information bit of C_1 and U_2^4 are the information bits of C_2. Consequently, the generator matrix of the 4-dimensional polar lattice is given by

$$\begin{bmatrix} 4\,0\,0\,0 \\ 2\,2\,0\,0 \\ 2\,0\,2\,0 \\ 1\,1\,1\,1 \end{bmatrix}.$$

A.2 Successive Cancellation Decoding Algorithm

We briefly describe the successive cancellation (SC) decoding algorithm in this section. For simplicity, we assume that a polar code is constructed according to the binary-input test channel $W : X \to Y$. Denote the source sequence by $y^{[m]}$ and its reconstruction by $x^{[m]}$. Let $x^{[m]} = u^{[m]}G_m$, with the frozen bits being fixed as $u^{\mathcal{F}}$. The task of the decoder is to determine the estimation $\widehat{u}^{[m]}$ based on $y^{[m]}$. Given the channel transition probability $P_{Y|X}$ of W, the i-th polarized subchannel channel is defined by the following probability

$$W_m^{(i)}(y^{[m]}, u^{1:i-1}|u^i) \triangleq \sum_{u^{i+1:m}} \frac{1}{2^{m-1}} W_m(y^{[m]}|u^{[m]}),$$

where $W_m(y^{[m]}|u^{[m]}) = P_{Y^{[m]}|X^{[m]}}(y^{[m]}|u^{[m]}G_m)$.

The SC decoder uses Log-Likelihood Ratio (LLR) $L_m^{(i)}\left(y^{[m]}, \widehat{u}_1^{i-1}\right) \triangleq \log_2 \frac{W_N^{(i)}\left(y^{[m]}, \hat{u}^{1:i-1}|0\right)}{W_N^{(i)}\left(y^{[m]}, \widehat{u}^{1:i-1}|1\right)}$ to decide the value of $\widehat{u}^i$ by the decision rule.

$$\widehat{u}^i = \begin{cases} u^i, & i \in \mathcal{F} \\ 0, & i \in \mathcal{I}, L_m^{(i)}\left(y^{[m]}, \widehat{u}^{1:i-1}\right) \geq 0 \\ 1, & i \in \mathcal{I}, L_m^{(i)}\left(y^{[m]}, \widehat{u}^{1:i-1}\right) < 0 \end{cases} \tag{35}$$

The LLR can be calculated recursively by

$$\begin{aligned} &L_m^{(2i-1)}\left(y^{[m]}, \widehat{u}^{1:2i-2}\right) \\ &= \mathsf{f}\Big(L_{m/2}^{(i)}(y^{[m/2]},\ \hat{u}_o^{1:2i-2} \oplus \hat{u}_e^{1:2i-2}),\ L_{m/2}^{(i)}(y^{m/2+1:m},\ \hat{u}_e^{1:2i-2})\Big) \end{aligned} \tag{36}$$

and

$$L_m^{(2i)}\left(y^{[m]}, \widehat{u}^{1:2i-1}\right) = \mathtt{g}\Big(L_{m/2}^{(i)}(y^{[m/2]},\ \hat{u}_o^{1:2i-2} \oplus \hat{u}_e^{1:2i-2}),\ L_{m/2}^{(i)}(y^{m/2+1:m},\ \hat{u}_e^{1:2i-2}),\ \hat{u}^{1:2i-1}\Big), \tag{37}$$

where $\hat{u}_o^{1:2i-2}$ and $\hat{u}_e^{1:2i-2}$ are subvectors of $\hat{u}^{1:2i}$ with odd and even indices respectively, $\mathtt{f}(a,b) \triangleq \log_2\left(\frac{e^{a+b}+1}{e^a+e^b}\right)$ and $\mathtt{g}(a,b,\lambda) \triangleq (-1)^\lambda a + b$. Note that the SC decision rule (35) is deterministic and it can be modified to the following random rounding version for more convenient analysis.

$$\widehat{u}^i = \begin{cases} u^i, & i \in \mathcal{F} \\ 0, & i \in \mathcal{I},\ \text{w. p. } \exp(L_m^{(i)})/\left(1+\exp(L_m^{(i)})\right) \\ 1, & i \in \mathcal{I},\ \text{w. p. } 1/\left(1+\exp(L_m^{(i)})\right) \end{cases} \tag{38}$$

The performance difference between the two decision rules is marginal, as observed in [26].

A.3 Quantization Based on Polar Lattices

Quantization and error correction are duals in the sense that: i) Error correction involves finding the closest lattice point to a noisy codeword, leveraging redundancy to correct errors. ii) Quantization involves mapping the input vector to the nearest lattice point, effectively reducing data resolution and removing redundancy. Consider error correction using Λ, generated by a basis matrix $\mathbf{B}$:

$$\Lambda = \{\mathbf{Bz} \mid \mathbf{z} \in \mathbb{Z}^m\}.$$

Error correction consists of two phases:

- *Encoding*: $\mathbf{c} = \mathbf{Bm}$ for message $\mathbf{m}$.
- *Decoding*: Given an additive noise channel $\mathbf{r} = \mathbf{c} + \mathbf{e}$, find $\mathbf{c} \in \Lambda$ such that $\|\mathbf{r} - \mathbf{c}\|$ is minimized.

Quantization also consists of two phases:

- *Quantizing*: Given $\mathbf{x} \in \mathbb{R}^n$, find $\mathbf{q} \in \Lambda$ such that $\|\mathbf{x} - \mathbf{q}\|$ is minimized.
- *Indexing*: $\mathbf{m} = \mathbf{B}^{-1}\mathbf{q}$.

Polar lattices [29] offer an efficient solution for achieving the rate-distortion bound for the i.i.d. Gaussian source. In essence, one constructs a polar lattice for the Gaussian source by utilizing a series of nested polar codes, as introduced by Forney *et al.* [22]. These polar codes compress the Gaussian source vector based on the characteristics of the test channel at each level. Moreover, research [30] indicates that employing a binary lattice partition keeps the number of levels

r relatively small ($r = O(\log_2 \log_2 m)$), yet still enables the attainment of the capacity $\frac{1}{2}\log(1+\mathsf{SNR})$ of the additive white Gaussian noise (AWGN) channel, where SNR represents the signal-to-noise ratio.

The concept of duality between source coding and channel coding allows us to interpret quantization polar lattices as analogous to a channel coding lattice constructed on the test channel [29]. In the scenario of a Gaussian source with variance σ_s^2 and an average distortion Δ, the test channel effectively becomes an AWGN channel with a noise variance of Δ. Consequently, the SNR of this test channel equals $\frac{\sigma_s^2-\Delta}{\Delta}$, while its capacity is $\frac{1}{2}\log_2\left(\frac{\sigma_s^2}{\Delta}\right)$. This insight suggests that the rate of the polar lattice quantizer can be finely adjusted to approach $\frac{1}{2}\log_2\left(\frac{\sigma_s^2}{\Delta}\right)$. Consequently, polar lattices demonstrate the capability to achieve the rate-distortion bound of Gaussian sources by employing discrete Gaussian distribution instead of continuous, offering a notable advancement in compression techniques.

A.4 Λ/Λ' Channel

A Λ/Λ' channel is defined according to a lattice partition Λ/Λ', as mentioned in Definition 12. Let $X \to Y$ denote the Λ/Λ' channel subject to Gaussian noise with variance σ^2. The input alphabet of X is restricted to the discrete set $(\Lambda+a)\cap\mathcal{P}_{\Lambda'}$, that is, the elements of a translate $\Lambda+a$ of the lattice Λ that fall in a fundamental region $\mathcal{P}_{\Lambda'}$ of Λ'. The specific choice of offset a does not affect the essence of the Λ/Λ' channel [22]. Then, the output Y is written as $Y = X + E \mod \Lambda'$, where E is the Gaussian noise with zero mean and variance σ^2. Since Λ is the union of the $|\Lambda/\Lambda'|$ cosets of Λ', the size of the input alphabet is $|\Lambda/\Lambda'|$. In particular, for the $\mathbb{Z}/q\mathbb{Z}$ channel, the partition order $|\mathbb{Z}/q\mathbb{Z}| = q$ and X can be chosen from any q distinct integers in $\mathcal{P}_{q\mathbb{Z}}$ when one sets $a = 0$.

By the regularity of the Λ/Λ' channel [22, Theorem 4], the Λ/Λ' channel is symmetric and the mutual information $I(X;Y)$ for both uniform X and Y is the capacity $C(\Lambda/\Lambda', \sigma^2)$ of the Λ/Λ' channel.

The capacity $C(\Lambda/\Lambda', \sigma^2)$ equals the gap between the capacity of the mod-Λ' channel and that of the mod-Λ channel [22], that is,

$$C(\Lambda/\Lambda', \sigma^2) = C(\Lambda', \sigma^2) - C(\Lambda, \sigma^2), \tag{39}$$

where $C(\Lambda, \sigma^2) \triangleq \log_2(V(\Lambda)) - h(\Lambda, \sigma^2)$ and $h(\Lambda, \sigma^2)$ is the entropy of the Λ-aliased Gaussian noise, i.e., [7]

$$h(\Lambda, \sigma^2) \triangleq -\int_{\mathcal{V}_\Lambda} f_{\sigma,\Lambda}(x) \log f_{\sigma,\Lambda}(x) dx, \tag{40}$$

where $\mathcal{V}_\Lambda$ denotes the Voronoi region of lattice Λ and the Λ-periodic function $f_{\sigma,\Lambda}(x)$ is defined as

$$f_{\sigma,\Lambda}(x) \triangleq \sum_{\lambda\in\Lambda} f_{\sigma,\lambda}(x) = \frac{1}{(\sqrt{2\pi}\sigma)^{n_\Lambda}} \sum_{\lambda\in\Lambda} e^{-\frac{\|x-\lambda\|^2}{2\sigma^2}} \tag{41}$$

[7] For discrete Gaussian noise, the integration in (40) is interpreted as summattion.

where n_Λ is the dimension of Λ. $C(\Lambda', \sigma^2)$ is defined in the same manner.

The capacity $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$ can be calculated by substituting $\Lambda = \mathbb{Z}$, $\Lambda' = q\mathbb{Z}$, and $n_\Lambda = 1$. For the parameters $q = \text{poly}(n)$ and $\sigma = \Omega(\sqrt{n})$ in LWE, the $\mathbb{Z}$-aliased Gaussian noise is almost uniform, while the $q\mathbb{Z}$-aliased Gaussian noise is almost Gaussian. Therefore, $C(\mathbb{Z}, \sigma^2) \approx 0$, while $C(q\mathbb{Z}, \sigma^2) \approx \log_2 q - \log_2\left(\sqrt{2\pi e} \cdot \sigma\right)$. As a result, $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2) \approx \log_2 q - \log_2\left(\sqrt{2\pi e} \cdot \sigma\right)$.

A.5 Polar Lattices Based on q-ary Polar Codes and Construction A

Here we provide an alternative construction of polar lattice constructed from q-ary polar codes and Construction A. Recall that $q = 2^r$, and we can construct q-ary polar codes based on the same generator matrix G_m in $GF(q)$ [46, 38]. The polarization effect remains when the underlying finite field moves from $GF(2)$ to $GF(q)$. As mentioned in Remark 7, the 2×2 polarization kernel $\left[\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right]$ can be replaced by a larger $\ell \times \ell$ kernel with elements in $GF(q)$ for better performance. Let $\mathcal{C}$ denote a q-ary polar code with block length m. The lattice Λ generated from $\mathcal{C}$ according to Constrction A is defined as

$$\Lambda \triangleq \{\lambda \in \mathbb{Z}^{[m]} : \lambda \bmod q\mathbb{Z} \in \mathcal{C}\}.$$

The decoding complexity of the Construction-A polar lattice is generally higher than the Construction-D polar lattice presented earlier. However, its construction is easier to understand as there are no nested codes involved.

B Hardness of LWQ with general quantization lattices

Beyond the specific case of polar lattices, we establish the following result for LWQ that applies to general quantization lattices, including hypercubic lattices $\mathbb{Z}^m$ and root lattices like the Gosset lattice E_8.

LWQ and Uniform-Noise LWE

Lemma 3 (Dithering lemma, adapted from [51]). *Let $\Lambda \subset \mathbb{R}^m$ be a lattice with fundamental cell $\mathcal{P}_\Lambda$ and Voronoi region $\mathcal{V}_\Lambda$.*

1. *Continuous Case: If $U \sim \mathcal{U}(\mathcal{P}_\Lambda)$ with density*

$$f_U(u) = \begin{cases} |\mathcal{P}_\Lambda|^{-1}, & u \in \mathcal{P}_\Lambda \\ 0, & \textit{otherwise} \end{cases},$$

   then the quantization error $\mathbf{e}_Q = Q_{\Lambda+U}(\mathbf{y}) - \mathbf{y}$ is uniformly distributed over $\mathcal{V}_\Lambda$, independent of $\mathbf{y} \in \mathbb{R}^m$.
2. *Discrete Case: If $U \sim \mathcal{U}(\mathbb{Z}^m/\Lambda)$, then $\mathbf{e}_Q = Q_{\Lambda+U}(\mathbf{y}) - \mathbf{y}$ is uniformly distributed over $\mathcal{V}_\Lambda \cap \mathbb{Z}^m$, independent of $\mathbf{y} \in \mathbb{Z}^m$.*

Proof. Employ the fact that $Q_{\Lambda+U}(y)$ is uniform over the coset $\Lambda + U$. We refer to [51] for a complete proof. □

Lemma 3 shows that uniform dithering over the fundamental cell "washes out" dependence on the input, and makes the quantization noise $\mathbf{e}_Q$ behave like a uniformly random error.

Theorem 9 (LWQ $\approx_s$ uniform noise LWE). *Let $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$ be a uniform dither. Then, the LWQ distribution is statistically indistinguishable from the LWE distribution with uniform noise over $\mathcal{V}_\Lambda \cap \mathbb{Z}^m$.*

Proof. By the dithering lemma, when $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$, the quantization error:

$$\mathbf{e}_Q = Q_\Lambda(\mathbf{As} - \mathbf{d}) - (\mathbf{As} - \mathbf{d})$$

is uniformly distributed over $\mathcal{V}_\Lambda \cap \mathbb{Z}^m$ and independent of $\mathbf{As}$.

Rewriting the LWQ sample as $(\mathbf{A}, \mathbf{b}) = Q_\Lambda(\mathbf{As} - \mathbf{d}) + \mathbf{d} = \mathbf{As} + \mathbf{e}_Q)$, this matches the LWE distribution $(\mathbf{A}, \mathbf{As} + \mathbf{e})$, where $\mathbf{e} \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$. Thus the LWQ distribution is identical to LWE with uniform noise and their statistical distance is zero:

$$\Delta\left(\mathrm{LWQ}_{\Lambda,\mathbf{d}}, \mathrm{LWE}_{\mathrm{Unif}(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)}\right) = 0.$$

□

The equivalence between LWQ and uniform-noise LWE holds for any quantization lattice Λ. A particularly important case arises when $\Lambda = \frac{q}{p}\mathbb{Z}^m$, as this generalizes the widely-used LWR problem while achieving stronger security guarantees. We formalize this special case below; the proof follows directly from Theorem 9 and is therefore omitted.

Corollary 2 (LWQ with $\Lambda = \frac{q}{p}\mathbb{Z}^m$). *Let $\Lambda = \frac{q}{p}\mathbb{Z}^m$ where $p \mid q$, and let $\mathbf{d} \leftarrow \mathbb{Z}^m_{q/p}$. The scalar-LWQ distribution*

$$\mathsf{LWQ}_{\frac{q}{p}\mathbb{Z}^m,\mathbf{d}} : (\mathbf{A}, Q_{\frac{q}{p}\mathbb{Z}^m}(\mathbf{As} - \mathbf{d}) + \mathbf{d})$$

is statistically indistinguishable from the LWE distribution with uniform noise over $\mathbb{Z}^m_{q/p}$.

This corollary highlights two key advantages of scalar-LWQ over LWR: Unlike LWR's deterministic rounding error, scalar-LWQ's dithering ensures independent noise uniformity over $\mathbb{Z}^m_{q/p}$, enabling direct security reductions to uniform noise LWE. Meanwhile, scalar-LWQ matches LWR's representation of $\log_2(p^m)$ bits for $\mathbf{b}$ when aided by using XOF for $(\mathbf{A}, \mathbf{d})$.

LWQ and Gaussian-Noise LWE

Definition 17 (Smoothing parameter [34]). *For any lattice Λ and positive real $\varepsilon > 0$, the smoothing parameter $\eta_\varepsilon(\Lambda)$ is the smallest real $\kappa > 0$ such that $\rho_{1/\kappa}(\tilde{\Lambda} \setminus \{0\}) \leq \varepsilon$ where $\tilde{\Lambda}$ is the dual lattice.*

The smoothing parameter $\eta_\varepsilon(\Lambda)$ quantifies the smallest Gaussian width κ needed to smooth out the discrete structure of a lattice Λ, making its discrete Gaussian distribution behave like a continuous one.

Lemma 4 (Gaussian mass bound, [34]). *For any lattice Λ and $\mathbf{c} \in \mathbb{R}^m$, $\varepsilon > 0$, and $\kappa \geq \eta_\varepsilon(\Lambda)$,*

$$\rho_r(\Lambda + \mathbf{c}) \in \kappa^m \det(\tilde{\Lambda})(1 \pm \varepsilon). \tag{42}$$

The discrete NSM is defined as: $\bar{G}(\Lambda) = \bar{\gamma}^2(\Lambda)/\det^{2/m}(\Lambda)$, where $\bar{\gamma}^2(\Lambda) = \frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m} \det(\Lambda)^{-1}\|\mathbf{x}\|^2$. Analogous to the continuous NSM $G(\Lambda)$, $\bar{G}(\Lambda)$ quantifies how "sphere-like" Λ is for quantization. From the high-resolution assumption [51], we have $G(\Lambda) \approx \bar{G}(\Lambda)$, and the approximation error can be arbitrarily small by increasing $|\mathcal{V}_\Lambda \cap \mathbb{Z}^m|$.

Theorem 10 (Distance to Gaussian noise LWE). *Define $\kappa = \sqrt{2\pi\bar{G}(\Lambda)}\det^{1/m}(\Lambda)$. If $\kappa \geq \eta_\varepsilon(\mathbb{Z}^m)$, the KL divergence between LWQ and LWE satisfies:*

$$D_{KL}((\mathbf{A}, \mathbf{As} + \mathbf{e}_Q)\|(\mathbf{A}, \mathbf{As} + \mathbf{e})) \in \frac{m}{2}\log_2(2\pi e\bar{G}(\Lambda)) + \log_2(1 \pm \varepsilon). \tag{43}$$

where the quantization noise $\mathbf{e}_Q \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$ and the discrete Gaussian noise $\mathbf{e} \sim \mathcal{D}_{\mathbb{Z}^m,\kappa}$.

Proof. From Lemma 4, we have

$$\rho_\kappa(\mathbb{Z}^m) \in \kappa^m(1 \pm \varepsilon). \tag{44}$$

Then we have

$$\frac{1}{m}D_{KL}(\mathbf{e}_Q\|\mathbf{e}) = \frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m} \det(\Lambda)^{-1}\log_2\frac{\rho_\kappa(\mathbb{Z}^m)}{\det(\Lambda)\rho_\kappa(\mathbf{x})} \tag{45}$$

$$\in \frac{1}{m}\log_2\frac{\kappa^m(1\pm\varepsilon)}{\det(\Lambda)} + \log_2 e\cdot\frac{\pi}{\kappa^2}\cdot\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m}\det(\Lambda)^{-1}\|\mathbf{x}\|^2 \tag{46}$$

$$= \frac{1}{2}\log_2\frac{\kappa^2}{\det(\Lambda)^{2/m}} + \log_2 e\cdot\frac{\pi}{\kappa^2}\cdot\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m}\det(\Lambda)^{-1}\|\mathbf{x}\|^2 + \frac{1}{m}\log_2(1\pm\varepsilon). \tag{47}$$

By setting the discrete un-normalized second moment as the Gaussian variance:

$$\kappa^2 = 2\pi\bar{\gamma}^2(\Lambda) = 2\pi\cdot\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m}\det(\Lambda)^{-1}\|\mathbf{x}\|^2, \tag{48}$$

where $\bar{\gamma}^2(\Lambda) = \bar{G}(\Lambda)\det^{2/m}(\Lambda)$, we obtain

$$D_{KL}(\mathbf{e}_Q\|\mathbf{e}) \in \frac{m}{2}\log_2(2\pi e\bar{G}(\Lambda)) + \log_2(1\pm\varepsilon). \tag{49}$$

Therefore, we can bound the divergence of LWQ and LWE by using

$$D_{KL}((\mathbf{A}, \mathbf{As} + \mathbf{e}_Q)\|(\mathbf{A}, \mathbf{As} + \mathbf{e})) = D_{KL}(\mathbf{e}_Q\|\mathbf{e}). \tag{50}$$

□

From Theorem 10, the distinguishing advantage between LWQ and LWE satisfies:

$$\mathsf{Adv}_{\mathsf{LWQ\text{-}LWE}} \le \sqrt{\frac{\ln 2}{2}\left(\frac{m}{2}\log_2(2\pi e\bar{G}(\Lambda)) + \log_2(1+\varepsilon)\right)}. \tag{51}$$

The NSM $\bar{G}(\Lambda)$ quantifies the security loss of LWQ relative to standard LWE. Lattices with $\bar{G}(\Lambda) \approx \frac{1}{2\pi e}$ achieve near-optimal security, as they induce small distinguishing advantages. However, this approach does not allow to show diminishing distinguishing advantages: since $\log_2(2\pi e\bar{G}(\Lambda)) = O(\log m/m)$ for optimal lattice quantizers [54, Lemma 1], we have $m\log_2(2\pi e\bar{G}(\Lambda)) = O(\log m) \to \infty$.

C Proof of Lemma 1

Proof. Using the telescoping expansion [27, Lemma 4]

$$B^{1:n} - A^{1:n} = \sum_{i=1}^{n}(B^i - A^i)A^{1:i-1}B^{i+1:n}, \tag{52}$$

$\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)$ can be decomposed as

$$\begin{aligned}
&2\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)\\
&= \sum_{u_1^{[m]},y^{[m]}} \left|\mathsf{Q}(u_1^{[m]},y^{[m]}) - \mathsf{P}(u_1^{[m]},y^{[m]})\right|\\
&= \sum_{u_1^{[m]},y^{[m]}} \left|\sum_i \left(\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]}) - \mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})\right)\right.\\
&\quad \left.\cdot \left(\prod_{j=1}^{i-1}\mathsf{P}(u_1^j|u_1^{1:j-1},y^{[m]})\right)\left(\prod_{j=i+1}^{m}\mathsf{Q}(u_1^j|u_1^{1:j-1},y^{[m]})\right)\mathsf{P}\left(y^{[m]}\right)\right| \qquad (53)
\end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{\le} \sum_{i\in\mathcal{F}_1}\sum_{u_1^{[m]},y^{[m]}} \left|\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]}) - \mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})\right| \left(\prod_{j=1}^{i-1}\mathsf{P}(u_1^j|u_1^{1:j-1},y^{[m]})\right)\\
&\qquad\qquad \cdot \left(\prod_{j=i+1}^{m}\mathsf{Q}(u_1^j|u_1^{1:j-1},y^{[m]})\right)\mathsf{P}\left(y^{[m]}\right)\\
&= \sum_{i\in\mathcal{F}_1}\sum_{u_1^{1:i},y^{[m]}} \left|\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]}) - \mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})\right| \left(\prod_{j=1}^{i-1}\mathsf{P}(u_1^j|u_1^{1:j-1},y^{[m]})\right)\mathsf{P}\left(y^{[m]}\right)\\
&= \sum_{i\in\mathcal{F}_1}\sum_{u_1^{1:i-1},y^{[m]}} 2\mathsf{P}\left(u_1^{1:i-1},y^{[m]}\right)\Delta\left(\mathsf{Q}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}, \mathsf{P}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}\right)\\
&\overset{(b)}{\le} \sum_{i\in\mathcal{F}_1}\sum_{u_1^{1:i-1},y^{[m]}} \mathsf{P}\left(u_1^{1:i-1},y^{[m]}\right)\sqrt{2\ln 2 D_{KL}\left(\mathsf{P}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}} || \mathsf{Q}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}\right)}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2 \sum_{u_1^{1:i-1},y^{[m]}} \mathsf{P}\left(u_1^{1:i-1},y^{[m]}\right) D_{KL}\left(\mathsf{P}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}||\mathsf{Q}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}\right)} \\
&= \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2 D_{KL}\left(\mathsf{P}_{U_1^i}||\mathsf{Q}_{U_1^i}|U_1^{1:i-1},Y^{[m]}\right)} \\
&\overset{(d)}{=} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2\left(1-H(U_1^i|U_1^{1:i-1},Y^{[m]})\right)} \\
&\overset{(e)}{\leq} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2\left(1-Z(U_1^i|U_1^{1:i-1},Y^{[m]})^2\right)} \\
&\overset{(f)}{\leq} m\sqrt{4\ln 2\cdot 2^{-m^\beta}}
\end{aligned}
$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence, and the equalities and the inequalities follow from

- (a) $\mathsf{Q}\left(u_1^i|u_1^{1:i-1},y^{[m]}\right)=\mathsf{P}\left(u_1^i|u_1^{1:i-1},y^{[m]}\right)$ for $i\in\mathcal{I}_1$.
- (b) Pinsker's inequality.
- (c) Jensen's inequality.
- (d) $\mathsf{Q}\left(u_1^i|u_1^{1:i-1}\right)=\frac{1}{2}$ for $i\in\mathcal{F}_1$.
- (e) $Z(X|Y)^2\leq H(X|Y)$.
- (f) Definition of $\mathcal{F}_1$.

□

D KL Divergence

Lemma 5. *Let $\mathsf{Q}_{U_1^{[m]},Y^{[m]}}$ denote the resulted joint distribution of $U_1^{[m]}$ and $Y^{[m]}$ according to the encoding rules* (26) *and* (27) *at the first partition level. Let $\mathsf{P}_{U_1^{[m]},Y^{[m]}}$ denote the joint distribution directly generated from $\mathsf{P}_{X_1,Y}$, i.e., U_1^i is generated according to the encoding rule* (26) *for all $i\in[m]$. The Kullback-Leibler divergence between $\mathsf{P}_{U_1^{[m]},Y^{[m]}}$ and $\mathsf{Q}_{U_1^{[m]},Y^{[m]}}$ is upper-bounded as follows:*

$$
D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)\leq 2\ln 2\cdot m2^{-m^\beta}. \tag{54}
$$

By induction, after the lattice quantization process with r sequential levels,

$$
\begin{aligned}
&D_{KL}\left(\mathsf{P}_{X^{[m]},Y^{[m]}}\|\mathsf{Q}_{X^{[m]},Y^{[m]}}\right) \\
&\quad= D_{KL}\left(\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}}\|\mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}}\right) &(55)\\
&\quad\leq 2\ln 2\cdot rm2^{-m^\beta}. &(56)
\end{aligned}
$$

Proof. For the 1st level,

$$\begin{aligned}
&D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log_2\frac{\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)}{\mathsf{Q}\left(u_1^{[m]},y^{[m]}\right)}\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log_2\frac{\mathsf{P}\left(u_1^{[m]}|y^{[m]}\right)}{\mathsf{Q}\left(u_1^{[m]}|y^{[m]}\right)}\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log_2\frac{\prod_{i=1}^m\mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})}{\prod_{i=1}^m\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]})} \qquad (57)\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\sum_{i\in\mathcal{F}_1}\log_2\frac{\mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})}{\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]})}\\
&\quad=\ln 2\cdot\sum_{i\in\mathcal{F}_1}(1-H(U_1^i|U_1^{1:i-1},Y^{[m]}))\\
&\quad\le\ln 2\cdot\sum_{i\in\mathcal{F}_1}(1-Z(U_1^i|U_1^{1:i-1},Y^{[m]})^2)\\
&\quad\le 2\ln 2\cdot m2^{-m^\beta},
\end{aligned}$$

where the second equality holds because $\mathsf{P}_Y=\mathsf{Q}_Y$, and the first inequality holds because $Z(X|Y)^2\le H(X|Y)$. The proof of the first part is completed.

For the second level, by the chain rule of the Kullback-Leibler divergence,

$$\begin{aligned}
&D_{KL}\left(\mathsf{P}_{U_{1:2}^{[m]},Y^{[m]}}\|\mathsf{Q}_{U_{1:2}^{[m]},Y^{[m]}}\right)\\
&\quad=D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}\|\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)\\
&\qquad+\mathsf{E}_{U_1^{[m]},Y^{[m]}}\left[\mathrm{D_{KL}}\left(\mathsf{P}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}\|\mathsf{Q}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}\right)\right] \qquad (58)\\
&\quad\le 2\ln 2\cdot m2^{-m^\beta}+2\ln 2\cdot m2^{-m^\beta},
\end{aligned}$$

where the first term holds because of the result for the 1st level, and the second term can be obtained by following the steps in (57) exactly, since it can be written as

$$\begin{aligned}
&\mathsf{E}_{U_1^{[m]},Y^{[m]}}\left[D_{KL}\left(\mathsf{P}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}\right)\right]\\
&\qquad=\ln 2\cdot\sum_{u_{1:2}^{[m]},y^{[m]}}\mathsf{P}\left(u_{1:2}^{[m]},y^{[m]}\right)\log_2\frac{\mathsf{P}\left(u_2^{[m]}|u_1^{[m]},y^{[m]}\right)}{\mathsf{Q}\left(u_2^{[m]}|u_1^{[m]},y^{[m]}\right)}. \qquad (59)
\end{aligned}$$

The proof of the second part of this lemma can be completed by induction. □

E Reducing the Size of Public Key in PKE

Frodo [11] is a widely studied post-quantum PKE scheme that employs LWE for both key generation and encryption. We propose a simple modification to Frodo by incorporating LWQ in the key generation step to reduce the public key size. We note that improvements to the encryption step of Frodo are also possible. However, such enhancements require a more involved analysis to carefully balance the parameters, and we leave this direction for future work. The proposed modification to the key generation step applies equally to any PKE scheme (e.g., Lizard [16]) that follows the LWE-based key generation template.

The Frodo key generation operates as follows:

- **Parameters**: Dimensions n, ℓ, modulus q, and noise width σ.
- **Key Construction**:
  1. Sample public matrix $\mathbf{A} \leftarrow \mathbb{Z}_q^{n \times n}$
  2. Sample secret key $\mathbf{S} \leftarrow \mathcal{D}_{\mathbb{Z}^{n \times \ell}, \sigma}$ and error $\mathbf{E} \leftarrow \mathcal{D}_{\mathbb{Z}^{n \times \ell}, \sigma}$
  3. Compute public matrix $\mathbf{B} = \mathbf{AS} + \mathbf{E}$
- **Output**:

$$\mathsf{pk} = (\mathbf{A} \mid \mathbf{B}), \quad \mathsf{sk} = \mathbf{S}$$

Using XOFs (AES128/SHAKE128) to generate $\mathbf{A}$, the public key size is:

$$|\mathsf{pk}| = |\mathsf{seed}(\mathbf{A})| + \log_2(q) \cdot n \cdot \ell \text{ bits.}$$

With $|\mathsf{seed}(\mathbf{A})| = 128$ bits, Frodo's parameters and baseline public key sizes are:

Variant	n	ℓ	q	$\lvert\mathsf{pk}\rvert$ (bytes)
Frodo-640	640	8	2^{15}	9,616
Frodo-976	976	8	2^{16}	15,632
Frodo-1344	1344	8	2^{16}	21,520

We replace LWE with LWQ in Frodo's key generation:

- **Parameters**: Dimensions n, ℓ, modulus q, and noise width σ, polar lattice Λ based on σ.
- **Key Construction**:
  1. Sample public matrix $\mathbf{A} \leftarrow \mathbb{Z}_q^{n \times n}$, $\mathbf{D} \leftarrow (\mathbb{Z}^n/\Lambda)^\ell$
  2. Sample secret key $\mathbf{S} \leftarrow \mathcal{D}_{\mathbb{Z}^{n \times \ell}, \sigma}$
  3. Compute public matrix by lattice quantization:

$$\mathbf{B} = Q_\Lambda(\mathbf{AS} - \mathbf{D}) + \mathbf{D}$$

- **Output**:

$$\mathsf{pk} = (\mathbf{A} \mid \mathbf{D} \mid \mathbf{B} - \mathbf{D}), \quad \mathsf{sk} = \mathbf{S}$$

By using XOFs to generate $(\mathbf{A}, \mathbf{D})$, the LWQ-based public key size becomes:

$$|\mathsf{pk}| \approx |\mathsf{seed}(\mathbf{A}, \mathbf{D})| + \log_2\left(\frac{q}{\sqrt{2\pi e}\cdot\sigma}\right)\cdot n\cdot\ell \text{ bits.}$$

The term $\log_2\left(\frac{q}{\sqrt{2\pi e}\cdot\sigma}\right)$ reflects the compression rate of $\mathbf{B}$, due to the use of polar lattices for quantization. Maintaining n, ℓ, q, σ and $|\mathsf{seed}(\mathbf{A}, \mathbf{D})| = |\mathsf{seed}(\mathbf{A})| = 128$ bits, LWQ achieves reduced public-key size:

Variant	Original $\lvert\mathsf{pk}\rvert$ (bytes)	LWQ $\lvert\mathsf{pk}\rvert$ (bytes)
Frodo-640	9,616	7,696 (-20%)
Frodo-976	15,632	12,704 (-19%)
Frodo-1344	21,520	18,832 (-13%)