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 high-information-rate encryption framework based on LWQ, demonstrating its advantage over constructions based on LWE and quantized LWE.

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

1 Introduction

Regev's Learning with Errors (LWE) problem [Reg05] 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 $\mathbf{A} \leftarrow \mathbb{Z}^{m\times n}$, 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 [Reg05,Pei09,BLP+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., [GPV08,MP12,GSW13,CKKS17,PS19], 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 [BPR12] 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 [BPR12], reusable randomness extractors [AKPW13], and public key encryption schemes such as Saber [DKRV18] and Lizard [CKLS18].

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 [BPR12].

Subsequent research [AKPW13,BGM+16,NR23] has further examined this area. The size of q was reduced to a polynomial by assuming it is a prime in [AKPW13,NR23], 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, [BGM+16] 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 [BPR12] 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 [AKPW13] and is applied using the Rényi divergence metric in [BGM+16].

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 [Reg05]. 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 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 special case of LWQ with $\Lambda = \frac{q}{p}\mathbb{Z}^m$; 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 (corresponding to Corollary 1 in Section 3.2):

Theorem 1 (Informal). *There exist a sequence of efficient lattice quantizers such that the LWQ distribution is computationally indistinguishable from the uniform distribution.*

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.

Intuitively, 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 prevent patterns in quantization errors. In general, dithering leads to $Q_\Lambda(\mathbf{As} - \mathbf{d})$ using a uniform $\mathbf{d}$. 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 pseudorandom generator, 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: $0.08333, 0.07660, 0.07168$ [CS99]. 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 [BDGM19]. Recently, Micciancio and Schultz [MS23] introduced a (quantized) LWE-based encryption framework to analyze the information rate of lattice-based encryption. In particular, their work [MS23, 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 $1 - o\left(\frac{1}{\log(q)}\right)$. 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.

1.2 Technical Overview

We show that the LWQ and LWE distributions are statistically indistinguishable (corresponding to Theorem 4 in Section 3.2):

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

In this work, we will adopt polar lattices to instantiate LWQ. 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 statistic relationship between the input and output for a target distortion; see [Cov99, 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 (see Theorem 7 in Section 4.2 for details):

Theorem 3 (Informal). *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 1. 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 2. 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 [STA09,Sas12].

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 [CKKS17] 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 [SAB+22]. 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. Additionally, the noise analysis presented in [SAB+22] is heuristic, as it assumes the Gaussianity of the quantization error and its independence from the LWE noise.

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 [Zam14]. Certain performance benefits of vector quantization have been justified in the secret-key encryption framework [MS23], and to reduce the ciphertext rate of CRYSTALS-Kyber [LS24].

The inquiry into optimal lattices for quantization, aiming for the smallest average distortion, is different from sphere packing [Via17,CKM+17]. The theoretical proof of optimal lattice quantizers has been limited to dimensions up to 3 (*i.e.*, $\mathbb{Z}$, A_2, A_3^*) [BS83], although efforts to identify good lattice quantizers have resulted in periodic updates of tables for small-dimensional lattices $n \leq 24$ [AA23]. 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 [ZF96b], 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 [Zam14], 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 [LSL21], where the computational complexity of the quantizer is $O(m \log m)$.

Polar codes and polar lattices The polar lattices investigated in this work originate from polar codes [Arı09]. 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 [Arı09]. 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 $O(m \log \log m)$ complexity for blocklength m [WD21]. Due to their outstanding performance, polar codes have been widely adopted in various practical

applications, including fifth-generation (5G) wireless communication networks [EXMH19]. 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" [CS99, 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 [LYLW19]. A follow-up work [LSL21] 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 [LYLW19,LSL21] 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 $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot)$ etc. We let λ denote the security parameter throughout the paper. All known valid attacks against the cryptographic scheme under consideration should take $\Omega(2^\lambda)$ bit operations. A function $\mathrm{negl} : \mathbb{N} \to \mathbb{R}^+$ is negligible if for every positive polynomial $p(\lambda)$, there exists $\lambda_0 \in \mathbb{N}$ such that $\mathrm{negl}(\lambda) < \frac{1}{p(\lambda)}$ 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 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 $\tilde{\Lambda}$, 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 tiers are broken in a systematic manner (such that $\mathbf{y} - Q_\Lambda(\mathbf{y}) \in \mathcal{V}_\Lambda$). The quantization can be implemented with polynomial-time algorithm by selecting fast-decodable lattices as Λ (e.g., $\mathbb{Z}^m$, the tensor product of Gosset lattice $\mathbb{Z}^{n/8} \otimes E_8$ and polar lattices).

In lossy source coding, the dithering technique is widely used along with nearest neighbor quantization:

Definition 3 (Dithered quantizer). *A dithered quantizer over lattice Λ is defined by sampling $\mathbf{d} \leftarrow \mathcal{P}_\Lambda$ and outputting $Q_\Lambda(\mathbf{y} - \mathbf{d})$.*

To compensate for the subtractive dither, the reconstructed vector is given by $Q_\Lambda(\mathbf{y}-\mathbf{d})+\mathbf{d}$. This amounts to quantizing an input vector to a coset $\Lambda+\mathbf{d}$ of Λ:

$$Q_{\Lambda+\mathbf{d}}(\mathbf{y}) = \arg\min_{\lambda\in\Lambda+\mathbf{d}} \|\mathbf{y}-\lambda\|. \tag{6}$$

It can be verified that

$$\begin{aligned} Q_{\Lambda+\mathbf{d}}(\mathbf{y}) &= \arg\min_{\lambda\in\Lambda} \|\mathbf{y}-(\lambda+\mathbf{d})\|+\mathbf{d} \\ &= \arg\min_{\lambda\in\Lambda} \|(\mathbf{y}-\mathbf{d})-\lambda\|+\mathbf{d} \\ &= Q_\Lambda(\mathbf{y}-\mathbf{d})+\mathbf{d}. \end{aligned}$$

For convenience, we will use the quantizer $Q_{\Lambda+\mathbf{d}}(\mathbf{y})$ in this paper. The quantization error is given by

$$\mathbf{e}_Q = Q_\Lambda(\mathbf{y}-\mathbf{d})+\mathbf{d}-\mathbf{y}. \tag{7}$$

Definition 4 (Second moment [Zam14]). *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.*

For the nearest neighbor dithered quantizer, $\mathbf{e}_Q$ is uniformly distributed over $\mathcal{V}_\Lambda$, so the averaged quantization error of the dithered quantizer can be quantified by $\gamma^2(\Lambda)$: for any distribution of $\mathbf{y}$, with $\mathbf{d}\sim U(\mathcal{P}_\Lambda)$, then

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{e}_Q\right\|^2 = \gamma^2(\Lambda). \tag{8}$$

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

$$G(\Lambda) = \frac{\gamma^2(\Lambda)}{\det^{2/n}(\Lambda)}. \tag{9}$$

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 [Zam14]). *A sequence of lattices* $\Lambda^{(n)}$ *with growing dimension is called good for mean squared error quantization if*

$$\lim_{n\to\infty} G(\Lambda^{(n)}) = \frac{1}{2\pi e}. \tag{10}$$

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 $\mathbf{c} \in \mathbb{R}^n$, $r > 0$,

$$\mathcal{D}_{\Lambda,r,\mathbf{c}}(\mathbf{x}) = \frac{\rho_r(\mathbf{x} - \mathbf{c})}{\rho_r(\Lambda - \mathbf{c})}, \quad \forall \mathbf{x} \in \Lambda \tag{11}$$

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:*

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

We analyze adversaries interacting within probabilistic experiments known as games. For an adversary $\mathcal{A}$ and two games $\mathcal{G}_0$ and $\mathcal{G}_1$ with which it can engage, the distinguishing advantage of $\mathcal{A}$ is given by

$$\mathrm{Adv}_{\mathcal{G}_0,\mathcal{G}_1}(\mathcal{A}) = \left| \Pr[A \text{ accepts in } \mathcal{G}_0] - \Pr[A \text{ accepts in } \mathcal{G}_1] \right|.$$

Definition 8 (Computational Indistinguishability). *Two games $\mathcal{G}_0$ and $\mathcal{G}_1$ are said to be computationally indistinguishable if, for every probabilistic polynomial-time distinguisher $\mathcal{A}$, there exists a negligible function* negl *such that $Adv_{\mathcal{G}_0,\mathcal{G}_1}(\mathcal{A}) \leq \mathrm{negl}(n)$.*

Similarly, we say that two probability distributions P and Q are statistically indistinguishable if their statistical distance $\Delta(P, Q)$ is negligible. By Pinsker's inequality, if the KL divergence $D_{\mathrm{KL}}(P \| Q)$ is negligible, then P and Q are statistically indistinguishable.

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 9 (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 over $\mathbb{Z}^m$, samples for the LWE/LWR/LWQ distributions are respectively generated by*

- LWE_{χ_e}*:* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{e} \leftarrow \chi_e$, *and output* $(\mathbf{A}, \mathbf{b} = \mathbf{A}\mathbf{s} + \mathbf{e}) \in \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$.
- LWR_p*:* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, *and output* $(\mathbf{A}, \mathbf{b} = \lfloor \mathbf{A}\mathbf{s} \rceil_p) \in \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_p^m$.
- LWQ_Λ*:* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$ *and output* $(\mathbf{A}, \mathbf{b} = Q_{\Lambda+\mathbf{d}}(\mathbf{A}\mathbf{s})) \in \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$.

Definition 10 (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 arbitrarily many samples from the LWE/LWR/LWQ distribution, where $\mathbf{s}$ is sampled from some distribution χ_s (fixed for all samples), the search problem asks to recover $\mathbf{s}$.*

Note that, in LWQ, we quantize to a coset $\Lambda + \mathbf{d}$ of the lattice for a random dither $\mathbf{d}$. 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, LWQ amounts to (dithered) LWR.

Remark 3. 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 $(\mathbf{A}, \mathbf{b} = Q_{\Lambda+\mathbf{d}}(\mathbf{A}\mathbf{s}) = Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d}) + \mathbf{d})$, we can identify the coset representative $\mathbf{d}$ in polynomial time (*e.g.*, using Babai's rounding off procedure [Bab86]) and rewrite it as

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

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, $Q_\Lambda(\mathbf{A}\mathbf{s} - \mathbf{d})$ is pseudorandom, while the matrix $\mathbf{A}$ and dither $\mathbf{d}$ can be transmitted as seeds of a pseudorandom number generator. These seeds, which do not need to remain secret, can effectively serve as the public information in practice.

Let $\mathcal{G} : \{0,1\}^k \to \mathbb{Z}_q^{m\times n} \times (\mathbb{Z}^m/\Lambda)$ be a pseudorandom number generator, and $(\mathbf{A}, \mathbf{d}) = \mathcal{G}(seed)$. Then storing an LWQ sample in the form of $(\mathbf{A}, Q_\Lambda(\mathbf{A}\mathbf{s}-\mathbf{d}), \mathbf{d})$ 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 [Reg05] and Peikert[Pei09], for any $m = n^{O(1)}$, any modulus $q \leq 2^{n^{O(1)}}$, and for a (discrete) Gaussian distribution χ_e with parameter $\sigma \geq 2\sqrt{n}$, 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 $\text{poly}(n)$ factor in the number of samples used. Although the primary hardness justification of LWE [Reg05] is based on continuous Gaussian errors, it also holds when the error distribution is a discrete Gaussian, $\chi_e = \mathcal{D}_{\mathbb{Z}^m,\sigma}$. This reduction can be proved by applying a randomized rounding algorithm to the $\mathbf{b}$ samples of $\mathsf{LWE}_{\chi_e=\rho_\sigma/\sigma^m}$ (cf. [Pei10, Theorem 3.1]). Throughout this paper, we refer to the hardness of the LWE assumption as

$$\mathsf{LWE}_{\chi_e=\mathcal{D}_{\mathbb{Z}^m,\sigma}} \approx_c U(\mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m). \tag{12}$$

3.2 Asymptotic Results of Hardness

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 a pseudorandom generator 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 7 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 $\mathsf{LWE}_{\chi_e=\mathcal{D}_{\mathbb{Z}^m,\sigma}}$: $\mathsf{P}_{\mathbf{A},\mathbf{b}}$ 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 LWQ_Λ: $\mathsf{Q}_{\mathbf{A},\mathbf{b}}$ where $\mathbf{b} = X^{[m]} = Q_{\Lambda+\mathbf{d}}(Y^{[m]})$ where $Y^{[m]} = \mathbf{As}$ and $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$, produced by a polar lattice quantizer.

Theorem 4 (LWQ $\approx_s$ LWE). *Let $m = m(\lambda)$, $n = n(\lambda)$, $q = p(\lambda)^r$ where λ is the security parameter, $p(\lambda)$ is a prime number and $r \in \mathbb{N}$. 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:*

$$\Delta\left(\mathsf{P}_{\mathbf{A},\mathbf{b}}, \mathsf{Q}_{\mathbf{A},\mathbf{b}}\right) = 2^{-\omega(\lambda^\beta)}, \quad \forall \ \ 0 < \beta < 1. \tag{13}$$

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

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

which is due to the Markov chain[4] $\mathbf{A} \to \mathbf{As} \to \mathbf{b}$. 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

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

Analogously, the LWQ distribution satisfies

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

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

Now we have

$$\begin{aligned}
&\Delta\left(\mathsf{P}_{\mathbf{A},\mathbf{b}}, \mathsf{Q}_{\mathbf{A},\mathbf{b}}\right) \\
&= \frac{1}{2}\sum_{\mathbf{A}} \mathsf{P}_{\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left|\mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot)\right| \\
&= \frac{1}{2}\sum_{\mathbf{As}} \mathsf{P}_{\mathbf{As}|\mathbf{A}}(\cdot) \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left|\mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot)\right| \\
&= \frac{1}{2}\sum_{\mathbf{As}} \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{As},\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left|\mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot)\right| \\
&= \frac{1}{2}\sum_{Y^{[m]}} \mathsf{P}_{Y^{[m]}}(\cdot) \sum_{X^{[m]}} \left|\mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot)\right| \\
&= \Delta\left(\mathsf{P}_{X^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right)
\end{aligned} \tag{14}$$

where the second equality of (14) holds since $\mathsf{P}_{\mathbf{As}|\mathbf{A}}$ is an indicator function when $\mathbf{s}$ is given. Proof is completed by using Theorem 7 and Remark 6 in Section 4.2, with an appropriate dimension m set according to λ. □

This theorem states that the LWE noise can be substituted with the quantization noise of LWQ while preserving security.

Remark 4. Theorem 4 holds under KL divergence too, by applying Lemma 7 in Appendix C.

Corollary 1 (LWQ $\approx_c$ Uniform). *Let $m = m(\lambda)$, $n = n(\lambda)$, $q = p(\lambda)^r$ where λ is the security parameter, $p(\lambda)$ is a prime number and $r \in \mathbb{N}$. Let $\mathbf{s} \in \mathbb{Z}_q^{n*}$. There exist a sequence of efficient derandomized quantization lattices $Q_{\Lambda+\mathbf{d}}$, indexed by dimension m, such that the LWQ distribution is computationally indistinguishable from a uniform distribution over $\mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$.*

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

[4] 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 $P(x, y, z) = P(x)P(y|x)P(z|y)$ [Cov99].

- $\mathcal{G}_0$: This is the real attack game against the LWQ distribution. That is, we choose $\mathbf{s}$ and upon request generate and give the attacker independent samples $c \leftarrow \mathsf{LWQ}_\Lambda(\mathbf{d}$ pseudorandom).
- $\mathcal{G}_1$: The attacker is against LWQ based on random quantization: $c \leftarrow \mathsf{LWQ}_\Lambda(\mathbf{d}$ random).
- $\mathcal{G}_2$: In this game, we give the attacker samples from LWE: $c \leftarrow \mathsf{LWE}_{\chi_e=\mathcal{D}_{\mathbb{Z}^m,\sigma}}$.
- $\mathcal{G}_3$: In this game, uniform samples are given: $c \leftarrow \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$.

$\mathcal{G}_1$ and $\mathcal{G}_2$ are statistically indistinguishable as $\mathsf{LWQ}_\Lambda(\mathbf{d}$ random) and $\mathsf{LWE}_{\chi_e=\mathcal{D}_{\mathbb{Z}^m,\sigma}}$ are statistically indistinguishable, thus

$$\begin{aligned}\mathsf{Adv}_{\mathcal{G}_1,\mathcal{G}_2}(\mathcal{A}) &= \left|\Pr(\mathcal{A}(\mathsf{LWQ}_\Lambda(\mathbf{d} \text{ random}) = 1) - \Pr(\mathcal{A}(\mathsf{LWE}_{\chi_e=\mathcal{D}_{\mathbb{Z}^m,\sigma}}) = 1\right| && (15)\\ &\le \Delta(L_\Lambda, L_{\mathcal{D}^m_{\mathbb{Z},\sigma}}) && (16)\\ &= 2^{-\omega(\lambda^\beta)}, \quad \forall\ 0<\beta<1 && (17)\end{aligned}$$

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

Since $\mathcal{G}_0$ and $\mathcal{G}_1$ are computationally indistinguishable, together with the hardness of LWE, we have

$$\mathsf{Adv}_{\mathcal{G}_0,\mathcal{G}_3}(\mathcal{A}) \le \mathsf{Adv}_{\mathcal{G}_0,\mathcal{G}_1}(\mathcal{A}) + \mathsf{Adv}_{\mathcal{G}_1,\mathcal{G}_2}(\mathcal{A}) + \mathsf{Adv}_{\mathcal{G}_2,\mathcal{G}_3}(\mathcal{A}) = \mathrm{negl}(\lambda). \quad (18)$$

□

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 $\mathsf{LWE}_{\chi_e=\mathcal{D}_{\mathbb{Z}^m,\sigma}} \approx_s \mathsf{LWQ}_\Lambda$. This is illustrated in Fig. 1 [5]. 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 $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2)$ [FTC00]. 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 [LSL21]. In fact, as $m \to \infty$,

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

Thus, the compression rate R_c is ultimately determined by the capacity of the LWE channel. For the parameters $q = \mathrm{poly}(n)$ and $\sigma = \Omega(\sqrt{n})$ in LWE, the capacity can be made explicit: it is possible to show $C(\mathbb{Z}/q\mathbb{Z}, \sigma^2) \approx \log_2\left(\frac{q}{\sqrt{2\pi e}\cdot\sigma}\right)$ [FTC00]. Intuitively, $\log_2\left(\sqrt{2\pi e}\cdot\sigma\right)$ of the $\log_2(q)$ bits are buried under noise.

[5] 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 $\chi_e = \mathcal{D}_{\mathbb{Z}^m,\sigma}$. Reversing the input/output is a common practice for the test channel in source coding theory [Cov99].

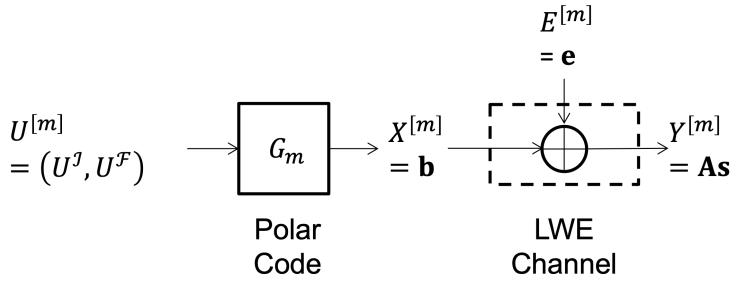

Fig. 1. Simulating the LWE channel using a polar lattice (which may be viewed as a q-ary polar code).

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.

3.3 Hardness of LWQ for an arbitrary quantization lattice

The asymptotic approach given above requires the lattice dimension m to grow, thus cannot be applied to quantization lattices of concatenated small dimensional lattices. In this case, we derive the following result for a given quantization dimension.

Lemma 1 (Dither lemma, adapted from [Zam14]). *If the dither U is uniformly distributed over the fundamental cell $\mathcal{P}_\Lambda$, i.e., with probability density function*

$$f_U(u) = \begin{cases} \frac{1}{|\mathcal{P}_\Lambda|}, & u \in \mathcal{P}_\Lambda, \\ 0, & u \notin \mathcal{P}_\Lambda, \end{cases}$$

then $\mathbf{e}_Q = Q_\Lambda(\mathbf{y} - U) + U - \mathbf{y}$ is uniformly distributed over the Voronoi region $\mathcal{V}_\Lambda$, independent of $\mathbf{y}$. And similarly, if U is uniformly distributed over $\mathbb{Z}^m/\Lambda$, $\mathbf{e}_Q$ is uniformly distributed over $\mathcal{V}_\Lambda \cap \mathbb{Z}^m$.

Theorem 5. *Let $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$. Then the LWQ distribution is equivalent to the LWE distribution with uniform noise $\mathbf{e}_Q \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$.*

Proof. By applying the dither lemma, given $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$, we have

$$Q_{\Lambda+\mathbf{d}}(\mathbf{As}) = \mathbf{As} + \mathsf{e}_Q, \tag{19}$$

where $\mathsf{e}_Q \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$ is independent of $\mathbf{As}$. Thus, the LWQ distribution takes the form $(\mathbf{A}, \mathbf{As} + \mathsf{e}_Q)$, which corresponds to LWE with noise term e_Q. □

We recall the definition of the smoothing parameter and a useful lemma.

Definition 11 (Smoothing parameter [MR04]). *For any lattice Λ and positive real $\varepsilon > 0$, the smoothing parameter $\eta_\varepsilon(\Lambda)$ is the smallest real $r > 0$ such that $\rho_{1/r}(\tilde{\Lambda} \setminus \{0\}) \leq \varepsilon$ where $\tilde{\Lambda}$ is the dual lattice.*

Lemma 2 ([MR04]). *For any lattice Λ and $\mathbf{c} \in \mathbb{R}^n$, $\varepsilon > 0$, and $r \geq \eta_\varepsilon(\Lambda)$,*

$$\rho_r(\Lambda + \mathbf{c}) \in r^n \det(\tilde{\Lambda})(1 \pm \varepsilon). \tag{20}$$

The discrete NSM is defined as: $\bar{G}(\Lambda) = \bar{\gamma}^2(\Lambda)/\det^{2/m}(\Lambda)$, where $\bar{\gamma}^2(\Lambda) = \frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m} \det(\Lambda)^{-1}\|\mathbf{x}\|^2$. From the high-resolution assumption [Zam14], 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 6. *Define $r = \sqrt{2\pi\bar{G}(\Lambda)}\det^{1/m}(\Lambda)$. If $r \geq \eta_\varepsilon(\mathbb{Z}^m)$, the KL divergence between LWQ and LWE satisfies:*

$$D_{KL}((\mathbf{A}, \mathbf{As}+\mathbf{e}_Q)\|(\mathbf{A}, \mathbf{As}+\mathbf{e})) \in \frac{m}{2}\ln(2\pi e\bar{G}(\Lambda)) + \ln(1 \pm \varepsilon). \tag{21}$$

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

Proof. From Lemma 2, we have

$$\rho_r(\mathbb{Z}^m) \in r^m(1 \pm \varepsilon). \tag{22}$$

Then we have

$$\frac{1}{m}D_{KL}(\mathbf{e}_Q\|\mathbf{e}) = \frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m} \det(\Lambda)^{-1}\ln\frac{\rho_r(\mathbb{Z}^m)}{\det(\Lambda)\rho_r(\mathbf{x})} \tag{23}$$

$$\in \frac{1}{m}\ln\frac{r^m(1\pm\varepsilon)}{\det(\Lambda)} + \frac{\pi}{r^2}\cdot\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m}\det(\Lambda)^{-1}\|\mathbf{x}\|^2 \tag{24}$$

$$= \frac{1}{2}\ln\frac{r^2}{\det(\Lambda)^{2/m}} + \frac{\pi}{r^2}\cdot\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m}\det(\Lambda)^{-1}\|\mathbf{x}\|^2 + \frac{1}{m}\ln(1\pm\varepsilon). \tag{25}$$

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

$$r^2 = 2\pi\bar{\gamma}^2(\Lambda) = 2\pi\cdot\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{V}_\Lambda\cap\mathbb{Z}^m}\det(\Lambda)^{-1}\|\mathbf{x}\|^2, \tag{26}$$

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

$$D_{KL}(\mathbf{e}_Q\|\mathbf{e}) \in \frac{m}{2}\ln(2\pi e\bar{G}(\Lambda)) + \ln(1\pm\varepsilon). \tag{27}$$

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

$$D_{KL}((\mathbf{A}, \mathbf{As}+\mathbf{e}_Q)\|(\mathbf{A}, \mathbf{As}+\mathbf{e})) = D_{KL}(\mathbf{e}_Q\|\mathbf{e}). \tag{28}$$

□

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 [LYLW19,LSL21] 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" [FTC00]. For convenience, we present Construction D using binary polar codes, whereas the extension to nonbinary codes is straightforward [CS99].

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 a reversed version of the test channel between X and Y in Fig. 3. 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. 3, 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 7 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.

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 an n-dimensional 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 $C_1 \subseteq C_2 \subseteq \cdots \subseteq C_r$.

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 $\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m$ as a basis of $\mathbb{F}_2^m$ such*

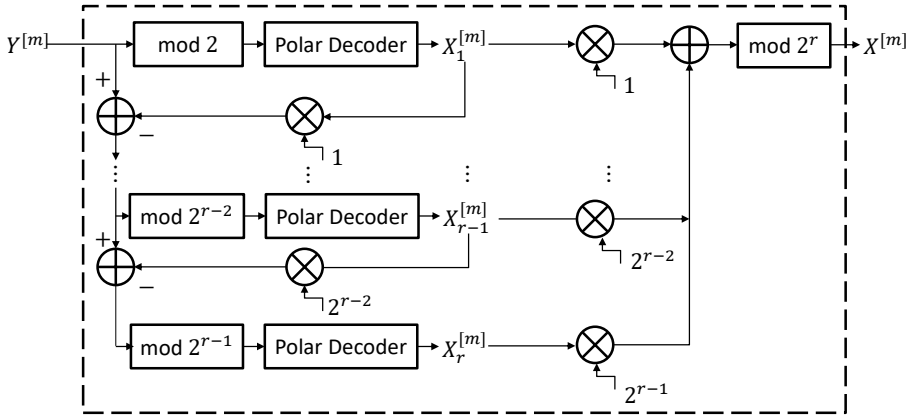

Fig. 2. The internal structure of a polar lattice quantizer.

that $\mathbf{d}_1, \cdots, \mathbf{d}_{k_\ell}$ *span* C_ℓ*. With* $n = 1$*, the binary lattice* L *of Construction D consists of all vectors of the form*

$$\sum_{\ell=1}^{r} 2^{\ell-1} \sum_{j=1}^{k_\ell} u_\ell^j \psi(\mathbf{d}_j) + 2^r z, \tag{29}$$

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*

$$Z(W) = Z(X|Y) \triangleq \sum_y \sqrt{\mathsf{P}_{Y|X}(y|0)\mathsf{P}_{Y|X}(y|1)}. \tag{30}$$

E.g., in [AT09], the rate of channel polarization is characterized in terms of the Bhattacharyya parameter as

$$\lim_{m\to\infty} \mathsf{Pr}\left(Z(W_m^{(i)}) < 2^{-m^\beta}\right) = C, \quad \text{for any } 0 < \beta < 0.5.$$

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 [TV13,PHTT11].

In the context of lossy compression, polar codes can achieve the rate-distortion bound for binary symmetric sources [KU10]. To achieve a target distortion:

Algorithm 1 Polar Lattice Quantization Algorithm

Require: Source Y uniformly random on $[-2^{r-1}, 2^{r-1})$
Ensure: Quantized output $X^{[m]}$
1: Build test channel $Y = X + E \mod q\mathbb{Z}$, where $q = 2^r$ and $E \sim \mathcal{D}_{\mathbb{Z},\sigma}$
2: Assume X uniformly random on $[-2^{r-1}, 2^{r-1})$
3: Construct polar lattice quantizer on test channel using binary partition chain $\mathbb{Z}/2\mathbb{Z}/\cdots/2^r\mathbb{Z}$
4: Assume r is large enough such that the modulo $2^r\mathbb{Z}$ operation is insignificant on E
5: Represent X as bit sequence $X_1, X_2, \ldots, X_r$, where X_ℓ specifies coset $2^{\ell-1}\mathbb{Z}/2^\ell\mathbb{Z}$
6: $X_1, \ldots, X_r$ uniquely describe cosets of $\mathbb{Z}/2^r\mathbb{Z}$
7: **for** $\ell = 1$ to r **do**
8: **if** $\ell = 1$ **then**
9: Execute SC decoding to obtain $X_1^{[m]}$ from $Y^{[m]}$ using statistic of partition channel $\mathsf{P}_{Y|X_1}$
10: **else**
11: Decode $X_\ell^{[m]}$ from $Y^{[m]}$ and $X_1^{[m]}, \ldots, X_{\ell-1}^{[m]}$ using $\mathsf{P}_{Y,X_1,\ldots,X_{\ell-1}|X_\ell}$
12: **end if**
13: **end for**
14: Return $X^{[m]} = X_1^{[m]} + 2X_2^{[m]} + \cdots + 2^{r-1}X_r^{[m]} \mod 2^r\mathbb{Z}$

- 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 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 decoding algorithm called Successive Cancellation (SC) decoding [Arı09], 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

$$\sum_{i \in \mathcal{I}} Z(W_m^{(i)}).$$

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 $m = 2^{20}$.

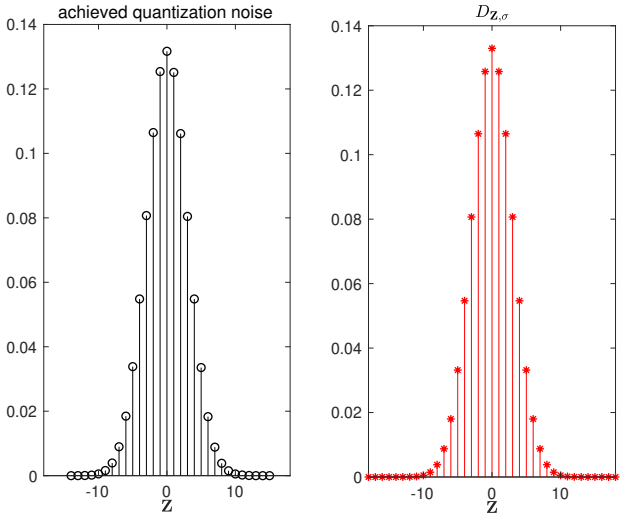

Fig. 3. A comparison between the distribution of quantization noise $Y - X$ and $D_{\mathbb{Z},\sigma=3}$.

Dithered quantization with polar lattices In the literature on traditional lattice quantization [ZF96a], 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

$$\mathbf{d} = \sum_{\ell=1}^{r} 2^{\ell-1} \sum_{j=k_\ell+1}^{N} u_\ell^j \psi(\mathbf{g}_j), \tag{31}$$

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 5. The dither $\mathbf{d}$ of LWQ is public, thus a pseudorandom number generator 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. 3, 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. [FTC00], 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 $\mathcal{I}_1 \triangleq \{i \in [m] : Z(U_1^i | U_1^{1:i-1}, Y^{[m]}) \leq 1 - 2^{-m^\beta}\}$ 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 3. *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* (32) *and* (33) *at the first partition level.*

$$U_1^i = \begin{cases} 0 \text{ w. p. } P_{U_1^i | U_1^{1:i-1}, Y^{[m]}} \left(0 | u_1^{1:i-1}, y^{[m]}\right) \\ 1 \text{ w. p. } P_{U_1^i | U_1^{1:i-1}, Y^{[m]}} \left(1 | u_1^{1:i-1}, y^{[m]}\right) \end{cases} \text{ if } i \in \mathcal{I}_1 \tag{32}$$

$$U_1^i = \begin{cases} 0 \text{ w. p. } \dfrac{1}{2} \\ 1 \text{ w. p. } \dfrac{1}{2}. \end{cases} \text{ if } i \in \mathcal{F}_1 \tag{33}$$

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 (32) *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:*

$$\Delta\left(\mathsf{P}_{U_1^{[m]}, Y^{[m]}}, \mathsf{Q}_{U_1^{[m]}, Y^{[m]}}\right) \leq m\sqrt{\ln 2 \cdot 2^{-m^\beta}}, \quad 0 < \beta < \frac{1}{2}. \tag{34}$$

Proof. See Appendix B.

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 4. *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* (32) *and* (33) *at the first partition*

level, and then rules (35) *and* (36) *at the second partition level.*

$$U_1^i = \begin{cases} 0 \ w.\ p.\ P_{U_2^i|U_2^{1:i-1},X_1^{[m]},Y^{[m]}}\left(0|u_2^{1:i-1},x_1^{[m]},y^{[m]}\right) \\ 1 \ w.\ p.\ P_{U_2^i|U_2^{1:i-1},X_1^{[m]},Y^{[m]}}\left(1|u_2^{1:i-1},x_1^{[m]},y^{[m]}\right) \end{cases} \text{if } i \in \mathcal{I}_2 \tag{35}$$

$$U_2^i = \begin{cases} 0 \ w.\ p.\ \dfrac{1}{2} \\ 1 \ w.\ p.\ \dfrac{1}{2}. \end{cases} \text{if } i \in \mathcal{F}_2 \tag{36}$$

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* (32) *and rule* (35) *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:*

$$\Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) \le 2m\sqrt{\ln 2 \cdot 2^{-m^\beta}}, \quad 0 < \beta < \frac{1}{2}. \tag{37}$$

Proof. Assume an auxiliary joint distribution $\mathsf{Q'}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ resulted from using the encoding rule (32) for all U_1^i with $i \in [m]$ at the first partition level, and rules (35) and (36) 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,

$$\begin{aligned} &\Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) \\ &\le \Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q'}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) + \Delta\left(\mathsf{Q'}_{U_1^{[m]},U_2^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right), \end{aligned} \tag{38}$$

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 3, and the second term is equal to $\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)$. □

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 $\Delta\left(\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}},\mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}}\right) \le rm\sqrt{\ln 2 \cdot 2^{-m^\beta}}$. Combining this result with Lemma **??**, 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.

For completeness, we also need the notation X', which is a reconstruction random variable defined over $\mathbb{Z}$ for the source Y, with the conditional probability $P_{X'|Y}$ defined by the relationship $X' = Y - E$, i.e., $X' - Y$ is a discrete Gaussian random variable independent of Y. The comparison between the two

test channels based on $\mathsf{P}_{X,Y}$ and $\mathsf{P}_{X',Y}$, respectively, is demonstrated in Fig. 4. As will be seen, the difference between the two channels, which is due to the modulo $q\mathbb{Z}$ operation, becomes negligible for large q. We remind the readers that the design target of our quantization lattice is to realize a quantization noise, whose distribution is indistinguishable from the lattice Gaussian distribution, as has been employed in LWE.

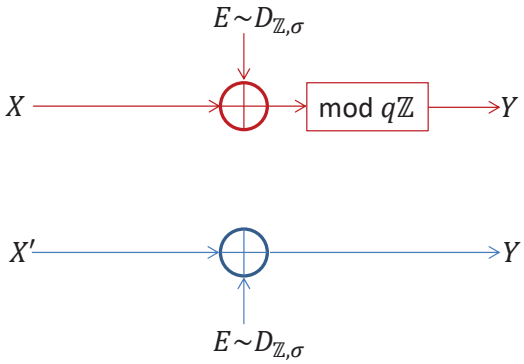

Fig. 4. A comparison between the two test channels based on $\mathsf{P}_{X,Y}$ and $\mathsf{P}_{X',Y}$, which are marked in red and blue, respectively.

Theorem 7. *The statistical distance between the joint distribution induced by the polar lattice and that by an i.i.d. $q\mathbb{Z}$-aliased discrete Gaussian distribution, i.e., the distribution of a discrete Gaussian after the modulo $q\mathbb{Z}$ operation, is bounded by*

$$\Delta\left(\mathsf{P}_{X^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right) \le r \cdot m\sqrt{\ln 2 \cdot 2^{-m^\beta}}, \quad 0 < \beta < \frac{1}{2}. \tag{39}$$

Moreover, when compared with the joint distribution induced by an i.i.d. discrete Gaussian distribution over $\mathbb{Z}$,

$$\begin{aligned}&\Delta\left(\mathsf{P}_{X'^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right)\\ &\quad\le r \cdot m\sqrt{\ln 2 \cdot 2^{-m^\beta}} + M \cdot m \cdot \exp\left(-\frac{(2^{r-1}-1)^2}{2\sigma^2}\right), \quad 0 < \beta < \frac{1}{2}.\end{aligned} \tag{40}$$

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

Recall that the test channel $X \underset{\mathsf{P}_{Y|X}}{\longrightarrow} Y$ is given by $Y = X + E \mod q\mathbb{Z}$, where $E \sim \mathcal{D}_{\mathbb{Z},\sigma}$. Suppose now P_Y is fixed, and $\mathsf{P}_{X|Y}$ is replaced with $\mathsf{P}_{X'|Y}$ by removing the modulo $q\mathbb{Z}$ operation, i.e., $X' = Y - E$. The statistical distance $\Delta(\mathsf{P}_{X'^{[m]},Y^{[m]}}, \mathsf{P}_{X^{[m]},Y^{[m]}})$ is equal to $\Delta(\mathsf{P}_{E'^{[m]}}, \mathsf{P}_{E^{[m]}})$ as shown in Lemma **??**,

where $E' = E \mod q\mathbb{Z}$. By using the telescoping expansion (43) and the triangle inequality again, the proof is completed. □

Remark 6. The restriction $0 < \beta < \frac{1}{2}$ in Theorem 7 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 [MT14]; thus we can obtain almost exponential decay of the statistical distance. Note that $q = p^r$ in this case. Although using nonbinary polar codes in Construction D will increase the computational complexity, it is still $O(m \log m)$.

Remark 7. Theorem 7 indicates that the performance of our lattice quantizer is determined by two parts. First, we need to ensure that $q = 2^r$ is large such that the modulo $q\mathbb{Z}$ operation has a little influence on the lattice Gaussian distribution, which is described by the second term on the right hand side of (40). Second, the dimension m of the lattice quantizer is required to be large to guarantee a sufficient polarization effect such that the quantization noise is close to the $q\mathbb{Z}$-aliased lattice Gaussian distribution, as described by the first term on the right hand side of (40). For completeness, since the two parts are both measured in the statistical distance, we also provide the counterparts of Lemma **??** and Lemma 3 with the measurement of the Kullback-Leibler divergence in Appendix C.

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 [FTC00].

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 $1 \leq \ell \leq r - 1$, 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 Quantized Encryption

This section introduces an LWQ-based encryption framework, denoted as $\mathsf{LWQ}_{E,\Lambda}$, and contrasts it with LWE and quantized LWE (LWEQ) based frameworks [MS23], noted as LWE_{E,χ_e} and $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$. It is important to note that the LWER problem in CRYSTALS-Kyber [SAB+22] represents a special case of LWEQ where the quantization is rounding. In comparison to quantized LWE [MS23], LWQ streamlines the processes of noise addition and quantization into a single step, where only quantization noise is present while ensuring security. This efficiency allows LWQ to replace LWE, LWR, or LWER in various cryptographic scenarios, resulting in higher information rate than LWE/LWER while providing enhanced security compared to LWR in general settings.

The presented LWE, LWQ, and LWEQ-based encryption schemes, each consists of a triplet $(\mathsf{KGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$. To enable comparisons, we let the schemes share a common key generation function (KGen) and nested lattice structures for error correction.

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 $\sigma \geq 2\sqrt{n}$, and a private key distribution χ_s over $\mathbb{Z}_q^{n*}$ with respect to the security parameter λ. Sample $\mathbf{s} \leftarrow \chi_s$ 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 $\mu \in \{0,1\}^{\log_2(q^m/\det(E))}$ 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 potentially 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}	$\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{e} \leftarrow \chi_e$ $\mathbf{b} = \mathbf{As} + \mathbf{e} + \mathsf{ec}_E(\mu)$ Return $(\mathbf{A}, \mathbf{b})$	Return $\mathsf{dc}_E(\mathbf{b} - \mathbf{As})$	$\mathbf{e}$	q^m
$\mathsf{LWEQ}_{E,\chi_e,\Lambda}$	$\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{e} \leftarrow \chi_e$ $\mathbf{b} = Q_\Lambda(\mathbf{As} + \mathbf{e}) + \mathsf{ec}_E(\mu)$ Return $(\mathbf{A}, \mathbf{b})$	Return $\mathsf{dc}_E(\mathbf{b} - \mathbf{As})$	$\mathbf{e} + \mathbf{e}_Q$	$\frac{q^m}{\det(\Lambda)}$
$\mathsf{LWQ}_{E,\Lambda}$	$\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{d} \leftarrow \mathbb{Z}^m/\Lambda$ $\mathbf{b} = Q_{\Lambda+\mathbf{d}}(\mathbf{As}) + \mathsf{ec}_E(\mu)$ Return $(\mathbf{A}, \mathbf{b})$	Return $\mathsf{dc}_E(\mathbf{b} - \mathbf{As})$	$\mathbf{e}_Q$	$\frac{q^m}{\det(\Lambda)}$

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

The $(\mathsf{Encrypt}, \mathsf{Decrypt})$ algorithms for $\mathsf{LWQ}_{E,\Lambda}$, LWE_{E,χ_e}, and $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$ 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 indicates that the effective ciphertext errors for these schemes are $\mathbf{e}$, $\mathbf{e} + \mathbf{e}_Q$, and $\mathbf{e}_Q$, respectively. Under the same conditions, LWQ produces smaller ciphertexts compared to LWE, and its effective ciphertext error is lower than that of LWEQ.

When measuring the size of the ciphertext, the cost of $\mathbf{A}$ and $\mathbf{d}$ can be excluded by using seeds as their generators. The information rate R of an encryption scheme is defined as the log size ratio of plaintext to ciphertext:

$$R = \frac{\log_2(q^m / \det(E))}{\log_2(|\mathsf{ct}|)}, \tag{41}$$

where $|\mathsf{ct}|$ denotes the size of the ciphertext. A scheme is said to achieve perfect rate when $R = 1$.

5.1 Security

In these encryption schemes, the pseudorandomness of the ciphertext, ensuring RND-CPA security [MS23], 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 8. *Under the LWE and LWQ indistinguishability assumptions, the schemes* LWE_{E,χ_e}*,* $\mathsf{LWEQ}_{E,\chi_e,\Lambda}$*, and* $\mathsf{LWQ}_{E,\Lambda}$ *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 $\mathsf{LWQ}_{E,\Lambda}$, 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 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 $\delta = 0$.

Lemma 5. *There exist a sequence of polar lattices Λ, indexed by dimension m, such that $\mathsf{LWQ}_{E=\Lambda,\Lambda}$ has perfect correctness, perfect rate, and is as secure as the LWE assumption with polynomial modulus $q = n^{O(1)}$ if $\det(\Lambda)^{1/m} = \sigma$.*

Proof. The δ-correctness condition can be evaluated by $\Pr(\tilde{\mathbf{e}} \notin \mathcal{V}_E) \leq \delta$. Note that $\mathbf{As} \sim U(\mathbb{Z}_q^m)$ since $\mathbf{s} \in \mathbb{Z}_q^{n^*}$. The effective decoding noise $\tilde{\mathbf{e}} = \mathbf{b} - \mathbf{As} - \mathsf{ec}_E(\mu)$ of $\mathsf{LWQ}_{E,\Lambda}$ can be expressed as:

$$\tilde{\mathbf{e}}_{\mathsf{LWQ}} = \mathbf{e}_Q, \tag{42}$$

where $\mathbf{e}_Q \sim U(\mathcal{V}_\Lambda \cap \mathbb{Z}^m)$. Thus, perfect correctness is guaranteed since $\Pr(\tilde{\mathbf{e}}_{\mathsf{LWQ}} \notin \mathcal{V}_E) = 0$. Lastly, perfect rate is guaranteed by its definition, and the condition $\det(\Lambda)^{1/m} = \sigma$ ensures that the underlying LWQ assumption meets the noise parameter requirement for a secure LWE assumption. □

This lemma breaks the rate-impossibility bound of the quantized LWE based encryption in [MS23, Bound 2], where $R = 1 - o\left(\frac{1}{\log(q)}\right)$ is impossible for the same level of quantization. The perfect rate of $\mathsf{LWQ}_{E=\Lambda,\Lambda}$ arises from the principle of noise matching in LWQ.

6 Conclusions and Open Questions

The paper has explored a novel hardness assumption termed LWQ, similar to the LWR assumption, but is parameterized by an arbitrary lattice Λ. By choosing Λ to be a near optimal lattice quantizer, one obtains a variant of LWR where the noise is Gaussian-like, rather than bounded over an ℓ_∞ ball (which is typical for LWR). The LWQ assumption, by leveraging the hardness of LWE and the

efficiency of vector quantization, enables the creation of cryptographic primitives that are not only secure but also more efficient and practical.

To reduce the bandwidth of LWE-based applications, algebraic variants of LWE has been developed, including Ring-LWE [LPR10], Module-LWE [LS15], Middle-Product-LWE [BBD+19], and Cyclic-LWE [GMLV22]. 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 [BLL+15]. 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.

The random dither $\mathbf{d}$ in LWQ is a minor drawback, since it needs to be shared between encryption and decryption. Would it be possible to derandomize LWQ for computational security, yielding a variant of LWQ in which the dither is not needed?

Our analysis of LWQ reveals an interesting phenomenon: in the security analysis of lattice-based cryptosystems involving LWR or LWER, such as CRYSTALS-Kyber [SAB+22], 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

[AA23] Erik Agrell and Bruce Allen. On the best lattice quantizers. *IEEE Transactions on Information Theory*, 69(12):7650–7658, 2023.

[AKPW13] Joël Alwen, Stephan Krenn, Krzysztof Pietrzak, and Daniel Wichs. Learning with rounding, revisited - new reduction, properties and applications. In Ran Canetti and Juan A. Garay, editors, *CRYPTO 2013, Part I*, volume 8042 of *LNCS*, pages 57–74, Santa Barbara, CA, USA, August 18–22, 2013. Springer, Berlin, Heidelberg, Germany.

[Arı09] E. Arıkan. 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.

[AT09] E Arıkan and IE. Telatar. On the rate of channel polarization. In *Proc. 2009 IEEE Int. Symp. Inform. Theory*, pages 1493–1495, Seoul, South Korea, June 2009.

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

[BBD+19] Shi Bai, Katharina Boudgoust, Dipayan Das, Adeline Roux-Langlois, Weiqiang Wen, and Zhenfei Zhang. Middle-product learning with rounding problem and its applications. In Steven D. Galbraith and Shiho Moriai, editors, *ASIACRYPT 2019, Part I*, volume 11921 of *LNCS*, pages 55–81, Kobe, Japan, December 8–12, 2019. Springer, Cham, Switzerland.

[BDGM19] Zvika Brakerski, Nico Döttling, Sanjam Garg, and Giulio Malavolta. Leveraging linear decryption: Rate-1 fully-homomorphic encryption and time-lock puzzles. In Dennis Hofheinz and Alon Rosen, editors, *TCC 2019, Part II*, volume 11892 of *LNCS*, pages 407–437, Nuremberg, Germany, December 1–5, 2019. Springer, Cham, Switzerland.

[BGM+16] Andrej Bogdanov, Siyao Guo, Daniel Masny, Silas Richelson, and Alon Rosen. On the hardness of learning with rounding over small modulus. In Eyal Kushilevitz and Tal Malkin, editors, *TCC 2016-A, Part I*, volume 9562 of *LNCS*, pages 209–224, Tel Aviv, Israel, January 10–13, 2016. Springer, Berlin, Heidelberg, Germany.

[BLL+15] Shi Bai, Adeline Langlois, Tancrède Lepoint, Damien Stehlé, and Ron Steinfeld. Improved security proofs in lattice-based cryptography: Using the rényi divergence rather than the statistical distance. In Tetsu Iwata and Jung Hee Cheon, editors, *Advances in Cryptology – ASIACRYPT 2015*, pages 3–24, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[BLP+13] Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé. Classical hardness of learning with errors. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *45th ACM STOC*, pages 575–584, Palo Alto, CA, USA, June 1–4, 2013. ACM Press.

[BPR12] Abhishek Banerjee, Chris Peikert, and Alon Rosen. Pseudorandom functions and lattices. In David Pointcheval and Thomas Johansson, editors, *EUROCRYPT 2012*, volume 7237 of *LNCS*, pages 719–737, Cambridge, UK, April 15–19, 2012. Springer, Berlin, Heidelberg, Germany.

[BS83] ES Barnes and NJA Sloane. The optimal lattice quantizer in three dimensions. *SIAM Journal on Algebraic Discrete Methods*, 4(1):30–41, 1983.

[CKKS17] Jung Hee Cheon, Andrey Kim, Miran Kim, and Yong Soo Song. Homomorphic encryption for arithmetic of approximate numbers. In Tsuyoshi Takagi and Thomas Peyrin, editors, *ASIACRYPT 2017, Part I*, volume 10624 of *LNCS*, pages 409–437, Hong Kong, China, December 3–7, 2017. Springer, Cham, Switzerland.

[CKLS18] Jung Hee Cheon, Duhyeong Kim, Joohee Lee, and Yongsoo Song. Lizard: Cut off the tail! A practical post-quantum public-key encryption from LWE and LWR. In Dario Catalano and Roberto De Prisco, editors, *SCN 18*, volume 11035 of *LNCS*, pages 160–177, Amalfi, Italy, September 5–7, 2018. Springer, Cham, Switzerland.

[CKM+17] Henry Cohn, Abhinav Kumar, Stephen Miller, Danylo Radchenko, and Maryna Viazovska. The sphere packing problem in dimension 24. *Annals of Mathematics*, 185(3):1017–1033, 2017.

[Cov99] Thomas M Cover. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 1999.

[CS99] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Springer, New York, 3 edition, 1999.

[DKRV18] Jan-Pieter D'Anvers, Angshuman Karmakar, Sujoy Sinha Roy, and Frederik Vercauteren. Saber: Module-LWR based key exchange, CPA-secure encryption and CCA-secure KEM. In Antoine Joux, Abderrahmane Nitaj, and Tajjeeddine Rachidi, editors, *AFRICACRYPT 18*, volume 10831 of *LNCS*, pages 282–305, Marrakesh, Morocco, May 7–9, 2018. Springer, Cham, Switzerland.

[EXMH19] Zeynep B Kaykac Egilmez, Luping Xiang, Robert G Maunder, and Lajos Hanzo. The development, operation and performance of the 5G polar codes. *IEEE Communications Surveys & Tutorials*, 22(1):96–122, 2019.

[FTC00] G.D. Forney, M.D. Trott, and Sae-Young Chung. Sphere-bound-achieving coset codes and multilevel coset codes. *IEEE Transactions on Information Theory*, 46(3):820–850, May 2000.

[GMLV22] Charles Grover, Andrew Mendelsohn, Cong Ling, and Roope Vehkalahti. Non-commutative ring learning with errors from cyclic algebras. *Journal of Cryptology*, 35(3):22, July 2022.

[GPV08] Craig Gentry, Chris Peikert, and Vinod Vaikuntanathan. Trapdoors for hard lattices and new cryptographic constructions. In Richard E. Ladner and Cynthia Dwork, editors, *40th ACM STOC*, pages 197–206, Victoria, BC, Canada, May 17–20, 2008. ACM Press.

[GSW13] Craig Gentry, Amit Sahai, and Brent Waters. Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In Ran Canetti and Juan A. Garay, editors, *CRYPTO 2013, Part I*, volume 8042 of *LNCS*, pages 75–92, Santa Barbara, CA, USA, August 18–22, 2013. Springer, Berlin, Heidelberg, Germany.

[KU10] S.B. Korada and R.L. Urbanke. Polar codes are optimal for lossy source coding. *IEEE Transactions on Information Theory*, 56(4):1751–1768, April 2010.

[LPR10] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. In Henri Gilbert, editor, *EUROCRYPT 2010*, volume 6110 of *LNCS*, pages 1–23, French Riviera, May 30 – June 3, 2010. Springer, Berlin, Heidelberg, Germany.

[LS15] Adeline Langlois and Damien Stehlé. Worst-case to average-case reductions for module lattices. *Designs, Codes and Cryptography*, 75(3):565–599, 2015.

[LS24] Shuiyin Liu and Amin Sakzad. Crystals-kyber with lattice quantizer. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2886–2891, 2024.

[LSL21] Ling Liu, Jinwen Shi, and Cong Ling. Polar lattices for lossy compression. *IEEE Transactions on Information Theory*, 67(9):6140–6163, 2021.

[LYLW19] Ling Liu, Yanfei Yan, Cong Ling, and Xiaofu Wu. Construction of capacity-achieving lattice codes: Polar lattices. *IEEE Trans. Commun.*, 67(2):915–928, Feb. 2019.

[MP12] Daniele Micciancio and Chris Peikert. Trapdoors for lattices: Simpler, tighter, faster, smaller. In David Pointcheval and Thomas Johansson, editors, *EUROCRYPT 2012*, volume 7237 of *LNCS*, pages 700–718, Cambridge, UK, April 15–19, 2012. Springer, Berlin, Heidelberg, Germany.

[MR04] Daniele Micciancio and Oded Regev. Worst-case to average-case reductions based on Gaussian measures. In *45th FOCS*, pages 372–381, Rome, Italy, October 17–19, 2004. IEEE Computer Society Press.

[MS23] Daniele Micciancio and Mark Schultz. Error correction and ciphertext quantization in lattice cryptography. In Helena Handschuh and Anna Lysyanskaya, editors, *CRYPTO 2023, Part V*, volume 14085 of *LNCS*, pages 648–681, Santa Barbara, CA, USA, August 20–24, 2023. Springer, Cham, Switzerland.

[MT14] Ryuhei Mori and Toshiyuki Tanaka. Source and channel polarization over finite fields and reed–solomon matrices. *IEEE Transactions on Information Theory*, 60(5):2720–2736, 2014.

[NR23] Parker Newton and Silas Richelson. A lower bound for proving hardness of learning with rounding with polynomial modulus. In Helena Handschuh and Anna Lysyanskaya, editors, *CRYPTO 2023, Part V*, volume 14085 of *LNCS*, pages 805–835, Santa Barbara, CA, USA, August 20–24, 2023. Springer, Cham, Switzerland.

[Pei09] Chris Peikert. Public-key cryptosystems from the worst-case shortest vector problem: extended abstract. In Michael Mitzenmacher, editor, *41st ACM STOC*, pages 333–342, Bethesda, MD, USA, May 31 – June 2, 2009. ACM Press.

[Pei10] Chris Peikert. An efficient and parallel Gaussian sampler for lattices. In Tal Rabin, editor, *CRYPTO 2010*, volume 6223 of *LNCS*, pages 80–97, Santa Barbara, CA, USA, August 15–19, 2010. Springer, Berlin, Heidelberg, Germany.

[PHTT11] R. Pedarsani, S.H. Hassani, I. Tal, and I.E. Telatar. On the construction of polar codes. In *Proc. 2011 IEEE Int. Symp. Inform. Theory*, pages 11–15, St. Petersburg, Russia, July 2011.

[PS19] Chris Peikert and Sina Shiehian. Noninteractive zero knowledge for NP from (plain) learning with errors. In Alexandra Boldyreva and Daniele Micciancio, editors, *CRYPTO 2019, Part I*, volume 11692 of *LNCS*, pages 89–114, Santa Barbara, CA, USA, August 18–22, 2019. Springer, Cham, Switzerland.

[Reg05] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In Harold N. Gabow and Ronald Fagin, editors, *37th ACM STOC*, pages 84–93, Baltimore, MA, USA, May 22–24, 2005. ACM Press.

[SAB+22] Peter Schwabe, Roberto Avanzi, Joppe Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, John M. Schanck, Gregor Seiler, Damien Stehlé, and Jintai Ding. CRYSTALS-KYBER. Technical report, National Institute of Standards and Technology, 2022. available at https://csrc.nist.gov/Projects/post-quantum-cryptography/selected-algorithms-2022.

[Sas12] Eren Sasoglu. Polar codes for discrete alphabets. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 2137–2141, 2012.

[STA09] Eren Sasoglu, Emre Telatar, and Erdal Arikan. Polarization for arbitrary discrete memoryless channels. In *2009 IEEE Information Theory Workshop*, pages 144–148, 2009.

[TV13] I Tal and A Vardy. How to construct polar codes. *IEEE Transactions on Information Theory*, 59(10):6562–6582, Oct. 2013.

[Via17] Maryna S Viazovska. The sphere packing problem in dimension 8. *Annals of Mathematics*, pages 991–1015, 2017.

[WD21] Hsin-Po Wang and Iwan M. Duursma. Log-logarithmic time pruned polar coding. *IEEE Transactions on Information Theory*, 67(3):1509–1521, 2021.

[Zam14] R. Zamir. *Lattice Coding for Signals and Networks*. Cambridge University Press, Cambridge, UK, 2014.
[ZF96a] R. Zamir and M. Feder. Information rates of pre/post-filtered dithered quantizers. *IEEE Transactions on Information Theory*, 42(5):1340–1353, 1996.
[ZF96b] Ram Zamir and Meir Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, 1996.

A Background of Polar Code/Lattice

Polar coding [Arı09] 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 $\begin{bmatrix} 4 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$.

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*: $\mathbf{m} = \mathbf{B}^{-1}\mathbf{q}$.

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 [LSL21]. 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 $\frac{1}{2}\log\left(\frac{\sigma_s^2}{\Delta}\right)$. This insight suggests that the rate of the polar lattice quantizer can be finely adjusted to approach $\frac{1}{2}\log\left(\frac{\sigma_s^2}{\Delta}\right)$. 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.

B Proof of Lemma 4

Proof. Using the telescoping expansion [KU10, Lemma 4]

$$B^{1:n} - A^{1:n} = \sum_{i=1}^{n} (B^i - A^i) A^{1:i-1} B^{i+1:n}, \tag{43}$$

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

$$\begin{aligned}
&2\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right) \\
&= \sum_{u_1^{[m]},y^{[m]}} \left|\mathsf{Q}(u_1^{[m]}, y^{[m]}) - \mathsf{P}(u_1^{[m]}, y^{[m]})\right| \\
&= \sum_{u_1^{[m]},y^{[m]}} \left| \sum_i \left(\mathsf{Q}(u_1^i | u_1^{1:i-1}, y^{[m]}) - \mathsf{P}(u_1^i | u_1^{1:i-1}, y^{[m]})\right) \right. \\
&\qquad \left. \cdot \left(\prod_{j=1}^{i-1} \mathsf{P}(u_1^j | u_1^{1:j-1}, y^{[m]})\right) \left(\prod_{j=i+1}^{m} \mathsf{Q}(u_1^j | u_1^{1:j-1}, y^{[m]})\right) \mathsf{P}\left(y^{[m]}\right) \right|
\end{aligned} \tag{44}$$

$$\begin{aligned}
&\overset{(a)}{\leq} \sum_{i\in\mathcal{F}_1} \sum_{u_1^{[m]},y^{[m]}} \left|\mathsf{Q}(u_1^i | u_1^{1:i-1}, y^{[m]}) - \mathsf{P}(u_1^i | u_1^{1:i-1}, y^{[m]})\right| \left(\prod_{j=1}^{i-1} \mathsf{P}(u_1^j | u_1^{1:j-1}, y^{[m]})\right) \\
&\qquad \cdot \left(\prod_{j=i+1}^{m} \mathsf{Q}(u_1^j | u_1^{1:j-1}, y^{[m]})\right) \mathsf{P}\left(y^{[m]}\right) \\
&= \sum_{i\in\mathcal{F}_1} \sum_{u_1^{1:i},y^{[m]}} \left|\mathsf{Q}(u_1^i | u_1^{1:i-1}, y^{[m]}) - \mathsf{P}(u_1^i | u_1^{1:i-1}, y^{[m]})\right| \left(\prod_{j=1}^{i-1} \mathsf{P}(u_1^j | u_1^{1:j-1}, y^{[m]})\right) \mathsf{P}\left(y^{[m]}\right) \\
&= \sum_{i\in\mathcal{F}_1} \sum_{u_1^{1:i-1},y^{[m]}} 2\mathsf{P}\left(u_1^{1:i-1}, y^{[m]}\right) \Delta\left(\mathsf{Q}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}, \mathsf{P}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}\right) \\
&\overset{(b)}{\leq} \sum_{i\in\mathcal{F}_1} \sum_{u_1^{1:i-1},y^{[m]}} \mathsf{P}\left(u_1^{1:i-1}, y^{[m]}\right) \sqrt{2\ln 2 D_{KL}\left(\mathsf{P}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}} || \mathsf{Q}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}\right)}
\end{aligned} \tag{45}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2 \sum_{u_1^{1:i-1},y^{[m]}} \mathsf{P}\left(u_1^{1:i-1},y^{[m]}\right) D_{KL}\left(\mathsf{P}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}||\mathsf{Q}_{U_1^i|U_1^{1:i-1}=u_1^{1:i-1},Y^{[m]}=y^{[m]}}\right)} \\
&= \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2 D_{KL}\left(\mathsf{P}_{U_1^i}||\mathsf{Q}_{U_1^i}|U_1^{1:i-1},Y^{[m]}\right)} \\
&\overset{(d)}{=} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2\left(1-H(U_1^i|U_1^{1:i-1},Y^{[m]})\right)} \\
&\overset{(e)}{\leq} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2\left(1-Z(U_1^i|U_1^{1:i-1},Y^{[m]})^2\right)} \\
&\overset{(f)}{\leq} m\sqrt{4\ln 2\cdot 2^{-m^\beta}}
\end{aligned} \tag{46}
$$

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 $\mathcal{F}_1$.

□

C KL Divergence

Lemma 6. *Let $E\sim\mathcal{D}_{\mathbb{Z},\sigma}$ be a discrete Gaussian random variable, and let $E'=E \mod q\mathbb{Z}$ be the residue in $[-2^{r-1},2^{r-1})$. The Kullback-Leibler divergence $D_{KL}(\mathsf{P}_{E'}||\mathsf{P}_E)$ between $\mathsf{P}_{E'}$ and P_E is upper-bounded as follows:*

$$
D_{KL}(\mathsf{P}_{E'}||\mathsf{P}_E)\triangleq\sum_{\lambda\in\mathbb{Z}}\mathsf{P}_{E'}(\lambda)\ln\frac{\mathsf{P}_{E'}(\lambda)}{\mathsf{P}_E(\lambda)}\leq\frac{20}{\sqrt{2\pi\sigma^2}}q\cdot\exp\left(-\frac{q^2}{8\sigma^2}\right), \tag{47}
$$

where $q=2^r$.

Proof. By the definition of $D_{\mathbb{Z},\sigma^2}$,

$$
\begin{aligned}
\mathsf{P}_{E'}(\lambda)&=\sum_{z\in\mathbb{Z}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\lambda+zq)^2}{2\sigma^2}} &(48)\\
&=\frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}}+\sum_{z\in\mathbb{Z}^+}e^{-\frac{(\lambda+zq)^2}{2\sigma^2}}+\sum_{z\in\mathbb{Z}^+}e^{-\frac{(\lambda-zq)^2}{2\sigma^2}}\right) &(49)\\
&\leq\frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}}+2\sum_{z\in\mathbb{Z}^+}e^{-\frac{(|\lambda|-zq)^2}{2\sigma^2}}\right), &(50)
\end{aligned}
$$

where $\mathbb{Z}^+$ denotes the set of positive integers.

Observe that $e^{-\frac{(|\lambda|-(i+1)q)^2}{2\sigma^2}}/e^{-\frac{(|\lambda|-iq)^2}{2\sigma^2}} = e^{\frac{(2|\lambda|-(2i+1)q)q}{2\sigma^2}} \leq e^{-\frac{q^2}{\sigma^2}}$ for $|\lambda| \leq \frac{q}{2}$ and $i \geq 1$. Therefore,

$$\mathsf{P}_{E'}(\lambda) \leq \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}} + 2\sum_{z\in\mathbb{Z}^+} e^{-\frac{(|\lambda|-zq)^2}{2\sigma^2}}\right) \tag{51}$$

$$\leq \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}} + 2\cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}}\cdot\frac{1}{1-e^{-\frac{q^2}{\sigma^2}}}\right) \tag{52}$$

$$\leq \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}} + 4\cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}}\right), \tag{53}$$

for large q such that $e^{-\frac{q^2}{\sigma^2}} \leq \frac{1}{2}$.

For the Kullback-Leibler divergence $D_{KL}(\mathsf{P}_{E'}||\mathsf{P}_E)$,

$$D_{KL}(\mathsf{P}_{E'}||\mathsf{P}_E) = \sum_{\lambda\in\mathbb{Z}} \mathsf{P}_{E'}(\lambda)\ln\frac{\mathsf{P}_{E'}(\lambda)}{\mathsf{P}_E(\lambda)} \tag{54}$$

$$\leq \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}} + 4\cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}}\right)\ln\left(1+4e^{\frac{\lambda^2-(|\lambda|-q)^2}{2\sigma^2}}\right) \tag{55}$$

$$\leq 4\cdot\sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{\lambda^2}{2\sigma^2}} + 4\cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}}\right)e^{\frac{2q|\lambda|-q^2}{2\sigma^2}} \tag{56}$$

$$= 4\cdot\sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}} + 16\cdot\sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(|\lambda|-2q)^2}{2\sigma^2}}e^{\frac{q^2}{\sigma^2}} \tag{57}$$

$$\leq 4\cdot\sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{q^2}{8\sigma^2}} + 16\cdot\sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{q^2}{8\sigma^2}} \tag{58}$$

$$= \frac{20}{\sqrt{2\pi\sigma^2}}q\cdot\exp\left(-\frac{q^2}{8\sigma^2}\right), \tag{59}$$

where we use the inequality $\ln(1+x) \leq x$ and the relationship $|\lambda| \leq \frac{q}{2}$ in the third and fifth steps, respectively.

Lemma 7. *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* (32) *and* (33) *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* (32) *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:*

$$D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right) \leq 2\ln 2\cdot m2^{-m^\beta}. \tag{60}$$

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

$$D_{KL}\left(\mathsf{P}_{X^{[m]},Y^{[m]}}||\mathsf{Q}_{X^{[m]},Y^{[m]}}\right) = D_{KL}\left(\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}}||\mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}}\right) \tag{61}$$

$$\leq 2\ln 2 \cdot rm2^{-m^\beta}. \tag{62}$$

Proof. For the 1st level,

$$\begin{aligned}
&D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right) \\
&= \ln 2 \cdot \sum_{u_1^{[m]},y^{[m]}} \mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log\frac{\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)}{\mathsf{Q}\left(u_1^{[m]},y^{[m]}\right)} \\
&= \ln 2 \cdot \sum_{u_1^{[m]},y^{[m]}} \mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log\frac{\mathsf{P}\left(u_1^{[m]}|y^{[m]}\right)}{\mathsf{Q}\left(u_1^{[m]}|y^{[m]}\right)} \\
&= \ln 2 \cdot \sum_{u_1^{[m]},y^{[m]}} \mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log\frac{\prod_{i=1}^m \mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})}{\prod_{i=1}^m \mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]})} \\
&= \ln 2 \cdot \sum_{u_1^{[m]},y^{[m]}} \mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\sum_{i\in\mathcal{F}_1}\log\frac{\mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})}{\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]})} \\
&= \ln 2 \cdot \sum_{i\in\mathcal{F}_1}(1-H(U_1^i|U_1^{1:i-1},Y^{[m]})) \\
&\leq \ln 2 \cdot \sum_{i\in\mathcal{F}_1}(1-Z(U_1^i|U_1^{1:i-1},Y^{[m]})^2) \\
&\leq 2\ln 2 \cdot m2^{-m^\beta},
\end{aligned} \tag{63}$$

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

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

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

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 (63) exactly, since it can be written as

$$\begin{aligned}
&\mathsf{E}_{U_1^{[m]},Y^{[m]}}\left[D_{KL}\left(\mathsf{P}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}\right)\right] \\
&\qquad= \ln 2 \cdot \sum_{u_{1:2}^{[m]},y^{[m]}} \mathsf{P}\left(u_{1:2}^{[m]},y^{[m]}\right)\log\frac{\mathsf{P}\left(u_2^{[m]}|u_1^{[m]},y^{[m]}\right)}{\mathsf{Q}\left(u_2^{[m]}|u_1^{[m]},y^{[m]}\right)}.
\end{aligned} \tag{64}$$

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