Learning with Quantization
Construction, Hardness, and Applications

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. This paper presents a generalization of the Learning With Rounding (LWR) problem, initially introduced by Banerjee, Peikert, and Rosen, by applying the perspective of vector quantization. In LWR, noise is induced by scalar quantization. By considering a new variant termed Learning With Quantization (LWQ), we explore large-dimensional fast-decodable lattices with superior quantization properties, aiming to enhance the compression performance over scalar quantization. We identify polar lattices as exemplary structures, effectively transforming LWQ into a problem akin to Learning With Errors (LWE), whose distribution of quantization error is statistically close to discrete Gaussian. We present two applications of LWQ: Lily, a smaller ciphertext public key encryption (PKE) scheme, and quancryption, a privacy-preserving secret-key encryption scheme. Lily achieves smaller ciphertext sizes without sacrificing security, while quancryption achieves a source-ciphertext ratio larger than 1.

Keywords: Lattice-Based Cryptography · Learning with Quantization · Polar Lattice · Ciphertext Compression.

1 Introduction

The Learning With Errors (LWE) problem [41] has been a versatile and foundational element in cryptography, providing strong security assurances and enabling a wide range of applications from basic encryption to cutting-edge privacy-preserving technologies [14, 40, 3, 42, 36]. At its core, LWE involves solving linear equations perturbed by small random errors, is provably as hard as certain worst-case lattice problems. One of the primary drawbacks of LWE-based applications is their inefficiency in terms of key and ciphertext sizes. Standard LWE-based schemes require large matrices and vectors to represent public keys and ciphertexts. This results in significant storage and bandwidth requirements, making conventional LWE less competitive.

A popular approach to addressing the inefficiency of LWE-based applications is to develop algebraic variants of LWE, including Ring-LWE [35], Module-LWE

[26], Middle-Product-LWE [8], and Cyclic-LWE [22]. These variants offer more compact representations and faster arithmetic operations, making them more suitable for practical implementations. However, the computational hardness of these algebraic variants is frequently challenged. A common method to attack these variants is by transforming them into worst-case lattice problems, such as the algebraic shortest vector problem (SVP) [2]. Over the past decade, significant advancements have been made in this area, including the development of algebraic LLL/BKZ algorithms [23, 27, 37, 24]. It is reasonable to conjecture that any algebraic simplification may lead to a certain reduction in the security level.

For long-term security considerations, a potentially fruitful avenue is to manipulate the error terms within the framework of LWE. One approach is to directly quantize the vector $\mathbf{As}$ in LWE, resulting in an observation term expressed as $\mathbf{As} + \mathbf{e}_Q$, where $\mathbf{e}_Q$ represents the error induced by quantization. This approach offers a dual benefit: firstly, it eliminates the conventional discrete Gaussian error $\mathbf{e}$, and secondly, it reduces the size of $\mathbf{As}$. Furthermore, a higher degree of quantization correlates with a heightened level of security. It's important to note that quantization entails reducing the information content of data, such as truncating a 32-bit datum to 8 bits. Quantization represents a narrower concept compared to compression, which involves reducing data size, either in a lossless manner (by entropy coding, the original data can be perfectly reconstructed from the compressed data) or in a lossy manner (by quantization, perfect reconstruction is not possible). An outstanding challenge remains in enhancing the efficiency of quantization, as well as investigating whether the associated problem is computationally as hard as LWE.

1.1 Our Results

This work formalize the above concept by defining the Learning With Quantization (LWQ) problem. The problem is based on an efficient lattice quantizer Q_Λ. By properly choosing a pair of nested lattices, the compression ratio can be flexibly designed, and the properties of the quantization errors can be easily analyzed. LWQ is a broader concept than Learning With Rounding (LWR) [9]: the quantization can be based on a higher dimensional lattice, rather than one-dimensional lattice/scalar quantization in LWR.

We prove that LWQ is as hard as LWE, if the quantization lattice Λ is a polar lattice. In a high level, it says that the statistical distance between $\mathbf{A}, Q_\Lambda(\mathbf{As})$ and $\mathbf{A}, \mathbf{As} + \mathbf{e}$ is vanishing with respect to the dimension m of quantization. We present an information-theoretic proof based on the properties of polar codes. This line of proof differs greatly from the hardness proof of LWR: they prove that the quantization of $\mathbf{As}$ and $\mathbf{As} + \mathbf{e}$ are the same with high probability. The takeaway of our proof is that, the quantization error of LWQ is close to a discrete Gaussian distribution, while that of LWR is close to a uniform distribution over a hypercube.

LWQ can be applied to replace LWE or LWR in many cryptography scenarios, either to reduce the size of ciphertext or to achieve a higher security level. We present two applications in this work. i) A PKE scheme called Lily.

It employs LWE in the key generation stage and LWQ in the encryption stage. Without compromising the security level, Lily has about 20% $\sim$ 30% smaller ciphertext size than Frodo-PKE. ii) A secret key encryption scheme referred to as quancryption. Its message space is designed by the quantization. As such, the source-ciphertext ratio of the scheme is larger than 1.

1.2 Related Work

Scalar quantization has been adapted to define LWR, a variant of LWE. Specifically, Banerjee, Peikert, and Rosen [9] introduced the LWR problem, serving as a derandomized version of LWE. By replacing Gaussian sampling in LWE with deterministic rounding, LWR samples can be generated faster and with less randomness. The hardness of LWR has been established only for restricted settings. Reference [9] demonstrated that if one can distinguish the LWR distribution from uniform distribution with advantage δ, then one can also distinguish LWE with advantage $\delta - O(mBp/q)$, where m defines the number of LWR samples, B defines a symmetric bounded interval, and q, p denote the original and reduced modulus, respectively. The size of q was reduced by assuming it is a prime in [5], while [11] showed that q can be polynomial when the given number of LWR samples is bounded. Restrictions on the number of samples were removed in [38], and a lower bound for proving the hardness of LWR with polynomial modulus was provided.

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. A prevalent compression technique is scalar quantization, also known as modulus switching/modulus reduction. For instance, CKKS homomorphic encryption [14] employs simple modulus reduction to a smaller modulus before computation on ciphertexts at different levels, while CRYSTALS-Kyber [42] utilizes it for ciphertext compression. By increasing the dimension of quantization, vector quantization can be expected to outperform scalar quantization [46]. Certain performance benefits of vector quantization have been justified in the secret-key ecnryption framework [36], and to reduce the ciphertext rate of CRYSTALS-Kyber [33].

The inquiry into optimal lattices for quantization, aiming for the smallest average distortion, is different from sphere packing [44, 16]. The theoretical proof of optimal lattice quantizers has been limited to dimensions up to 3 (*i.e.*, $\mathbb{Z}$, A_2, A_3^*) [10], although efforts to identify good lattice quantizers have resulted in periodic updates of tables for small-dimensional lattices $n \leq 24$ [4, 1]. Closely related research focuses on the pursuit of optimal quantization lattices in the information theory community, aiming to achieve the optimal rate-distortion bound [47]. In this context, dithered quantization has been under development for decades [21, 48], 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 [46], decoding a high-dimensional random lattice poses challenges, albeit mitigated by the law

of large numbers. For a continuous Gaussian source, an explicit construction of polar lattices to achieve the rate-distortion bound has been presented in [31], where the computational complexity of the quantizer is $O(m \log m)$.

The polar lattices investigated in this work originate from polar codes. These lattices are derived using Construction D, which means combining multiple binary codes to construct lattices. 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 [6]. 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 [45]. Due to their outstanding performance, polar codes have been widely adopted in various practical applications, including fifth-generation (5G) wireless communication networks [19].

2 Preliminaries

Table 1 summarizes a few important notations in this paper for easy reference. 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 $\text{negl} : \mathbb{N} \to \mathbb{R}^+$ is negligible if for every positive polynomial $p(\lambda)$, there exists $\lambda_0 \in \mathbb{N}$ such that $\text{negl}(\lambda) < \frac{1}{p(\lambda)}$ for all $\lambda > \lambda_0$.

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 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 is a discrete subgroup $\Lambda \subseteq \mathbb{R}^n$. 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.

Definition 1 (Fundamental Cell). *A fundamental cell 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 cells intersect, then $\mathbf{v}$ must equal $\mathbf{w}$.*

For instance, the half-open Voronoi cell $\mathcal{V}_\Lambda$ is a fundamental cell. This cell encompasses the set of points in $\mathbb{R}^n$ that are closer to a specific lattice point (referred to as the generating lattice point) within Λ than to any other lattice point. Essentially, it defines the region surrounding each generating lattice point where it is the closest lattice point.

By referring to a quantizer Q_Λ, it refers to a function that maps a vector $\mathbf{y} \in \mathbb{R}^n$ to the nearest lattice point in Λ. This is formulated as:

$$Q_\Lambda(\mathbf{y}) = \arg\min_{\lambda\in\Lambda} \|\mathbf{y}-\lambda\|. \tag{1}$$

This problem is in the form of solving a closest vector problem (CVP). But a few special features of a quantizer should be bear in mind: i) The lattice Λ is not random, but rather fast decodable (e.g., the hypercube lattice: $\Lambda = \mathbb{Z}^n$, or the tensor produce of E_8: $\Lambda = \mathbb{Z}^{n/8} \otimes E_8$). ii) The implicit selection of a half-open Voronoi cell is crucial, as it allows Q_Λ to consistently choose a single representative when multiple lattice points are equidistant from $\mathbf{y}$. iii) It can be think of as the decoding step of error correction. The duality of error correction codes and quantizers is presented in Appendix A.

Definition 2 (Dithered quantizer). *A dithered quantizer over lattice Λ is defined by sampling $\mathbf{g} \leftarrow \mathcal{V}_\Lambda$ and outputting*

$$Q_{\Lambda+\mathbf{g}}(\mathbf{y}) = \mathbf{g} + Q_\Lambda(\mathbf{y}-\mathbf{g}). \tag{2}$$

Definition 3 (Second moment). *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 fundamental Voronoi cell $\mathcal{V}$:*

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

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

For a dithered quantizer, $\mathbf{y} - Q_{\Lambda+\mathbf{g}}(\mathbf{y})$ is uniformly distributed over $\mathcal{V}_\Lambda$, so the averaged quantization error of the dithered quantizer can be quantified by $\tilde{\sigma}^2(\Lambda)$: for any distribution of $\mathbf{y}$, with $\mathbf{g} \leftarrow \mathcal{V}_\Lambda$, then

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{y} - Q_{\Lambda+\mathbf{g}}(\mathbf{y})\right\|^2 = \tilde{\sigma}^2(\Lambda). \tag{3}$$

The normalized second moment (NSM), i.e., the second-moment to volume ratio, is defined as

$$G(\Lambda) = \frac{\tilde{\sigma}^2(\Lambda)}{\det^{2/n}(\Lambda)}. \tag{4}$$

The minimum possible value of $G(\Lambda)$ over all lattices in $\mathbb{R}^n$ is denoted by G_n.

Definition 4 (Quantization-good). *A sequence of lattices $\Lambda^{(n)}$ with growing dimension is called good for mean squared error quantization if*

$$\lim_{n\to\infty} G(\Lambda^{(n)}) = \frac{1}{2\pi e}. \tag{5}$$

The integer lattice $\mathbb{Z}$, checker-board lattice D_4, Gosset lattice E_8, and Leech lattice Λ_{24} have the best reported NSMs in their respective dimensions [1]:

$$0.08333, 0.07660, 0.07168, 0.06577.$$

2.2 Statistics and Privacy

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

Definition 5 (Statistical Distance). *The statistical distance between two probability distributions P and Q over the same sample space $\mathcal{X}$ is defined as:*

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

We will also make use of the Rényi divergence as an alternative to the statistical distance to measure the similarity between two distributions.

Definition 6 (Rényi Divergence). *Rényi divergence of order α between two distributions P and Q over the same sample space $\mathcal{X}$ is defined as:*

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\sum_{x\in\mathcal{X}} P(x)\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}. \tag{6}$$

The limit as $\alpha \to 1$ amounts to KL Divergence.

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:*

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

One of our applications is related to showing the quantization error can induce differential privacy. A generalized notion of differential privacy is given as follows.

Definition 8 (KL Differential Privacy, [29]). *For $t \in \mathbb{R}_{\geq 0}$, let $\mathcal{M}_t : \mathcal{B} \to \mathcal{C}$ be a family of randomized algorithms, where $\mathcal{B}$ is a normed space with norm $\|\cdot\| : \mathcal{B} \to \mathbb{R}_{\geq 0}$. Let $\rho \in \mathbb{R}$ be a privacy bound. We say that the family $\mathcal{M}_t$ is ρ-differentially private (ρ-DP) if, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{B}$ with $\|\mathbf{x} - \mathbf{x}'\| \leq t$,*

$$D_{KL}(\mathcal{M}_t(\mathbf{x}) || \mathcal{M}_t(\mathbf{x}')) \leq \rho.$$

3 LWQ

3.1 Definition and Representation

This section reviews the definitions of LWE [41] and LWR [9], and presents our generalization called LWQ.

Definition 9 (LWE/LWR/LWQ distributions). *Let $n, m, q \in \mathbb{N}$, $p \mid q$ and $p \geq 2$, and Λ be an m-dimensional integer lattice satisfying $q\mathbb{Z}^m \subset \Lambda \subset \mathbb{Z}^m$. For a "secret" $\mathbf{s} \in \mathbb{Z}_q^n$, and an error distribution $\mathcal{D}_e$ over $\mathbb{Z}^m$, samples for the LWE/LWR/LWQ distributions are respectively generated by*

- *LWE distribution $L_{\mathcal{D}_e}(\mathbf{s})$: $\mathbf{A} \leftarrow \mathbb{Z}_q^{m \times n}$, $\mathbf{e} \leftarrow \mathcal{D}_e$, and output $(\mathbf{A}, \mathbf{As} + \mathbf{e}) \in \mathbb{Z}_q^{m \times n} \times \mathbb{Z}_q^m$.*
- *LWR distribution $L_{q/p\mathbb{Z}^m}(\mathbf{s})$: $\mathbf{A} \leftarrow \mathbb{Z}_q^{m \times n}$, and output $(\mathbf{A}, Q_{\frac{q}{p}\mathbb{Z}^m}(\mathbf{As})) \in \mathbb{Z}_q^{m \times n} \times (\mathbb{Z}_q^m \cap \frac{q}{p}\mathbb{Z}^m)$ [4].*
- *LWQ distribution $L_\Lambda(\mathbf{s})$: $\mathbf{A} \leftarrow \mathbb{Z}_q^{m \times n}$, and output $(\mathbf{A}, Q_\Lambda(\mathbf{As})) \in \mathbb{Z}_q^{m \times n} \times \mathbb{Z}_q^m$.*

Definition 10 (Decision LWE/LWR/LWQ problems). *It challenges an adversary to distinguish between*

- *LWE: $L_{\mathcal{D}_e}(\mathbf{s})$ where $\mathbf{s} \leftarrow \mathcal{D}_s$, and a uniform distribution over $\mathbb{Z}_q^{m \times n} \times \mathbb{Z}_q^m$.*
- *LWR: $L_{q/p\mathbb{Z}^m}(\mathbf{s})$ where $\mathbf{s} \leftarrow \mathcal{D}_s$, and a uniform distribution over $\mathbb{Z}_q^{m \times n} \times (\mathbb{Z}_q^m \cap \frac{q}{p}\mathbb{Z}^m)$.*
- *LWQ: $L_\Lambda(\mathbf{s})$ where $\mathbf{s} \leftarrow \mathcal{D}_s$, and a uniform distribution over $\mathbb{Z}_q^{m \times n} \times \mathbb{Z}_q^m$.*

This paper only investigate the decision LWE/LWR/LWQ problems. The search problems are defined in the conventional manner: Given arbitrarily many samples of the LWE/LWR/LWQ distribution where $\mathbf{s} \leftarrow \mathcal{D}_s$, the search problem of LWE/LWR/LWQ asks to recover $\mathbf{s}$.

When the quantizer Q_Λ is deterministic, we have $\mathbf{b} = Q_\Lambda(\mathbf{As}) \in \mathbb{Z}_q^m \cap \Lambda$. In the more general case of a randomized quantizer (can be think of as dithered quantizer), we have $\mathbf{b} = Q_\Lambda(\mathbf{As}) \in \mathbb{Z}_q^m$, but the number of bits required for storage remains $\log_2(q^m / \det(\Lambda))$.

[4] $\frac{p}{q}(\mathbb{Z}_q^m \cap \frac{q}{p}\mathbb{Z}^m) = \mathbb{Z}_p^m$, so this definition is the same as the conventional LWR definition.

We require a method to convert between lattice cosets (the output of a lattice quantizer) and bit streams. For the sake of notational simplicity, we will omit the explicit steps of converting lattice cosets to bit streams during transmission and the reverse process from bit streams to lattice cosets during reception.

Theorem 1. *With $\mathbb{Q}^m \subset \Lambda \subset q\mathbb{Z}^m$, $\mathbf{g} \in \mathbb{Z}_q^m$, there exist a bijection between $\Lambda + \mathbf{g} \cap \mathbb{Z}_q^m$ and $\{0,1\}^{q^m/\det(\Lambda)}$.*

Proof. Denote the full-rank lattice basis of Λ as $\mathbf{B}^* \in \mathbb{Q}^{m\times m}$, and the least common multiplier of denominators in the entries of $\mathbf{B}^*$ as s_{LCM}. Via the Smith Normal Form (SNF) factorization, we have

$$\mathbf{B}^* = \mathbf{USV}, \tag{7}$$

where $\mathbf{S} = s_{LCM}^{-1}\mathrm{diag}(\pi_1, \ldots, \pi_m) = \mathrm{diag}(s_1, \ldots, s_m)$, $\mathbf{U}$ and $\mathbf{V}$ are unimodular matrices. Thus lattice Λ can be generated by basis $\mathbf{B} = \mathbf{B}^*\mathbf{V}^{-1} = \mathbf{US}$, which is referred to as a rectangular form in [34].

The bit-decomposition function BD of $\mathbf{v} \in \Lambda = L(\mathbf{B}) + \mathbf{g}$ is given by

$$\mathrm{BD}_{\mathbf{B},\mathbf{S},q}(\mathbf{v}) = (\mathbf{u}_1, \ldots, \mathbf{u}_m) \in \{0,1\}^{q^m/\det(\Lambda)} \tag{8}$$

where

$$(\mathbf{B}^{-1}(\mathbf{v} - \mathbf{g}))_i/s_i \mod q/s_i = \sum_{k=1}^{\log 2(q/s_i)} 2^k u_{i,k}. \tag{9}$$

It can be verified that BD is a bijection function. □

In summary, both LWR and LWQ sidestep Gaussian sampling by discarding the $\mathbf{e}$ term of LWE. Thanks to the advantage of vector quantization over scalar quantization, the statistics of the LWQ quantization error is more favarable over LWR: it has a smaller second moment with the same information rate, and it can be Gaussian-like.

3.2 Asymptotic Results of Hardness

We prove the asymptotic hardness of LWQ by showing the distributions of LWQ and LWE are information theoretically indistinguishable, for carefully designed polar lattice quantizers. Our main result is to derive the following bound on the statistical distance between the LWQ and LWE distributions. We rewrite the following distributions given earlier in this subsection to serve our purpose.

- Consider the LWE distribution $L_{D^m_{\mathbb{Z},\sigma}}(\mathbf{s})$: $\mathsf{P}_{\mathbf{A},\mathbf{b}}$ where $\mathbf{b} = X^{[m]} = Y^{[m]} + \mathbf{e}$ mod $q\mathbb{Z}$ where $Y^{[m]} = \mathbf{As}$ and $e_i \sim D_{\mathbb{Z},\sigma}$.
- Consider the LWQ distribution $L_\Lambda(\mathbf{s})$: $\mathsf{Q}_{\mathbf{A},\mathbf{b}}$ where $\mathbf{b} = X^{[m]} = Q_\Lambda(Y^{[m]})$ where $Y^{[m]} = \mathbf{As}$, produced by a polar lattice quantizer.

Theorem 2. *There exist a sequence of efficient quantizers Q_{Λ^m}, such that the statistical distance between the LWE distribution $\mathsf{P}_{\mathbf{A},\mathbf{b}}$ and the LWQ distribution $\mathsf{Q}_{\mathbf{A},\mathbf{b}}$ is negligible:*

$$\Delta\left(\mathsf{P}_{\mathbf{A},\mathbf{b}}, \mathsf{Q}_{\mathbf{A},\mathbf{b}}\right) = \mathrm{negl}(\lambda) \tag{10}$$

where λ is the security parameter.

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 $\mathbf{A} \to \mathbf{As} \to \mathbf{b}$. 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 $\mathbf{b} = X^{[m]}$,

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

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

$$\begin{aligned}
&\Delta\left(\mathsf{P}_{\mathbf{A},\mathbf{b}}, \mathsf{Q}_{\mathbf{A},\mathbf{b}}\right) \\
&= \frac{1}{2} \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \frac{1}{2} \sum_{\mathbf{As}} \mathsf{P}_{\mathbf{As}|\mathbf{A}}(\cdot) \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \frac{1}{2} \sum_{\mathbf{As}} \sum_{\mathbf{A}} \mathsf{P}_{\mathbf{As},\mathbf{A}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \frac{1}{2} \sum_{Y^{[m]}} \mathsf{P}_{Y^{[m]}}(\cdot) \sum_{X^{[m]}} \left| \mathsf{P}_{X^{[m]}|Y^{[m]}}(\cdot) - \mathsf{Q}_{X^{[m]}|Y^{[m]}}(\cdot) \right| \\
&= \Delta\left(\mathsf{P}_{X^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right)
\end{aligned} \tag{11}$$

where the second equality of (11) holds since $\mathsf{P}_{\mathbf{As}|\mathbf{A}}$ is an indicator function when $\mathbf{s}$ is fixed. Proof is completed by using the first part of Theorem 6 with an appropriate dimension m set according to λ. □

Theorem 3. *Let $m = m(\lambda)$, $n = n(\lambda)$, $q = q(\lambda)$ with λ being the security parameter. There exist a sequence of efficient quantizers Q_{Λ^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.

- G_0: The challenge flips a coin $b \in \{0,1\}$, based on which either $c \leftarrow L_\Lambda$ or $c \leftarrow \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$ is sent to the adversary $\mathcal{A}$. The adversary outputs a guess $b' \in \{0,1\}$.
- G_1: The challenge flips a coin $b \in \{0,1\}$, based on which either $c \leftarrow L_{D^m_{\mathbb{Z},\sigma}}$ or $c \leftarrow \mathbb{Z}_q^{m\times n} \times \mathbb{Z}_q^m$ is sent to the adversary $\mathcal{A}$. The adversary outputs a guess $b' \in \{0,1\}$.

G_0 and G_1 are computationally indistinguishable as

$$\begin{aligned}\mathsf{Adv}_{G_0,G_1}(\mathcal{A}) &= \left|\Pr(\mathcal{A}(L_\Lambda)=1) - \Pr(\mathcal{A}(L_{D^m_{\mathbb{Z},\sigma}}=1)\right| &(12)\\ &\le \Delta(L_\Lambda, L_{D^m_{\mathbb{Z},\sigma}}) &(13)\\ &= \mathrm{negl}(\lambda). &(14)\end{aligned}$$

where the inequality is due to the data processing inequality of distributions, and the last equality is due to Theorem 2.

As the advantage of winning G_1 is also negligible due to the hardness of LWE, we have

$$\mathsf{Adv}_{G_0}(\mathcal{A}) \le \mathsf{Adv}_{G_0,G_1}(\mathcal{A}) + \mathsf{Adv}_{G_1}(\mathcal{A}) = \mathrm{negl}(\lambda). \quad (15)$$

□

Hardness Proof for General Quantizers The asymptotic approach given above requires the lattice dimension m to grow, thus cannot be applied to quantization lattices of small dimension (e.g., E_8). In this case, one may use the following result for a given dimension m:

Theorem 4 (Divergence from white Gaussianity, Thm 7.3.3, [46]). *If the dither U is uniform over the fundamental Voronoi cell $\mathcal{V}_\Lambda$ of a lattice $\Lambda \in \mathbb{Z}^m$, then its (continuous) KL divergence from white Gaussianity is given by*

$$\frac{1}{m} D_{KL}(U||W) = \frac{1}{2}\log(2\pi e G(\Lambda)),$$

where W is the corresponding white-Gaussian noise $\mathcal{N}(0, G(\Lambda)\det^{2/m}(\Lambda))^m$ with the same variance.

For discrete uniform distributions over $\mathcal{V}_\Lambda$, denoted as $\mathsf{P}_{X^{[m]}|Y^{[m]}}$, and discrete Gaussian distribution $D_{\mathbb{Z},\sigma}$, denoted as $\mathsf{Q}_{X^{[m]}|Y^{[m]}}$, one can show that for some small constant ε,

$$\frac{1}{m} D_{KL}(\mathsf{P}_{X^{[m]}|Y^{[m]}}||\mathsf{Q}_{X^{[m]}|Y^{[m]}}) \le \frac{1}{2}\log(2\pi e G(\Lambda)) + \varepsilon. \quad (16)$$

Together with the data processing inequality and similar techniqes in the proof of Theorem 2, we can show that the KL divergence between LWE and LWQ distortions is bounded.

It is therefore desirable to adopt a quantizer with $G(\Lambda)$ as small as possible. The following lemma shows that we cannot obtain a quantization-good lattice by stacking low-dimensional quantizers, i.e., $G(\Lambda') = G(\mathbb{Z}^{m/k} \otimes \Lambda')$. Its proof is trivial and omitted.

Lemma 1. *Assume that $k \mid m$, $k \geq 2$. If Λ is constructed from the m/k-fold Cartesian product of Λ', i.e., $\Lambda = \mathbb{Z}^{m/k} \otimes \Lambda' \subset \mathbb{R}^m$, then the lattices Λ' and Λ have the same NSM, i.e., $G(\Lambda) = G(\Lambda')$.*

3.3 Alternative Hardness Proof

Here, we quantify the loss of security of LWQ by computing the Rényi divergence between two distributions of finite dimension. This method has its upside and downside. On one hand, Rényi divergence can be more useful than the aforegoing statistical distance in cryptography. On the other hand, the following analysis has the disadvantage in that it only compares with a quantized version of LWE samples, not with LWE samples themselves.

For the sake of theoretical analysis, this section considers $\mathbf{s} \in \mathbb{Z}_q^{n*}$ such that $\mathbf{As}$ admits a uniform distribution in $\mathbb{Z}_q^m$. This adaption is minor as the probability of $\mathbf{s} \in \mathbb{Z}_q^{n*}$ is at least $1 - O(1/2^n)$ for $\mathbf{s} \in \mathbb{Z}_q^n$. We set $\alpha = 2$ in the Rényi divergence in the following analysis.

Lemma 2. *Let $\mathbf{A} \leftarrow \mathbb{Z}_q^{m \times n}$, $\mathbf{s} \leftarrow \mathbb{Z}_q^{n*}$, $q\mathbb{Z}^m \subset \Lambda \subset \mathbb{Z}^m$. For any $\mathbf{s} \in \mathbb{Z}_q^{n*}$, let $\mathsf{X}_\mathbf{s}$ be the distribution of m LWQ samples $(\mathbf{A}, Q_\Lambda(\mathbf{As}))$, and $\mathsf{Y}_\mathbf{s}$ be the distribution of m quantized LWE samples $(\mathbf{A}, Q_\Lambda(\mathbf{As} + \mathbf{e}))$. For a small noise in the form of $\mathbf{e} \sim \mathcal{V}_{\frac{1}{p}\Lambda}$, $p > 2$, $p \in \mathbb{Z}$, we have*

$$e^{D_2(\mathsf{X}_\mathbf{s} || \mathsf{Y}_\mathbf{s})} \leq \frac{(p-1)^m}{p^m} + \frac{2^m}{p^m}. \tag{17}$$

Proof. Since $\mathbf{s} \in \mathbb{Z}^{n*}$, $\mathbf{As}$ admits a uniform distribution in $\mathbb{Z}_q^m$. Using the definition of Rényi divergence, we have

$$\begin{aligned}
e^{D_2(\mathsf{X}_\mathbf{s} || \mathsf{Y}_\mathbf{s})} &= \mathsf{E}_{\mathsf{X}_\mathbf{s}} \frac{\mathsf{Pr}(\mathsf{X}_\mathbf{s} = (\mathbf{A}, Q_\Lambda(\mathbf{As})))}{\mathsf{Pr}(\mathsf{Y}_\mathbf{s} = (\mathbf{A}, Q_\Lambda(\mathbf{As}))} && (18)\\
&= \mathsf{E}_{\mathbf{A} \leftarrow \mathbb{Z}_q^{m \times n}} \frac{1}{\mathsf{Pr}_\mathbf{e}(Q_\Lambda(\mathbf{As}) = Q_\Lambda(\mathbf{As} + \mathbf{e}) \mod q\mathbb{Z}^m)} && (19)\\
&= \mathsf{E}_{\mathbf{u} \leftarrow \mathbb{Z}_q^m} \frac{1}{\mathsf{Pr}_\mathbf{e}(Q_\Lambda(\mathbf{u}) = Q_\Lambda(\mathbf{u} + \mathbf{e}) \mod q\mathbb{Z}^m)} && (20)\\
&= \mathsf{E}_{\mathbf{u} \leftarrow \mathcal{V}_\Lambda} \frac{1}{\mathsf{Pr}_\mathbf{e}(Q_\Lambda(\mathbf{u} + \mathbf{e}) = \mathbf{0} \mod q\mathbb{Z}^m)} && (21)\\
&\leq \mathsf{E}_{\mathbf{u} \sim \mathcal{V}_\Lambda} \frac{1}{\mathsf{Pr}_\mathbf{e}(Q_\Lambda(\mathbf{u} + \mathbf{e}) = \mathbf{0})} && (22)
\end{aligned}$$

If $\mathbf{e} \sim \mathcal{V}_{\frac{1}{p}\Lambda}$, this Voronoi region can be partitioned into two parts, U_1 and U_2: the first part that corresponds to $\mathsf{Pr}_{\mathbf{e} \sim U_1}(Q_\Lambda(\mathbf{e}) = \mathbf{0}) = 1$ has probability

$\left(\frac{p-1}{p}\right)^m$ over $\mathbf{u}$, while the second part that corresponds to $\mathsf{Pr}_{\mathbf{e}\sim U_2}(Q_\Lambda(\mathbf{u}+\mathbf{e}) = \mathbf{0}) \geq \frac{1}{2^m}$ has probability $\frac{1}{p^m}$ over $\mathbf{u}$. Summarizing the above, we have

$$\mathsf{E}_{\mathbf{u}\sim\mathcal{V}_\Lambda} \frac{1}{\mathsf{Pr}_{\mathbf{e}}(Q_\Lambda(\mathbf{u}+\mathbf{e})=\mathbf{0})} \leq \frac{(p-1)^m}{p^m} + \frac{2^m}{p^m}. \tag{23}$$

□

Remark 1. Our proof works on m dimensional inputs directly, rather than bounding one dimensional distributions and getting their products. E.g., with $\Lambda = \frac{q}{p}\mathbb{Z}^m$, and $\mathbf{e}$ admitting i.i.d. symmetric bounded noise in $\{-B, ..., B\}$, with $1 \leq B \leq \frac{q}{2p}$, [11, Lemma 1] yields a divergent bound:

$$e^{D_2(\mathsf{X_s}||\mathsf{Y_s})} \leq \left(1 + \frac{2Bp}{q}\right)^m. \tag{24}$$

This is in sharp contrast to our convergent bound in Lemma 2.

Lemma 3 ([11]). *For any two distributions* X *and* Y*, for any event* E*,*

$$\mathsf{Pr}(\mathsf{Y}\in E) \geq \mathsf{Pr}(\mathsf{X}\in E)^2 / e^{D_2(\mathsf{X}||\mathsf{Y})}. \tag{25}$$

Combining the above lemmas, we arrive at the following theorem. The proof is straightforward and omitted. It says that, for small noises of LWE in the form of the Voronoi region of the quantizer, LWQ is as hard as LWE.

Theorem 5. *Let* $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$, $\mathbf{s} \leftarrow \mathbb{Z}_q^{n*}$, $\mathbb{Z}_q^m \subset \Lambda \subset \mathbb{Z}^m$. *For every adversary* $\mathcal{A}$*, if* $\mathbf{e} \sim \mathcal{V}_{\frac{1}{p}\Lambda}$, $p \geq 2$, $p \in \mathbb{Z}$*, we have*

$$\begin{aligned} &\mathsf{Pr}_{\mathbf{A},\mathbf{s},\mathbf{e}}\left(\mathcal{A}(\mathbf{A}, Q_\Lambda(\mathbf{A}\mathbf{s}+\mathbf{e})) = \mathbf{s}\right) \\ &\geq \frac{p^m}{(p-1)^m + 2^m} \mathsf{Pr}_{\mathbf{A},\mathbf{s},\mathbf{e}}\left(\mathcal{A}(\mathbf{A}, Q_\Lambda(\mathbf{A}\mathbf{s})) = \mathbf{s}\right)^2. \end{aligned} \tag{26}$$

4 Polar Lattice for Quantization

In this section, we will adopt polar lattices to compress the discrete sources. The technical novelty is to prove the quantization noise 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.

Polar lattices are an instance of the well-known "Construction D" [17, 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 [32]. A follow-up work [31] 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 [32, 31] 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. To help readers understand polar quantizers, an overview of polar codes is provided in Appendix A.

4.1 Duality Between Quantization and Error Correction

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 a continuous signal to the nearest lattice point, effectively reducing data resolution and removing redundancy. Consider error correction using Λ, generated by a basis matrix $\mathbf{B}$:

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

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*: $\mathbf{m} = \mathbf{B}^{-1}\mathbf{q}$. In this work, the uniquess of indexing is guaranteed by Theorem 1.

The test channel is a hypothetical communication channel used to model the quantization process, analogous to how error correction is typically framed.

Definition 11 (Test Channel). *The test channel is formulated as* $\mathbf{x} = \mathbf{q} + \mathbf{e}_Q$*, where* $\mathbf{e}_Q = \mathbf{x} - Q_\Lambda(\mathbf{x})$*. Here,* $\mathbf{q} \in \Lambda$ *is referred to as the codeword, and* $\mathbf{e}_Q$ *is referred to as the noise.*

4.2 Polar Quantizer: Construction

In this subsection, we present an explicit construction of polar lattices for the quantization of random integers, which produces Gaussian-like quantization errors. In a nutshell, the quantizer consists of a series of decoders for binary polar codes according to the multilevel structure of "Construction D". We begin by some preliminaries on the lattice structure based on multi-level codes [20].

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. 1, whose task is to mimic a reversed version of the test channel between X and Y in Fig. 2. 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. 2, 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 as in Definition 11. However, Theorem 6 illustrates that the difference between these two test channels can be negligible, which confirms the motivation of introducing lattice quantization in our LWQ scheme. Moreover, the relationship between the lattice quantization from $Y^{[m]}$ to $X^{[m]}$ and the lattice construction based on the test channel from $X^{[m]}$ to $Y^{[m]}$ will be explained in Remark 4.

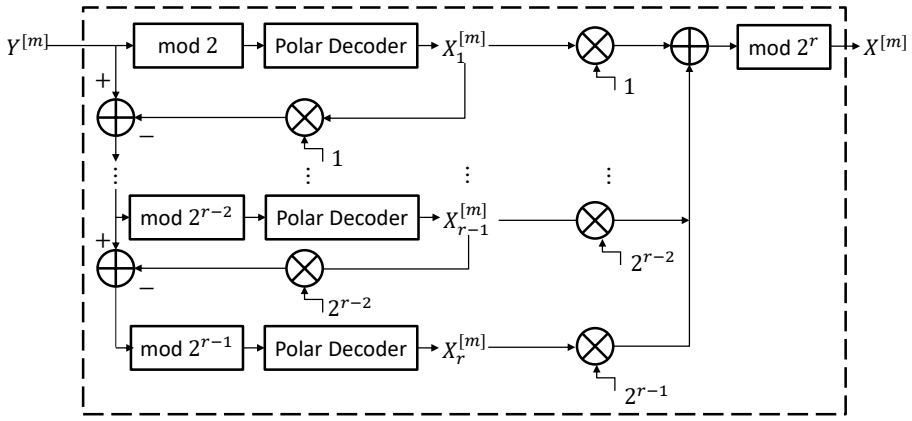

Fig. 1. The internal structure of a polar lattice quantizer.

Definition 12 (Partition Chain). *A sublattice $\Lambda' \subset \Lambda$ induces a partition (denoted by Λ/Λ') of Λ into equivalence groups modulo Λ'. The order of the partition is denoted by $|\Lambda/\Lambda'|$, which is equal to the number of cosets. If $|\Lambda/\Lambda'| = 2$, this is called a binary partition. A lattice partition chain, which is denoted by $\Lambda(\Lambda_0)/\Lambda_1/\cdots/\Lambda_{r-1}/\Lambda'(\Lambda_r)$ for $r \geq 1$, is 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 $\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_m$ as a basis of $\mathbb{F}_2^m$ such that $\mathbf{g}_1, \cdots, \mathbf{g}_{k_\ell}$ span C_ℓ. With $n = 1$, the binary lattice L of Construction D*

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 $E \sim D_{\mathbb{Z},\sigma}$
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 $X^{[m]} = X_1^{[m]} + 2X_2^{[m]} + \cdots + 2^{r-1}X_r^{[m]} \mod 2^r\mathbb{Z}$

consists of all vectors of the form

$$\sum_{\ell=1}^{r} 2^{\ell-1} \sum_{j=1}^{k_\ell} u_\ell^j \psi(\mathbf{g}_j) + 2^r z, \tag{27}$$

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

Pseudo-codes of the polar lattice quantization algorithm are given in Algorithm 1. 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 very high dimensional lattices, e.g. $m = 2^{20}$, 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. 2 shows a comparison between the distribution of quantization noise $Y - X$ achieved by the polar lattice quantizer and the genuine discrete Gaussian distribution $D_{\mathbb{Z},\sigma}$ with parameters $\sigma = 3$, $r = 8$ and $m = 2^{20}$.

4.3 Polar Quantizer: Performance Analysis

Because the quantization noise is represented by $q = 2^r$ integers in $[-2^{r-1}, 2^{r-1})$ but not exactly in $\mathbb{Z}$, the following lemma shows that a discrete Gaussian distribution after the modulo $q\mathbb{Z}$ operation behaves similarly to the standard one.

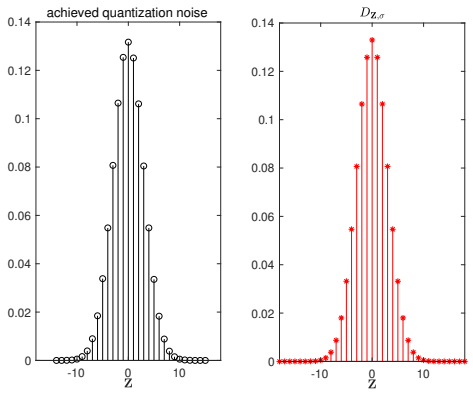

Fig. 2. A comparison between the distribution of quantization noise $Y-X$ and $D_{\mathbb{Z},\sigma_s=3}$.

Lemma 4. *Let $E \sim D_{\mathbb{Z},\sigma}$ 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 statistical distance $\Delta(\mathsf{P}_E, \mathsf{P}_{E'})$ between P_E and $\mathsf{P}_{E'}$ is upper-bounded as follows:*

$$\Delta(\mathsf{P}_E, \mathsf{P}_{E'}) \triangleq \frac{1}{2}\sum_{e\in\mathbb{Z}} |\mathsf{P}_E(e) - \mathsf{P}_{E'}(e)| \le M \cdot \exp\left(-\frac{(2^{r-1}-1)^2}{2\sigma^2}\right), \tag{28}$$

where $M = 2/\big(\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{\lambda\in\mathbb{Z}} \exp\big(-\frac{\lambda^2}{2\sigma^2}\big)\big)$.

Proof.

$$\Delta(\mathsf{P}_E, \mathsf{P}_{E'}) = \min \sum_{L\subset\mathbb{Z}/q\mathbb{Z}} \mathsf{Pr}(E\in L) - \mathsf{Pr}(E = \text{coset leader of } L) \tag{29}$$

$$= \sum_{\lambda=\{-2^{r-1},\ldots,2^{r-1}-1\}} \mathsf{P}_E(\lambda + q\mathbb{Z}) - \mathsf{P}_E(\lambda) \tag{30}$$

$$= \sum_{\lambda=-2^{r-1}-1}^{-\infty} \frac{\exp(-\frac{\lambda^2}{2\sigma^2})}{\sum_{\lambda'\in\mathbb{Z}}(\exp(-\frac{\lambda'^2}{2\sigma^2}))} + \sum_{\lambda=2^{r-1}}^{\infty} \frac{\exp(-\frac{\lambda^2}{2\sigma^2})}{\sum_{\lambda'\in\mathbb{Z}}(\exp(-\frac{\lambda'^2}{2\sigma^2}))} \tag{31}$$

$$\le 2\sum_{\lambda=2^{r-1}}^{\infty} \frac{\exp(-\frac{\lambda^2}{2\sigma^2})}{\sum_{\lambda'\in\mathbb{Z}}(\exp(-\frac{\lambda'^2}{2\sigma^2}))} \tag{32}$$

$$\le 2\frac{\int_{2^{r-1}-1}^{\infty} \exp(-\frac{t^2}{2\sigma^2})dt}{\sum_{\lambda'\in\mathbb{Z}}(\exp(-\frac{\lambda'^2}{2\sigma^2}))} \tag{33}$$

$$= M \cdot Q\left(\frac{2^{r-1}-1}{\sigma}\right) \tag{34}$$

$$\le M \cdot \exp\left(-\frac{(2^{r-1}-1)^2}{2\sigma^2}\right), \tag{35}$$

where $Q(x) = 1 - \Phi(x)$ is the Q-function of a standard normal distribution, and we use $Q(x) \leq \exp(-\frac{x^2}{2})$ in the last inequality. □

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 $E^i \sim D_{\mathbb{Z},\sigma}$. 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 $1 \leq \ell \leq r$. 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. 2, 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. [20], the test channel can be partitioned into r $2^{\ell-1}\mathbb{Z}/2^\ell\mathbb{Z}$ binary-input channels with $1 \leq \ell \leq r$, 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 5. *Let* $\mathsf{Q}_{U_1^{[m]},Y^{[m]}}$ *denote the resulted joint distribution of* $U_1^{[m]}$ *and* $Y^{[m]}$ *according to the encoding rules* (36) *and* (37) *at the first partition level.*

$$U_1^i = \begin{cases} 0 \text{ w. p. } P_{U_1^i|U_1^{1:i-1},Y^{[m]}}\left(0|u_1^{1:i-1},y^{[m]}\right) \\ 1 \text{ w. p. } P_{U_1^i|U_1^{1:i-1},Y^{[m]}}\left(1|u_1^{1:i-1},y^{[m]}\right) \end{cases} \text{if } i \in \mathcal{I}_1 \tag{36}$$

$$U_1^i = \begin{cases} 0 \text{ w. p. } \dfrac{1}{2} \\ 1 \text{ w. p. } \dfrac{1}{2}. \end{cases} \text{if } i \in \mathcal{F}_1 \tag{37}$$

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* (36) *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:*

$$\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}},\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right) \le m\sqrt{\ln 2 \cdot 2^{-m^\beta}}. \tag{38}$$

Proof. See Appendix B.

Remark 2. It seems that the encoding rules (36) and (37) are not deterministic. We note that the randomized forms in (36) and (37) are just for convenience of proof. By the symmetry of the $\mathbb{Z}/2\mathbb{Z}$ channel, it can be shown that any fixed realization $U_1^{\mathcal{F}_1} = u_1^{\mathcal{F}_1}$ causes the same statistical distance [25], meaning that one can safely choose all-zero frozen bits in practice. Similarly, by the polarization effect, the bit U_1^i for $i \in \mathcal{I}_1$ has conditional entropy $H(U_1^i|U_1^{1:i-1},Y^{[m]}) \to 0$ almost surely as $m \to \infty$. The rule (36) can be replaced with a deterministic MAP rule.

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 $\mathcal{I}_2 \triangleq \{i \in [m] : Z(U_2^i|U_2^{1:i-1}, X_1^{[m]}, Y^{[m]}) \le 1 - 2^{-m^\beta}\}$, and the frozen set is defined as $\mathcal{F}_2 \triangleq \mathcal{I}_2^c$.

Lemma 6. *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* (36) *and* (37) *at the first partition level, and then rules* (39) *and* (40) *at the second partition level.*

$$U_1^i = \begin{cases} 0 \text{ w. p. } P_{U_2^i|U_2^{1:i-1},X_1^{[m]},Y^{[m]}}\left(0|u_2^{1:i-1},x_1^{[m]},y^{[m]}\right) \\ 1 \text{ w. p. } P_{U_2^i|U_2^{1:i-1},X_1^{[m]},Y^{[m]}}\left(1|u_2^{1:i-1},x_1^{[m]},y^{[m]}\right) \end{cases} \text{if } i \in \mathcal{I}_2 \tag{39}$$

$$U_2^i = \begin{cases} 0 \ w.\ p.\ \dfrac{1}{2} \\ 1 \ w.\ p.\ \dfrac{1}{2}. \end{cases} \quad if\ i \in \mathcal{F}_2 \tag{40}$$

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* (36) *and rule* (39) *for all* $i \in [m]$*, respectively. The statistical distance between* $\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ *and* $\mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ *is upper-bounded as follows:*

$$\Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) \leq 2m\sqrt{\ln 2 \cdot 2^{-m^\beta}}. \tag{41}$$

Proof. Assume an auxiliary joint distribution $\mathsf{Q}'_{U_1^{[m]},U_2^{[m]},Y^{[m]}}$ resulted from using the encoding rule (36) for all U_1^i with $i \in [m]$ at the first partition level, and rules (39) and (40) 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,

$$\begin{aligned} &\Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) \\ &\leq \Delta\left(\mathsf{P}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}, \mathsf{Q}'_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right) + \Delta\left(\mathsf{Q}'_{U_1^{[m]},U_2^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},U_2^{[m]},Y^{[m]}}\right), \end{aligned} \tag{42}$$

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 5, 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 $\Delta\left(\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}}, \mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}}\right) \leq rm\sqrt{\ln 2 \cdot 2^{-m^\beta}}$. Combining this result with Lemma 4, 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. 3. 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.

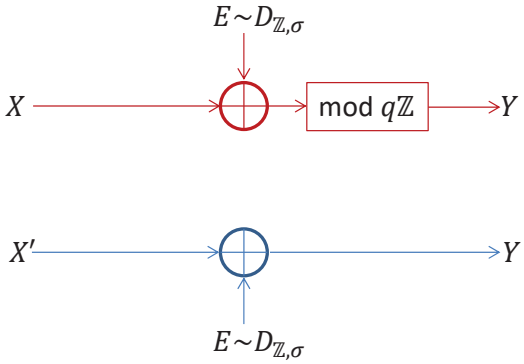

Fig. 3. A comparison between the two test channels based on $\mathsf{P}_{X,Y}$ and $\mathsf{P}_{X',Y}$, which are marked in red and blue, respectively.

Theorem 6. *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*

$$\Delta\left(\mathsf{P}_{X^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right) \le r \cdot m\sqrt{\ln 2 \cdot 2^{-m^\beta}}. \tag{43}$$

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

$$\begin{aligned}&\Delta\left(\mathsf{P}_{X'^{[m]},Y^{[m]}}, \mathsf{Q}_{X^{[m]},Y^{[m]}}\right)\\ &\quad\le r \cdot m\sqrt{\ln 2 \cdot 2^{-m^\beta}} + M \cdot m \cdot \exp\left(-\frac{(2^{r-1}-1)^2}{2\sigma^2}\right).\end{aligned} \tag{44}$$

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 $E \sim D_{\mathbb{Z},\sigma}$. 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 4, where $E' = E \mod q\mathbb{Z}$. By using the telescoping expansion (58) and the triangle inequality again, the proof is completed. □

Remark 3. Theorem 6 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 (44). 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 (44). For completeness, since the two parts are both measured in the statistical distance, we also provide the counterparts of Lemma 4 and Lemma 5 with the measurement of the Kullback-Leibler divergence in Appendix C.

Remark 4. 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 [20].

Remark 5. We note that the validity of polar lattice structure can be easily guaranteed. Taking the above simulation as an example, when constructing multilevel polar codes along the binary partition chain $\mathbb{Z}/2\mathbb{Z}/\cdots/2^r\mathbb{Z}$ for the additive discrete Gaussian test channel ($\sigma = 3$), the capacities of the partition channels from $\ell = 1$ to r are given by 0, 3.2732×10^{-10}, 0.0056, 0.3933, 0.9690, 1.0000 and 1.0000, respectively. The size of the information set is chosen as $|\mathcal{I}_\ell| = \lceil m \cdot C(W_\ell)\rceil$, where $C(W_\ell)$ denotes the capacity of the ℓ-th partition channel. As a result, the component polar codes are consecutively nested by ensuring $\mathcal{I}_\ell \subseteq \mathcal{I}_{\ell+1}$ for $1 \le \ell \le r-1$, and we have an ascertained polar lattice quantizer. Moreover, the constructed polar lattice is roughly sphere-bound achieving, by the capacity-achieving property of polar codes for all partition levels.

5 Application 1: PKE

Following the PKE structure outlined by Lindner and Peikert [30], we can develop PKE schemes via LWQ. This section presents a PKE scheme named Lily, which is based on both the plain (non-algebraic) Learning with Errors (LWE) and LWQ problems. Compared to existing PKE schemes, such as Frodo (based on LWE) and Lizard (based on LWE+LWQ), Lily offers the advantages of reduced ciphertext size and enhanced security.

5.1 Lily: PKE via LWQ

The public key consists of a number m of n-dimensional LWE samples, and encryptions of zero form $(n+\ell)$ samples of m-dimensional LWQ where ℓ is the dimension of plaintext vectors. The scheme is described as follows:

- **Lily.Setup**(1^λ): Choose positive integers n, q, ℓ and p_0, p_1, p_2. Choose private key distribution $\mathcal{D}_s$ over $\mathbb{Z}^n$, ephemeral secret distribution D_r over $\mathbb{Z}^n$, and parameter σ for discrete Gaussian distribution DG_σ. Select a lattice encoder $\mathsf{ec}_{\Lambda_0} : \{0,1\}^{\ell^2 B} \to \Lambda_0 \cap \mathbb{Z}_q^m$, s.t. $q\mathbb{Z}^{\ell^2} \subset \Lambda_0 \subset \mathbb{Z}^{\ell^2}$, and two lattice quantizers Q_{Λ_1} and Q_{Λ_2}, s.t. $q\mathbb{Z}^{\ell n} \subset \Lambda_1 \subset \mathbb{Z}^{\ell n}$, $q\mathbb{Z}^{\ell^2} \subset \Lambda_2 \subset \mathbb{Z}^{\ell^2}$. Output
$$params \leftarrow (n, \ell, q, \mathcal{D}_s, \mathcal{D}_r, \sigma, p_0, p_1, p_2, \Lambda_0, \Lambda_1, \Lambda_2). \tag{45}$$
- **Lily.KeyGen**(params): Generate a random matrix $\mathbf{A} \leftarrow \mathbb{Z}_q^{n\times n}$. Choose a secret matrix $\mathbf{S} = (\mathbf{s}_1 \mid \cdots \mid \mathbf{s}_\ell)$ by sampling column vectors $\mathbf{s}_i \in \mathbb{Z}^n$ independently from the distribution $\mathcal{D}_s$. Generate an error matrix $\mathbf{E} = (\mathbf{e}_1 \mid \cdots \mid \mathbf{e}_\ell)$ from $DG_\sigma^{n\times\ell}$ and let $\mathbf{B} = \mathbf{AS} + \mathbf{E} \in \mathbb{Z}_q^{n\times\ell}$ where the operations are held modulo q. Output the public key $\mathrm{pk} = (\mathbf{A} \mid \mathbf{B}) \in \mathbb{Z}_q^{n\times(n+\ell)}$ and the secret key $\mathrm{sk} = \mathbf{S} \in \mathbb{Z}_q^{n\times\ell}$.
- **Lily.Enc**$_{\mathrm{pk}}(m)$: For a plaintext $\mu \in \mathcal{M} = \{0,1\}^{\ell^2 B} \cong \mathbb{Z}_{p_0}^{\ell^2}$, choose an ephemeral key matrix $\mathbf{R} = (\mathbf{r}_1 \mid \cdots \mid \mathbf{r}_\ell)$ from $DG_r^{n\times\ell}$. Output
$$\mathbf{c} = (Q_{\Lambda_1}(\mathbf{R}^T\mathbf{A}), Q_{\Lambda_2}(\mathbf{R}^T\mathbf{B} + \mathsf{ec}_{\Lambda_0}(\mu))) \cong (\mathbb{Z}_{p_1}^{\ell\times n}, \mathbb{Z}_{p_2}^{\ell\times\ell}). \tag{46}$$
- **Lily.Dec**$_{\mathrm{sk}}(c)$: For a ciphertext $\mathbf{c} = (\mathbf{C}_1, \mathbf{C}_2)$, compute and output the estimated message
$$\hat{\mu} = \mathsf{ec}_{\Lambda_0}^{-1}(\mathbf{C}_2 - \mathbf{C}_1\mathbf{S}), \tag{47}$$
where $\mathsf{ec}_{\Lambda_0}^{-1}$ standards for the lattice decoding function.

5.2 Correctness

To ensure the correctness of our PKE scheme, the following lemma establishes a necessary condition for the parameter setup.

Lemma 7 (Correctness). *The PKE scheme Lily is correct if the following inequality holds for the security parameter λ:*
$$\Pr\left[Q_{\Lambda_0}(\mathrm{vec}(\mathbf{R}^T\mathbf{E} + \mathbf{F}_2 - \mathbf{F}_1^T\mathbf{S})) \neq q\mathbb{Z}^{\ell^2}\right] < \mathrm{negl}(\lambda),$$
where the randomness is taken over $\mathbf{E}, \mathbf{R}, \mathbf{S}$.

Proof. In decryption, we have
$$\mathbf{C}_2 - \mathbf{C}_1\mathbf{S} = \mathbf{R}^T\mathbf{E} + \mathbf{F}_2 - \mathbf{F}_1^T\mathbf{S} + \mathsf{ec}_{\Lambda_0}(\mu). \tag{48}$$
In the modulo lattice additive noise model of [34], we have $\mathbf{y} = \mathbf{x} + \mathbf{n} \mod q\mathbb{Z}^{\ell^2}$, where $\mathbf{n} = \mathrm{vec}(\mathbf{R}^T\mathbf{E} + \mathbf{F}_2 - \mathbf{F}_1^T\mathbf{S}) \in \mathbb{Z}^{\ell^2}$. The find lattice for error correction is Λ_0, and the coarse lattice for shaping is $q\mathbb{Z}^{\ell^2}$. Thus the decryption failure probability can be analyzed through $\Pr(Q_{\Lambda_0}(\mathbf{n}) \neq q\mathbb{Z}^{\ell^2})$. □

Explicit formulas for the decryption error rate can be evaluated through sub-Gaussian.

Definition 14. *A random variable X is sub-Gaussian with parameter $\sigma > 0$ if for all $t \in \mathbb{R}$, the tails of X are dominated by a Gaussian of parameter σ, i.e., the moment-generating function satisfies $\mathbb{E}[e^{tX}] \leq e^{\sigma^2 t^2/2}$. Thus, $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/(2\sigma^2)}$ for all $t \geq 0$. More generally, we say that a random vector $\mathbf{x}$ is sub-Gaussian (of parameter σ) if all its one-dimensional marginals $\mathbf{u}^\top \mathbf{x}$ for a unit vector $\mathbf{u}$ are sub-Gaussian (of parameter σ).*

It can be verified that $\mathbf{f}_1 = \mathrm{vec}(\mathbf{R}^T\mathbf{A} \mod \Lambda_1)$, $\mathbf{f}_2 = \mathrm{vec}(\mathbf{R}^T\mathbf{B} \mod \Lambda_2)$ are sub-Gaussians of parameter

$$\sigma_{f_1}^2 = G(\Lambda_1)\det(\Lambda_1)^{2/\ell n}, \tag{49}$$

$$\sigma_{f_2}^2 = G(\Lambda_2)\det(\Lambda_2)^{2/\ell^2}. \tag{50}$$

Thus the effective noise $\mathbf{n} = \mathrm{vec}(\mathbf{R}^T\mathbf{E} + \mathbf{F}_2 - \mathbf{F}_1^T\mathbf{S})$ is of parameter

$$\bar{\sigma}^2 = n(\sigma_r^2\sigma^2 + \sigma_{f_1}^2\sigma_s^2) + \sigma_{f_2}^2. \tag{51}$$

The decryption failure probability is then upper bounded by $e^{-\ell^2\lambda_1^2/(8\bar{\sigma}^2)}$, where λ_1 is the length of the shortest vector of Λ_0.

5.3 Cryptographic Properties

A PKE scheme is IND-CPA (Indistinguishability under Chosen Plaintext Attack) secure if for all probabilistic polynomial-time adversaries $\mathcal{A}$, the probability that $\mathcal{A}$ correctly guesses b is at most negligibly better than random guessing. Mathematically, this can be expressed as:

$$\left|\Pr[\mathcal{A}(\mathrm{pk}, c^*) = b] - \frac{1}{2}\right| \leq \mathrm{negl}(\lambda)$$

where the probability is taken over the random choices of the key generation algorithm, the encryption algorithm, and the adversary's internal randomness, and $\mathrm{negl}(\lambda)$ denotes a negligible function in the security parameter λ.

We demonstrate that the proposed encryption scheme is IND-CPA secure under the hardness assumptions of the LWE problem and the LWQ problem. The following theorem provides an explicit proof of our security argument.

Theorem 7. *The PKE scheme Lily is IND-CPA secure under the hardness assumption of ℓ-secret $LWE_{n,n,q,D_{G_\sigma}}(\mathcal{D}_s)$, $LWQ_{n,n,q,p_1}(\mathcal{D}_r)$ and $LWQ_{n,\ell,q,p_2}(\mathcal{D}_r)$.*

Proof. It suffices to show that the pair of public information $\mathrm{pk} = (\mathbf{A} \mid \mathbf{B}) \leftarrow \mathrm{Lily.KeyGen}(\textit{params})$ and encryption of zero $c \leftarrow \mathrm{Lily.Enc}_{\mathrm{pk}}(0)$ is computationally indistinguishable from the uniform distribution over $\mathbb{Z}_q^{m\times(n+\ell)} \times (\mathbb{Z}_{p_1}^{\ell\times n} \times \mathbb{Z}_{p_2}^{\ell\times\ell})$ for a parameter set $\textit{params} \leftarrow \mathrm{Lily.Setup}(1^\lambda)$. We construct three distributions as follows:

- $\mathcal{D}_0 = \{(\mathrm{pk}, c) : \mathrm{pk} \leftarrow \mathrm{Lily.KeyGen}(\mathrm{params}), c \leftarrow \mathrm{Lily.Enc}_{\mathrm{pk}}(0)\}$.

– $\mathcal{D}_1 = \{(\mathrm{pk}, c) : \mathrm{pk} \leftarrow \mathbb{Z}_q^{n\times(n+\ell)}, c \leftarrow \mathrm{Lily.Enc}_{\mathrm{pk}}(0)\}$.
– $\mathcal{D}_2 = \{(\mathrm{pk}, c) : \mathrm{pk} \leftarrow \mathbb{Z}_q^{n\times(n+\ell)}, c \leftarrow (\mathbb{Z}_{p_1}^{\ell\times n}, \mathbb{Z}_{p_2}^{\ell\times\ell})\}$.

The public key $\mathrm{pk} = (\mathbf{A} \mid \mathbf{B}) \leftarrow \mathrm{Lily.KeyGen}(\mathit{params})$ is generated by sampling n instances of the LWE problem with ℓ independent secret vectors $s_1, \ldots, s_\ell \leftarrow \mathcal{D}_s$. In addition, the multi-secret LWE problem is no easier than the ordinary LWE problem. Therefore, distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ are computationally indistinguishable under the ℓ-secret $\mathrm{LWE}_{n,n,q,D_{G_\sigma}}(\mathcal{D}_s)$ assumption.

Now assume that pk is uniformly random over $\mathbb{Z}_q^{n\times(n+\ell)}$. Then pk and $c \leftarrow \mathrm{Lily.Enc}_{\mathrm{pk}}(0)$ together form two ℓ-secret LWQ problems over quantization lattices Λ_1 and Λ_2. Therefore, distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ are computationally indistinguishable under the assumptions of ℓ-secret $\mathrm{LWQ}_{n,n,q,p_1}(\mathcal{D}_r)$ and ℓ-secret $\mathrm{LWQ}_{n,\ell,q,p_2}(\mathcal{D}_r)$. As a result, distributions $\mathcal{D}_0$ and $\mathcal{D}_2$ are computationally indistinguishable under the hardness assumptions of ℓ-secret $\mathrm{LWE}_{n,n,q,D_{G_\sigma}}(\mathcal{D}_s)$, $\mathrm{LWQ}_{\ell,n,q,p_1}(\mathcal{D}_r)$ and $\mathrm{LWQ}_{\ell,\ell,q,p_2}(\mathcal{D}_r)$, which demonstrates the IND-CPA security of the PKE scheme. □

5.4 Reduced size of ciphertext

The main difference of Lily against Frodo is the replacement of Gaussian errors with quantization errors in the phase of encryption. Targeting Frodo-640, 976, 1344, without compromising security level or decryption error rate, as well as excluding the gains of better error correction due to Λ_0, we show the reduce size of Lily in the Table 2. The Lily scheme demonstrates significant savings in ciphertext size compared to the Frodo schemes across various parameter sets. Specifically, Lily-640 achieves a reduction of approximately 26.66%, Lily-976 saves about 18.91%, and Lily-1344 reduces the ciphertext size by around 18.73%.

Scheme	n	ℓ	σ	q	$\lvert c\rvert$ (bytes)	quantizer of $\mathbf{C}_1$	quantizer of $\mathbf{C}_2$
Frodo-640	640	8	2.75	2^{15}	9720	-	-
Frodo-976	976	8	2.3	2^{16}	15744	-	-
Frodo-1344	1344	8	1.4	2^{16}	21632	-	-
Lily-640	640	8	2.75	2^{15}	7128	$2^4\mathbb{Z}^{640} \otimes E_8$	$2^4\mathbb{Z}^8 \otimes E_8$
Lily-976	976	8	2.3	2^{16}	12792	$2^3\mathbb{Z}^{976} \otimes E_8$	$2^3\mathbb{Z}^8 \otimes E_8$
Lily-1344	1344	8	1.4	2^{16}	17576	$2^3\mathbb{Z}^{1344} \otimes E_8$	$2^4\mathbb{Z}^8 \otimes E_8$

Table 2. Summary of Frodo and Lily parameters

6 Application 2: Privacy-Preserving Secret Key Encryption

Differential privacy has been quite important in many fields, such as securing CKKS [29]. In this section, we design a secret-key encryption scheme that simultaneously supports differential privacy.

The quancryption scheme is parameterized by the source dimension m, secret dimension n, modulus q, and a quantization lattice Λ, where $q\mathbb{Z}^m \subset \Lambda \subset \mathbb{Z}^m$. It consists of three components: key generation (KGen), encryption (Enc), and decryption (Dec). The space of plaintext is extrmely large: $\mathbf{m} \in \mathbb{Z}_q^m$.

In the encryption process, let $\mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}$ and $\mathbf{s} \leftarrow \mathcal{D}_s$. The ciphertext is generated as follows:

$$\mathsf{Enc}_{\mathbf{s}}(\mathbf{m}) = (\mathbf{A}, \mathbf{b} = Q_\Lambda(\mathbf{As} + \mathbf{m})) \in \mathbb{Z}_q^{m\times n} \times (\mathbb{Z}_q^m \cap \Lambda). \tag{52}$$

Denote the ciphertext as $\mathbf{c}$, the decryption function is given by:

$$\hat{\mathbf{m}} = \mathsf{Dec}_{\mathbf{s}}(\mathbf{c}) = \mathbf{c} \cdot \left[-\mathbf{s}^T, 1\right]^T \in \mathbb{Z}_q^m. \tag{53}$$

Denote $-\mathbf{e}_q = \mathbf{As}+\mathbf{m}-Q_\Lambda(\mathbf{As}+\mathbf{m})$. Then we have $\hat{\mathbf{m}} = \mathbf{m}+\mathbf{e}_q$. The decrypted message has been enabled with differential privacy thanks to the perturbation of $\mathbf{e}_q$.

Some features of quancryption are:

- **Source to ciphertext ratio**: This parameter is defined as

$$r_{SC} = \frac{\log q^m}{\log q^m - \log\det\Lambda}. \tag{54}$$

  Obviously, $r_{SC} > 1$ for quancryption. It surpasses those in existing encryption schemes [36, Table 1] that excludes privacy. The interesting feature of quancryption is that, higher security level of LWQ also leads to higher secrecy level.
- **Correctness**: Quancryption is an *approximate* encryption scheme, where the decrypted message is not identical to the plaintext, but is rather a function of the plaintext. An adversary is said to obtain the privacy level of the legitimate receiver if a high-quality $\hat{\mathbf{m}}$ can be guessed:

$$Q_{\Lambda+\mathbf{m}}(\hat{\mathbf{m}}) = \mathbf{0}, \tag{55}$$

  which implies that $\mathbf{m} - \hat{\mathbf{m}} \in \mathcal{V}_\Lambda$.

6.1 Cryptographic Properties

We use the security notion of RND-CPA, which is better suited to lattice-based primitives, as RND-CPA security implies IND-CPA security [36].

Definition 15 (RND-CPA). *An encryption scheme* $(\mathsf{KGen}, \mathsf{Enc}, \mathsf{Dec})$ *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. $\mathsf{sk} \leftarrow \mathsf{KGen}(1^n)$,

2. $b' \leftarrow \mathcal{A}^{O_b(\cdot)}$ *where* $O_b(m)$ *returns either an encryption* $\mathsf{Enc}_{\mathsf{sk}}(m)$ *of the message* m *under the key* sk *if* $b = 0$*, or a sample from a distribution that has support* $\{\mathsf{Enc}_{\mathsf{sk}}(m) \mid \mathsf{sk} \in \mathrm{supp}(\mathsf{KGen}(1^n)), m \in \mathcal{M}\}$ *if* $b = 1$.

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

Theorem 8. *The quancryption scheme* $\mathsf{LWQ}_{\Lambda}^{m,n,q}$ *is RND-CPA secure if the LWQ problem is hard.*

Proof. In the RND-CPA game of quancryption, the support of is $\mathbb{Z}_q^{m\times n} \times (\mathbb{Z}_q^m \cap \Lambda)$. The hardness of LWQ implies that $\mathcal{A}^{O_b(\mathbf{m}=\mathbf{0})}$ is negligible. For the set of vectors $\mathbf{v}_0, ..., \mathbf{v}_i \in \mathcal{V}_\Lambda \cap \mathbb{Z}^m$, there is a bijection to

$$\mathbf{v}_0 + \mathbf{m} \mod \Lambda, ..., \mathbf{v}_i + \mathbf{m} \mod \Lambda \in \mathcal{V}_\Lambda \cap \mathbb{Z}^m \tag{56}$$

for $\mathbf{m} \in \mathbb{Z}_q^m$. Then computationally $Q_\Lambda(\mathbf{As})$ and $Q_\Lambda(\mathbf{As} + \mathbf{m})$ are indistinguishable. Thus $\mathcal{A}^{O_b(\mathbf{m})} = \mathcal{A}^{O_b(\mathbf{0})}$ for $\mathbf{m} \in \mathbb{Z}_q^m$, and the RND-CPA security of quancryption can be built upon decisional-LWQ. □

It is noteworthy that the proposed quancryption scheme is not IND-CPA^D secure. Specifically, recently Li and Micciancio [28] introduced a model for the passive security of incorrect encryption schemes (IND-CPA^D). In effect, it allows an adversary to decrypt honestly generated ciphertexts, so that a scheme that somehow leaks sensitive information during honest decryption is not seen as secure. If needed, a simple solution to enable IND-CPA^D security is to add discrete Gaussian noises to $\mathbf{m}$ before encryption.

6.2 Differential Privacy

Proposition 1 (Prop. 5, [13]). *Let* $\sigma \in \mathbb{R}_{\geq 0}$*, and let* $\mu, \nu \in \mathbb{Z}^m$*. Then:*

$$D_{KL}(D^m_{\mathbb{Z},\mu,\sigma} \| D^m_{\mathbb{Z},\nu,\sigma}) = (\nu - \mu)^2.$$

Regarding the term $\mathbf{m} + \mathbf{e}_q$ at the receiver, the privacy level is controlled by $\mathbf{e}_q$, which is a discre Gaussian moodulo q. Together to the data processing inequality, we have

$$D_{KL}(D^m_{\mathbb{Z},\mu,\sigma} \bmod q \| D^m_{\mathbb{Z},\nu,\sigma} \bmod q) \leq (\nu - \mu)^2.$$

The above proposition provides a bound on KL divergence, showing our quancryption scheme is ρ-differentially private.

7 Conclusions and Future Works

The paper has explored a novel hardness assumption termed LWQ, similar to the LWR assumption, but is parameterized by an arbitrary lattice Λ (where setting $\Lambda = \frac{q}{p}\mathbb{Z}^m$ recovers LWR). 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. Future works may explore using LWQ for private stream aggregation, pseudorandom functions, pseudorandom number generators, homomorphic encryption, etc.

Acknowledgment

The authors would like to sincerely thank Baoming Bai, Hao Chen, Zhiyong Zheng, Chuanming Zong, Mark Schultz, and Yang Yu for the enlightening discussions.

References

1. Agrell, E., Allen, B.: On the best lattice quantizers. IEEE Transactions on Information Theory **69**(12), 7650–7658 (2023). https://doi.org/10.1109/TIT.2023.3291313
2. Albrecht, M.R., Deo, A., Paterson, K.G.: Cold boot attacks on ring and module LWE keys under the NTT. IACR TCHES **2018**(3), 173–213 (2018). https://doi.org/10.13154/tches.v2018.i3.173-213, https://tches.iacr.org/index.php/TCHES/article/view/7273
3. Alkim, E., Ducas, L., Pöppelmann, T., Schwabe, P.: Post-quantum key exchange - A new hope. In: Holz, T., Savage, S. (eds.) USENIX Security 2016. pp. 327–343. USENIX Association, Austin, TX, USA (Aug 10–12, 2016)
4. Allen, B., Agrell, E.: The optimal lattice quantizer in nine dimensions. Annalen der Physik **533**(12), 2100259 (2021)
5. Alwen, J., Krenn, S., Pietrzak, K., Wichs, D.: Learning with rounding, revisited - new reduction, properties and applications. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013, Part I. LNCS, vol. 8042, pp. 57–74. Springer, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 18–22, 2013). https://doi.org/10.1007/978-3-642-40041-4_4
6. Arıkan, E.: Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Transactions on Information Theory **55**(7), 3051–3073 (July 2009). https://doi.org/10.1109/TIT.2009.2021379
7. Arıkan, E., Telatar, I.: On the rate of channel polarization. In: Proc. 2009 IEEE Int. Symp. Inform. Theory. pp. 1493–1495. Seoul, South Korea (June 2009)
8. Bai, S., Boudgoust, K., Das, D., Roux-Langlois, A., Wen, W., Zhang, Z.: Middle-product learning with rounding problem and its applications. In: Galbraith, S.D., Moriai, S. (eds.) ASIACRYPT 2019, Part I. LNCS, vol. 11921, pp. 55–81. Springer, Heidelberg, Germany, Kobe, Japan (Dec 8–12, 2019). https://doi.org/10.1007/978-3-030-34578-5_3
9. Banerjee, A., Peikert, C., Rosen, A.: Pseudorandom functions and lattices. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 719–737. Springer, Heidelberg, Germany, Cambridge, UK (Apr 15–19, 2012). https://doi.org/10.1007/978-3-642-29011-4_42

10. Barnes, E., Sloane, N.: The optimal lattice quantizer in three dimensions. SIAM Journal on Algebraic Discrete Methods **4**(1), 30–41 (1983)
11. Bogdanov, A., Guo, S., Masny, D., Richelson, S., Rosen, A.: On the hardness of learning with rounding over small modulus. In: Kushilevitz, E., Malkin, T. (eds.) TCC 2016-A, Part I. LNCS, vol. 9562, pp. 209–224. Springer, Heidelberg, Germany, Tel Aviv, Israel (Jan 10–13, 2016). https://doi.org/10.1007/978-3-662-49096-9_9
12. Boutros, J., Viterbo, E., Rastello, C., Belfiore, J.: Good lattice constellations for both Rayleigh fading and Gaussian channels. IEEE Transactions on Information Theory **42**(2), 502–518 (1996). https://doi.org/10.1109/18.485720, https://doi.org/10.1109/18.485720
13. Canonne, C.L., Kamath, G., Steinke, T.: The discrete gaussian for differential privacy. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Annual Conference on Neural Information Processing Systems, NeurIPS 2020, December 6-12, 2020 (2020)
14. Cheon, J.H., Kim, A., Kim, M., Song, Y.S.: Homomorphic encryption for arithmetic of approximate numbers. In: Takagi, T., Peyrin, T. (eds.) ASIACRYPT 2017, Part I. LNCS, vol. 10624, pp. 409–437. Springer, Heidelberg, Germany, Hong Kong, China (Dec 3–7, 2017). https://doi.org/10.1007/978-3-319-70694-8_15
15. Cheon, J.H., Kim, D., Lee, J., Song, Y.: Lizard: Cut off the tail! A practical post-quantum public-key encryption from LWE and LWR. In: Catalano, D., De Prisco, R. (eds.) SCN 18. LNCS, vol. 11035, pp. 160–177. Springer, Heidelberg, Germany, Amalfi, Italy (Sep 5–7, 2018). https://doi.org/10.1007/978-3-319-98113-0_9
16. Cohn, H., Kumar, A., Miller, S., Radchenko, D., Viazovska, M.: The sphere packing problem in dimension 24. Annals of Mathematics **185**(3), 1017–1033 (2017). https://doi.org/10.4007/annals.2017.185.3.8
17. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups. Springer, New York, 3 edn. (1999). https://doi.org/10.1007/978-1-4757-6568-7
18. D'Anvers, J.P., Karmakar, A., Roy, S.S., Vercauteren, F.: Saber: Module-LWR based key exchange, CPA-secure encryption and CCA-secure KEM. In: Joux, A., Nitaj, A., Rachidi, T. (eds.) AFRICACRYPT 18. LNCS, vol. 10831, pp. 282–305. Springer, Heidelberg, Germany, Marrakesh, Morocco (May 7–9, 2018). https://doi.org/10.1007/978-3-319-89339-6_16
19. Egilmez, Z.B.K., Xiang, L., Maunder, R.G., Hanzo, L.: The development, operation and performance of the 5g polar codes. IEEE Communications Surveys & Tutorials **22**(1), 96–122 (2019)
20. Forney, G., Trott, M., Chung, S.Y.: Sphere-bound-achieving coset codes and multilevel coset codes. IEEE Transactions on Information Theory **46**(3), 820–850 (May 2000). https://doi.org/10.1109/18.841165
21. Gray, R.M., Stockham, T.G.: Dithered quantizers. IEEE Transactions on Information Theory **39**(3), 805–812 (1993)
22. Grover, C., Mendelsohn, A., Ling, C., Vehkalahti, R.: Non-commutative ring learning with errors from cyclic algebras. Journal of Cryptology **35**(3), 22 (Jul 2022). https://doi.org/10.1007/s00145-022-09430-6
23. Kim, T., Lee, C.: Lattice reductions over euclidean rings with applications to cryptanalysis. In: IMA International Conference on Cryptography and Coding. pp. 371–391. Springer (2017)
24. Kirchner, P., Espitau, T., Fouque, P.A.: Fast reduction of algebraic lattices over cyclotomic fields. In: Micciancio, D., Ristenpart, T. (eds.) CRYPTO 2020, Part II. LNCS, vol. 12171, pp. 155–185. Springer, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 17–21, 2020). https://doi.org/10.1007/978-3-030-56880-1_6

25. Korada, S., Urbanke, R.: Polar codes are optimal for lossy source coding. IEEE Transactions on Information Theory **56**(4), 1751–1768 (April 2010). https://doi.org/10.1109/TIT.2010.2040961
26. Langlois, A., Stehlé, D.: Worst-case to average-case reductions for module lattices. Designs, Codes and Cryptography **75**(3), 565–599 (2015)
27. Lee, C., Pellet-Mary, A., Stehlé, D., Wallet, A.: An LLL algorithm for module lattices. In: Galbraith, S.D., Moriai, S. (eds.) ASIACRYPT 2019, Part II. LNCS, vol. 11922, pp. 59–90. Springer, Heidelberg, Germany, Kobe, Japan (Dec 8–12, 2019). https://doi.org/10.1007/978-3-030-34621-8_3
28. Li, B., Micciancio, D.: On the security of homomorphic encryption on approximate numbers. In: Canteaut, A., Standaert, F.X. (eds.) EUROCRYPT 2021, Part I. LNCS, vol. 12696, pp. 648–677. Springer, Heidelberg, Germany, Zagreb, Croatia (Oct 17–21, 2021). https://doi.org/10.1007/978-3-030-77870-5_23
29. Li, B., Micciancio, D., Schultz, M., Sorrell, J.: Securing approximate homomorphic encryption using differential privacy. In: Dodis, Y., Shrimpton, T. (eds.) CRYPTO 2022, Part I. LNCS, vol. 13507, pp. 560–589. Springer, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 15–18, 2022). https://doi.org/10.1007/978-3-031-15802-5_20
30. Lindner, R., Peikert, C.: Better key sizes (and attacks) for LWE-based encryption. In: Kiayias, A. (ed.) CT-RSA 2011. LNCS, vol. 6558, pp. 319–339. Springer, Heidelberg, Germany, San Francisco, CA, USA (Feb 14–18, 2011). https://doi.org/10.1007/978-3-642-19074-2_21
31. Liu, L., Shi, J., Ling, C.: Polar lattices for lossy compression. IEEE Transactions on Information Theory **67**(9), 6140–6163 (2021), https://doi.org/10.1109/TIT.2021.3097965
32. Liu, L., Yan, Y., Ling, C., Wu, X.: Construction of capacity-achieving lattice codes: Polar lattices. IEEE Trans. Commun. **67**(2), 915–928 (Feb 2019)
33. Liu, S., Sakzad, A.: Crystals-kyber with lattice quantizer. arXiv preprint arXiv:2401.15534 (2024)
34. Lyu, S., Liu, L., Ling, C., Lai, J., Chen, H.: Lattice codes for lattice-based PKE. Des. Codes Cryptogr. **92**(4), 917–939 (2024). https://doi.org/10.1007/S10623-023-01321-6
35. Lyubashevsky, V., Peikert, C., Regev, O.: On ideal lattices and learning with errors over rings. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 1–23. Springer, Heidelberg, Germany, French Riviera (May 30 – Jun 3, 2010). https://doi.org/10.1007/978-3-642-13190-5_1
36. Micciancio, D., Schultz, M.: Error correction and ciphertext quantization in lattice cryptography. In: Handschuh, H., Lysyanskaya, A. (eds.) CRYPTO 2023, Part V. LNCS, vol. 14085, pp. 648–681. Springer, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 20–24, 2023). https://doi.org/10.1007/978-3-031-38554-4_21
37. Mukherjee, T., Stephens-Davidowitz, N.: Lattice reduction for modules, or how to reduce ModuleSVP to ModuleSVP. In: Micciancio, D., Ristenpart, T. (eds.) CRYPTO 2020, Part II. LNCS, vol. 12171, pp. 213–242. Springer, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 17–21, 2020). https://doi.org/10.1007/978-3-030-56880-1_8
38. Newton, P., Richelson, S.: A lower bound for proving hardness of learning with rounding with polynomial modulus. In: Handschuh, H., Lysyanskaya, A. (eds.) CRYPTO 2023, Part V. LNCS, vol. 14085, pp. 805–835. Springer, Heidelberg, Germany, Santa Barbara, CA, USA (Aug 20–24, 2023). https://doi.org/10.1007/978-3-031-38554-4_26

39. Pedarsani, R., Hassani, S., Tal, I., Telatar, I.: On the construction of polar codes. In: Proc. 2011 IEEE Int. Symp. Inform. Theory. pp. 11–15. St. Petersburg, Russia (July 2011). https://doi.org/10.1109/ISIT.2011.6033724
40. Peikert, C.: Public-key cryptosystems from the worst-case shortest vector problem: extended abstract. In: Mitzenmacher, M. (ed.) 41st ACM STOC. pp. 333–342. ACM Press, Bethesda, MD, USA (May 31 – Jun 2, 2009). https://doi.org/10.1145/1536414.1536461
41. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. In: Gabow, H.N., Fagin, R. (eds.) 37th ACM STOC. pp. 84–93. ACM Press, Baltimore, MA, USA (May 22–24, 2005). https://doi.org/10.1145/1060590.1060603
42. Schwabe, P., Avanzi, R., Bos, J., Ducas, L., Kiltz, E., Lepoint, T., Lyubashevsky, V., Schanck, J.M., Seiler, G., Stehlé, D., Ding, J.: CRYSTALS-KYBER. Tech. rep., National Institute of Standards and Technology (2022), available at https://csrc.nist.gov/Projects/post-quantum-cryptography/selected-algorithms-2022
43. Tal, I., Vardy, A.: How to construct polar codes. IEEE Transactions on Information Theory **59**(10), 6562–6582 (Oct 2013). https://doi.org/10.1109/TIT.2013.2272694
44. Viazovska, M.S.: The sphere packing problem in dimension 8. Annals of Mathematics pp. 991–1015 (2017). https://doi.org/10.4007/annals.2017.185.3.7
45. Wang, H.P., Duursma, I.M.: Log-logarithmic time pruned polar coding. IEEE Transactions on Information Theory **67**(3), 1509–1521 (2021). https://doi.org/10.1109/TIT.2020.3041523
46. Zamir, R.: Lattice Coding for Signals and Networks. Cambridge University Press, Cambridge, UK (2014)
47. Zamir, R.: The rate loss in the wyner-ziv problem. IEEE Transactions on Information Theory **42**(6), 2073–2084 (1996). https://doi.org/10.1109/18.556597
48. Zamir, R., Feder, M.: On lattice quantization noise. IEEE Transactions on Information Theory **42**(4), 1152–1159 (1996). https://doi.org/10.1109/18.508838

A Polar Codes

Polar coding [6] 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.

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

Definition 16. *Given a BMSC W with transition probability $\mathsf{P}_{Y|X}$, the Bhattacharyya parameter $Z \in [0,1]$ is defined as*

$$Z(W) = Z(X|Y) \triangleq \sum_y \sqrt{\mathsf{P}_{Y|X}(y|0)\mathsf{P}_{Y|X}(y|1)}. \tag{57}$$

E.g., in [7], 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 [43, 39].

The channel splitting process also leads to a simple decoding algorithm called Successive Cancellation (SC) decoding [6], 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

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

In the context of lossy compression, polar codes can achieve the rate-distortion bound for binary symmetric sources [25]. To achieve a target distortion:

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

By the duality between channel coding and source coding, the 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 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. 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

$$B^{1:n} - A^{1:n} = \sum_{i=1}^{n} (B^i - A^i) A^{1:i-1} B^{i+1:n}, \tag{58}$$

$\Delta\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}, \mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)$ 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| \qquad (59) \\
&\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) \qquad (60) \\
&= \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)} \\
&\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)} \qquad (61) \\
&\overset{(e)}{\leq} \sum_{i\in\mathcal{F}_1} \sqrt{2\ln 2\left(1 - Z(U_1^i|U_1^{1:i-1},Y^{[m]})^2\right)} \\
&\overset{(f)}{\leq} m\sqrt{4\ln 2 \cdot 2^{-m^\beta}}
\end{aligned}
$$

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

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

□

C KL Divergence

Lemma 8. *Let $E \sim D_{\mathbb{Z},\sigma}$ 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:*

$$D_{KL}(\mathsf{P}_{E'}||\mathsf{P}_E) \triangleq \sum_{\lambda\in\mathbb{Z}} \mathsf{P}_{E'}(\lambda) \ln \frac{\mathsf{P}_{E'}(\lambda)}{\mathsf{P}_E(\lambda)} \leq \frac{20}{\sqrt{2\pi\sigma^2}} q \cdot \exp\left(-\frac{q^2}{8\sigma^2}\right), \tag{62}$$

where $q = 2^r$.

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

$$\mathsf{P}_{E'}(\lambda) = \sum_{z\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\lambda+zq)^2}{2\sigma^2}} \tag{63}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + \sum_{z\in\mathbb{Z}^+} e^{-\frac{(\lambda+zq)^2}{2\sigma^2}} + \sum_{z\in\mathbb{Z}^+} e^{-\frac{(\lambda-zq)^2}{2\sigma^2}}\right) \tag{64}$$

$$\leq \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + 2\sum_{z\in\mathbb{Z}^+} e^{-\frac{(|\lambda|-zq)^2}{2\sigma^2}}\right), \tag{65}$$

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,

$$\mathsf{P}_{E'}(\lambda) \leq \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + 2\sum_{z\in\mathbb{Z}^+} e^{-\frac{(|\lambda|-zq)^2}{2\sigma^2}}\right) \tag{66}$$

$$\leq \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + 2\cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}} \cdot \frac{1}{1-e^{-\frac{q^2}{\sigma^2}}}\right) \tag{67}$$

$$\leq \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + 4\cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}}\right), \tag{68}$$

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

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

$$D_{KL}(\mathsf{P}_{E'}||\mathsf{P}_E) = \sum_{\lambda\in\mathbb{Z}} \mathsf{P}_{E'}(\lambda) \ln \frac{\mathsf{P}_{E'}(\lambda)}{\mathsf{P}_E(\lambda)} \tag{69}$$

$$\leq \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + 4 \cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}} \right) \ln \left(1 + 4e^{\frac{\lambda^2-(|\lambda|-q)^2}{2\sigma^2}} \right) \tag{70}$$

$$\leq 4 \cdot \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{\lambda^2}{2\sigma^2}} + 4 \cdot e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}} \right) e^{\frac{2q|\lambda|-q^2}{2\sigma^2}} \tag{71}$$

$$= 4 \cdot \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(|\lambda|-q)^2}{2\sigma^2}} + 16 \cdot \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(|\lambda|-2q)^2}{2\sigma^2}} e^{\frac{q^2}{\sigma^2}} \tag{72}$$

$$\leq 4 \cdot \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{q^2}{8\sigma^2}} + 16 \cdot \sum_{\lambda\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{q^2}{8\sigma^2}} \tag{73}$$

$$= \frac{20}{\sqrt{2\pi\sigma^2}} q \cdot \exp\left(-\frac{q^2}{8\sigma^2} \right), \tag{74}$$

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 9. *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* (36) *and* (37) *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* (36) *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:*

$$D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}} || \mathsf{Q}_{U_1^{[m]},Y^{[m]}} \right) \leq 2\ln 2 \cdot m 2^{-m^\beta}. \tag{75}$$

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

$$D_{KL}\left(\mathsf{P}_{X^{[m]},Y^{[m]}} || \mathsf{Q}_{X^{[m]},Y^{[m]}} \right) = D_{KL}\left(\mathsf{P}_{U_{1:r}^{[m]},Y^{[m]}} || \mathsf{Q}_{U_{1:r}^{[m]},Y^{[m]}} \right) \tag{76}$$

$$\leq 2\ln 2 \cdot rm 2^{-m^\beta}. \tag{77}$$

Proof. For the 1st level,

$$\begin{aligned}
&D_{KL}\left(\mathsf{P}_{U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_1^{[m]},Y^{[m]}}\right)\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log\frac{\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)}{\mathsf{Q}\left(u_1^{[m]},y^{[m]}\right)}\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log\frac{\mathsf{P}\left(u_1^{[m]}|y^{[m]}\right)}{\mathsf{Q}\left(u_1^{[m]}|y^{[m]}\right)}\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\log\frac{\prod_{i=1}^m\mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})}{\prod_{i=1}^m\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]})} \qquad (78)\\
&\quad=\ln 2\cdot\sum_{u_1^{[m]},y^{[m]}}\mathsf{P}\left(u_1^{[m]},y^{[m]}\right)\sum_{i\in\mathcal{F}_1}\log\frac{\mathsf{P}(u_1^i|u_1^{1:i-1},y^{[m]})}{\mathsf{Q}(u_1^i|u_1^{1:i-1},y^{[m]})}\\
&\quad=\ln 2\cdot\sum_{i\in\mathcal{F}_1}(1-H(U_1^i|U_1^{1:i-1},Y^{[m]}))\\
&\quad\le\ln 2\cdot\sum_{i\in\mathcal{F}_1}(1-Z(U_1^i|U_1^{1:i-1},Y^{[m]})^2)\\
&\quad\le 2\ln 2\cdot m2^{-m^\beta},
\end{aligned}$$

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

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

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

where the first term holds because of the result for the 1st level, and the second term can be obtained by following the steps in (78) exactly, since it can be written as

$$\begin{aligned}
&\mathsf{E}_{U_1^{[m]},Y^{[m]}}\left[D_{KL}\left(\mathsf{P}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}||\mathsf{Q}_{U_2^{[m]}|U_1^{[m]},Y^{[m]}}\right)\right]\\
&\qquad=\ln 2\cdot\sum_{u_{1:2}^{[m]},y^{[m]}}\mathsf{P}\left(u_{1:2}^{[m]},y^{[m]}\right)\log\frac{\mathsf{P}\left(u_2^{[m]}|u_1^{[m]},y^{[m]}\right)}{\mathsf{Q}\left(u_2^{[m]}|u_1^{[m]},y^{[m]}\right)}. \qquad (79)
\end{aligned}$$

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