Error Estimation of Practical Convolution Discrete Gaussian Sampling

Zhongxiang Zheng[1], Xiaoyun Wang[2,3*], Guangwu Xu[4*], Chunhuan Zhao[2]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] Institute for Advanced Study, Tsinghua University, Beijing 100084, China
[3] Key Laboratory of Cryptologic Technology and Information Security, Ministry of Education, Shandong Universtiy, Jinan 250100, China
[4] Department of Electrical Engineering and Computer Sciences, University of Wisconsin, Milwaukee, WI 53201, USA
[*] Corresponding authors

xiaoyunwang@mail.tsinghua.edu.cn

Abstract. Discrete Gaussian Sampling is a fundamental tool in lattice cryptography which has been used in digital signatures, identify-based encryption, attribute-based encryption, zero-knowledge proof and fully homomorphic cryptosystem. As a subroutine of lattice-based scheme, a high precision sampling usually leads to a high security level and also brings large time and space complexity. In order to optimize security and efficiency, how to achieve a higher security level with a lower precision becomes a widely studied open question. A popular method for addressing this question is to use different metrics other than statistical distance to measure errors. The proposed metrics include KL-divergence, Rényi-divergence, and Max-log distance, and these techniques are supposed to achieve 2^p security with $\frac{p}{2}$ precision or even less. However, large gaps of errors between previous results and practical experiments have been spotted. In this paper, we make two novel observations about practical errors. These observations reveal missing keys that link theoretical and practical results and also provide an answer to the open question. As an application of the observations, we revisit convolution theorem of discrete Gaussian sampling and reformulate it into a practical one with much more accurate error bounds than existing results. We describe a rigorous proof of it and demonstrate that the bounds are tightly matched by our experiments. We also theoretically and experimentally show that the sets of parameters of several existing practical convolution sampling schemes need to be modified in order to achieve their preset goals. Some technical tools including several improved inequalities for discrete Gaussian measure are developed. Finally, we would like to point out that the successful analysis of a couple of previous practical convolution theorems using our observations makes us to believe that a large number of existing practical schemes and designs concerning discrete Gaussian sampling may need to be reconsidered.

Key words: Discrete Gaussian Sampling, convolution theorem, lattice, error estimation

1 Introduction

In recent years, research in lattice-based cryptography has attracted considerable attention. This is mainly because mathematical and computational properties of lattices provide basis for advanced schemes, such as digital signatures, identity-based and attribute-based encryption, zero-knowledge proof and fully homomorphic schemes, and some of the lattice-based cryptosystems are likely to be effective against quantum computing attacks in the future. Many of these lattice-based schemes rely on a polynomial-time algorithm which samples from a discrete Gaussian distribution over a lattice. Thus discrete Gaussian sampling is one of the fundamental tools of lattice cryptography.

Discrete Gaussian over lattices has been well studied in mathematics [1, 2] and becomes an exceedingly useful analytical tool in discussing the computational complexity of lattice problems [5, 6, 13]. A discrete Gaussian sampling algorithm takes a basis of the lattice Λ, a vector $\mathbf{c} \in \mathbb{R}^n$, and a *width* parameter $s > 0$ as inputs, and outputs a vector $\mathbf{v}$ that obeys the distribution $D_{\Lambda+\mathbf{c},s}$ which assigns a probability proportional to $e^{-\pi\|\mathbf{v}-\mathbf{c}\|^2/s^2}$. Two of the most influential discrete Gaussian sampling algorithms are Babai's nearest-plane algorithm [7] and the sampling algorithm of Gentry, Peikert and Vaikuntanathan [11]. Babai's algorithm was proposed in 1986 and Gentry, Peikert and Vaikuntanathan improved it by replacing the deterministic rounding process in each iteration by a probabilistic rounding process which is determined by its distance from the target point [7]. The work [11] also provided an analysis of the sample distribution using smoothing parameter of Micciancio and Regev [8], in terms of statistical distance. A further improvement and extension of the sampling algorithm of [11] was obtained by Peikert [14] in 2010, where a parallelizable Gaussian sampling algorithm is established based on the famous *convolution theorem* of discrete Gaussian as well as its theoretical bound. The convolution theorem of discrete Gaussian allows the generation of a sample with relatively large standard deviation s by combining results of different samples with small standard deviation s'. This technique greatly improves the efficiency for sampling with large standard deviation. Many practical improvements about Gaussian sampling have been made based on the convolution theorem. For example, Pöppelmann, Ducas and Güneysu proposed a highly efficient lattice-based signatures on reconfigurable hardware in 2014 [15] and Micciancio and Walter provided a generic Gaussian sampling algorithm with high efficiency and constant-time in 2017 [10]. Improvements have been reported in recent work [12, 16], where results of [10] were further utilized and expanded.

The error estimation of discrete Gaussian sampling is one of the key issues. The influence of float-point errors has reveived special attention because precision directly decides the time and space complexity of practical sampling. In 2010, Peikert [14] gave a theoretical error estimation under statistical distance, with an error bound 2ε (where $\varepsilon \leq 1/2$ is with respect to the smoothing parameter) without considering floating-point errors and truncation errors. Pöppelmann, Ducas and Güneysu [15] adapted Peikert's analysis by scaling the standard deviation s' of one of the base samplers by a factor of 11 and provided an error estimation about the convolution as $32\varepsilon^2$ under the Kullback-Leibler divergence. With this bound, [15] reduces the precision by half and claims the same size of errors. Furthermore, Micciancio and Walter [10] improved the analysis about error estimation of convolution theorem by using a novel notion of "max-log" distance. Another kind of approach [16] is based on Rényi divergence which improved the result of [3, 4]. All of the above work using a metric other than statistical distance claim to achieve 2^p security with $p/2$ precision or even less. In other words, $p/2$ precision are claimed to be sufficient to keep the errors smaller than 2^{-p}.

In this paper, after spotting that the above claims fail to fit with experiments, we make two important observations (Propositions 1 and 2) in the area of practical analysis and propose a new practical convolution theorem based on these observations, under sum-like metrics (i.e. statistical distance, KL-divergence, Rényi-divergence) and max-like metrics (i.e. relative difference, max-log distance) which are well consist with experiment results. The new convolution also leads to a modification of parameter sets in several existing practical discrete Gaussian sampling schemes to achieve their preset goals. Our modification is justified and validated by experiments. Besides, this paper also contains some new technical frameworks and improved inequalities concerning discrete Gaussian measure. Furthermore, from the observations, we naturally get a conclusion that the question of achieving higher security level with much lower precision seems to be impossible to solve.[1] It should also be noted that an implication of the observations is that, for several practical designs and schemes that are based on low precision discrete Gaussian sampling, revisiting is necessary.

The rest of the paper is organized as follows. In section 2, we introduce some background about lattice, discrete Gaussian sampling, as well as error estimation results for convolution theorem from [10, 14, 15]. Our observations and their proofs are are presented in section 3. In section 4, we use the observations to revisit convolution theorem of discrete Gaussian sampling. Then we discuss applications of the new practical convolution theorem and provide experiments. Finally, a conclusion is given in section 5.

[1] Specifically, constant improvements are available, i.e. 2^p security level with p+O(1) precision, but exponential improvements are not, i.e. 2^p security level with p/2 precision.

2 Preliminaries

2.1 Error Estimation

Statistical Distance. Statistical distance is defined as the sum of absolute errors, let P and Q be two distributions over a common countable set S, the statistical distance between distributions P and Q, denoted as Δ_{SD}, is

$$\Delta_{SD}(P, Q) = \frac{1}{2} \sum_{x \in S} |P(x) - Q(x)|$$

Relative Difference. The relative difference is defined as the maximum ratio between absolute error and corresponding probability, let P and Q be two distributions over a common countable set S, the relative difference, denoted as Δ_{RE} between distributions P and Q, is

$$\Delta_{RE}(P, Q) = \max_{x \in S} \delta_{RE}(P(x), Q(x))$$

where $\delta_{RE}(P(x), Q(x)) = \frac{|P(x)-Q(x)|}{P(x)}$.
Kullback-Leibler Divergence. Let P and Q be two distributions over a common countable set Ω, and let $S \subset \Omega$ be the strict support of P ($P(i) > 0$ iff $i \in S$). The Kullback-Leibler divergence, denoted as Δ_{KL} of Q from P, is defined as

$$\Delta_{KL}(P \| Q) = \sum_{x \in S} P(x) \ln \frac{P(x)}{Q(x)}$$

where $\ln(x/0) = +\infty$ for any $x > 0$.
Max-log Distance. This metric was first introduced in [10]. Given two distributions P and Q over a common countable set S, their max-log distance Δ_{ML} is defined as

$$\Delta_{ML}(P, Q) = \max_{x \in S} \delta_{ML}(P(x), Q(x))$$

where $\delta_{ML}(P(x), Q(x)) = |\ln P(x) - \ln Q(x)|$.
Rényi-divergence. Given two distributions P and Q over a common countable set S, for $\alpha \in (1, +\infty)$, their Rényi-divergence is defined in [16] as

$$\Delta_{RD_\alpha}(P \| Q) = \Big(\sum_{x \in S} \frac{P(x)^\alpha}{Q(x)^{\alpha-1}} \Big)^{\frac{1}{\alpha-1}}$$

and for $\alpha = +\infty$ Rényi-divergence is defined as

$$\Delta_{RD_\infty}(P \| Q) = \max_{x \in S} \frac{P(x)}{Q(x)}$$

Relationships between Metrics. For a real number x and its p-bit approximation $\bar{x}$ which stores the p most significant bits of x in binary. More specifically, if $x = 2^k \sum_{i=1}^{+\infty} x_i 2^{-i}$ with $x_i \in \{0, 1\}$ and $x_1 = 1$, we denote the rounding function with precision p as Rd_p whose evaluation at x is $Rd_p(x) = 2^k(\sum_{i=1}^{p} x_i 2^{-i} + x_{i+1} 2^{-p})$. We write $\bar{x} = Rd_p(x)$ and obtain

$$\delta_{RE}(x, \bar{x}) < 2^{-p}$$

by computing $|\bar{x}-x|/\bar{x} = 2^k|x_{i+1}2^{-p-1} - \sum_{i=p+2}^{+\infty} x_i 2^{-i}|/(2^k(\sum_{i=1}^{p} x_i 2^{-i} + x_{i+1}2^{-p})) < 2^{-p-1}/2^{-1} = 2^{-p}$.

A relation that links statistical distance and Δ_{KL} is described by the following Pinsker's inequality

$$\Delta_{KL}(P\|Q) \geq 2\Delta_{SD}^2(P,Q).$$

For Δ_{KL} and Δ_{RE}, the inequality

$$\Delta_{KL}(P\|Q) \leq 2\Delta_{RE}^2(P\|Q)$$

was proved in [15] under the condition that $\Delta_{RE}(P,Q) < 1/4$. Actually, this argument is a special case of a general result: assume that for any $i \in S$, there exists some $\delta(i) \in (0,1/4)$ such that $|P(x)-Q(x)| \leq \delta(x)P(x)$, then $\Delta_{KL}(P\|Q) \leq 2\sum_{x\in S}\delta^2(x)P(x)$ holds. The relationship between Δ_{KL} and Δ_{RE} follows by setting $\delta(i) = \Delta_{RE}(P,Q)$.

Recently in [10], the above relation was further improved to

$$\Delta_{KL}(P\|Q) \leq (8/9)\Delta_{RE}^2(P\|Q).$$

In fact, [10] established a more general inequality $\Delta_{KL}(P\|Q) \leq \frac{\Delta_{RE}^2(P\|Q)}{2(1-\Delta_{RE}(P,Q))^2}$ for the case $\Delta_{RE}(P,Q) < 1$.

The following relationship between Rényi-divergence and relative difference is given in [16].

$$\Delta_{RD_\alpha}(P\|Q) \leq \left(1 + \frac{\alpha(\alpha-1)\Delta_{RE}^2}{2(1-\Delta_{RE})^{\alpha+1}}\right)^{\frac{1}{\alpha-1}}.$$

Lemma 4.2 of [10] sets up a relation between Δ_{ML} and Δ_{RE}, however the statement of the lemma contains an error because $-\ln(1-x) \leq x$ is not true for $x \in (0,1)$, and the proof needs to be taken care of. Here we establish a slightly more precise inequality for these two quantities with a rigorous proof. It should be pointed out that we assume that P and Q share exactly the same strict support S. This is always true if the condition $\Delta_{RE}(P,Q) < 1$ holds.

Lemma 2.1 *If $\Delta_{RE}(P,Q) < 1$, then*

$$\big|\Delta_{ML}(P,Q) - \Delta_{RE}(P,Q)\big| \leq \frac{\Delta_{RE}^2(P,Q)}{2(1-\Delta_{RE}(P,Q))}.$$

Proof. Note that for $|t| < 1$, we have $|\ln(1-t)| = \left|t + \frac{t^2}{2} + \frac{t^3}{3} + \cdots\right|$. For $x \in S$, we set $t_x = \frac{P(x)-Q(x)}{P(x)}$. On the one hand, we have

$$\begin{aligned}\left|\ln\frac{Q(x)}{P(x)}\right| &= \big|\ln(1-t_x)\big| = \big|t_x + \frac{t_x^2}{2} + \frac{t_x^3}{3} + \cdots\big| \leq |t_x| + \frac{|t_x|^2}{2} + \frac{|t_x|^3}{3} + \cdots \\ &\leq \Delta_{RE}(P,Q) + \frac{\Delta_{RE}^2(P,Q)}{2} + \frac{\Delta_{RE}^3(P,Q)}{3} + \cdots \\ &\leq \Delta_{RE}(P,Q) + \frac{\Delta_{RE}^2(P,Q)}{2(1-\Delta_{RE}(P,Q))}.\end{aligned}$$

This gives $\Delta_{ML}(P,Q) \leq \Delta_{RE}(P,Q) + \frac{\Delta^2_{RE}(P,Q)}{2(1-\Delta_{RE}(P,Q))}$.

On the other hand, $\left|\ln \frac{Q(x)}{P(x)}\right| = \left|\ln(1-t_x)\right| \geq |t_x| - \left|\frac{t_x^2}{2} + \frac{t_x^3}{3} + \cdots\right|$. So

$$\begin{aligned}|t_x| &\leq \left|\ln \frac{Q(x)}{P(x)}\right| + \left|\frac{t_x^2}{2} + \frac{t_x^3}{3} + \cdots\right| \leq \max_{x\in S}\left|\ln \frac{Q(x)}{P(x)}\right| + \frac{\Delta^2_{RE}(P,Q)}{2} + \frac{\Delta^3_{RE}(P,Q)}{3} + \cdots \\ &\leq \Delta_{ML}(P,Q) + \frac{\Delta^2_{RE}(P,Q)}{2(1-\Delta_{RE}(P,Q))}.\end{aligned}$$

This yields $\Delta_{RE}(P,Q) \leq \Delta_{ML}(P,Q) + \frac{\Delta^2_{RE}(P,Q)}{2(1-\Delta_{RE}(P,Q))}$ and the lemma is proved. □

It can be easily verified that the result of the lemma is also true if we use δ_{RE} and δ_{ML}.

For distribution P_i and Q_i over support $\prod_i S_i$, [10] also proved that if $\Delta_{ML}(P_i|a_i, Q_i|a_i) \leq 1/3$ for all i and $a_i \in \prod_{j<i} S_j$, then

$$\Delta_{SD}((P_i)_i, (Q_i)_i) \leq \|(\max_{a_i} \Delta_{ML}(P_i|a_i, Q_i|a_i))_i\|_2 \tag{1}$$

2.2 Discrete Gaussian Sampling

Given $x \in \mathbb{R}^n$ and and a countable set $A \subset \mathbb{R}^n$, we define the Gaussian function $\rho_{s,c}(x) = e^{-\pi\frac{\|x-c\|^2}{s^2}}$ and Gaussian sum $\rho_{s,c}(A) = \sum_{x\in A} \rho_{s,c}(x)$, then $Pr(x) = \frac{\rho_{s,c}(x)}{\rho_{s,c}(A)}$ gives a discrete (Gaussian) probability distribution on A which we call $D_{A,c,s}$. The subindexes c or/and s are omitted if $c = 0$ or/and $s = 1$. Gaussian function can be defined in terms of a positive definite matrix instead of s.

The insight-conveying concept of smoothing parameter of Micciancio and Regev [8] for an n-dimensional lattice Λ is with respect to an $\varepsilon > 0$ and given by $\eta_\varepsilon(\Lambda) = \min\{r : \rho_{1/r}(\Lambda^*) \leq 1+\varepsilon\}$. One of the bounds given in [8] states

$$\eta_\varepsilon(\Lambda) \leq \sqrt{\ln(2n(1+1/\varepsilon))/\pi} \cdot \lambda_n(\Lambda).$$

For the special case of $\Lambda = \mathbb{Z}$, we have

$$\eta_\varepsilon(\mathbb{Z}) \leq \sqrt{\ln 2(1+1/\varepsilon)/\pi}.$$

This, together with the fact that $2e^{-\pi\eta_\varepsilon(\mathbb{Z})^2} < \rho_{\frac{1}{\eta_\varepsilon(\mathbb{Z})}}(\mathbb{Z}\setminus\{0\}) \leq \varepsilon$, yields

$$\frac{2}{e^{\pi(\eta_\varepsilon(\mathbb{Z}))^2}} < \varepsilon \leq \frac{2}{e^{\pi(\eta_\varepsilon(\mathbb{Z}))^2} - 2}.$$

We shall assume that $\eta_\varepsilon(\mathbb{Z}) \geq 1$ since ε is small. Note that $\rho(\mathbb{Z}) < 1.086435$, it is thus meaningful to choose $\varepsilon < 0.086$ in the rest of our discussion.

Next, we will prove a little tighter tail bound about discrete Gaussian probability which improves Lemma 4.1 of [11]. To this end, we also need to develop a slightly more precise estimation over Banaszczyk lemma [2] for the case of $\mathbb{Z}$.

Lemma 2.2 *Let s, t be positive numbers such that $ts \geq 1$ and $c \in [0,1)$. We have*

1.

$$\sum_{\substack{k\in\mathbb{Z} \\ |k-c|\geq ts}} \rho_s(k-c) \leq 2e^{-\pi t^2}\left(1+\frac{e^{-\frac{2\pi t}{s}}}{2}(\rho_s(\mathbb{Z})-1)\right). \quad (2)$$

2. If $s \geq \eta_\varepsilon(\mathbb{Z})$, then

$$\sum_{\substack{x\in\mathbb{Z} \\ |x-c|\geq t\cdot s}} Pr_{x\leftarrow D_{\mathbb{Z},c,s}}(x) \leq 2e^{-\pi t^2}\cdot\frac{1+\varepsilon}{1-\varepsilon}\left(\frac{1+\frac{e^{-\frac{2\pi t}{s}}}{2}(\rho_s(\mathbb{Z})-1)}{\rho_s(\mathbb{Z})}\right). \quad (3)$$

Remarks.

1. We include a proof of the lemma in the appendix.
2. We remark that the proof of equation (2) can be easily extended to get an alternative proof of Banaszczyk lemma (Lemma 2.4 in [2]) for a general lattice $L \subset \mathbb{R}^n$.
3. Our new bound (2) improves the original bound $2e^{-\pi t^2}\rho_s(\mathbb{Z})$ to $Ce^{-\pi t^2}\rho_s(\mathbb{Z})$ with $C = \frac{2}{\rho_s(\mathbb{Z})} + e^{-\frac{2\pi t}{s}}\left(1-\frac{1}{\rho_s(\mathbb{Z})}\right)$. Obviously under the natural condition $s \geq 1$ we have that $C \leq 2$ [2]. This C can be much smaller. For example, in our later application, we will choose $s = 34, t = 6$, so $C < 0.38$.

2.3 Convolution Theorem and its Improvements

In 2010, a convolution theorem for discrete Gaussian was formulated and proved by Peikert [14] which utilizes smoothing parameter. The convolution theorem states

Theorem 2.3 *(Convolution Theorem [14]) Let $\Sigma_1, \Sigma_2 > \mathbf{0}$ be positive definite matrices and set $\Sigma = \Sigma_1 + \Sigma_2$ and $\Sigma_3^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1}$. Let Λ_1, Λ_2 be lattices such that $\sqrt{\Sigma_1} \geq \eta_\varepsilon(\Lambda_1)$ and $\sqrt{\Sigma_3} \geq \eta_\varepsilon(\Lambda_2)$ for some positive $\varepsilon \leq 1/2$, and let $\mathbf{c_1}, \mathbf{c_2} \in \mathbb{R}^n$ be arbitrary. Choose $\mathbf{x_2} \leftarrow D_{\Lambda_2+\mathbf{c_2},\sqrt{\Sigma_2}}$ and $\mathbf{x_1} \leftarrow \mathbf{x_2} + D_{\Lambda_1+\mathbf{c_1}-\mathbf{x_2},\sqrt{\Sigma_1}}$. If $\tilde{D}_{\mathbf{c_1}+\Lambda_1,\sqrt{\Sigma}}$ is the distribution of $\mathbf{x_1}$, then*

$$\delta_{RE}(Pr_{\tilde{D}_{\mathbf{c_1}+\Lambda_1,\sqrt{\Sigma}}}[x=\bar{x}], Pr_{D_{\mathbf{c_1}+\Lambda_1,\sqrt{\Sigma}}}[x=\bar{x}]) \leq (\frac{1+\varepsilon}{1-\varepsilon})^2 - 1.$$

This convolution theorem was strengthened by Micciancio and Peikert in 2013 (Theorem 3.3 of [9]). We observe that the proof in [9] can be modified so that an improved version of Theorem 3.3 of [9] can be stated. For a vector $\mathbf{z} \in \mathbb{Z}^m$, we denote $z_{\max}$ and $z_{\min}$ to be the largest and smallest components (in absolute values) of $\mathbf{z}$ respectively, then our form of the theorem is

[2] Note that by the Poisson Summation formula we get $s < \rho_s(\mathbb{Z}) < s + \frac{2se^{-\pi s^2}}{1-e^{-3\pi s^2}}$.

Theorem 2.4 *Let Λ be an n-dimensional lattice, $\mathbf{z} \in \mathbb{Z}^m$ a nonzero integer vector, $\mathbf{s} \in \mathbb{R}^m$ with $s_i \geq \sqrt{z_{\max}^2 + z_{\min}^2}\eta_\varepsilon(\mathbb{Z})$ for all $i \leq m$ and $\mathbf{c_i} + \mathbf{\Lambda}$ arbitrary cosets. Let $\mathbf{y}_i$ be independent samples from $D_{\mathbf{c_i}+\Lambda,s_i}$, respectively. Let $Y = \sum_i z_i\mathbf{c_i} + \gcd(\mathbf{z})\Lambda$ and $s = \sqrt{\sum_i (z_i s_i)^2}$. Then $\tilde{D}_{Y,s}$, the distribution of $\mathbf{y} = \sum z_i\mathbf{y_i}$, is close to $D_{Y,s}$. More precisely,*

$$\delta_{RE}(Pr_{\tilde{D}_{Y,s}}[x = \bar{x}], Pr_{D_{Y,s}}[x = \bar{x}]) \leq \frac{1+\varepsilon}{1-\varepsilon} - 1.$$

Remark. We note that the assumption of Theorem 3.3 of [9] was $s_i \geq \sqrt{2}\|\mathbf{z}\|_\infty \eta_\varepsilon(\mathbb{Z})$. Our version is more efficient as $\sqrt{z_{\max}^2 + z_{\min}^2} \leq \sqrt{2}\|\mathbf{z}\|_\infty$. Notice that in applications, one often requires $\gcd(\mathbf{z}) = 1$, so $z_{\max} > z_{\min}$ and hence $\sqrt{z_{\max}^2 + z_{\min}^2} < \sqrt{2}\|\mathbf{z}\|_\infty$. The improvement has a significant impact on estimating ε with respect to the smoothing parameter η_ε. To illustrate simply, for a fixed s, the original result gives an error $\varepsilon_{old} \leq 2e^{-\pi\frac{s^2}{2z_{\max}^2}}$ while ours shows $\varepsilon_{new} \leq 2e^{-\pi\frac{s^2}{z_{\max}^2+z_{\min}^2}}$. So for some choices of parameters (e.g. $z_{\max}$ is much larger than $z_{\min}$), our estimated error may be as finer as the square of the previous one, i.e., $\varepsilon_{new} \approx \varepsilon_{old}^2$. The proof of the current version of the theorem just modifies the last part of that given in [9] and we include that part in the appendix.

Pöppelmann, Ducas and Güneysu considered one-dimensional case in [15]. Using Δ_{KL} instead of Δ_{SD} and with one lattice being sampled to be $k\mathbb{Z}$, their improved convolution theorem states

Theorem 2.5 *(Convolution Theorem [15]) Let $x_1 \leftarrow D_{\mathbb{Z},s_1}$, $x_2 \leftarrow D_{k\mathbb{Z},s_2}$ for some positive reals s_1, s_2, and let $s_3^{-2} = s_1^{-2} + s_2^{-2}$ and $s^2 = s_1^2 + s_2^2$. For any $\varepsilon \in (0, 1/2)$ if $s_1 \geq \eta_\varepsilon(\mathbb{Z})$ and $s_3 \geq \eta_\varepsilon(k\mathbb{Z})$, to the distribution of $x = x_1 + x_2$, denoted as D_x, is close to $D_{\mathbb{Z},s}$ under KL-divergence*

$$\Delta_{KL}(D_x \| D_{\mathbb{Z},s}) \leq 2(1 - (\frac{1+\varepsilon}{1-\varepsilon})^2)^2.$$

Another useful bound in studying error estimation of convolution theorem is also proposed in [10] which describes errors when continuously using approximated output results as inputs of the next round.

Theorem 2.6 *Let Δ be a metric. Let A^P be an algorithm querying a distribution ensemble P_θ at most q times. Then*

$$\Delta(A^Q, R) \leq \Delta(A^P, R) + q \cdot \Delta(P_\theta, Q_\theta)$$

for any distribution R and any ensemble Q_θ.

Micciancio and Walter are the first to analyze error estimation of convolution discrete Gaussian sampling using the metric Δ_{ML} by combining equation (1), theorem 2.3, theorem 2.4, and theorem 2.6. Their result is also the first practical refinement of convolution theorem that takes floating-point errors into account. The following two corollaries from [10] give error estimation under max-log distance.

Corollary 2.7 *(Corollary 4.1 of [10]) Let* $\mathbf{z} \in \mathbb{Z}^m$ *be a nonzero integer vector with* $gcd(\mathbf{z}) = 1$ *and* $\mathbf{s} \in \mathbb{R}^m$ *with* $s_i \geq \sqrt{2}\|\mathbf{z}\|_\infty \eta_\varepsilon(\mathbb{Z})$ *for all* $i \leq m$. *Let* y_i *be independent samples from* $\tilde{D}_{\mathbb{Z},s_i}$, *respectively, with* $\Delta_{ML}(D_{\mathbb{Z},s_i}, \tilde{D}_{\mathbb{Z},s_i}) \leq \mu_i$ *for all* i. *Let* $\tilde{D}_{\mathbb{Z},s}$ *be the distribution of* $y = \sum z_i y_i$ *and* $s^2 = \sum s_i^2$. *Then* $\Delta_{ML}(D_{\mathbb{Z},s}, \tilde{D}_{\mathbb{Z},s}) \lesssim 2\varepsilon + \sum_i \mu_i$.

Remark. The assumption of $s_i \geq \sqrt{2}\|\mathbf{z}\|_\infty \eta_\varepsilon(\mathbb{Z})$ can be replaced by $s_i \geq \sqrt{z_{\max}^2 + z_{\min}^2}\,\eta_\varepsilon(\mathbb{Z})$ according to Theorem 2.4.

Corollary 2.8 *(Corollary 4.2 of [10]) Let* $s_1, s_2 > 0$ *with* $s^2 = s_1^2 + s_2^2$ *and* $s_3^{-2} = s_1^{-2} + s_2^{-2}$. *Let* $\Lambda = K\mathbb{Z}$ *be a copy of the integer lattice* $\mathbb{Z}$ *scaled by a constant* K. *For any* c_1 *and* $c_2 \in \mathbb{R}$, *denote the distribution of* $x_1 \leftarrow x_2 + \tilde{D}_{c_1 - x_2 + \mathbb{Z}, s_1}$, *where* $x_2 \leftarrow \tilde{D}_{c_2+\Lambda,s_2}$, *by* $\tilde{D}_{c_1+\mathbb{Z},s}$. *If* $s_1 \geq \eta_\varepsilon(\mathbb{Z})$, $s_3 \geq \eta_\varepsilon(\Lambda) = K\eta_\varepsilon(\mathbb{Z})$, $\Delta_{ML}(D_{c_2+\Lambda,s_2}, \tilde{D}_{c_2+\Lambda,s_2}) \leq \mu_1$ *and* $\Delta_{ML}(D_{c+\mathbb{Z},s_1}, \tilde{D}_{c+\mathbb{Z},s_1}) \leq \mu_2$ *for any* $c \in \mathbb{R}$, *then* $\Delta_{ML}(D_{c_1+\mathbb{Z},s}, \tilde{D}_{c_1+\mathbb{Z},s}) \lesssim 4\varepsilon + \mu_1 + \mu_2$.

3 Two Critical Observations about Practical Errors

In this section, we make two novel observations about practical errors. These two observations are the keys to more precisely determine the dominant term of practical errors in discrete Gaussian sampling. We first define two bounds for practical errors: $\varepsilon_t = \rho_{1/t}(\mathbb{Z}) - 1$ and $\mu = 2^{-p}$. Note that for $t > 1, \varepsilon_t = 2\sum_{i=1}^{+\infty} e^{-\pi t^2 i^2} \in (2e^{-\pi t^2}, \frac{2e^{-\pi t^2}}{1-e^{-3\pi t^2}})$. We will use ε_t to control the truncation error with respect to t, and μ to control float-point errors.

Our first observation indicates that, in general, the sum of the stored probabilities cannot be close to 1 by the order of μ^2.

Proposition 1 *Let* $P_1, ..., P_n$ *be a finite probability distribution and* $\bar{P}_1, ..., \bar{P}_n$ *the corresponding p-bits approximations (i.e.* $\bar{P}_i = Rd_p(P_i)$*). We have*

$$\left|1 - \sum_{i=1}^{n} \bar{P}_i\right| \leq \mu.$$

Moreover, this bound is sharp in the sense that it cannot be improved to $< \frac{\mu}{2}$.

Proof. Write $P_i = 2^{k_i}\sum_{j=1}^{\infty} x_{i,j}2^{-j}$ with $x_{i,1} = 1$ and $x_{i,j} \in \{0,1\}$ for $j = 2, 3, \cdots$. Since $\sum_{i=1}^n P_i = 1$, i.e., $\sum_{i=1}^n 2^{k_i-1} + 2^{-1}\sum_{i=1}^n x_{i,2}2^{k_i-1} + 2^{-2}\sum_{i=1}^n x_{i,3}2^{k_i-2} + \cdots = 1$, we see that

$$\frac{1}{2} \leq \sum_{i=1}^{n} 2^{k_i-1} \leq 1.$$

Note that $\bar{P}_i = Rd_p(P_i) = 2^{k_i}(\sum_{j=1}^{p} x_{i,j}2^{-j} + x_{i,p+1}2^{-p})$, we get

$$\begin{aligned}
\sum_{i=1}^{n} \bar{P}_i &= \sum_{i=1}^{n} 2^{k_i}(\sum_{j=1}^{p} x_{i,j}2^{-j} + x_{i,p+1}2^{-p}) \\
&= \sum_{i=1}^{n} 2^{k_i}(\sum_{j=1}^{\infty} x_{i,j}2^{-j} - \sum_{j=p+1}^{\infty} x_{i,j}2^{-j} + x_{i,p+1}2^{-p}) \\
&= 1 + \sum_{i=1}^{n} 2^{k_i} x_{i,p+1}2^{-p-1} - \sum_{i=1}^{n} 2^{k_i} x_{i,p+2}2^{-p-2} - \cdots \\
&= 1 + 2^{-p}\big(\sum_{i=1}^{n} 2^{k_i-1}x_{i,p+1} - 2^{-1}\sum_{i=1}^{n} 2^{k_i-1}x_{i,p+2} - 2^{-2}\sum_{i=1}^{n} 2^{k_i-1}x_{i,p+3} - \cdots\big)
\end{aligned}$$

Therefore

$$\sum_{i=1}^{n} \bar{P}_i \le 1 + 2^{-p}\sum_{i=1}^{n} 2^{k_i-1}x_{i,p+1} \le 1 + 2^{-p}\sum_{i=1}^{n} 2^{k_i-1} \le 1 + 2^{-p},$$

and

$$\begin{aligned}
\sum_{i=1}^{n} \bar{P}_i &\ge 1 - 2^{-p}\big(2^{-1}\sum_{i=1}^{n} 2^{k_i-1}x_{i,p+2} + 2^{-2}\sum_{i=1}^{n} 2^{k_i-1}x_{i,p+3} + \cdots\big) \\
&\ge 1 - 2^{-p}\big(2^{-1}\sum_{i=1}^{n} 2^{k_i-1} + 2^{-2}\sum_{i=1}^{n} 2^{k_i-1} + \cdots\big) \\
&= 1 - 2^{-p}\sum_{i=1}^{n} 2^{k_i-1} \ge 1 - 2^{-p}.
\end{aligned}$$

Thus

$$\big|1 - \sum_{i=1}^{n} \bar{P}_i\big| \le \mu.$$

Next, we shall construct a counterexample to show that $\big|1 - \sum_{i=1}^{n} \bar{P}_i\big| < \frac{\mu}{2}$ is false. There are many such examples. We give a simple one: Let $P_1 = 2^{-1} + 2^{-p-2}$ and $P_2 = 1 - P_1$. Then $\bar{P}_1 = 2^{-1}$ and $\bar{P}_2 = 2^{-1}(2^{-1} + 2^{2} + \cdots + 2^{-p})$. So

$$\bar{P}_1 + \bar{P}_2 = 2^{-1} + 2^{-2} + \cdots + 2^{-p-1} = 1 - 2^{-p-1} = 1 - \frac{\mu}{2}.$$

Remarks

1. We should remark that there are more cases for $\big|1 - \sum_{i=1}^{n} \bar{P}_i\big| \ge \frac{\mu}{4}$. However, the probability for $\big|1 - \sum_{i=1}^{n} \bar{P}_i\big| \le \mu^2$, contrary to what was believed previously, is extremely small.
2. Letting $P'_i = \frac{\bar{P}_i}{\sum_{i=1}^{n} \bar{P}_i}$, then $\frac{\bar{P}_i}{1+2^{-p}} \le P'_i \le \frac{\bar{P}_i}{1-2^{-p}}$. We observe that in most cases, $Rd_p(P'_i) = \bar{P}_i$ for all $i = 1, 2, \cdots, n$. In other words, for many cases, normalizing the stored probabilities achieves nothing in terms of storage. We shall call this anti-intuitive phenomena the *Distribution Precision Paradox*.
3. We also have $\sum_{i=1}^{n} Rd_p(P'_i) = 1 + O(\mu)$. This can be seen from the fact that $\big|Rd_p(P'_i) - \bar{P}_i\big| \le 2^{-p+1}\bar{P}_i$.

4. The proposition naturally leads to a result that floating-point errors are mostly around $O(\mu)$ rather than $O(\mu^2)$ under all metrics mentioned above. Given a fixed precision p, one may try to redistribute $1 - \sum_{i=1}^{n} Rd_p(P_i')$ to make the sum of the stored probabilities get closer to 1. However this does not seem to change the situation because the method also introduces a larger relative error for corresponding $Rd_p(P_i')$. These expanded relative errors finally cause non-decreasing floating-point errors under all metrics mentioned above.
5. When adding truncation errors into consideration, we can get a similar result that normalization process will not efficiently remove the influence of truncation errors on the sum of probabilities because of the limitation of the storage space. As a result, in a base sampler of discrete Gaussian sampling, we always have
$$\sum_{i=1}^{n} Rd_p(P_i') = 1 + O(\mu) + O(\varepsilon_t).$$

Possibly due to the extremely small tail bound of the discrete Gaussian measure, truncation errors are often ignored in the previous considerations. Our second observation reveals a contrary result for the case of convolution of two discrete Gaussian variables. We actually show that during the process of convolution, the ignored part may contribute significantly and become the main term with respect to several metrics. Though we will discuss this issue in great detail later, we describe its conclusion here.

Let x_1, x_2 be sampled from D_s independently and be restricted on the truncation ranges $S_1 = [-ts, ts]$ (namely, supports of x_1 and x_2 are all in S_1). Let a, b be positive integers with $\gcd(a, b) = 1$ and $x = ax_1 + bx_2$. The probability of x is computed by the convolution, denoted by $P'(x)$. We also restrict the support of x to $S = [-t\sqrt{a^2+b^2}s, t\sqrt{a^2+b^2}s]$. Setting $\eta = \frac{\sqrt{a^2+b^2}}{s}, \psi = \min\{\frac{\sqrt{a^2+b^2}-a}{b}, \frac{\sqrt{a^2+b^2}-b}{a}\}$ and $\omega = 1 - \frac{\eta}{\psi t}$, we can state our observation as

Proposition 2 *Let $P(x)$ be the probability of x for the ideal discrete Gaussian distribution $D_{\sqrt{a^2+b^2}s}$. If $st \geq \frac{\sqrt{a^2+b^2}}{\psi}$, then*
$$\Delta_{RE}(P', P) \leq \varepsilon_t^{\omega^2\psi^2}.$$
Moreover, this bound is sharp in the sense that it cannot be improved to $< \varepsilon_t^{(\sqrt{2}-1)^2}$.

Remarks

1. The proof of this result will be given in the next section (Lemma 4.2). It can be seen that $\omega^2\psi^2 \leq (\sqrt{2}-1)^2$ and our experiments (Table 2 and Figure 4) in next section show that our bound is sharp and $\Delta_{RE}(P', P) \leq \varepsilon_t^{(\sqrt{2}-1)^2}$ is false. However, the inequality $\Delta_{RE}(P', P) \leq O(\varepsilon_t)$ was assumed previously,

2. It is also interesting to note if $st < \frac{\sqrt{a^2+b^2}}{\psi}$, $\Delta_{RE}(P', P)$ can be close to 1.

The facts described in the propositions have not been previously noted. More precise error estimations are given for computations based on practical sampling which will lead to spotting gaps in the existing estimations under sum-like metrics (e.g., KL-divergence, Rényi-divergence) and max-like metrics (e.g., relative difference, max-log distance). Thus the two observations (i.e., propositions 1 and 2) are the keys to transforming theoretical results into practical ones, we will take practical convolution theorem as an example to show how they work in the next section.

4 Refinement of Practical Convolution Theorem and Its Appplication

This section will be divided into two parts to consider practical issues of convolution of discrete Gaussian samplings. In the first part we perform analysis of some existing results by using our observations. The second part devotes to a derivation of convolution theorem with more accurate bounds. We would like to remark that using the two observations of the previous section is the key to determine the dominant term.

4.1 Existing Convolution Theorems Revisit

In this part, we revisit Convolution Theorems reported in [15, 16, 10] and conduct analysis by using our observations (Propositions 1 and 2).

4.1.1 Analysis under KL-divergence

Let us review the analysis under KL-divergence given in [15]. The relationship between Δ_{KL} and Δ_{RE} is proved by

$$\begin{aligned}\Delta_{KL}(P\|Q) &= \sum_{x_i\in S} \ln\left(\frac{P(x_i)}{Q(x_i)}\right)\cdot P(x_i)\\ &= \sum_{x_i\in S} \ln\Big(1-\delta_{RE}(Q(x_i),P(x_i))\Big)\cdot P(x_i)\\ &\le \sum_{x_i\in S}\Big(\delta_{RE}(Q(x_i),P(x_i))+2\delta_{RE}^2(Q(x_i)\|P(x_i))\Big)\cdot P(x_i)\end{aligned}$$

In [15], $\Delta_{KL}(P\|Q) \le 2\Delta_{RE}^2(Q\|P)$ is reached by assuming $\sum_{x\in S}\delta_{RE}(Q(x_i),P(x_i))P(x_i) = 0$ and $\delta_{RE}(Q(x_i),P(x_i)) \le \Delta_{RE}(Q,P)$. When integrating truncation errors and floating-point errors, the following bounds are obtained

$$\Delta_{RE} \le \mu+\varepsilon_t,\ \ \Delta_{KL}(P\|Q) \le 2(\mu+\varepsilon_t)^2 \approx 2\mu^2+2\varepsilon_t^2+O(\mu\varepsilon_t).$$

However, our Proposition 1 as well as our experiment (Table 1 and Figure 2) suggest that $\sum_{x\in S}\delta_{RE}(Q(x_i),P(x_i))P(x_i)$ should not be ignored.

We should have

$$\begin{aligned}\Delta_{KL}(P\|Q) &\leq \sum_{x_i\in S}\Big(\delta_{RE}(Q(x_i),P(x_i))+2\delta_{RE}^2(Q(x_i)\|P(x_i))\Big)\cdot P(x_i)\\ &= \sum_{x_i\in S}P(x_i)-\sum_{x_i\in S}Q(x_i)+\sum_{x_i\in S}2\delta_{RE}^2(Q(x_i)\|P(x_i))\cdot P(x_i)\\ &= O(\mu)+O(\varepsilon_t)+O(\mu^2+\mu\varepsilon_t+\varepsilon_t^2)\end{aligned}$$

This bound is much larger than that of [15], but is consistent with our experiments.

4.1.2 Analysis under Rényi-divergence

When it comes to Rényi-divergence, let us look at the proof in [16] first. The following inequality is obtained for each i by using Taylor bounds, $\frac{P(x_i)^\alpha}{Q(x_i)^{\alpha-1}}\leq P(x_i)+(1-\alpha)\delta_{RE}(P(x_i),Q(x_i))\cdot P(x_i)+\frac{\alpha(\alpha-1)\delta_{RE}^2(P(x_i)\|Q(x_i))}{2(1-\delta_{RE}(P(x_i),Q(x_i)))^{\alpha+1}}\cdot P(x_i)$ Assuming $\sum_{x_i\in S}\delta_{RE}(P(x_i),Q(x_i))\cdot P(x_i)=0$ and $\sum_{x_i\in S}P(x_i)=1$, [16] obtains

$$\begin{aligned}\Delta_{RD_\alpha}(P\|Q) &= \Big(\sum_{x_i\in S}\frac{P(x_i)^\alpha}{Q(x_i)^{\alpha-1}}\Big)^{\frac{1}{\alpha-1}}\leq 1+\Big(\frac{\alpha(\alpha-1)\delta_{RE}^2(P(x_i)\|Q(x_i))}{2(1-\delta_{RE}(P(x_i),Q(x_i)))^{\alpha+1}}\Big)^{\frac{1}{\alpha-1}}\\ &\approx 1+\frac{\alpha\delta_{RE}^2(P(x_i)\|Q(x_i))}{2}\approx 1+\frac{\alpha}{2}\mu^2+\frac{\alpha}{2}\varepsilon_t^2+O(\mu\varepsilon_t)\end{aligned}$$

Similar to the discussion earlier, when we use Proposition 1 to analyze, we get

$$\begin{aligned}\Delta_{RD_\alpha}(P\|Q) &\leq \Big(1+O(\mu)+O(\varepsilon_t)+(1-\alpha)\cdot(O(\mu)+O(\varepsilon_t))+(O(\mu)+O(\varepsilon_t))^2\Big)^{\frac{1}{\alpha-1}}\\ &= 1+O(\mu)+O(\varepsilon_t)+O(\mu\varepsilon_t)\end{aligned}$$

Again, our experiments (Table 1 and Figure 3) support this new bound and indicate that the bound of [16] cannot be always achieved.

4.1.3 Analysis under Max-log Distance

From Lemma 2.1 and Lemma 4.2 of [10], we see that

$$\Delta_{KL}(P\|Q)\leq\frac{\Delta_{RE}^2(P\|Q)}{2(1-\Delta_{RE}(P,Q))^2},\ \Delta_{ML}(P,Q)\approx\Delta_{RE}(P,Q).$$

While according to our result in analysing Δ_{KL}, the relationship between Δ_{KL} and Δ_{RE} becomes invalid in practice and as a result, $\Delta_{ML}(P,Q)$ does not satisfy pythagorean probability preservation defined in [10]. Let P denote the probability distribution obtained by convolution and Q denote the ideal discrete Gaussian distribution, by our second observation (Proposition 2), we have

$$\Delta_{ML}(P,Q)\leq O(\mu)+O(\varepsilon_t^c)+O(\varepsilon)+O(\varepsilon_t^{1+c}+\mu\varepsilon+\varepsilon_t^c\varepsilon+\varepsilon_t^c\mu).$$

where $c \in (0, (\sqrt{2}-1)^2]$. So $O(\varepsilon_t^c)$ is a dominant term which is much bigger than $O(\varepsilon_t)$ as in the previous work [10].

As a practical sampling procedure may need several iterative convolution processes, it will bring a rapid growth of $\Delta_{ML}(P, Q)$ and consequently bring a huge challenge in choosing security parameters.

4.2 Practical Convolution Theorem by Derivation

In this part, we provide a step-by-step derivation of the practical convolution theorem in detail, with a more precise coefficients for the dominant terms. We will use the revised version of Convolution Theorem (Theorem 2.4) for dealing with two random variables, and analyse three types of errors, namely, convolution errors, truncation errors, as well as floating-point errors. The effectiveness of convolutions are evaluated by statistical distance, KL-divergence, Rényi-divergence, relative difference and max-log distance. We will use the tail bound from Lemma 2.2 to control truncation errors.

Recall that for a real number $t > 1$, we use $\varepsilon_t = \rho_{1/t}(\mathbb{Z}) - 1$ to control the truncation error with respect to t. For positive integers a, b be positive and real number s_1, we also defined $\eta = \frac{\sqrt{a^2+b^2}}{s_1}$, $\psi = \frac{\sqrt{a^2+b^2}-a}{b}$ and $\omega = 1 - \frac{\eta}{\psi t}$ in last section.

Now, we state our version of convolution theorem.

Theorem 4.1 *Let $a > b \in \mathbb{Z}$ be nonzero integers with $gcd(a, b) = 1$ and $\mathbf{s} \in \mathbb{R}^2$ with $s_1 = s_2 \geq \sqrt{a^2+b^2}\eta_\varepsilon(\mathbb{Z})$ [3] . Let $x_i \in [-ts_i, ts_i]$ be independent samples from $D_{\mathbb{Z},s_i}$ respectively, with floating-point error $\mu_i \leq \mu$ for $i = 1, 2$. Let $\tilde{D}_{\mathbb{Z},s}$ be the distribution of $x = ax_1 + bx_2 \in S = [-ts, ts]$ where $s = \sqrt{a^2 s_1^2 + b^2 s_2^2}$. Then*

$$
\begin{aligned}
&\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \leq C_1\varepsilon_t + \mu + \varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\\
&\Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \leq C_3\varepsilon_t^{\omega^2\psi^2} + 2\mu + 2\varepsilon + O(\varepsilon_t^{1+\omega^2\psi^2} + \mu\varepsilon + \varepsilon_t^{\psi^2}\varepsilon + \varepsilon_t^{\psi^2}\mu)\\
&\Delta_{ML}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \leq C_3\varepsilon_t^{\omega^2\psi^2} + 2\mu + 2\varepsilon + O(\varepsilon_t^{2\omega^2\psi^2} + \mu^2 + \varepsilon^2 + \mu\varepsilon + \varepsilon_t^{\psi^2}\varepsilon + \varepsilon_t^{\psi^2}\mu)\\
&\Delta_{KL}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \leq (2C_1 + C_4)\varepsilon_t + 2\mu + 2\varepsilon^2 + O(\varepsilon_t^2 + \mu^2 + \varepsilon^3 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\\
&\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \leq 1 + (2C_1 + C_4)\varepsilon_t + 2\mu + \frac{\alpha}{2}\varepsilon^2 + O(\varepsilon_t^2 + \mu^2 + \varepsilon^3 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu).
\end{aligned}
$$

[3] It is note that our discussion can be extended to the case of $s_1 \neq s_2$. We choose $s_1 = s_2$ is for the purpose of simplifying our discussion. This is also a very common setup in practice.

where $C_1 = \frac{1-\frac{1}{2}e^{-\frac{2\pi t}{s_1}}}{s_1} + \frac{1}{2}e^{\frac{-2\pi t}{s_1}}$, $C_3 = \frac{2}{\left(1-e^{-\pi(2\omega\psi\eta t+\eta^2)}\right)\left(1+e^{-2\pi\eta^2}(1+e^{-4\pi\eta^2})\right)}$, *and* $C_4 = \frac{1-\frac{1}{2}e^{-\frac{2\pi t}{s}}}{s} + \frac{1}{2}e^{\frac{-2\pi t}{s}}$. *In particular, when* $t = \eta_\varepsilon(\mathbb{Z})$ *and* $\varepsilon_t = \varepsilon$,

$$
\begin{aligned}
&\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le (C_1+1)\varepsilon + \mu + O(\varepsilon^2 + \varepsilon\mu)\\
&\Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le C_3\varepsilon^{\omega^2\psi^2} + 2\mu + O(\varepsilon + \varepsilon^{\omega^2\psi^2}\mu)\\
&\Delta_{ML}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le C_3\varepsilon^{\omega^2\psi^2} + 2\mu + O(\varepsilon^{2\omega^2\psi^2} + \mu^2 + \varepsilon^{\omega^2\psi^2}\mu)\\
&\Delta_{KL}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s}) \le (2C_1+C_4)\varepsilon + 2\mu + O(\varepsilon^2 + \mu^2 + \varepsilon\mu)\\
&\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s}) \le 1 + (2C_1+C_4)\varepsilon + 2\mu + O(\varepsilon^2 + \mu^2 + \varepsilon\mu).
\end{aligned}
$$

We would like to remark that C_1, C_3, C_4 are considered as constants because the parameters of convolution theorem are selected as $s_1 = s_2 \ge \sqrt{a^2+b^2}\eta_\varepsilon(\mathbb{Z}) \gg 1, t \ge \eta_\varepsilon(\mathbb{Z}) \gg 1$. It is obvious that $C_1, C_4 \in (0,1)$ and $C_1 = O(e^{-2\pi t/s_1}), C_4 = O(e^{-2\pi t/s})$. We also have $\varepsilon_t = O(e^{-2\pi t^2}) \le \varepsilon = O(e^{-2\pi\eta_\varepsilon^2(\mathbb{Z})})$ and $\mu \le 2^{-p}$ ($p \in [53, 200]$), note that $e^{-2\pi t^2} \le e^{-2\pi\eta_\varepsilon^2(\mathbb{Z})} \ll e^{-2\pi t} \ll e^{-2\pi t/s} \le e^{-2\pi t/s_1}$ (i.e when takes $s_1 = 34, t = \eta_\varepsilon(\mathbb{Z}) = 6$, $\varepsilon_t = \varepsilon \approx 2^{-160}$ and $C_1 \ge \varepsilon^{1/(ts_1)} \approx 2^{-0.78}$ and there are similar cases for C_3 and C_4). So C_1, C_3, C_4 can be viewed as constants that do not affect the analysis of $\varepsilon_t, \varepsilon$ and μ.

We would also like to remark that our result is quite different from the existing ones, even compared with the practical result of [10]. It is noted that the relationships between Δ_{ML} and other metrics are discussed in [10], but the influence of truncation error, which acts as a dominant term in computing Δ_{ML}, seems to be ignored.

Our analysis of practical convolution theorem can be divided into three parts by the nature of errors, i.e., convolution errors, floating-point errors and truncation errors. Details of our analysis will be given in the following subsections. Our version of convolution theorem (Theorem 2.4) will be used.

4.3 Error Analysis–Proof of Theorem 4.1

We will denote a distribution with practical errors as $\tilde{D}$ and an ideal distribution as D in this section. We start the analysis by considering two base samplers which samples $x_1 \leftarrow \tilde{D}_{c_1,s_1}$ and $x_2 \leftarrow \tilde{D}_{c_2,s_2}$ respectively. As the practical precision as well as the set of x_1, x_2 can not be infinite, there exists both truncation errors and floating-point errors for base samplers. Without loss of generality, we assume $c = c_1 = c_2 = 0, s_1 = s_2$. The truncation ranges for x_1 and x_2 are denoted by $S_1 = [-ts_1, ts_1]$ and $S_2 = [-ts_2, ts_2]$ respectively. As mentioned earlier, we set $\varepsilon_t = 2\sum_{i=1}^{+\infty} e^{-\pi t^2 i^2}$ to be the truncation error and we know that $\varepsilon_t < 0.086463$ for all $t > 1$. Denote floating-point errors as μ_1, μ_2 with $\mu_1 \le \mu, \mu_2 \le \mu$. We first treat truncation errors

$$Pr_{\tilde{D}_{s_1}}(x = x_1) = \frac{\rho_{s_1}(x_1)}{\sum_{x \in S_1} \rho_{s_1}(x)}$$
$$Pr_{D_{s_1}}(x = x_1) = \frac{\rho_{s_1}(x_1)}{\sum_{x \in \mathbb{Z}} \rho_{s_1}(x)}$$

From Lemma 2.2 and the fact that $\rho_{s_1}(\mathbb{Z}) > s_1$ we get

$$\sum_{\substack{x_1 \in \mathbb{Z} \\ |x_1| \geq t s_1}} \rho_{s_1}(x) \leq 2e^{-\pi t^2}\left(1 + \frac{1}{2}e^{-\frac{2\pi t}{s_1}}(\rho_{s_1}(\mathbb{Z}) - 1)\right) \leq \varepsilon_t \left(1 + \frac{1}{2}e^{-\frac{2\pi t}{s_1}}(\rho_{s_1}(\mathbb{Z}) - 1)\right)$$
$$\leq \varepsilon_t \left(\frac{1 - \frac{1}{2}e^{-\frac{2\pi t}{s_1}}}{s_1} + \frac{1}{2}e^{\frac{-2\pi t}{s_1}}\right)\rho_{s_1}(\mathbb{Z}) = C_1 \cdot \varepsilon_t \rho_{s_1}(\mathbb{Z})$$

where $C_1 = \frac{1 - \frac{1}{2}e^{-\frac{2\pi t}{s_1}}}{s_1} + \frac{1}{2}e^{\frac{-2\pi t}{s_1}}$.

From the fact that $\rho_{s_1}(\mathbb{Z}) < s_1 + \frac{2s_1 e^{-\pi s_1^2}}{1 - e^{-3\pi s_1^2}}$ and $\varepsilon_t \in (2e^{-\pi t^2}, \frac{2e^{-\pi t^2}}{1 - e^{-3\pi t^2}})$, we get

$$\sum_{\substack{x_1 \in \mathbb{Z} \\ |x_1| \geq t s_1}} \rho_{s_1}(x) \geq \frac{2e^{-\pi t^2}}{1 - e^{-\frac{2\pi t}{s_1}}} \geq \varepsilon_t \frac{1 - e^{-3\pi t^2}}{1 - e^{-\frac{2\pi t}{s_1}}}$$
$$\geq \varepsilon_t \left(\frac{\frac{1 - e^{-3\pi t^2}}{1 - e^{-\frac{2\pi t}{s_1}}}}{s_1 + \frac{2s_1 e^{-\pi s_1^2}}{1 - e^{-3\pi s_1^2}}}\right)\rho_{s_1}(\mathbb{Z}) = C_2 \cdot \varepsilon_t \rho_{s_1}(\mathbb{Z})$$

where $C_2 = \frac{\frac{1 - e^{-3\pi t^2}}{1 - e^{-\frac{2\pi t}{s_1}}}}{s_1 + \frac{2s_1 e^{-\pi s_1^2}}{1 - e^{-3\pi s_1^2}}}$.

These yield that for all $x_1 \in S_1$

$$\frac{Pr_{D_{s_1}}(x = x_1)}{1 - C_2\varepsilon_t} \leq Pr_{\tilde{D}_{s_1}}(x = x_1) \leq \frac{Pr_{D_{s_1}}(x = x_1)}{1 - C_1\varepsilon_t}$$

Since the probabilities of base samplers are stored with finite precision p which may introduce relative errors as large as $\mu \leq 2^{-p}$, for a base sampler which samples $x_1 \leftarrow \tilde{D}_{s_1}$ (or $x_2 \leftarrow \tilde{D}_{s_2}$), we have

$$\frac{Pr_{D_{s_1}}(x = x_1)}{1 - C_2\varepsilon_t} \leq [1 - \mu, 1 + \mu] \cdot Pr_{\tilde{D}_{s_1}}(x = x_1) \leq \frac{Pr_{D_{s_1}}(x = x_1)}{1 - C_1\varepsilon_t}$$
$$\frac{Pr_{D_{s_1}}(x = x_1)}{1 - C_2\varepsilon_t + \mu + O(\varepsilon_t\mu)} \leq Pr_{\tilde{D}_{s_1}}(x = x_1) \leq \frac{Pr_{D_{s_1}}(x = x_1)}{1 - C_1\varepsilon_t - \mu + O(\varepsilon_t\mu)} \tag{4}$$

As $C_1 > C_2$, the relative error is bounded by

$$\delta_{RE}(Pr_{\tilde{D}_{s_1}}(x = x_1), Pr_{D_{s_1}}(x = x_1)) \leq C_1\varepsilon_t + \mu + O(\varepsilon_t\mu)$$

Next, let us analyze the joint distribution of the two independent base samplers. Recall that we set $s_1 = s_2$ and $c = c_1 = c_2 = 0$, and $S_1 = [-ts_1, ts_1]$, $S_2 = [-ts_2, ts_2]$, $S = [-ts, ts]$ with $s = \sqrt{a^2 s_1^2 + b^2 s_2^2}$.

The Convolution Theorem (Thoerem 2.5) proves that

$$\delta_{RE}(Pr_{\tilde{D}_{Y,s}}[x = \bar{x}], Pr_{D_{Y,s}}[x = \bar{x}]) \leq \frac{1+\varepsilon}{1-\varepsilon} - 1.$$

It should be noted that Theorem 2.5 applies to the ideal situation where we can obtain all possibilities with neither truncation errors nor floating-point errors, thus for all $x_c \in S = [-ts, ts]$, $Pr_{\tilde{D}_{Y,s}}(x = x_c) = \sum_{\substack{x_1 \in \mathbb{Z}, x_2 \in \mathbb{Z} \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2)$. As a result, we have

$$Pr_{D_s}(x = x_c) = (1 + a(x_c)) \cdot \sum_{\substack{x_1 \in \mathbb{Z}, x_2 \in \mathbb{Z} \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2)$$

where $(1 + a(x_c)) \in [\frac{1-\varepsilon}{1+\varepsilon}, \frac{1+\varepsilon}{1-\varepsilon}]$ for all $x_c \in \mathbb{Z}$. On the other hand, we know the probability of convolution of two base samples is given by

$$Pr_{\tilde{D}_s}(x = x_c) = \sum_{\substack{x_1 \in S_1, x_2 \in S_2 \\ x_c = ax_1 + bx_2}} Pr_{\tilde{D}_{s_1}}(x = x_1) \cdot Pr_{\tilde{D}_{s_2}}(x = x_2)$$

According to the previous analysis about relative error of base samplers, it is clear that for some $C_t \in \left[\frac{1}{(1-C_2\varepsilon_t+\mu)^2}, \frac{1}{(1-C_1\varepsilon_t-\mu)^2}\right]$, where C_1, C_2 as previously defined. We have

$$\begin{aligned}
& Pr_{\tilde{D}_s}(x = x_c) \\
&= \sum_{\substack{x_1 \in S_1, x_2 \in S_2 \\ x_c = ax_1 + bx_2}} Pr_{\tilde{D}_{s_1}}(x = x_1) \cdot Pr_{\tilde{D}_{s_2}}(x = x_2) \\
&= C_t \cdot \sum_{\substack{(x_1, x_2) \in S_1 \times S_2 \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2) \\
&= C_t \cdot \left(\sum_{\substack{(x_1, x_2) \in \mathbb{Z}^2 \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2) - \sum_{\substack{(x_1, x_2) \notin S_1 \times S_2 \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2) \right) \\
&= C_t \cdot (1 - b(x_c)) \cdot \sum_{\substack{(x_1, x_2) \in \mathbb{Z}^2 \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2) \\
&= C_t \cdot \frac{1 - b(x_c)}{1 + a(x_c)} \cdot Pr_{D_s}(x = x_c) \\
&= g(x_c) \cdot Pr_{D_s}(x = x_c)
\end{aligned}$$

where $g(x_c) = C_t \cdot \frac{1-b(x_c)}{1+a(x_c)}$, with $b(x_c) = \frac{\beta(x_c)}{\alpha(x_c)} = \frac{\sum_{\substack{(x_1, x_2) \notin S_1 \times S_2 \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x=x_1) \cdot Pr_{D_{s_2}}(x=x_2)}{\sum_{\substack{(x_1, x_2) \in \mathbb{Z}^2 \\ x_c = ax_1 + bx_2}} Pr_{D_{s_1}}(x=x_1) \cdot Pr_{D_{s_2}}(x=x_2)}$.

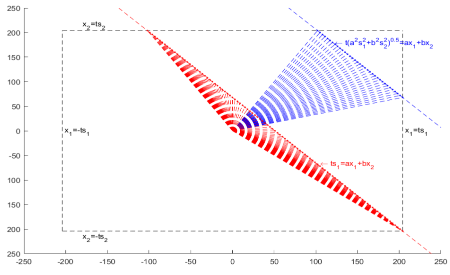

Figure 1. $x_c = ax_1 + bx_2$ in the plane of (x_1, x_2), the Horizontal Axis is for x_1 and the Vertical Axis for x_2

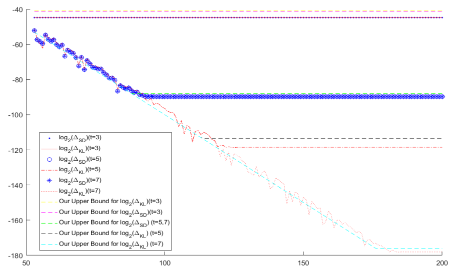

Figure 2. Relationship between Bounds and Practical Errors Measured by Δ_{SD}, Δ_{KL} with Different Precisions, the Horizontal Axis is for Precisions and the Vertical Axis for log_2 of the Errors

Now we shall analyse $b(x_c)$. Given x_c, we denote ℓ_{x_c} the line defined by the equation $x_c = ax_1 + bx_2$ in the (x_1, x_2)-plane. We are concerning with the integral point $(x_1, x_2) \in \mathbb{Z}^2$ on the line ℓ_{x_c}. Note that

$$Pr_{D_{s_1}}(x = x_1) \cdot Pr_{D_{s_2}}(x = x_2) = \frac{(e^{-\pi/s_1^2})^{(x_1^2+x_2^2)}}{(\rho_{s_1}(\mathbb{Z}))^2} = \frac{1}{\rho_{s_1}^2(\mathbb{Z})} e^{-\pi \frac{x_1^2+x_2^2}{s_1^2}}.$$

So we can connect the convolution probabilities with the distances from the origin of the (x_1, x_2)-plane, as it is shown in Fig 1.

Since $\gcd(a, b) = 1$, we may assume $a > b$ without loss of generality. By the extended Euclidean algorithm, there are positive integers $u < b, v < a$ such that

$$au - bv = 1.$$

Let $ST_0 = \{k(b, -a) : k \in \mathbb{Z}\}$ denote the set of integral solutions of $ax_1+bx_2 = 0$. Then the set of integral solutions of $\ell_{x_c} : ax_1 + bx_2 = x_c$ is

$$ST_{x_c} = x_c(u, -v) + ST_0.$$

This means that a point in ST_x is of the form : $(x_c u + kb, -x_c v - ka)$.

The point on ℓ_{x_c} that is closest to the origin is

$$P = (\frac{ax_c}{a^2 + b^2}, \frac{bx_c}{a^2 + b^2}) = (x_c u + \xi b, -x_c v - \xi a)$$

with $\xi = -\frac{ub+va}{a^2+b^2} x_c$. So the two possible shortest vectors in ST_{x_c} are

$$P_0 = (x_c u + \lfloor \xi \rfloor b, -x_c v - \lfloor \xi \rfloor a) \text{ and } P_1 = (x_c u + \lceil \xi \rceil b, -x_c v - \lceil \xi \rceil a).$$

Consider a vector $(x_cu + kb, -x_cv - ka) \in ST_{x_c}$. Its norm relates the norms of P_0, P_1 through the following[4] [5]

$$\|(x_cu + kb, -x_cv - ka)\|^2 = \begin{cases} \|P_0\|^2 + (i^2 - 2i(\xi - \lfloor\xi\rfloor)), & \text{if } i = k - \lfloor\xi\rfloor, \\ \|P_1\|^2 + (i^2 + 2i(\lceil\xi\rceil - \xi)), & \text{if } i = k - \lceil\xi\rceil. \end{cases} \quad (5)$$

The relation (5) will be used in deriving explicit formulas of $\alpha(x_c)$ and $\beta(x_c)$. These formulas enable us to establish a key estimation for the relative convolution error. More precisely, this estimation states

Lemma 4.2 *If* $s_1t \geq \frac{\sqrt{a^2+b^2}}{\psi}$,

$$b(x_c) = \frac{\beta(x_c)}{\alpha(x_c)} \leq C_3 e^{-\pi\omega^2\psi^2t^2}.$$

We include a proof of the lemma in the appendix due to space limitations, which also gives a proof Proposition 2.

According to Lemma 4.2, we see that

$$\begin{aligned} g(x_c) &= C_t \cdot \frac{1 - b(x_c)}{1 + a(x_c)} \\ &\geq \frac{1}{(1 - C_2\varepsilon_t + \mu)^2}(1 - 2\varepsilon)(1 - C_3e^{-\pi\omega^2\psi^2t^2}) \\ &\geq \frac{1}{(1 - C_2\varepsilon_t + \mu)^2}(1 - 2\varepsilon)(1 - C_3\varepsilon_t^{\omega^2\psi^2}) \\ &= 1 - C_3\varepsilon_t^{\omega^2\psi^2} - 2\mu - 2\varepsilon + O(\varepsilon_t^{1+\omega^2\psi^2}) + O(\varepsilon_t^{\omega^2\psi^2}\mu) + O(\varepsilon_t^{\omega^2\psi^2}\varepsilon) + O(\mu\varepsilon) \end{aligned}$$

To analysis Δ_{RE}, we have

$$\begin{aligned} &\Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \\ &= \max_{x\in S} \delta_{RE}(Pr_{\tilde{D}_s}(x), Pr_{D_s}(x)) \\ &= \max_{x\in S} \frac{|Pr_{\tilde{D}_s}(x) - Pr_{D_s}(x)|}{Pr_{\tilde{D}_s}(x)} \\ &= \max_{x\in S} |g(x) - 1| \\ &\leq C_3\varepsilon_t^{\omega^2\psi^2} + 2\mu + 2\varepsilon + O(\varepsilon_t^{1+\omega^2\psi^2}) + O(\varepsilon_t^{\omega^2\psi^2}\mu) + O(\varepsilon_t^{\omega^2\psi^2}\varepsilon) + O(\mu\varepsilon). \end{aligned}$$

---

[4] Here we just verify the second relation of (5), and the other is similar. $\|(x_cu + kb, -x_cv - ka)\|^2 - \|P_1\|^2 = (k - \lceil\xi\rceil)(2x_cub + 2x_cva) + (k^2 - \lceil\xi\rceil^2)(a^2 + b^2) = (a^2+b^2)(k-\lceil\xi\rceil)\left(2\frac{ub+va}{a^2+b^2}x_c+k+\lceil\xi\rceil\right) = (a^2+b^2)i\big(-2\xi+i+2\lceil\xi\rceil\big) = (i^2+2i(\lceil\xi\rceil-\xi))$.

[5] It should be noted that the result of (5) is obtained under the condition $s_1 = s_2$, for the case when $s_1 \neq s_2$, a similar result can also be derived with a small difference as $\xi = -\frac{us_2^2b+vs_1^2a}{s_1^2a^2+s_2^2b^2}x_c$.

And from lemma 2.1, we also have

$$\left|\Delta_{ML}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) - \Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s})\right| \leq \frac{\Delta_{RE}^2(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s})}{2(1-\Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}))}.$$

So

$$\begin{aligned}\Delta_{ML}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) &\leq \Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \cdot \left(1 + \frac{1}{2}\Delta_{RE}^2(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) + \Delta_{RE}^3(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s})\right) \\ &= C_3\varepsilon_t^{\omega^2\psi^2} + 2\mu + 2\varepsilon + O(\varepsilon_t^{2\omega^2\psi^2} + \mu^2 + \varepsilon^2 + \mu\varepsilon + \varepsilon_t^{\psi^2}\varepsilon + \varepsilon_t^{\psi^2}\mu).\end{aligned}$$

It is seen that the truncations in the base samplers bring an extra error for the joint distribution after convolution. More specifically, the extra error is negligible when x is close to the center, but it acts as the dominant term when x is close to the edges. This error has a profound effect in computing max-like divergences, such as Δ_{ML} and Δ_{RE}, however, when considering sum-like divergences, such as Δ_{SD}, Δ_{KL} and Δ_{RD}, it contributes little because the corresponding probability is very small. So we use a general bound $Pr_{\tilde{D}_s}(x) \leq \left(1 + 2C_1\varepsilon_t + 2\mu + 2\varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\right) \cdot Pr_{D_s}(x)$ (obtained by ignoring $b(x_c)$) to make following analysis about Δ_{SD}, Δ_{KL}

$$\begin{aligned}\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) &= \frac{1}{2}\sum_{x\in S}|Pr_{\tilde{D}_s}(x) - Pr_{D_s}(x)| \\ &\leq \frac{1}{2}\cdot\left(2C_1\varepsilon_t + 2\mu + 2\varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\right)\sum_{x\in S} Pr_{D_s}(x) \\ &\leq \frac{1}{2}\cdot\left(2C_1\varepsilon_t + 2\mu + 2\varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\right) \\ &= C_1\varepsilon_t + \mu + \varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\end{aligned}$$

For Δ_{KL}, we let $Pr_{\tilde{D}_s}(x) = (1 + c(x))Pr_{D_s}(x)$ with $|c(x)| \leq 2C_1\varepsilon_t + 2\mu + 2\varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)$, we have

$$\begin{aligned}\Delta_{KL}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s}) &= \sum_{x\in S}\ln\left(\frac{Pr_{\tilde{D}_s}(x)}{Pr_{D_s}(x)}\right)\cdot Pr_{\tilde{D}_s}(x) \\ &= \sum_{x\in S}\ln\left(1 + c(x)\right)\cdot(1 + c(x))Pr_{D_s}(x) \\ &= \sum_{x\in S}\left(c(x) + \frac{1}{2}c^2(x) + O(c^3(x))\right)\cdot Pr_{D_s}(x) \\ &\leq \sum_{x\in S}c(x)Pr_{D_s}(x) + \frac{1}{2}\left(2C_1\varepsilon_t + 2\mu + 2\varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)\right)^2\sum_{x\in S}Pr_{D_s}(x) \\ &\quad + O\left((2C_1\varepsilon_t + 2\mu + 2\varepsilon)^3\right)\end{aligned}$$

It is also noted that, according to Lemma 2.2

$$\sum_{x\in S}Pr_{D_s}(x) = \sum_{x\in\mathbb{Z}}Pr_{D_s}(x) - \sum_{x\notin S}Pr_{D_s}(x) \geq 1 - \varepsilon_t\cdot\frac{1+\varepsilon}{1-\varepsilon}\left(\frac{1 + \frac{e^{-\frac{2\pi t}{s}}}{2}(\rho_s(\mathbb{Z}) - 1)}{\rho_s(\mathbb{Z})}\right)$$

According to an early analysis of Equation (4), $\sum_{x_1\in S_1} Pr_{\tilde{D}_{s_1}}(x_1) \le 1 + C_1\varepsilon_t + \mu$ (and $\sum_{x_2\in S_2} Pr_{\tilde{D}_{s_2}}(x_2) \le 1 + C_1\varepsilon_t + \mu$), we see that

$$\begin{aligned}\sum_{x\in S} Pr_{\tilde{D}_s}(x) &= \sum_{x\in S}\Big(1+c(x)\Big)Pr_{D_s}(x)\\ &= \sum_{x\in S} Pr_{D_s}(x) + \sum_{x\in S} c(x)Pr_{D_s}(x)\end{aligned}$$

From Proposition Proposition 1, we get

$$\begin{aligned}\sum_{x\in S} Pr_{\tilde{D}_s}(x) &= \sum_{x_1\in S_1, x_2\in S_2} Pr_{\tilde{D}_{s_1}}(x_1)\cdot Pr_{\tilde{D}_{s_2}}(x_2) - \sum_{\substack{x_1\in S_1, x_2\in S_2\\ x=ax_1+bx_2\ \notin S}} Pr_{\tilde{D}_{s_1}}(x_1)\cdot Pr_{\tilde{D}_{s_2}}(x_2)\\ &\le \sum_{x_1\in S_1, x_2\in S_2} Pr_{\tilde{D}_{s_1}}(x_1)\cdot Pr_{\tilde{D}_{s_2}}(x_2)\\ &\le 1 + 2\mu + 2C_1\varepsilon_t + O(\mu^2) + O(\varepsilon_t^2) + O(\mu\varepsilon_t).\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{x\in S} c(x)Pr_{D_s}(x) &= \sum_{x\in S} Pr_{\tilde{D}_s}(x) - \sum_{x\in S} Pr_{D_s}(x)\\ &\le 2\mu + 2C_1\varepsilon_t + \varepsilon_t\cdot\frac{1+\varepsilon}{1-\varepsilon}\left(\frac{1+\frac{e^{-\frac{2\pi t}{s}}}{2}(\rho_s(\mathbb{Z})-1)}{\rho_s(\mathbb{Z})}\right) + O(\mu^2+\varepsilon_t^2+\mu\varepsilon_t)\\ &\le (2C_1+C_4)\varepsilon_t + 2\mu + O(\mu^2+\varepsilon_t^2+\mu\varepsilon_t+\varepsilon_t\varepsilon).\end{aligned}$$

where $C_4 = \frac{1-\frac{1}{2}e^{-\frac{2\pi t}{s}}}{s} + \frac{1}{2}e^{\frac{-2\pi t}{s}}$.

This yields

$$\begin{aligned}&\Delta_{KL}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s})\\ &\le \sum_{x\in S} c(x)Pr_{D_s}(x) + \frac{1}{2}(2C_1\varepsilon_t+2\mu+2\varepsilon)^2\sum_{x\in S} Pr_{D_s}(x) + O\left((2C_1\varepsilon_t+2\mu+2\varepsilon)^3\right)\\ &\le (2C_1+C_4)\varepsilon_t + 2\mu + 2\varepsilon^2 + O(\varepsilon_t^2+\mu^2+\varepsilon^3+\mu\varepsilon+\varepsilon_t\varepsilon+\varepsilon_t\mu).\end{aligned}$$

Now we analyse Δ_{RD}. Let $Pr_{\tilde{D}_s}(x) = (1+c(x))Pr_{D_s}(x)$ where $|c(x)| \le 2C_1\varepsilon_t + 2\mu + 2\varepsilon + O(\varepsilon_t^2 + \mu\varepsilon + \varepsilon_t\varepsilon + \varepsilon_t\mu)$. By the Taylor bound give in [16], we have

$$\sum_{x\in S}\frac{Pr_{\tilde{D}_s}(x)^\alpha}{Pr_{D_s}(x)^{\alpha-1}} \le \sum_{x\in S}\left((1+c(x))Pr_{D_s}(x) + (1-a)c(x)Pr_{D_s}(x) + \frac{\alpha(\alpha-1)c^2(x)}{2(1-c(x))^{\alpha+1}}\cdot Pr_{\tilde{D}_s}(x)\right)$$

As $\sum_{x\in S} c(x)Pr_{D_s}(x) \le (2C_1+C_4)\varepsilon_t + 2\mu + O(\mu^2+\varepsilon_t^2+\mu\varepsilon_t+\varepsilon_t\varepsilon)$, we get

$$\sum_{x\in S}\frac{Pr_{\tilde{D}_s}(x)^\alpha}{Pr_{D_s}(x)^{\alpha-1}} \le 1 + (2C_1+C_4)(\alpha-1)\varepsilon_t + 2(\alpha-1)\mu + \frac{\alpha(\alpha-1)}{2}\varepsilon^2 + O(\varepsilon_t^2+\mu^2+\varepsilon^3+\mu\varepsilon+\varepsilon_t\varepsilon+\varepsilon_t\mu).$$

and hence

$$\begin{aligned}\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s}) &= \Big(\sum_{x\in S}\frac{Pr_{\tilde{D}_s}(x)^\alpha}{Pr_{D_s}(x)^{\alpha-1}}\Big)^{\frac{1}{\alpha-1}}\\ &\le 1 + (2C_1+C_4)\varepsilon_t + 2\mu + \frac{\alpha}{2}\varepsilon^2 + O(\varepsilon_t^2+\mu^2+\varepsilon^3+\mu\varepsilon+\varepsilon_t\varepsilon+\varepsilon_t\mu).\end{aligned}$$

4.4 Experiment Results

In this subsection, we describe our experiments about the practical errors of convolution discrete Gaussian sampling, followed by an analysis about experiments results.

4.4.1 Convolution Errors, Truncation Errors and Floating-point Errors

Our first experiment is to exhibit the influences of convolution errors, truncation errors and floating-point errors respectively. More specifically, we choose $s_1 = s_2$ and compute the probability distributions for $x_1 \leftarrow D_{\mathbb{Z},s_1}$ and $x_2 \leftarrow D_{\mathbb{Z},s_2}$ under different precisions where $x_1 \in [-ts_1, ts_1], x_2 \in [-ts_2, ts_2]$. Then we compute the probability distribution of the variable $\tilde{x} = ax_1 + bx_2$, denoted as $\tilde{D}_{\mathbb{Z},s=\sqrt{a^2s_1^2+b^2s_2^2}}$, and compare it with a pre-computed and much more accurate probability distribution for $x \leftarrow D_{\mathbb{Z},s=\sqrt{a^2s_1^2+b^2s_2^2}}$ (i.e the probability distribution is computed with a much larger precision and t) to get a result of output errors.

The detailed parameters are selected as $s_1 = s_2 = 19.53\sqrt{2\pi}$, $a = 11, b = 1$, $s = \sqrt{a^2s_1^2 + b^2s_2^2}$, $x_1 \in [-ts_1, ts_1], x_2 \in [-ts_2, ts_2]$, the experiment is conducted with t varying from 3 to 8 and precision varying from 53 to 200. For the contrast probability distribution, the precision is selected as 500 and $t = 10$ which make truncation errors and floating-point errors as small as possible. An overview result is displayed in Table 1[6], and we will make further analysis for Δ_{SD} and Δ_{KL}.

We have obtained bounds $\Delta_{SD} \le C_1\varepsilon_t+\mu+\varepsilon$, $\Delta_{KL} \le (2C_1+C_4)\varepsilon_t+2\mu+2\varepsilon^2$ and $\Delta_{RD_\alpha} \le 1 + (2C_1 + C_4)\varepsilon_t + 2\mu + \frac{\alpha}{2}\varepsilon^2$ in Section 4, where $\mu \le 2^{-p}$. As the bound ε is determined by the relation $s_1 = s_2 \ge \sqrt{a^2 + b^2}\eta_\varepsilon(\mathbb{Z})$ according to Theorem 4.1, we have

$$\eta_\varepsilon(\mathbb{Z}) \approx \frac{19.53\sqrt{2\pi}}{\sqrt{11^2+1^2}} \Rightarrow \varepsilon \le 2^{-88.02}$$
$$\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le C_1\varepsilon_t + \mu + 2^{-88.02}$$
$$\Delta_{KL}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s}) \le (2C_1 + C_4)\varepsilon_t + 2\mu + 2^{-175.05}$$
$$\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s}) \le (2C_1 + C_4)\varepsilon_t + 2\mu + \alpha \cdot 2^{-177.05}.$$

[6] Due to the space limit, Table 1 only lists about 0.3% of the total results, to obtain the complete results, one can access the public codes of our experiments from https://github.com/zhengzx/Gsample or run a program by oneself. It also should be noted that Δ_{KL} and Δ_{RE} are not symmetric metrics, and different input orders lead to different results, however, the difference is quite small, i.e. $\left|\log_2(\frac{\Delta_{KL}(D_{\mathbb{Z},s}\|\tilde{D}_{\mathbb{Z},s})}{\Delta_{KL}(\tilde{D}_{\mathbb{Z},s}\|D_{\mathbb{Z},s})})\right| \le O(1)$.

Table 1. Experiment1: Practical Errors with Different Precisions and t

t	$log_2(\varepsilon_t)$	Precisions	$log_2(\mu)$	$log_2(\Delta_{SD})$	$log_2(\Delta_{KL})$	$log_2(\Delta_{RD_2})$	$log_2(\Delta_{RD_{200}})$
3	-39.79	53	-54	-44.74	-44.70	-45.68	-44.43
		73	-74	-44.74	-44.71	-45.74	-44.44
		93	-94	-44.74	-44.71	-45.74	-44.44
		113	-114	-44.74	-44.71	-45.74	-44.44
		133	-134	-44.74	-44.71	-45.74	-44.44
		153	-154	-44.74	-44.71	-45.74	-44.44
		173	-174	-44.74	-44.71	-45.74	-44.44
		193	-194	-44.74	-44.71	-45.74	-44.44
5	-112.31	53	-54	-52.11	-51.11	-50.11	-51.10
		73	-74	-70.97	-69.97	-68.96	-69.95
		93	-94	-89.69	-91.11	-90.11	-91.10
		113	-114	-89.69	-111.36	-110.36	-111.3
		133	-134	-89.69	-118.49	-120.83	-118.07
		153	-154	-89.69	-118.49	-120.83	-118.07
		173	-174	-89.69	-118.49	-120.83	-118.07
		193	-194	-89.69	-118.49	-120.83	-118.07
7	-221.09	53	-54	-52.11	-51.11	-50.11	-51.10
		73	-74	-70.97	-69.97	-68.96	-69.95
		93	-94	-89.69	-91.11	-90.11	-91.10
		113	-114	-89.69	-111.41	-110.41	-111.40
		133	-134	-89.69	-129.41	-128.40	-129.40
		153	-154	-89.69	-149.16	-148.15	-149.15
		173	-174	-89.69	-171.41	-170.41	-169.84
		193	-194	-89.69	-178.09	-177.08	-170.44

Take $t = 3, 5, 7$ as examples with precisions vary from 53 to 200, then
When $t = 3$, $\varepsilon_t \le 2^{-39.79}, C_1\varepsilon_t \le 2^{-41.29}, (2C_1 + C_4)\varepsilon_t \le 2^{-39.38}$, we have

$$\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le 2^{-41.29}$$
$$\Delta_{KL}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \le 2^{-39.38}$$
$$\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \le 2^{-39.38}.$$

When $t = 5$, $\varepsilon_t \le 2^{-112.31}, C_1\varepsilon_t \le 2^{-114.16}, (2C_1 + C_4)\varepsilon_t \le 2^{-112.25}$, we have

$$\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le 2^{-88.02} + \mu$$
$$\Delta_{KL}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \le 2^{-112.25} + 2\mu$$
$$\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \le 2^{-112.25} + 2\mu.$$

And when $t = 7$, $\varepsilon_t \le 2^{-221.09}, C_1\varepsilon_t \le 2^{-223.28}, (2C_1 + C_4)\varepsilon_t \le 2^{-221.37}$, we have

$$\Delta_{SD}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \le 2^{-88.02} + \mu$$
$$\Delta_{KL}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \le 2^{-175.05} + 2\mu$$
$$\Delta_{RD_\alpha}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \le \alpha \cdot 2^{-177.05} + 2\mu.$$

From Fig 2 and 3, we find our theoretical bounds for Δ_{SD}, Δ_{RD} and Δ_{KL} fit well with practical results.

As for Δ_{RE} and Δ_{ML}, we select following parameters to conduct experiments: $s_1 = s_2 = 34$, $a = 4, b = 3$, $s = \sqrt{a^2 s_1^2 + b^2 s_2^2}$, $x_1 \in [-ts_1, ts_1], x_2 \in [-ts_2, ts_2]$,

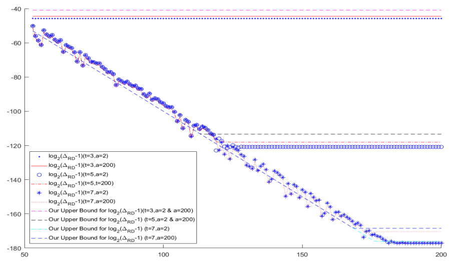

Figure 3. Relationship between Bounds and Practical Errors Measured by Δ_{RD_2} and $\Delta_{RD_{200}}$ with Different Precisions, the Horizontal Axis is for Precisions and the Vertical Axis for $log_2(|\Delta_{RD_\alpha} - 1|)$

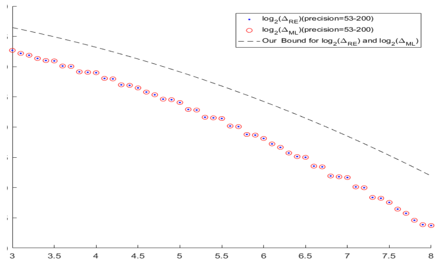

Figure 4. Relationship between theretical bounds and practical errors measured by Δ_{RE}, Δ_{ML} with different t. For each case, the result remains unchanged with precisions varying from 53 to 200, the Horizontal Axis is for t and the Vertical Axis for log_2 of the Errors

with t varying from 3 to 8 and precisions varying from 53 to 200. For the contrast probability distribution, the precision is selected as 500 and $t = 10$ which make truncation errors and floating-point errors as small as possible. An overview of the result is shown in Table 2 and the details can be found in Fig 4. Our analysis of Δ_{RE} and Δ_{ML} gives

$$\Delta_{ML}(D_{\mathbb{Z},s}, \tilde{D}_{\mathbb{Z},s}) \approx \Delta_{RE}(D_{\mathbb{Z},s}, \tilde{D}_{\mathbb{Z},s}) \leq C_3 \varepsilon_t^{\omega^2 \psi^2} + 2\mu + 2\varepsilon$$

where $\psi = (\sqrt{4^2 + 3^2} - 4)/3 \approx 0.3333, \omega \approx 0.9265, C_3 \approx 0.9466.$
As $C_3 \varepsilon_t^{\omega^2 \psi^2} \gg \max(2\mu, 2\varepsilon)$, our estimation indicates that the practical errors may not change when the precisions varies from 53 to 200 which seems to be well supported by the experiment.

Table 2. Experiment2: Practical Errors with Different Precisions and t

t	$log_2(\varepsilon_t)$	Precisions	$log_2(\Delta_{RE})$	$log_2(\Delta_{ML})$
3.0	-39.79	53-200	-7.30	-7.30
4.0	-71.52	53-200	-11.01	-11.01
5.0	-112.31	53-200	-15.93	-15.93
6.0	-162.17	53-200	-21.88	-21.88
7.0	-221.09	53-200	-28.33	-28.33
8.0	-289.07	53-200	-36.27	-36.27

4.5 Application of Our Practical Bound

We have established a new bound for convolution discrete Gaussian sampling which fits well with experiments. In this section, we use this bound to reanalyse

Gaussian sampling schemes using convolution discrete Gaussian theorem, such as the sampling schemes of Pöppelmann, Ducas and Güneysu and of Micciancio and Walter, and try to give suggested modified parameters based on our analysis.

4.5.1 Revisit the Sampling Scheme of Pöppelmann, Ducas and Güneysu

In Pöppelmann, Ducas and Güneysu's sampling scheme [15], the parameters are selected as: $s_1 = s_2 \approx 19.53 \cdot \sqrt{2\pi}, a = 11, b = 1, \eta_\varepsilon(\mathbb{Z}) \leq 3.860$, $t \approx 5.35$ to ensure $\varepsilon_t \leq 2^{-128}$ and precision is set to 72 with $\mu \leq 2^{-71}$ (in [15], a technique was used to store probabilities with different precisions vary from 16 to 72, here we take the fixed precision of 72 which leads to a smaller practical errors than using the original setting in [15]), the goal of the design in [15] was to ensure that

$$\Delta_{KL} \leq 2^{-128}. \tag{6}$$

Now let us analyse if the scheme can achieve this goal, the convolution theorem demands for $s_1 = s_2 \geq \sqrt{a^2 + b^2}\eta_\varepsilon(\mathbb{Z})$, ε is bounded by

$$\eta_\varepsilon(\mathbb{Z}) \approx \frac{19.53\sqrt{2\pi}}{\sqrt{11^2 + 1^2}} \Rightarrow \varepsilon \leq 2^{-88.02}.$$

This gives

$$\Delta_{KL}(\tilde{D}_{\mathbb{Z},s} \| D_{\mathbb{Z},s}) \leq (2C_1 + C_4)\varepsilon_t + 2\mu + 2\varepsilon^2 \approx 2^{-70}.$$

Our bound is much bigger than the requirement in (6). However, the experiment results shown in Fig 5 indicate that the error fits well with our estimation. To make the convolution of sampling to satisfy equation (6), one possible way with a minor modification is to set precision larger than 128, which can be naturally obtained from observation 1.

To validate this analysis, we also conduct experiments with the suggested modification. The experimental results, as shown in Fig 4, are consist with our estimations.

4.5.2 Revisit the Sampling Scheme of Micciancio and Walter

In Micciancio and Walter's sampling scheme, the parameters are selected iteratively, takes the example in [10].

The parameters of the first round sampling, known as the base sampling, are selected as: $s_1 = s_2 = 34, a = \lfloor \frac{s_1}{\sqrt{2}\eta_{\varepsilon(\mathbb{Z})}} \rfloor = 4, b = \max(1, a-1) = 3, t = \eta_\varepsilon(\mathbb{Z}) = 6$, $\varepsilon_t = \varepsilon \leq 2^{-160}$ with precision set to 60 bits, and the goal of the design in [10] was to ensure that

$$\Delta_{ML} \leq 2^{-55}. \tag{7}$$

However, based on our analysis

$$\Delta_{ML}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \approx \Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \leq C_3\varepsilon_t^{\omega^2\psi^2} + 2\mu + 2\varepsilon,$$

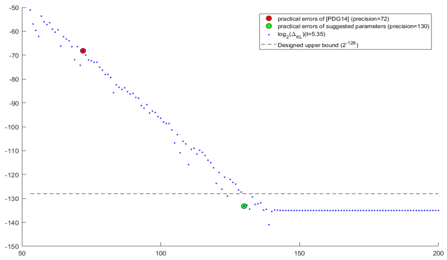

Figure 5. Δ_{KL} of Revised Parameters for [15]'s Scheme with Different Precisions, the Horizontal Axis is for Precisions and the Vertical Axis for log_2 of the Errors

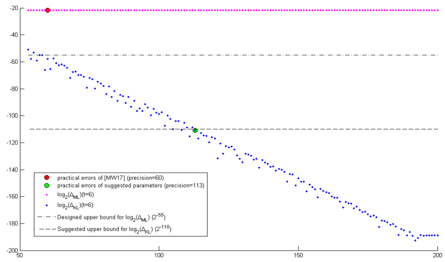

Figure 6. Δ_{ML} and Δ_{KL} of Revised Parameters for [10]'s Scheme with Different Precisions, the Horizontal Axis is for Precisions and the Vertical Axis for log_2 of the Errors

where $\omega \approx 0.9265, \psi = (\sqrt{4^2+3^2}-4)/3 \approx 0.3333$, $C_3 \approx 0.0632$. Thus we have the following bound

$$\Delta_{ML}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \approx \Delta_{RE}(\tilde{D}_{\mathbb{Z},s}, D_{\mathbb{Z},s}) \leq 2^{-19.45}.$$

Our experimental results shown in Fig 6 seem to be consistent with our analysis. To make Δ_{ML} satisfying the original goal of (7), one should ask $\varepsilon_t \leq 2^{-538}$ according to our result. This would require t to be very large (i.e., $t \geq 10.90$) and also a large corresponding precision. Besides, as mentioned above, the convolution process in a practical sampling may be used several times, which will lead to a huge error for Δ_{ML} according to the Proposition 2. Moreover, as we have proved before, the relationship between Δ_{KL} and Δ_{ML} becomes invalid in practice thus ensuring $\Delta_{ML} \leq 2^{-55}$ generally means about 2^{55} security level. As a result, we suggest to set $\Delta_{KL} \leq 2^{-110}$ instead of $\Delta_{ML} \leq 2^{-55}$ to be the goal and set the precision larger than 110.

To validate our suggested parameters, experiments are conducted and the results support our analysis. See Fig 5.

5 Conclusion

In this paper, we make two critical observations about practical errors and take the practical error estimation for convolution theorem with respect to discrete Gaussian sampling as an example to show how to use these observations to more precisely determine the dominate term of practical errors. We start our analysis with a rather brief description and then approach the problems with a detailed and accurate analysis. where extensive experiments have been conducted and the results highly agree with our derived bound. In addition,, we revisit two previous practical convolution based sampling schemes under our new error estimation. It is observed that with the parameters originally proposed, their preset goals may not be achievable due to the absence of truncation error. Our bound suggests modified sets of parameters that ensure their goals to be met. Our experiments results also support the correctness of new sets of parameters. The results in this paper also answers a widely studied open question about achieving much higher security level with much lower precision.

Moreover, it is believed that our observations will have impacts on related areas, because estimating the boundary of probabilities is a very common way to analyse security level. However, the observations show theoretical analysis may not be consist with what really happen in practice. The successful application of the observations indicates that a large number of previous results that are used as theoretical basis in practical schemes may need to be revised.

Acknowledgment

The authors would like to thank Dr.s Daniele Micciancio and Michael Walter for their help in explaining the idea of [10].

References

1. Banaszczyk W. New bounds in some transference theorems in the geometry of numbers. Mathematische Annalen, 296(4):625-635, 1993.
2. Banaszczyk W. Inequalites for convex bodies and polar reciprocal lattices in Rn. Discrete & Computational Geometry, 13:217-231, 1995.
3. Bai S, Langlois A, Lepoint T, et al. Improved Security Proofs in Lattice-Based Cryptography: Using the Rényi Divergence Rather Than the Statistical Distance[C]//International Conference on the Theory and Application of Cryptology and Information Security. Springer, Berlin, Heidelberg, 2015: 3-24.
4. Bai S, Lepoint T, Langlois A, et al. Improved security proofs in lattice-based cryptography: using the Rényi divergence rather than the statistical distance[J]. Journal of Cryptology, 2018, 31(2): 610-640.
5. Aharonov D, Regev O. A lattice problem in quantum NP. In FOCS, pages 210-219, 2003.
6. Aharonov D, Regev O. Lattice problems in NP / coNP. J. ACM, 52(5):749-765, 2005.
7. László Babai. On Lovász' lattice reduction and the nearest lattice point problem. Combinatorica, 6(1):1-13, 1986.
8. Micciancio D, Regev O. Worst-case to average-case reductions based on Gaussian measures. SIAM J. Comput., 37(1):267-302, 2007. Preliminary version in FOCS 2004.
9. Micciancio D, Peikert C. Hardness of SIS and LWE with small parameters[M]//Advances in Cryptology-CRYPTO 2013. Springer, Berlin, Heidelberg, 2013: 21-39.

10. Micciancio D, Walter M. Gaussian Sampling over the Integers: Efficient, Generic, Constant-Time[J]. In proc. CRYPTO 2017, pages 455-485, 2017.
11. Gentry C, Peikert C, Vaikuntanathan V. How to Use a Short Basis: Trapdoors for Hard Lattices and New Cryptographic Constructions, 2008[J].
12. Gür K D, Polyakov Y, Rohloff K, et al. Implementation and Evaluation of Improved Gaussian Sampling for Lattice Trapdoors[J]. IACR Cryptology ePrint Archive 2017: 285 (2017).
13. Peikert C. Limits on the hardness of lattice problems in l_p norms. In IEEE Conference on Computational Complexity, 17(2): 300-351, 2008.
14. Peikert C. An efficient and parallel Gaussian sampler for lattices[C]//Annual Cryptology Conference. Springer, Berlin, Heidelberg, 2010: 80-97.
15. Pöppelmann T, Ducas L, Güneysu T. Enhanced lattice-based signatures on reconfigurable hardware[C]//International Workshop on Cryptographic Hardware and Embedded Systems. Springer, Berlin, Heidelberg, 2014: 353-370.
16. Prest T. Sharper bounds in lattice-based cryptography using the Rényi divergence[C]//In proc. ASIACRYPTO 2017, pages 347-374, 2017.

Appendix I: Proof of Lemma 2.2

Note that

$$\begin{aligned}\sum_{\substack{k\in\mathbb{Z}\\|k-c|\geq ts}} \rho_s(k-c) &= e^{-\pi t^2} \sum_{\substack{k\in\mathbb{Z}\\|k-c|\geq ts}} e^{-\pi\frac{(k-c)^2-s^2t^2}{s^2}} \\ &= e^{-\pi t^2} \sum_{k\geq c+ts} e^{-\frac{\pi}{s^2}(|k-c|-ts)(|k-c|+ts)} \\ &\quad + e^{-\pi t^2} \sum_{k\leq c-ts} e^{-\frac{\pi}{s^2}(|k-c|-ts)(|k-c|+ts)}.\end{aligned}$$

Since

$$\begin{aligned}\sum_{k\geq c+ts} e^{-\frac{\pi}{s^2}(|k-c|-ts)(|k-c|+ts)} &= \sum_{k\geq\lceil c+ts\rceil} e^{-\frac{\pi}{s^2}(k-(c+ts))^2} e^{-\frac{2\pi}{s}(k-(c+ts))t} \\ &\leq 1 + e^{-\frac{2\pi t}{s}} \sum_{k=\lceil c+ts\rceil+1}^{\infty} e^{-\frac{\pi}{s^2}(k-(c+ts))^2} \\ &\leq 1 + e^{-\frac{2\pi t}{s}} \sum_{k=1}^{\infty} e^{-\pi\frac{k^2}{s^2}},\end{aligned}$$

and

$$\begin{aligned}\sum_{k\leq c-ts} e^{-\frac{\pi}{s^2}(|k-c|-ts)(|k-c|+ts)} &\leq \sum_{k\leq\lfloor c-ts\rfloor} e^{-\frac{\pi}{s^2}(k-(c-ts))^2} e^{-\frac{2\pi}{s}|k-(c-ts)|t} \\ &\leq 1 + e^{-\frac{2\pi t}{s}} \sum_{k=\lfloor c-ts\rfloor-1}^{-\infty} e^{-\frac{\pi}{s^2}(k-(c-ts))^2} \\ &\leq 1 + e^{-\frac{2\pi t}{s}} \sum_{k=-1}^{-\infty} e^{-\pi\frac{k^2}{s^2}}.\end{aligned}$$

So we get an improved Banaszczyk bound

$$\sum_{|k-c|\geq ts} \rho_s(k-c) \leq 2e^{-\pi t^2}\left(1+\frac{e^{-\frac{2\pi t}{s}}}{2}(\rho_s(\mathbb{Z})-1)\right).$$

□

Appendix II: Proof of Theorem 2.4

We just include the modification part here. Readers are referred to the proof Theorem 3.2 of [9] for necessary notations.

Proof. Our goal is to show that the result holds for a larger scope of s_i where $s_i \geq \sqrt{z_{max}^2+z_{min}^2}\eta_\varepsilon(\mathbb{Z})$.

When bounding the smoothing parameter of L in [9]

$$\eta(L) \leq \eta((S')^{-1}\cdot(Z\otimes \Lambda)) \leq \eta_\varepsilon(\mathbb{Z})\cdot \tilde{bl}(Z)/\min(s_i)$$

where $Z=\mathbb{Z}^m \cap ker(\mathbf{z}^T)=\{\mathbf{v}\in\mathbb{Z}^m : \langle \mathbf{z},\mathbf{v}\rangle=0\}$ and $\tilde{bl}(\Lambda)$ represents the Gram-Schmidt minimum of a lattice Λ where $\tilde{bl}(\Lambda)=\min_{\mathbf{B}}\|\tilde{\mathbf{B}}\|$, $\|\tilde{\mathbf{B}}\|=\max_i\|\tilde{\mathbf{b}_i}\|$ and the minimum is taken over all bases $\mathbf{B}$ of Λ.

Micciancio and Peikert bound $\tilde{bl}(Z)\leq \min(\|\mathbf{z}\|,\sqrt{2}\|\mathbf{z}\|_\infty)$ because there exist a full-rank set of vectors $z_i\cdot\mathbf{e_j}-z_j\cdot\mathbf{e_i}\in Z$ where z_i has the minimal $|z_i|\neq 0$ and $j\neq i\in[1,..,m]$. Among this set of vectors, we have $\max_i\|\tilde{\mathbf{b}_i}\|=\sqrt{z_{max}^2+z_{min}^2}$ where $\sqrt{z_{max}^2+z_{min}^2}\leq\|\mathbf{z}\|$ when $m=2$ it takes equality and $\sqrt{z_{max}^2+z_{min}^2}\leq \sqrt{2}\|\mathbf{z}\|_\infty$ when $z_{max}=z_{min}$ it takes equality.

And by bounding $\tilde{bl}(Z)\leq\sqrt{z_{max}^2+z_{min}^2}$, we have $\eta(L)\leq\eta_\varepsilon(\mathbb{Z})\cdot\tilde{bl}(Z)/\min(s_i)\leq \eta_\varepsilon(\mathbb{Z})\cdot\sqrt{z_{max}^2+z_{min}^2}/\min(s_i)$. And for $s_i\geq\sqrt{z_{max}^2+z_{min}^2}\eta_\varepsilon(\Lambda)$, it is seen that $\eta_\varepsilon(\mathbb{Z})\cdot\sqrt{z_{max}^2+z_{min}^2}/\min(s_i)\leq 1$. □

Appendix III: Proof of Lemma 4.2

Recall that we use the following notations: $\eta=\frac{\sqrt{a^2+b^2}}{s_1}, \psi=\frac{\sqrt{a^2+b^2}-a}{b}$ and $\omega=1-\frac{\eta}{\psi t}$. Our goal is to show that under the condition of $s_1=s_2$ and $s_1 t\geq\frac{\sqrt{a^2+b^2}}{\psi}$, we have

$$b(x_c)=\frac{\beta(x_c)}{\alpha(x_c)}\leq Ce^{-\pi\omega^2\psi^2t^2}.$$

where $C=\frac{2}{\left(1-e^{-\pi(2\omega\psi\eta t+\eta^2)}\right)\left(1+e^{-2\pi\eta^2}(1+e^{-4\pi\eta^2})\right)}$.

We first analyse $\alpha(x_c)$

$$\alpha(x_c)=\frac{1}{\rho_{s_1}^2(\mathbb{Z})}\sum_{(x_1,x_2)\in S_{x_c}} e^{-\pi\frac{x_1^2+x_2^2}{s_1^2}}.$$

Note that $\xi = -\frac{ub+va}{a^2+b^2}x_c$. By (5), we know that

$$\sum_{k=\lceil\xi\rceil+1}^{\infty} e^{-\pi\frac{(x_c u+kb)^2+(x_c v+ka)^2}{s_1^2}} = e^{-\pi\frac{\|P_1\|^2}{s_1^2}} \sum_{i=1}^{\infty} e^{-\pi\eta^2(i^2+2i(\lceil\xi\rceil-\xi))},$$

and

$$\sum_{k=\lfloor\xi\rfloor-1}^{-\infty} e^{-\pi\frac{(x_c u+kb)^2+(x_c v+ka)^2}{s_1^2}} = e^{-\pi\frac{\|P_0\|^2}{s_1^2}} \sum_{i=1}^{\infty} e^{-\pi\eta^2(i^2+2i(\xi-\lfloor\xi\rfloor))}.$$

Thus

$$\alpha(x_c) = \begin{cases} \frac{1}{\rho_{s_1}^2(\mathbb{Z})}\left(e^{-\pi\frac{\|P_1\|^2}{s_1^2}} \sum_{i=0}^{\infty} e^{-\pi\eta^2(i^2+2i(\lceil\xi\rceil-\xi))} + e^{-\pi\frac{\|P_0\|^2}{s_1^2}} \sum_{i=0}^{\infty} e^{-\pi\eta^2(i^2+2i(\xi-\lfloor\xi\rfloor))}\right), & \text{if } \xi \notin \mathbb{Z}, \\ \frac{1}{\rho_{s_1}^2(\mathbb{Z})}\left(e^{-\pi\frac{\|P_0\|^2}{s_1^2}} + 2e^{-\pi\frac{\|P_0\|^2}{s_1^2}} \sum_{i=1}^{\infty} e^{-\pi\eta^2 i^2}\right), & \text{if } \xi \in \mathbb{Z}. \end{cases}$$

Let

$$d_0 = e^{-\pi\eta^2(1+2(\xi-\lfloor\xi\rfloor))}(1+e^{-\pi\eta^2(3+2(\xi-\lfloor\xi\rfloor))}),$$
$$d_1 = e^{-\pi\eta^2(1+2(\lceil\xi\rceil-\xi))}(1+e^{-\pi\eta^2(3+2(\lceil\xi\rceil-\xi))}).$$

We have

$$1+d_0 \le \sum_{i=0}^{\infty} e^{-\pi\eta^2(i^2+2i(\xi-\lfloor\xi\rfloor))},$$
$$1+d_1 \le \sum_{i=0}^{\infty} e^{-\pi\eta^2(i^2+2i(\lceil\xi\rceil-\xi))}.$$

These yield an estimation of $\alpha(x)$, if $\xi \notin \mathbb{Z}$

$$\frac{1}{\rho_{s_1}^2(\mathbb{Z})}\left(e^{-\pi\frac{\|P_1\|^2}{s_1^2}}(1+d_1) + e^{-\pi\frac{\|P_0\|^2}{s_1^2}}(1+d_0)\right) \le \alpha(x);$$

if $\xi \in \mathbb{Z}$

$$\frac{(1+2d_0)e^{-\pi\frac{\|P_0\|^2}{s_1^2}}}{\rho_{s_1}^2(\mathbb{Z})} \le \alpha(x).$$

And as for $\beta(x_c)$, we have

$$\beta(x_c) = \frac{1}{\rho_{s_1}^2(\mathbb{Z})} \sum_{\substack{(x_1,x_2)\in S_{x_c} \\ |x_1|\ge s_1 t \text{ or } |x_2|\ge s_1 t}} e^{-\pi\frac{x_1^2+x_2^2}{s_1^2}}.$$

where $|x_c| \le \sqrt{a^2+b^2}s_1 t$.

Three cases shall be discussed separately

1. $(a-b)s_1t \le x_c \le \sqrt{a^2+b^2}s_1t$;
2. $-(a-b)s_1t < x_c < (a-b)s_1t$;
3. and $-\sqrt{a^2+b^2}s_1t \le x_c \le -(a-b)s_1t$.

Case I: $(a-b)s_1t \le x_c \le \sqrt{a^2+b^2}s_1t$.

In this case, condition $|x_1| \ge s_1t$ or $|x_2| \ge s_1t$ corresponds to $k \le \left\lfloor \frac{-s_1t-xv}{a} \right\rfloor$ or $k \ge \left\lceil \frac{s_1t-xu}{b} \right\rceil$. So by (5),

$$\begin{aligned}\beta(x_c) &= \frac{1}{\rho_{s_1}^2(\mathbb{Z})}\left(\sum_{k=\left\lceil \frac{s_1t-x_cu}{b} \right\rceil}^{\infty} e^{-\pi\frac{(x_cu+kb)^2+(x_cv+ka)^2}{s_1^2}} + \sum_{k=\left\lfloor \frac{-s_1t-xv}{a} \right\rfloor}^{-\infty} e^{-\pi\frac{(xu+kb)^2+(xv+ka)^2}{s_1^2}}\right)\\ &= \frac{1}{\rho_{s_1}^2(\mathbb{Z})}\left(e^{-\pi\frac{\|P_1\|^2}{s_1^2}} \sum_{i=\left\lceil \frac{s_1t-x_cu}{b} \right\rceil-\lceil\xi\rceil}^{\infty} e^{-\pi\eta^2(i^2+2i(\lceil\xi\rceil-\xi))}\right)+\\ &\quad \frac{1}{\rho_{s_1}^2(\mathbb{Z})}\left(e^{-\pi\frac{\|P_0\|^2}{s_1^2}} \sum_{i=\left\lfloor \frac{-s_1t-x_cv}{a} \right\rfloor-\lfloor\xi\rfloor}^{-\infty} e^{-\pi\eta^2(i^2-2i(\xi-\lfloor\xi\rfloor))}\right)\end{aligned}$$

Note that $(a-b)s_1t \le x_c \le \sqrt{a^2+b^2}s_1t$, we see that

$$\left\lceil \frac{s_1t-x_cu}{b} \right\rceil - \lceil\xi\rceil \ge \frac{s_1t-x_cu}{b} - \xi - 1 \ge \frac{\sqrt{a^2+b^2}-a}{b\sqrt{a^2+b^2}}s_1t - 1$$

Obviously, $\frac{\sqrt{a^2+b^2}-b}{a\sqrt{a^2+b^2}} \ge \frac{\sqrt{a^2+b^2}-a}{b\sqrt{a^2+b^2}}$ as $a > b > 0$, we get

$$\left\lfloor \frac{-s_1t-x_cv}{a} \right\rfloor - \lfloor\xi\rfloor \le \frac{-s_1t-x_cv}{a} - \xi + 1 \le -\frac{\sqrt{a^2+b^2}-b}{a\sqrt{a^2+b^2}}s_1t + 1 \le -(\frac{\sqrt{a^2+b^2}-a}{b\sqrt{a^2+b^2}}s_1t - 1).$$

Case III: $-\sqrt{a^2+b^2}s_1t \le x_c \le -(a-b)s_1t$.

In this case, condition $|x_1| \ge s_1t$ or $|x_2| \ge s_1t$ corresponds to $k \le \left\lfloor \frac{-s_1t-x_cu}{b} \right\rfloor$ or $k \ge \left\lceil \frac{s_1t-x_cv}{a} \right\rceil$. So similarly with Case I, we see that

$$\left\lceil \frac{s_1t-x_cv}{a} \right\rceil - \lceil\xi\rceil \ge \frac{s_1t-x_cv}{a} - \xi - 1 \ge \frac{\sqrt{a^2+b^2}-b}{a\sqrt{a^2+b^2}}s_1t - 1 \ge \frac{\sqrt{a^2+b^2}-a}{b\sqrt{a^2+b^2}}s_1t - 1.$$

Also

$$\left\lfloor \frac{-s_1t-x_cu}{b} \right\rfloor - \lfloor\xi\rfloor \le \frac{-s_1t-x_cu}{b} - \xi + 1 \le -(\frac{\sqrt{a^2+b^2}-a}{b\sqrt{a^2+b^2}}s_1t - 1).$$

Case II: $-(a-b)s_1t < x_c < (a-b)s_1t$.

In this case, condition $|x_1| \geq s_1 t$ or $|x_2| \geq s_1 t$ corresponds to $k \leq \left\lfloor \frac{-s_1 t - x_c v}{a} \right\rfloor$ or $k \geq \left\lceil \frac{s_1 t - x_c v}{a} \right\rceil$. So by (5),

$$\begin{aligned}
\beta(x) &= \frac{1}{\rho^2_{s_1}(\mathbb{Z})} \left(\sum_{k=\left\lceil \frac{s_1 t - x_c v}{a} \right\rceil}^{\infty} e^{-\pi \frac{(x_c u + kb)^2 + (xv + ka)^2}{s_1^2}} + \sum_{k=\left\lfloor \frac{-s_1 t - x_c v}{a} \right\rfloor}^{-\infty} e^{-\pi \frac{(x_c u + kb)^2 + (x_c v + ka)^2}{s_1^2}} \right) \\
&= \frac{1}{\rho^2_{s_1}(\mathbb{Z})} \left(e^{-\pi \frac{\|P_1\|^2}{s_1^2}} \sum_{i=\left\lceil \frac{s_1 t - x_c v}{a} \right\rceil - \lceil \xi \rceil}^{\infty} e^{-\pi \eta^2 (i^2 + 2i(\lceil \xi \rceil - \xi))} \right) + \\
&\quad \frac{1}{\rho^2_{s_1}(\mathbb{Z})} \left(e^{-\pi \frac{\|P_0\|^2}{s_1^2}} \sum_{i=\left\lfloor \frac{-s_1 t - x_c v}{a} \right\rfloor - \lfloor \xi \rfloor}^{-\infty} e^{-\pi \eta^2 (i^2 - 2i(\xi - \lfloor \xi \rfloor))} \right)
\end{aligned}$$

Obviously, $\frac{a^2+2b^2-ab}{a(a^2+b^2)} \geq \frac{\sqrt{a^2+b^2}-a}{b\sqrt{a^2+b^2}}$ as $a > b > 0$, we have

$$\left\lceil \frac{s_1 t - x_c v}{a} \right\rceil - \lceil \xi \rceil \geq \frac{s_1 t - x_c v}{a} - \xi - 1 \geq \frac{a^2 + 2b^2 - ab}{a(a^2+b^2)} s_1 t - 1 \geq \frac{\sqrt{a^2+b^2} - a}{b\sqrt{a^2+b^2}} s_1 t - 1.$$

and

$$\left\lfloor \frac{-s_1 t - x_c v}{a} \right\rfloor - \lfloor \xi \rfloor \leq \frac{-s_1 t - x_c v}{a} - \xi + 1 \leq -\frac{a^2 + 2b^2 - ab}{a(a^2+b^2)} s_1 t + 1 \leq -(\frac{\sqrt{a^2+b^2} - a}{b\sqrt{a^2+b^2}} s_1 t - 1).$$

When $s_1 t \geq \frac{\sqrt{a^2+b^2}}{\psi}$, we have $\omega \geq 0$ and $\frac{\psi}{\eta} t - 1 \geq \frac{\omega\psi}{\eta} t$. As a result, for all $x_c \in [-\sqrt{a^2+b^2} s_1 t, \sqrt{a^2+b^2} s_1 t]$, we have

$$\sum_{i=\left\lfloor \frac{-s_1 t - x_c v}{a} \right\rfloor - \lfloor \xi \rfloor}^{-\infty} e^{-\pi \eta^2 (i^2 - 2i(\xi - \lfloor \xi \rfloor))} \leq D_0, \text{ and } \sum_{i=\left\lceil \frac{s_1 t - x_c v}{a} \right\rceil - \lceil \xi \rceil}^{\infty} e^{-\pi \eta^2 (i^2 + 2i(\lceil \xi \rceil - \xi))} \leq D_0.$$

where $D_0 = \frac{e^{-\pi \omega^2 \psi^2 t^2}}{1 - e^{-\pi(2\omega\psi\eta t + \eta^2)}}$.

So

$$\begin{aligned}
\beta(x) &\leq \frac{1}{\rho^2_{s_1}(\mathbb{Z})} \left(e^{-\pi \frac{\|P_1\|^2}{s_1^2}} \sum_{i=\left\lceil \frac{s_1 t - x_c v}{a} \right\rceil - \lceil \xi \rceil}^{\infty} e^{-\pi \eta^2 (i^2 + 2i(\lceil \xi \rceil - \xi))} \right) + \\
&\quad \frac{1}{\rho^2_{s_1}(\mathbb{Z})} \left(e^{-\pi \frac{\|P_0\|^2}{s_1^2}} \sum_{i=\left\lfloor \frac{-s_1 t - x_c v}{a} \right\rfloor - \lfloor \xi \rfloor}^{-\infty} e^{-\pi \eta^2 (i^2 - 2i(\xi - \lfloor \xi \rfloor))} \right) \\
&\leq \frac{1}{\rho^2_{s_1}(\mathbb{Z})} D_0 \left(e^{-\pi \frac{\|P_1\|^2}{s_1^2}} + e^{-\pi \frac{\|P_0\|^2}{s_1^2}} \right)
\end{aligned}$$

Thus when $\xi \in \mathbb{Z}$

$$\begin{aligned}
b(x_c) = \frac{\beta(x_c)}{\alpha(x_c)} &\leq \frac{D_0\left(e^{-\pi\frac{\|P_1\|^2}{s_1^2}} + e^{-\pi\frac{\|P_0\|^2}{s_1^2}}\right)}{e^{-\pi\frac{\|P_0\|^2}{s_1^2}}(1+2d_0)} \\
&= \frac{(1+e^{-\pi/s_1^2})e^{-\pi\psi^2\omega^2t^2}}{(1-e^{-\pi(2\omega\psi\eta t+\eta^2)})(1+2e^{-\pi\eta^2}(1+e^{-3\pi\eta^2}))} \\
&\leq D_1 e^{-\pi\omega^2\psi^2t^2}
\end{aligned}$$

where $D_1 = \frac{1+e^{-\pi/s_1^2}}{(1-e^{-\pi(2\omega\psi\eta t+\eta^2)})(1+2e^{-\pi\eta^2}(1+e^{-3\pi\eta^2}))}$.

And when $\xi \notin \mathbb{Z}$, assume $\|P_1\|^2 \geq \|P_0\|^2$ without loss of generality, we have

$$\begin{aligned}
b(x_c) = \frac{\beta(x_c)}{\alpha(x_c)} &\leq \frac{D_0 e^{-\pi\frac{\|P_1\|^2}{s_1^2}} + D_0 e^{-\pi\frac{\|P_0\|^2}{s_1^2}}}{e^{-\pi\frac{\|P_1\|^2}{s_1^2}}(1+d_1) + e^{-\pi\frac{\|P_0\|^2}{s_1^2}}(1+d_0)} \leq \frac{2D_0 e^{-\pi\frac{\|P_0\|^2}{s_1^2}}}{e^{-\pi\frac{\|P_0\|^2}{s_1^2}}(1+d_0)} \\
&\leq \frac{2e^{-\pi\omega^2\psi^2t^2}}{(1-e^{-\pi(2\omega\psi\eta t+\eta^2)})(1+2e^{-2\pi\eta^2}(1+e^{-4\pi\eta^2}))} \\
&\leq D_2 e^{-\pi\omega^2\psi^2t^2}
\end{aligned}$$

where $D_2 = \frac{2}{(1-e^{-\pi(2\omega\psi\eta t+\eta^2)})(1+2e^{-2\pi\eta^2}(1+e^{-4\pi\eta^2}))}$.

Let $C = D_2 > D_1$, for all ξ, we have

$$b(x_c) \leq C e^{-\pi\omega^2\psi^2t^2}$$

□

It should be noted that to ensure $\omega = 1 - \frac{\sqrt{a^2+b^2}}{\psi s_1 t} \geq 0$, $s_1 t \geq \frac{\sqrt{a^2+b^2}}{\psi}$ is required. Without this requirement, $b(x_c)$ could be very close to 1 and the discussion would not be meaningful.

Besides, theorem 4.1 demands $s_1 \geq \sqrt{a^2+b^2}\eta_\varepsilon(\mathbb{Z})$, $t \geq \eta_\varepsilon(\mathbb{Z})$ and $\eta_\varepsilon(\mathbb{Z})$ can be regarded as a constant because it is controlled by ε which is related to the designed errors. We have

$$\omega \geq 1 - \frac{1}{\psi\eta_\varepsilon^2(\mathbb{Z})}$$

It is seen that a larger a/b leads to smaller ψ as well as ω and turns out to be a much larger $b(x_c)$.