Quantum algorithms for computing general discrete logarithms and orders with tradeoffs

Martin Ekerå[1,2]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Swedish NCSA, Swedish Armed Forces, Stockholm, Sweden

August 30, 2018

Abstract

In this paper, we generalize and bridge Shor's groundbreaking works [13, 14] on computing orders and general discrete logarithms, our earlier works [3, 4, 5] on computing short discrete logarithms with tradeoffs and Seifert's work [12] on computing orders with tradeoffs.

In particular, we demonstrate how the idea of enabling tradeoffs may be extended to the case of computing general discrete logarithms.

This yields a reduction by up to a factor of two in the number of group operations that need to be computed in each run of the quantum algorithm at the expense of having to run the algorithm multiple times.

Combined with our earlier works [4, 5] this implies that the number of group operations that need to be computed in each run of the quantum algorithm equals the number of bits in the logarithm times a small constant factor that depends on the tradeoff factor.

We comprehensively analyze the probability distribution induced by our quantum algorithm, describe how the algorithm may be simulated, and estimate the number of runs required to compute logarithms with a given minimum success probability for different tradeoff factors.

Our algorithm does not require the group order to be known. If the order is unknown, it may be computed at no additional quantum cost from the same set of outputs as is used to compute the logarithm.

1 Introduction

Let $\mathbb{G}$ under $\odot$ be a finite cyclic group of order r generated by g, and let

$$x = [\,d\,]\,g = \underbrace{g \odot g \odot \cdots \odot g \odot g}_{d \text{ times}}.$$

Given x, a generator g, and a description of $\mathbb{G}$ and $\odot$, the discrete logarithm problem is to compute $d = \log_g x$. In cryptographic applications, the group $\mathbb{G}$ is typically a subgroup of $\mathbb{F}_p^*$ for some prime p or an elliptic curve group.

In the general discrete logarithm problem $0 \leq d < r$, whereas d is smaller than r by some order of magnitude in the short discrete logarithm problem.

1.1 Earlier works

In a groundbreaking paper [13] from 1994, subsequently extended and revised in a later publication [14], Shor introduced polynomial time quantum computer algorithms for factoring integers and for computing discrete logarithms in $\mathbb{F}_p^*$.

Although Shor's algorithm for computing discrete logarithms was originally described for $\mathbb{F}_p^*$, it may be generalized to any finite cyclic group, provided the group operation may be implemented efficiently using quantum circuits.

More recently, Ekerå [3] introduced a modified version of Shor's algorithm for computing discrete logarithms that is more efficient than Shor's original algorithm when the logarithm is short. This work was motivated by the use of short discrete logarithms in some instantiations of cryptographic schemes based on the computational intractability of the discrete logarithm problem in $\mathbb{F}_p^*$. A concrete example is the use of short exponents in the Diffie-Hellman key exchange protocol when instantiated with safe-prime groups.

The work was subsequently generalized by Ekerå and Håstad in [4] to enable tradeoffs between the number of times that the algorithm has to be executed and the requirements that it imposes on the quantum computer. These ideas parallel ideas by Seifert [12] for making tradeoffs in Shor's order finding algorithm; the quantum part of Shor's general integer factoring algorithm.

Ekerå and Håstad furthermore explained how the RSA integer factoring problem may be expressed as a short discrete logarithm problem, giving rise to a new algorithm for factoring RSA integers that imposes less requirements on the quantum computer than Shor's general integer factoring algorithm even when taking Seifert's ideas for making tradeoffs into account.

As it is seemingly difficult to construct and operate large-scale quantum computers, any reduction in the requirements imposed on the computer by the algorithm when solving cryptographically relevant problems is potentially important and merits study.

1.2 Our contributions

In this paper, we show how the idea of enabling tradeoffs may be extended to the case of computing general discrete logarithms in finite cyclic groups.

This yields a reduction by up to a factor of two in the number of group operations that need to be computed in each run of the quantum algorithm at the expense of having to run the algorithm multiple times.

Combined with our earlier works [4, 5] this implies that the number of group operations that need to be computed in each run of the quantum algorithm equals the number of bits in the logarithm times a small constant factor that depends on the tradeoff factor.

We comprehensively analyze the probability distribution induced by our quantum algorithm, describe how the algorithm may be simulated classically, and estimate the number of runs required to compute logarithms with a given minimum success probability for different tradeoff factors.

Our algorithm does not require the group order to be known. If the order is unknown, it may be computed from the same set of outputs as is used to compute the discrete logarithm.

1.3 Overview of this paper

Our algorithm for computing discrete logarithms consists of two algorithms;

- a quantum algorithm, that upon input of a generator g of order r, and an element $x = [\,d\,]\,g$ where $0 \leq d < r$, outputs a pair (j, k), and
- a classical probabilistic post-processing algorithm, that upon input of a set of n pairs (j, k), produced by n runs of the quantum algorithm, computes the discrete logarithm d.

In addition to the above post-processing algorithm, we furthermore specify

- a classical probabilistic post-processing algorithm, that upon input of a set of n pairs (j, k), computes the order r. The same set of pairs may be used as input to both this and the above post-processing algorithm.

The quantum algorithm is parameterized under a tradeoff factor s. This factor controls the tradeoff between the requirements that the algorithm imposes on the quantum computer, and the number of runs, n, required to attain a given minimum probability q of recovering d and r when post-processing the outputs.

One of our contributions in this paper is to show how to estimate n for a specific problem instance, represented by d and r, and fixed s and q. To compute the estimate, we simulate the quantum algorithm; that is we generate pairs (j, k) that very closely approximate the pairs that would be output by the quantum algorithm if it was to be executed on a quantum computer with respect to the given problem instance and tradeoff factor. We first use the simulated output to heuristically estimate n, and then verify the estimate by executing the two post-processing algorithms with respect to simulated output.

In practice, the simulator is based on a high-resolution two-dimensional histogram of the probability distribution induced by the quantum algorithm. By sampling the histogram, we generate pairs (j, k) that very closely approximate the output that would be produced by the quantum algorithm if it was to be executed on a quantum computer.

To construct the histogram, we first derive a closed form expression that approximates the probability of the quantum algorithm yielding (j, k) as output, and an upper bound on the error in the approximation. We then integrate this expression and the error bound numerically in different regions of the plane.

Note that the simulator requires d and r to be explicitly known; it cannot be used for problem instances represented by group elements g and $x = [\,d\,]\,g$.

1.3.1 Structure of this paper

The remainder of this paper is structured as follows:

The quantum algorithm is introduced in section 2. In sections 3, 4 and 5, we analyze the probability distribution induced by the quantum algorithm. In particular, we derive a closed form expression approximating the probability of obtaining (j, k) as output from the algorithm, and an associated error bound.

In section 6, we describe how the high-resolution histogram is constructed by numerically integrating the closed form expression. We furthermore describe how the histogram is sampled to simulate the quantum algorithm.

In section 7, we introduce the two post-processing algorithms for recovering d and r from a set of n pairs (j, k). In section 8, we use the simulator to estimate the number of runs n required to solve a given problem instance for d and r, with minimum success probability q, as a function of the tradeoff factor s.

We summarize past and new results, and discuss related applications, such as order finding and integer factoring, in sections 9 and 10, and in the appendices.

1.4 Notation

Before proceeding to introduce the quantum algorithm, we first introduce the below notation that is used throughout this paper:

- $u \bmod n$ denotes u reduced modulo n constrained to $0 \leq u \bmod n < n$.
- $\{u\}_n$ denotes u reduced modulo n constrained to $-n/2 \leq \{u\}_n < n/2$.
- $\lceil u \rceil$ denotes $u \in \mathbb{R}$ rounded upwards to the closest integer.
- $\lfloor u \rceil$ denotes $u \in \mathbb{R}$ rounded to the closest integer.
- $\lfloor u \rfloor$ denotes $u \in \mathbb{R}$ rounded downwards to the closest integer.
- $|\, a + ib \,| = \sqrt{a^2 + b^2}$ where $a, b \in \mathbb{R}$ denotes the Euclidean norm of $a + ib$.
- $|\, \mathbf{u} \,|$ denotes the Euclidean norm of the vector $\mathbf{u} = (u_0, \, \ldots, \, u_{n-1}) \in \mathbb{R}^n$.
- $\operatorname{sgn}(u) \in \{-1, \, 1\}$ denotes the sign of $u \in \mathbb{R}$.

1.5 Randomization

Before proceeding to introduce the quantum algorithm, we furthermore remark that any instance of the general discrete logarithm problem may be randomized.

Hence, it may be assumed, without loss of generality, that d is selected uniformly at random on $0 \leq d < r$. Given two group elements g and $x' = [\, d' \,]\, g$ to be solved for d', the problem instance may be randomized as follows:

1. Select a random integer t. Let $x = x' \odot [\, t \,]\, g = [\, d \,]\, g$.
2. Solve g and x for $d \equiv d' + t \pmod r$ and optionally for r.
3. Compute and return $d' \equiv d - t \pmod r$.

If r is known, t should be selected uniformly at random on $0 \leq t < r$, otherwise on $0 \leq t < 2^{m+c}$ for c a sufficiently large integer constant for the selection of x to be indistinguishable from a uniform selection from $\mathbb{G}$. Solving for r in step 2 is only necessary if r is unknown and d' must be on $0 \leq d' < r$ when returned.

2 The quantum algorithm

In this section we describe the quantum algorithm that upon input of a generator g and an element $x = [\, d \,]\, g$ where $0 \leq d < r$ outputs a pair (j, k).

The algorithm is parameterized under a small integer constant $s \geq 1$ that controls the tradeoff between the number of times that the algorithm needs to be executed and the requirements it imposes on the quantum computer.

1. Let m be the integer such that $2^{m-1} \leq r < 2^m$, let $\ell = \lceil m/s \rceil$, and let

$$| \, \Psi \, \rangle = \frac{1}{\sqrt{2^{m+2\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} \sum_{b\,=\,0}^{2^\ell-1} | \, a \, \rangle \, | \, b \, \rangle \, | \, 0 \, \rangle \, .$$

2. Compute $[a] \, g \odot [-b] \, x$ and store the result in the third register

$$\begin{aligned} | \, \Psi \, \rangle &= \frac{1}{\sqrt{2^{m+2\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} \sum_{b\,=\,0}^{2^\ell-1} | \, a, b, [a] \, g \odot [-b] \, x \, \rangle \\ &= \frac{1}{\sqrt{2^{m+2\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} \sum_{b\,=\,0}^{2^\ell-1} | \, a, b, [a \, - \, bd] \, g \, \rangle \, . \end{aligned}$$

3. Compute QFTs of size $2^{m+\ell}$ and 2^ℓ of the first two registers to obtain

$$| \, \Psi \, \rangle = \frac{1}{\sqrt{2^{m+2\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} \sum_{b\,=\,0}^{2^\ell-1} | \, a, b, [a \, - \, bd] \, g \, \rangle \quad \xrightarrow{\text{QFT}}$$

$$\frac{1}{2^{m+2\ell}} \sum_{a\,=\,0}^{2^{m+\ell}-1} \sum_{b\,=\,0}^{2^\ell-1} \sum_{j\,=\,0}^{2^{m+\ell}-1} \sum_{k\,=\,0}^{2^\ell-1} \mathrm{e}^{\,2\pi i \, (aj + 2^m bk)/2^{m+\ell}} \, | \, j, k, [a \, - \, bd] \, g \, \rangle \, .$$

4. Observe the system to obtain (j, k) and $y = [e] \, g$ where $e = (a - bd) \bmod r$.

The above steps may be interleaved rather than executed sequentially so as to allow the qubits in the first two registers to be recycled [10]. A single control qubit then suffices to implement the first two control registers. This is possible as the qubits in the control registers are not initially entangled; the registers are initialized to uniform superpositions of $2^{m+\ell}$ and 2^ℓ values, respectively.

In Shor's algorithm for computing discrete logarithms in $\mathbb{G}$, both control registers are of length m qubits. Both registers are initialized to uniform superpositions of r values. This makes the single control qubit optimization less straightforward to apply and the initial superpositions harder to induce.

In practice, the exponentiation of group elements is performed by computing a group operation conditioned on each bit in the exponent. Hence, a total of $2m$ group operations are performed in Shor's algorithm, compared to $m + 2m/s$ in our algorithm. As s increases, this tends to m operations, providing an advantage over Shor's original algorithm by up to a factor of two at the expense of having to execute the algorithm multiple times. This reduction in the number of group operations translates into a corresponding reduction in the coherence time and circuit depth requirements of our quantum algorithm.

Our algorithm does not require the order r to be known; it suffices that the size of r is known and that group operations and inverses may be computed.

3 The probability of observing (j, k) and y

Above, the pair (j, k) and element $y = [e] \, g$ are obtained with probability

$$\frac{1}{2^{2(m+2\ell)}} \left| \sum_a \sum_b \exp \left[\frac{2\pi i}{2^{m+\ell}} \, (aj + 2^m bk) \right] \right|^2 \qquad (1)$$

where the sum is over all pairs (a, b), such that $0 \leq a < 2^{m+\ell}$ and $0 \leq b < 2^{\ell}$, respecting the condition $e \equiv a - bd \pmod{r}$. In this section, we seek a closed form error-bounded approximation to (1) summed over all $y = [e]\, g \in \mathbb{G}$.

To this end, we first perform a variable substitution to obtain contiguous summation intervals. As $a = e + bd + n_r r$ for n_r an integer, the index a is a function of b and n_r, where $0 \leq a = e + bd + n_r r < 2^{m+\ell}$, so

$$\lceil -(e + bd)/r \rceil \leq n_r < \left\lceil (2^{m+\ell} - (e + bd))/r \right\rceil . \tag{2}$$

Substituting a for $e + bd + n_r r$ in (1) and adjusting the phase therefore yields

$$\frac{1}{2^{2(m+2\ell)}} \left| \sum_{b\,=\,0}^{2^{\ell}-1} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \exp\left[\frac{2\pi i}{2^{m+\ell}} \left(n_r r j + b(dj + 2^m k) \right) \right] \right|^2 . \tag{3}$$

By introducing arguments α_d and α_r, and corresponding angles θ_d and θ_r, where

$$\alpha_d = \{dj + 2^m k\}_{2^{m+\ell}} \quad \alpha_r = \{rj\}_{2^{m+\ell}} \quad \theta_d = \theta(\alpha_d) = \frac{2\pi\alpha_d}{2^{m+\ell}} \quad \theta_r = \theta(\alpha_r) = \frac{2\pi\alpha_r}{2^{m+\ell}}$$

we may write (3) as a function of α_d and α_r, and e, as

$$\frac{1}{2^{2(m+2\ell)}} \left| \sum_{b\,=\,0}^{2^{\ell}-1} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \exp\left[\frac{2\pi i}{2^{m+\ell}} \left(n_r \alpha_r + b\alpha_d \right) \right] \right|^2 \tag{4}$$

or of θ_d and θ_r, and e, as

$$\rho(\theta_d, \theta_r, e) = \frac{1}{2^{2(m+2\ell)}} \left| \sum_{b\,=\,0}^{2^{\ell}-1} \mathrm{e}^{i\theta_d b} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 . \tag{5}$$

This implies that the probability of observing the pair (j, k) and $y = [e]\, g$ depends only on (α_d, α_r) and e, or equivalently on (θ_d, θ_r) and e. The probability is virtually independent of e in practice, as e can at most shift the endpoints of the summation interval in the inner sums in (4) and (5) by one step.

As stated above, we seek a closed-form approximation to $\rho(\theta_d, \theta_r, e)$ summed over all r group elements $y = [e]\, g \in \mathbb{G}$. Hereinafter, we denote this probability

$$\begin{aligned} P(\theta_d, \theta_r) &= \sum_{e\,=\,0}^{r-1} \rho(\theta_d, \theta_r, e) \\ &= \frac{1}{2^{2(m+2\ell)}} \sum_{e\,=\,0}^{r-1} \left| \sum_{b\,=\,0}^{2^{\ell}-1} \mathrm{e}^{i\theta_d b} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 , \end{aligned} \tag{6}$$

and we furthermore use angles and arguments interchangeably, depending on which representation best lends itself to analysis in each step of the process.

3.1 Preliminaries

To gain some intuition, we write $\rho(\theta_d, \theta_r, e)$ as

$$\frac{1}{2^{2(m+2\ell)}} \left| \sum_{b=0}^{2^\ell-1} \mathrm{e}^{i(\theta_d b + \theta_r \lceil -(e+bd)/r \rceil)} \sum_{n_r=0}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - \lceil -(e+bd)/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2$$

and note that there are two obstacles to placing this expression on closed form:

Firstly, the summation interval in the inner sum over n_r depends on the summation variable b of the outer sum. Secondly, the exponent of the summand in the outer sum over b contains a rounding operation that depends on b.

By using that $\lceil (2^{m+\ell} - (e+bd))/r \rceil - \lceil -(e+bd)/r \rceil \approx \lceil 2^{m+\ell}/r \rceil$ we may remove the dependency between the inner and outer sums, and by using that $\lceil -(e+bd)/r \rceil \approx -(e+bd)/r$ we may remove the rounding operation.

By making these approximations, and adjusting the phase, we may hence derive an approximation to $\rho(\theta_d, \theta_r, e)$ that is independent of e, enabling us to sum $\rho(\theta_d, \theta_r, e)$ over the r values of e, corresponding to the r group elements $y = [e]\, g \in \mathbb{G}$, simply by multiplying by r. This yields

$$\begin{aligned} P(\theta_d, \theta_r) &\approx \frac{r}{2^{2(m+2\ell)}} \left| \sum_{b=0}^{2^\ell-1} \mathrm{e}^{i(\theta_d - \theta_r d/r)b} \right|^2 \left| \sum_{n_r=0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 \\ &= \frac{r}{2^{2(m+2\ell)}} \left| \frac{\mathrm{e}^{i2^\ell(\theta_d - \theta_r d/r)} - 1}{\mathrm{e}^{i(\theta_d - \theta_r d/r)} - 1} \right|^2 \left| \frac{\mathrm{e}^{i\lceil 2^{m+\ell}/r \rceil \theta_r} - 1}{\mathrm{e}^{i\theta_r} - 1} \right|^2 \qquad (7) \end{aligned}$$

where we assume in (7) that $\theta_d - \theta_r d/r \neq 0$ and $\theta_r \neq 0$.

This closed form approximation captures the general characteristics of the probability distribution induced by the quantum algorithm. However, it is non-trivial to derive a good bound for the error in this approximation.

In what follows, we use techniques similar to those employed above to derive an error-bounded closed form approximation to $\rho(\theta_d, \theta_r, e)$ such that the error is negligible in the regions of the plane where the probability mass is concentrated.

As was the case above, we will find that the error-bounded approximation of $\rho(\theta_d, \theta_r, e)$ is independent of e, enabling us to approximate $P(\theta_d, \theta_r)$ simply by multiplying the closed form approximation to $\rho(\theta_d, \theta_r, e)$ by r.

3.1.1 Constructive interference

Before we proceed to develop the closed form approximation, we note that for a fixed problem instance and fixed e, the sums in $\rho(\theta_d, \theta_r, e)$ are of a constant number of unit vectors in the complex plane. For such a sum, constructive interference arises when all vectors point in approximately the same direction.

In regions of the plane where θ_r and $\theta_d - d/r\,\theta_r$ are both small, we hence expect constructive interference to arise. The probability mass is expected to concentrate in regions where constructive interference arises, and where the concentration of pairs (θ_d, θ_r) yielded by the integers pairs (j, k) is great.

In what follows, we therefore seek to derive a closed form approximation to $\rho(\theta_d, \theta_r, e)$, and an associated bound on the error in the approximation, such that the error is small when θ_d and $\theta_d - d/r\,\theta_r$ are small.

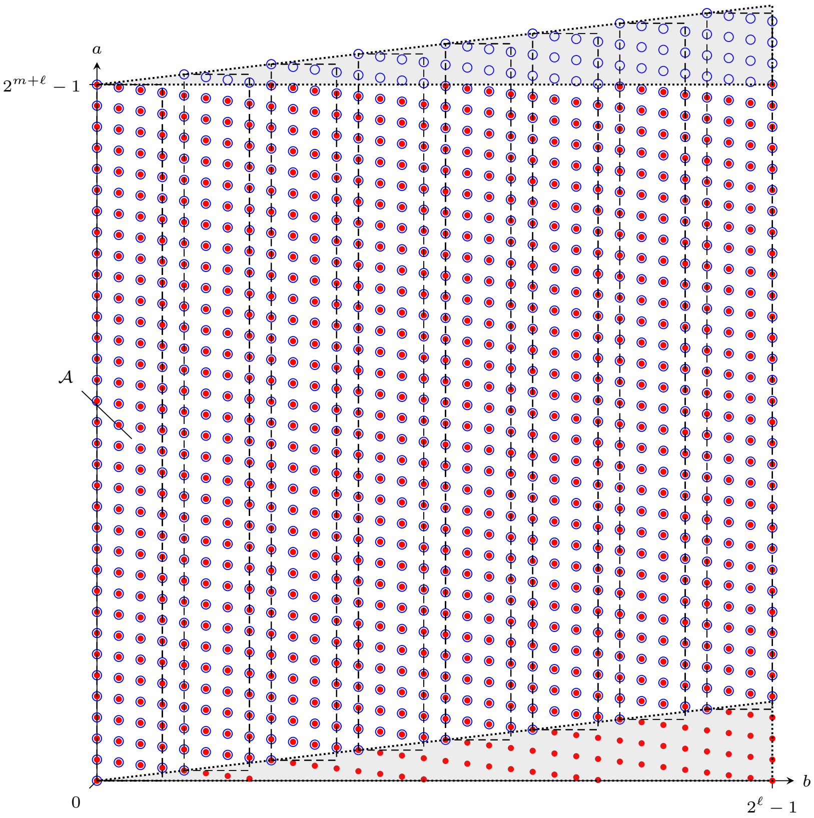

Fig. 1: The lattice $L^{a,b}$ for $\sigma = 2$ introduced in section 3.2.3. All red filled points are in $\mathcal{R}$, introduced below. The region $\mathcal{A}$, introduced in section 3.2.3, and its translated replicas are drawn as dashed rectangles. All blue outlined points are in A or in one of its translated replicas. The gray triangles outline the points that are in A or one of its replicas, but not in $\mathcal{R}$, and vice versa.

3.2 Closed form approximation with error bounds

To derive a closed form approximation to $\rho(\theta_d, \theta_r, e)$, we first observe that the sums in the expression for $\rho(\theta_d, \theta_r, e)$ may be regarded as sums over the points in a region $\mathcal{R}$ in a lattice $L^{a,b}$, as defined and shown formally below:

Definition 3.1. *Let $L^{a,b}$ be the lattice spanned by $(d, 1)$ and $(r, 0)$ so that the set of points in $L^{a,b}$ is given by $(a, b) = b(d, 1) + n_r(r, 0)$ for integers b and n_r.*

Definition 3.2. *Let $\mathcal{R}$ be the region in $L^{a,b}$ where $0 \le a < 2^{m+\ell}$ and $0 \le b < 2^{\ell}$.*

The lattice $L^{a,b}$ and region $\mathcal{R}$ are illustrated in Fig. 1 alongside additional elements to which we will return as the analysis progresses.

Definition 3.3. *Let*

$$S_{\mathcal{R}} = \frac{|\, s_{\mathcal{R}} \,|^2}{2^{2(m+2\ell)}} \qquad \textit{where} \qquad s_{\mathcal{R}} = \sum_{(a,b)\,\in\,\mathcal{R}} \exp\left[\frac{2\pi i}{2^{m+\ell}}(aj + 2^m bk)\right].$$

Claim 3.1. *The probability* $\rho(\theta_d, \theta_r, e) = S_{\mathcal{R}}$ *as*

$$s_{\mathcal{R}} = \sum_{b\,=\,0}^{2^\ell-1} \mathrm{e}^{i\theta_d b} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r}.$$

Proof. The points in $\mathcal{R}$ are given by $(a, b) = b(d, 1) + n_r(r, 0)$ for $0 \le b < 2^\ell$ and n_r on (2) so that $0 \le a = e + bd + n_r r < 2^{m+\ell}$ which implies that

$$\begin{aligned} S_{\mathcal{R}} &= \frac{1}{2^{2(m+2\ell)}} \left| \sum_{b\,=\,0}^{2^\ell-1} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \exp\left[\frac{2\pi i}{2^{m+\ell}} \left(n_r r j + b(dj + 2^m k) \right) \right] \right|^2 \\ &= \frac{1}{2^{2(m+2\ell)}} \left| \sum_{b\,=\,0}^{2^\ell-1} \mathrm{e}^{i\theta_d b} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 = \rho(\theta_d, \theta_r, e) \end{aligned}$$

by the preliminary analysis in section 3 and so the claim follows. ■

It follows from Claim 3.1 that we may derive a closed form approximation of $\rho(\theta_d, \theta_r, e)$ by deriving a closed form approximation of $S_{\mathcal{R}}$. In what follows, we derive such an approximation, and an associated error bound, in three steps.

3.2.1 Preliminaries

Before proceeding as outlined above, we first introduce some preliminary claims.

Claim 3.2. *For* $u, v \in \mathbb{C}$ *and* $\Delta = u - v$ *it holds that*

$$\big|\, |\,u\,|^2 - |\,v\,|^2 \,\big| \le 2\,|\,u\,|\,|\,\Delta\,| + |\,\Delta\,|^2.$$

Proof. First verify that

$$\begin{aligned} |\,u\,|^2 - |\,v\,|^2 = |\,u\,|^2 - |\,u - \Delta\,|^2 &= u\overline{u} - (u - \Delta)\overline{(u - \Delta)} \\ &= u\overline{u} - (u - \Delta)(\overline{u} - \overline{\Delta}) = u\overline{\Delta} + \overline{u}\Delta - |\,\Delta\,|^2 \end{aligned}$$

where the overlines denote complex conjugates. This implies that

$$\big|\, |\,u\,|^2 - |\,v\,|^2 \,\big| \le |\,u\,|\,|\,\overline{\Delta}\,| + |\,\overline{u}\,|\,|\,\Delta\,| + |\,\Delta\,|^2 = 2\,|\,u\,|\,|\,\Delta\,| + |\,\Delta\,|^2$$

and so the claim follows. ■

Claim 3.3. $|\,\mathrm{e}^{i\phi} - 1\,| \le |\,\phi\,|$ *for any* $\phi \in \mathbb{R}$.

Proof. It suffices to show that $|\,\mathrm{e}^{i\phi} - 1\,|^2 = 2(1 - \cos\phi) \le \phi^2$ from which the claim follows as $\cos\phi \ge 1 - \phi^2/2$ for any $\phi \in \mathbb{R}$. ■

3.2.2 Bounding $|\, s_{\mathcal{R}} \,|$

Before proceeding to the first approximation step, we furthermore bound $|\, s_{\mathcal{R}} \,|$ in this section, as this bound is needed in the following analysis.

Lemma 3.1. *The sum $s_{\mathcal{R}}$ is bounded by $|\, s_{\mathcal{R}} \,| \leq 2^{2\ell+1}$.*

Proof. By Claim 3.1 the sum

$$s_{\mathcal{R}} = \sum_{b\,=\,0}^{2^\ell-1} \mathrm{e}^{i\theta_d b} \sum_{n_r\,=\,\lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r}$$

where the outer sum over b is over 2^ℓ values and the inner sum over n_r is over at most $2^{\ell+1}$ values by Claim 3.4 below. As $s_{\mathcal{R}}$ is a sum of at most $2^{2\ell+1}$ complex unit vectors, it follows that $|\, s_{\mathcal{R}} \,| \leq 2^{2\ell+1}$, and so the lemma follows. ■

Claim 3.4. *It holds that*

$$\left\lfloor 2^{m+\ell}/r \right\rfloor \leq \left\lceil (2^{m+\ell} - (e+bd))/r \right\rceil - \lceil -(e+bd)/r \rceil \leq \left\lfloor 2^{m+\ell}/r \right\rfloor + 1$$

and furthermore $\left\lceil (2^{m+\ell} - (e+bd))/r \right\rceil - \lceil -(e+bd)/r \rceil \leq 2^{\ell+1}$.

Proof. The claim counts all integers $a \in [0, 2^{m+\ell})$ such that $a \equiv e + bd \pmod r$ where $2^{m-1} \leq r < 2^m$, from which the claim follows. More formally, for f_1, f_2 appropriately selected fractions on $0 \leq f_1, f_2 < 1$, it holds that

$$\begin{aligned}
&\left\lceil (2^{m+\ell} - (e+bd))/r \right\rceil - \lceil -(e+bd)/r \rceil \\
&= \left\lceil \left\lfloor 2^{m+\ell}/r \right\rfloor + f_1 - \lfloor (e+bd)/r \rfloor - f_2 \right\rceil - \lceil -(e+bd)/r \rceil \\
&= \left\lfloor 2^{m+\ell}/r \right\rfloor - \lfloor (e+bd)/r \rfloor + \lceil f_1 - f_2 \rceil + \lfloor (e+bd)/r \rfloor \\
&= \left\lfloor 2^{m+\ell}/r \right\rfloor + \lceil f_1 - f_2 \rceil
\end{aligned}$$

where we use that $\lceil x \rceil = - \lfloor -x \rfloor$. As $-1 < f_1 - f_2 < 1$ we have $\lceil f_1 - f_2 \rceil \in \{0, 1\}$ and so the first part of the claim follows.

For the second part, recall that the order r is an integer on the interval $2^{m-1} \leq r < 2^m$. For $r > 2^{m-1}$ the claim is true as $\left\lfloor 2^{m+\ell}/r \right\rfloor < 2^{\ell+1}$ and $\lceil f_1 - f_2 \rceil \in \{0, 1\}$. For $r = 2^{m-1}$, we have $\left\lfloor 2^{m+\ell}/r \right\rfloor = 2^{\ell+1}$ and $f_1 = 0$, which implies that $\lceil f_1 - f_2 \rceil = \lceil -f_2 \rceil = - \lfloor f_2 \rfloor = 0$ and so the claim follows. ■

3.2.3 Approximating $S_{\mathcal{R}}$ by $S_{\mathcal{A}} T_{\mathcal{A}}$

In the first step of the approximation, we approximate $S_{\mathcal{R}}$ by summing the points in a small region $\mathcal{A}$ in $\mathcal{R}$, and then replicating and translating the points in $\mathcal{A}$, and the associated sum over these points, so as to approximately cover $\mathcal{R}$.

By bounding the number of points that are thus erroneously excluded from or included in the sum, we bound the error in the approximation. This yields an approximation to $S_{\mathcal{R}}$ that is not on closed form, but that may be placed on closed form by means of two additional approximation steps.

It should be noted that in the preliminary analysis in section 3.1 we simply removed a rounding operation. This corresponds to moving the points in $\mathcal{R}$. In this section, we exclude or include points, but we do not move any points, and this enables us to bound the error simply by counting the erroneous points.

To get started, we define the region $\mathcal{A}$, and the sum $S_{\mathcal{A}}$ over the points in $\mathcal{A}$, in analogy with the sum $S_{\mathcal{R}}$ over $\mathcal{R}$ previously introduced, as follows:

Definition 3.4. *Let $\mathcal{A}$ be the region in $L^{a,b}$ where $0 \leq a < 2^{m+\ell}$ and $0 \leq b < 2^{\sigma}$ for σ an integer parameter selected on $0 < \sigma < \ell$.*

Definition 3.5. *Let*

$$S_{\mathcal{A}} = \frac{\mid s_{\mathcal{A}} \mid^2}{2^{2(m+2\ell)}} \qquad where \qquad s_{\mathcal{A}} = \sum_{(a,b) \, \in \, \mathcal{A}} \exp\left[\frac{2\pi i}{2^{m+\ell}}(aj + 2^m bk)\right].$$

Claim 3.5.

$$S_{\mathcal{A}} = \frac{1}{2^{2(m+2\ell)}} \left| \sum_{b \, = \, 0}^{2^{\sigma}-1} \mathrm{e}^{i\theta_d b} \sum_{n_r \, = \, \lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2.$$

Proof. The points in $\mathcal{A}$ are given by $(a,b) = b(d,1) + n_r(r,0)$ for $0 \leq b < 2^{\sigma}$ and n_r on (2) so that $0 \leq a = e + bd + n_r r < 2^{m+\ell}$ which implies that

$$\begin{aligned} S_{\mathcal{A}} &= \frac{1}{2^{2(m+2\ell)}} \left| \sum_{b \, = \, 0}^{2^{\sigma}-1} \sum_{n_r \, = \, \lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \exp\left[\frac{2\pi i}{2^{m+\ell}} \left(n_r rj + b(dj + 2^m k)\right)\right] \right|^2 \\ &= \frac{1}{2^{2(m+2\ell)}} \left| \sum_{b \, = \, 0}^{2^{\sigma}-1} \mathrm{e}^{i\theta_d b} \sum_{n_r \, = \, \lceil -(e+bd)/r \rceil}^{\lceil (2^{m+\ell}-(e+bd))/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 \end{aligned}$$

in analogy with the analysis in section 3, but with b on $0 \leq b < 2^{\sigma}$ as opposed to $0 \leq b < 2^{\ell}$, and so the claim follows. ■

To replicate and translate the points in $\mathcal{A}$ so as to approximately cover $\mathcal{R}$, we furthermore introduce $t_{\mathcal{A}}$ and $T_{\mathcal{A}}$, as defined below:

Definition 3.6. *Let*

$$T_{\mathcal{A}} = \mid t_{\mathcal{A}} \mid^2 \qquad where \qquad t_{\mathcal{A}} = \sum_{t \, = \, 0}^{2^{\ell-\sigma}-1} \mathrm{e}^{i(\theta_d 2^{\sigma} + \theta_r \lceil -2^{\sigma} d/r \rceil) \, t}.$$

The error when approximating $S_{\mathcal{R}}$ by $S_{\mathcal{A}} T_{\mathcal{A}}$ may now be bounded as follows:

Lemma 3.2. *The error when approximating $s_{\mathcal{R}}$ by $s_{\mathcal{A}} t_{\mathcal{A}}$ is bounded by*

$$\mid s_{\mathcal{R}} - s_{\mathcal{A}} t_{\mathcal{A}} \mid \leq 2^{2\ell - \sigma + 1}.$$

Proof. The exponential sum $t_{\mathcal{A}}$ replicates and translates the partial sum over $\mathcal{A}$ so as to approximately cover $\mathcal{R}$ as is illustrated in Fig. 1. Every time the region is replicated it is translated by $\mathrm{e}^{i(\theta_d 2^{\sigma} + \theta_r \lceil -2^{\sigma} d/r \rceil)}$ and this exponential function may be easily shown to correspond to a vector in $L^{a,b}$.

The error in this approximation is due to points that are in $\mathcal{R}$ but excluded from the sum, and conversely to points not in $\mathcal{R}$ that are erroneously included in the sum. Hereinafter these points will be referred to as the erroneous points.

The erroneous points fall within the two gray triangles in Fig. 1. Both triangles are of horizontal length 2^{ℓ} and vertical side length $2^{\ell-\sigma}(2^{\sigma} d \bmod r)$, as the region $\mathcal{A}$ is replicated and translated $2^{\ell-\sigma}$ times in total, and as it is shifted vertically by $2^\sigma d \bmod r$ every time it is translated.

To upper-bound the number of lattice points in each triangle, note that the lattice points are on 2^ℓ vertical lines evenly separated horizontally by a distance of one. The points on each vertical line are evenly separated vertically by a distance of r with varying starting positions on each line.

Let $f(b)$ denote the height of each triangle at b. At most

$$N(b) = 1 + \lfloor f(b)/r \rfloor \leq 1 + f(b)/r = 1 + \frac{2^{\ell-\sigma}(2^\sigma d \text{ mod r})}{r} \frac{b}{2^\ell} \leq 1 + \frac{b}{2^\sigma}.$$

lattice points are then on the vertical line that cuts through the triangles at b are then within each of the two triangles, as may be seen by maximizing over all possible starting points.

By summing $N(b)$ over all 2^ℓ lines, we thus obtain an upper bound of

$$\sum_{b=0}^{2^\ell-1} N(b) \leq 2^\ell + \frac{1}{2^\sigma} \sum_{b=0}^{2^\ell-1} b = 2^\ell + \frac{1}{2^\sigma} \frac{2^\ell(2^\ell-1)}{2} \leq 2^{2\ell-\sigma}$$

on the number of points in each triangle, where we have used that $2^{2\ell-\sigma-1} \geq 2^\ell$ as σ is an integer on $0 < \sigma < \ell$. As there are two triangles, the total number of erroneous points is upper-bounded by $2^{2\ell-\sigma+1}$ points. Each point corresponds to a unit vector in the complex sum $s_\mathcal{R} - s_\mathcal{A} t_\mathcal{A}$ which implies $|\, s_\mathcal{R} - s_\mathcal{A} t_\mathcal{A} \,| \leq 2^{2\ell-\sigma+1}$ and so the lemma follows. ■

Lemma 3.3. *The error when approximating $S_\mathcal{R}$ by $S_\mathcal{A} T_\mathcal{A}$ is bounded by*

$$|\, S_\mathcal{R} - S_\mathcal{A} T_\mathcal{A} \,| \leq 2^{-2m-\sigma+4}.$$

Proof. By Definition 3.3, 3.5 and 3.6, we now have that

$$S_\mathcal{R} = \frac{|\, s_\mathcal{R} \,|^2}{2^{2(m+2\ell)}} \quad S_\mathcal{A} T_\mathcal{A} = \frac{|\, s_\mathcal{A} t_\mathcal{A} \,|^2}{2^{2(m+2\ell)}} \quad |\, S_\mathcal{R} - S_\mathcal{A} T_\mathcal{A} \,| = \frac{\big|\, |\, s_\mathcal{R} \,|^2 - |\, s_\mathcal{A} t_\mathcal{A} \,|^2 \,\big|}{2^{2(m+2\ell)}}.$$

where $|\, s_\mathcal{R} - s_\mathcal{A} t_\mathcal{A} \,| \leq 2^{2\ell-\sigma+1}$ by Lemma 3.2 and $|\, s_\mathcal{R} \,| \leq 2^{2\ell+1}$ by Lemma 3.1.

By Claim 3.2 it therefore follows that

$$\begin{aligned} \big|\, |\, s_\mathcal{R} \,|^2 - |\, s_\mathcal{A} t_\mathcal{A} \,|^2 \,\big| &\leq 2 \,|\, s_\mathcal{R} \,|\, |\, s_\mathcal{R} - s_\mathcal{A} t_\mathcal{A} \,| + |\, s_\mathcal{R} - s_\mathcal{A} t_\mathcal{A} \,|^2 \\ &\leq 2 \cdot 2^{2\ell+1} \cdot 2^{2\ell-\sigma+1} + 2^{4\ell-2\sigma+2} \\ &= 2 \cdot 2^{4\ell-\sigma+2} + 2^{4\ell-2\sigma+2} \\ &\leq 3 \cdot 2^{4\ell-\sigma+2} \leq 2^{4(\ell+1)-\sigma} \end{aligned}$$

from which it follows that

$$|\, S_\mathcal{R} - S_\mathcal{A} T_\mathcal{A} \,| = \frac{\big|\, |\, s_\mathcal{R} \,|^2 - |\, s_\mathcal{A} t_\mathcal{A} \,|^2 \,\big|}{2^{2(m+2\ell)}} \leq \frac{2^{4(\ell+1)-\sigma}}{2^{2(m+2\ell)}} = 2^{-2m-\sigma+4}$$

and so the lemma follows. ■

As $t_\mathcal{A}$ is a geometric series $T_\mathcal{A} = |\, t_\mathcal{A} \,|^2$ may be placed on closed form. It remains to derive a closed form approximation to $S_\mathcal{A}$ in two more steps.

3.2.4 Approximating $S_{\mathcal{A}}$ by $S'_{\mathcal{A}}$

In the second step of the approximation, we derive a closed form approximation to $S_{\mathcal{A}}$, by first approximating $S_{\mathcal{A}}$ by the product $S'_{\mathcal{A}}$ of two sums, such that the leading sum may be placed on closed form, and such that the trailing sum may be placed on closed form by means of a third approximation step.

Definition 3.7. *Let*

$$S'_{\mathcal{A}} = \frac{|\, s'_{\mathcal{A}} \,|^2}{2^{2(m+2\ell)}} \qquad \textit{where} \qquad s'_{\mathcal{A}} = \sum_{b\,=\,0}^{2^\sigma-1} \mathrm{e}^{i(\theta_d b + \theta_r \lceil -(e+bd)/r \rceil)} \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r}.$$

Lemma 3.4. *The error when approximating $s_{\mathcal{A}}$ by $s'_{\mathcal{A}}$ is bounded by*

$$|\, s_{\mathcal{A}} - s'_{\mathcal{A}} \,| \leq 2^\sigma.$$

Proof. As $s_{\mathcal{A}}$ and $s'_{\mathcal{A}}$ are sums of complex unit vectors, and as the sums differ by at most 2^σ vectors, as may be seen by comparing the summation intervals using Claim 3.4, it follows that $|\, s_{\mathcal{A}} - s'_{\mathcal{A}} \,| \leq 2^\sigma$, and so the lemma follows. ■

Lemma 3.5. *The sum $s'_{\mathcal{A}}$ is bounded by $|\, s'_{\mathcal{A}} \,| \leq 2^{\ell+\sigma+1}$.*

Proof. By Definition 3.7

$$s'_{\mathcal{A}} = \sum_{b\,=\,0}^{2^\sigma-1} \mathrm{e}^{i(\theta_d b + \theta_r \lceil -(e+bd)/r \rceil)} \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r}$$

where the outer sum over b assumes 2^σ values and the inner sum over n_r assumes at most $2^{\ell+1}$ values as the order $r \geq 2^{m-1}$.

Hence $s'_{\mathcal{A}}$ is a sum of at most $2^{\ell+\sigma+1}$ complex unit vectors, from which it follows that $|\, s'_{\mathcal{A}} \,| \leq 2^{\ell+\sigma+1}$, and so the lemma follows. ■

Lemma 3.6. *The error when approximating $S_{\mathcal{A}}$ by $S'_{\mathcal{A}}$ is upper-bounded by*

$$|\, S_{\mathcal{A}} - S'_{\mathcal{A}} \,| \leq 2^{-2m-3\ell+2\sigma+3}.$$

Proof. By Definition 3.5 and 3.7, and Claim 3.5, we have

$$S_{\mathcal{A}} = \frac{|\, s_{\mathcal{A}} \,|^2}{2^{2(m+2\ell)}} \qquad S'_{\mathcal{A}} = \frac{|\, s'_{\mathcal{A}} \,|^2}{2^{2(m+2\ell)}} \qquad |\, S'_{\mathcal{A}} - S_{\mathcal{A}} \,| \leq \frac{\big|\, |\, s_{\mathcal{A}} \,|^2 - |\, s'_{\mathcal{A}} \,|^2 \,\big|}{2^{2(m+2\ell)}}$$

where $|\, s_{\mathcal{A}} - s'_{\mathcal{A}} \,| \leq 2^\sigma$ by Lemma 3.4 and $|\, s'_{\mathcal{A}} \,| \leq 2^{\ell+\sigma+1}$ by Lemma 3.5.

By Claim 3.2 it therefore follows that

$$\begin{aligned} \big|\, |\, s_{\mathcal{A}} \,|^2 - |\, s'_{\mathcal{A}} \,|^2 \,\big|^2 &\leq 2\,|\, s'_{\mathcal{A}} \,|\,|\, s_{\mathcal{A}} - s'_{\mathcal{A}} \,| + |\, s_{\mathcal{A}} - s'_{\mathcal{A}} \,| \\ &\leq 2 \cdot 2^{\ell+\sigma+1} \cdot 2^\sigma + 2^{2\sigma} \\ &= 2 \cdot 2^{\ell+2\sigma+1} + 2^{2\sigma} \\ &\leq 3 \cdot 2^{\ell+2\sigma+1} \leq 2^{\ell+2\sigma+3} \end{aligned}$$

from which it follows that

$$|\, S_{\mathcal{A}} - S'_{\mathcal{A}} \,| \leq \frac{\big|\, |\, s_{\mathcal{A}} \,|^2 - |\, s'_{\mathcal{A}} \,|^2 \,\big|}{2^{2(m+2\ell)}} \leq \frac{2^{\ell+2\sigma+3}}{2^{2(m+2\ell)}} = 2^{-2m-3\ell+2\sigma+3}$$

and so the lemma follows. ■

As the trailing sum in $S'_{\mathcal{A}}$ is the square norm of a geometric series it may be placed on closed form. Due to the rounding operation in the exponent, this approach does not work for the leading sum. In the third step of the approximation, we take a different approach to placing it on closed form.

3.2.5 Approximating $S'_{\mathcal{A}}$ by $S''_{\mathcal{A}}$

For θ_d and θ_r such that the angles

$$\theta_d b + \theta_r \left\lceil -(e + bd)/r \right\rceil \approx (\theta_d - \theta_r d/r)\, b$$

in the leading sum $S_{\mathcal{A}}$ are small for all b on $0 \leq b < 2^\sigma$, all 2^σ terms in the leading sum in $S'_{\mathcal{A}}$ are approximately one. In the third step of the approximation, we show that the error when approximating all terms in this sum by one, may be bounded as described below:

Definition 3.8. *Let*

$$S''_{\mathcal{A}} = \frac{|\, s''_{\mathcal{A}} \,|^2}{2^{2(m+2\ell)}} \qquad \textit{where} \qquad s''_{\mathcal{A}} = 2^\sigma \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r}.$$

Lemma 3.7. *The difference between $s'_{\mathcal{A}}$ and $s''_{\mathcal{A}}$ is upper-bounded by*

$$|\, s'_{\mathcal{A}} - s''_{\mathcal{A}} \,| \leq 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,| \right) |\, s''_{\mathcal{A}} \,|.$$

Proof. First observe that

$$|\, s'_{\mathcal{A}} - s''_{\mathcal{A}} \,| = \underbrace{\left| \sum_{b\,=\,0}^{2^\sigma - 1} \left(\mathrm{e}^{i(\theta_d b + \theta_r \lceil -(e+bd)/r \rceil)} - 1 \right) \right|}_{|\,\Delta\,|} \left| \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|.$$

By using Claim 3.3 below and the triangle inequality it follows that

$$\begin{aligned} |\,\Delta\,| &= \left| \sum_{b\,=\,0}^{2^\sigma - 1} \left(\mathrm{e}^{i(\theta_d b + \theta_r \lceil -(e+bd)/r \rceil)} - 1 \right) \right| \leq \sum_{b\,=\,0}^{2^\sigma - 1} \left| \mathrm{e}^{i(\theta_d b + \theta_r \lceil -(e+bd)/r \rceil)} - 1 \right| \\ &\leq \sum_{b\,=\,0}^{2^\sigma - 1} |\, \theta_d b + \theta_r \left\lceil -(e+bd)/r \right\rceil \,| = \sum_{b\,=\,0}^{2^\sigma - 1} |\, \theta_d b - \theta_r \left\lfloor (e+bd)/r \right\rfloor \,| \\ &\leq (|\,\theta_d\,| + |\,\theta_r\,|) \sum_{b\,=\,0}^{2^\sigma - 1} b \leq (|\,\theta_d\,| + |\,\theta_r\,|)\ \frac{2^\sigma(2^\sigma - 1)}{2} \leq 2^{2\sigma-1} \left(|\,\theta_d\,| + |\,\theta_r\,| \right) \end{aligned}$$

where we use that $\lceil -x \rceil = -\lfloor x \rfloor$ and $\lfloor (e+bd)/r \rfloor \leq b$. To verify the latter claim, note that $0 \leq f_1 = e/r < 1$ and $0 \leq f_2 = bd/r < b$ as $0 \leq e, d < r$. This implies that $0 \leq \lfloor (e+bd)/r \rfloor = \lfloor f_1 + f_2 \rfloor \leq b$ as $0 \leq f_1 + f_2 < b + 1$.

By combining the above results we now have

$$\begin{aligned} |\, s'_{\mathcal{A}} - s''_{\mathcal{A}} \,| &\leq 2^{2\sigma-1} \left(|\,\theta_d\,| + |\,\theta_r\,| \right) \left| \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right| \\ &= 2^{\sigma-1} \left(|\,\theta_d\,| + |\,\theta_r\,| \right) |\, s''_{\mathcal{A}} \,| \end{aligned}$$

and so the lemma follows. ■

Lemma 3.8. *The error when approximating $S'_{\mathcal{A}}$ by $S''_{\mathcal{A}}$ is upper-bounded by*

$$|\, S'_{\mathcal{A}} - S''_{\mathcal{A}} \,| \leq 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right) \left(2 + 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right)\right) S''_{\mathcal{A}}.$$

Proof. By Definitions 3.7 and 3.8 we have

$$S'_{\mathcal{A}} = \frac{|\, s'_{\mathcal{A}} \,|^2}{2^{2(m+2\ell)}} \qquad S''_{\mathcal{A}} = \frac{|\, s''_{\mathcal{A}} \,|^2}{2^{2(m+2\ell)}} \qquad \big|\, |\, S'_{\mathcal{A}} \,|^2 - |\, S''_{\mathcal{A}} \,|^2 \,\big| = \frac{\big|\, |\, s'_{\mathcal{A}} \,|^2 - |\, s''_{\mathcal{A}} \,|^2 \,\big|}{2^{2(m+2\ell)}}$$

where $|\, s'_{\mathcal{A}} - s''_{\mathcal{A}} \,| \leq 2^{\sigma-1}(|\, \theta_d \,| + |\, \theta_r \,|)\, |\, s''_{\mathcal{A}} \,|$ by Lemma 3.7.

By Claim 3.2 it therefore follows that

$$\begin{aligned}
\big|\, |\, s'_{\mathcal{A}} \,|^2 - |\, s''_{\mathcal{A}} \,|^2 \,\big| &\leq 2\, |\, s''_{\mathcal{A}} \,|\, |\, s'_{\mathcal{A}} - s''_{\mathcal{A}} \,| + |\, s'_{\mathcal{A}} - s''_{\mathcal{A}} \,|^2 \\
&\leq 2 \cdot 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right) |\, s''_{\mathcal{A}} \,|^2 + 2^{2(\sigma-1)} \left(|\, \theta_d \,| + |\, \theta_r \,|\right)^2 |\, s''_{\mathcal{A}} \,|^2 \\
&\leq 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right) \left(2 + 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right)\right) |\, s''_{\mathcal{A}} \,|^2
\end{aligned}$$

from which it follows that

$$\begin{aligned}
|\, S'_{\mathcal{A}} - S''_{\mathcal{A}} \,| &\leq \frac{\big|\, |\, s'_{\mathcal{A}} \,|^2 - |\, s''_{\mathcal{A}} \,|^2 \,\big|}{2^{2(m+2\ell)}} \\
&\leq 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right) \left(2 + 2^{\sigma-1} \left(|\, \theta_d \,| + |\, \theta_r \,|\right)\right) S''_{\mathcal{A}}
\end{aligned}$$

and so the lemma follows. ∎

This yields an approximation $S''_{\mathcal{A}}$ to $S'_{\mathcal{A}}$ that may be placed on closed form.

3.2.6 Main approximability result

By combining the results in the above lemmas, the main result below follows:

Theorem 3.1. *The probability $P(\theta_d, \theta_r)$ of observing a specific pair (j, k) with angle pair (θ_d, θ_r), summed over all $y \in \mathbb{G}$, may be approximated by*

$$\begin{aligned}
\widetilde{P}(\theta_d, \theta_r) &= \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \left|\, \sum_{t=0}^{2^{\ell-\sigma}-1} \mathrm{e}^{i(\theta_d 2^\sigma + \theta_r \lceil -2^\sigma d/r \rceil)\, t} \,\right|^2 \left|\, \sum_{n_r=0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \,\right|^2 \\
&= \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \left|\, \frac{\mathrm{e}^{i(\theta_d 2^\sigma + \theta_r \lceil -2^\sigma d/r \rceil)\, 2^{\ell-\sigma}} - 1}{\mathrm{e}^{i(\theta_d 2^\sigma + \theta_r \lceil -2^\sigma d/r \rceil)} - 1} \,\right|^2 \left|\, \frac{\mathrm{e}^{i\theta_r \lceil 2^{m+\ell}/r \rceil} - 1}{\mathrm{e}^{i\theta_r} - 1} \,\right|^2
\end{aligned}$$

assuming $\theta_d 2^\sigma + \theta_r \lceil -2^\sigma d/r \rceil \neq 0$ and $\theta_r \neq 0$ when placing the expression on closed form. The approximation error $|\, P(\theta_d, \theta_r) - \widetilde{P}(\theta_d, \theta_r) \,| \leq \tilde{e}(\theta_d, \theta_r)$ where

$$\tilde{e}(\theta_d, \theta_r) \leq \frac{2^4}{2^{m+\sigma}} + \frac{2^3}{2^{m+\ell}} + \frac{2^\sigma}{2}(|\, \theta_d \,| + |\, \theta_r \,|) \left(2 + \frac{2^\sigma}{2}(|\, \theta_d \,| + |\, \theta_r \,|)\right) \widetilde{P}(\theta_d, \theta_r).$$

Proof. The probability $\rho(\theta_d, \theta_r, e)$ of observing a specific pair (j, k), with angle pair (θ_d, θ_r), and some group element $y = [e]\, g \in \mathbb{G}$, is $S_{\mathcal{R}}$ by Claim 3.1.

The error when approximating $S_{\mathcal{R}}$ by $S_{\mathcal{A}} T_{\mathcal{A}}$ is bounded by

$$|\, S_{\mathcal{R}} - S_{\mathcal{A}} T_{\mathcal{A}} \,| \leq 2^{-2m-\sigma+4}$$

by Lemma 3.3. The error when approximating $S_{\mathcal{A}}T_{\mathcal{A}}$ by $S'_{\mathcal{A}}T_{\mathcal{A}}$ is bounded by

$$| \, S_{\mathcal{A}}T_{\mathcal{A}} - S'_{\mathcal{A}}T_{\mathcal{A}} \, | \leq 2^{-2m-3\ell+2\sigma+3} \, T_{\mathcal{A}}$$

by Lemma 3.6. The error when approximating $S'_{\mathcal{A}}T_{\mathcal{A}}$ by $S''_{\mathcal{A}}T_{\mathcal{A}}$ is bounded by

$$| \, S'_{\mathcal{A}}T_{\mathcal{A}} - S''_{\mathcal{A}}T_{\mathcal{A}} \, | \leq 2^{\sigma-1}(| \, \theta_d \, | + | \, \theta_r \, |) \, (2 + 2^{\sigma-1}(| \, \theta_d \, | + | \, \theta_r \, |)) \, S''_{\mathcal{A}}T_{\mathcal{A}}$$

by Lemma 3.8. As neither $S''_{\mathcal{A}}T_{\mathcal{A}}$ nor the error terms depend on e, we may sum over all r elements $y = [e] \, g \in \mathbb{G}$ by multiplying by r. Hence $\widetilde{P}(\theta_d, \theta_r) = rS''_{\mathcal{A}}T_{\mathcal{A}}$ is an approximation to $P(\theta_d, \theta_r)$. The error that arises in this approximation is given by $| \, P(\theta_d, \theta_r) - \widetilde{P}(\theta_d, \theta_r) \, | \leq \tilde{e}(\theta_d, \theta_r)$ where

$$\begin{aligned}
\tilde{e}(\theta_d, \theta_r) &\leq r \, | \, S_{\mathcal{R}} - S''_{\mathcal{A}}T_{\mathcal{A}} \, | \\
&= r \, | \, (S_{\mathcal{R}} - S_{\mathcal{A}}T_{\mathcal{A}}) + (S_{\mathcal{A}}T_{\mathcal{A}} - S'_{\mathcal{A}}T_{\mathcal{A}}) + (S'_{\mathcal{A}}T_{\mathcal{A}} - S''_{\mathcal{A}}T_{\mathcal{A}}) \, | \\
&\leq r \, | \, S_{\mathcal{R}} - S_{\mathcal{A}}T_{\mathcal{A}} \, | + rT_{\mathcal{A}} \, | \, S_{\mathcal{A}} - S'_{\mathcal{A}} \, | + rT_{\mathcal{A}} \, | \, S'_{\mathcal{A}} - S''_{\mathcal{A}} \, | \\
&\leq 2^{-2m-\sigma+4} \, r + 2^{-2m-3\ell+2\sigma+3} \, rT_{\mathcal{A}} + \\
&\qquad 2^{\sigma-1}(| \, \theta_d \, | + | \, \theta_r \, |) \, (2 + 2^{\sigma-1}(| \, \theta_d \, | + | \, \theta_r \, |)) \, rS''_{\mathcal{A}}T_{\mathcal{A}} \\
&\leq \frac{2^4}{2^{m+\sigma}} + \frac{2^3}{2^{m+\ell}} + \frac{2^\sigma}{2}(| \, \theta_d \, | + | \, \theta_r \, |) \left(2 + \frac{2^\sigma}{2}(| \, \theta_d \, | + | \, \theta_r \, |) \right) \widetilde{P}(\theta_d, \theta_r)
\end{aligned}$$

by the triangle inequality, where we have used that $r < 2^m$ and $T_{\mathcal{A}} \leq 2^{2(\ell-\sigma)}$, and so the theorem follows. ■

In section 5 we demonstrate the soundness of this approximation. Before proceeding to the soundness analysis, however, we first need to understand how the pairs (θ_d, θ_r) are distributed in the plane.

4 The distribution of pairs (α_d, α_r)

In this section, we first identify and count all integers j that yield α_r and all pairs (j, k) that yield α_d. We then identify and count all pairs (j, k) that yield (α_d, α_r) and analyze the distribution and density of pairs (α_d, α_r) in the plane.

4.1 Pairs (j, k) yielding α_d

In this section we identify and count all pairs (j, k) that yield α_d.

Definition 4.1. *Let κ_d denote the greatest integer such that 2^{κ_d} divides d.*

Definition 4.2. *An argument α_d is said to be admissible if there exists a pair $(j, k) \in \mathbb{Z}^2$ where $0 \leq j < 2^{m+\ell}$ and $0 \leq k < 2^\ell$ such that $\alpha_d = \{dj + 2^m k\}_{2^{m+\ell}}$.*

Claim 4.1. *All admissible $\alpha_d = \{dj + 2^m k\}_{2^{m+\ell}}$ are multiples of 2^{κ_d}.*

Proof. As $2^{\kappa_d} \mid d < 2^m$ and the modulus is a power of two the claim follows. ■

Lemma 4.1. *The set of integer pairs (j, k) on $0 \leq j < 2^{m+\ell}$ and $0 \leq k < 2^\ell$ that yield the admissible argument α_d is given by*

$$j = \left(\frac{\alpha_d - 2^m k}{2^{\kappa_d}} \left(\frac{d}{2^{\kappa_d}} \right)^{-1} + 2^{m+\ell-\kappa_d} \, t_d \right) \mod 2^{m+\ell}$$

as t_d runs trough all integers on $0 \leq t_d < 2^{\kappa_d}$ and k runs trough all integers on $0 \leq k < 2^{\ell}$. Each admissible argument α_d hence occurs with multiplicity $2^{\ell+\kappa_d}$.

Proof. As $\alpha_d \equiv dj + 2^m k \mod 2^{m+\ell}$, it follows by solving for j that

$$j = ((\alpha_d - 2^m k)d^{-1} + 2^{m+\ell-\kappa_d}\, t_d) \mod 2^{m+\ell}$$

for k on $0 \leq k < 2^{\ell}$ and t_d on $0 \leq t_d < 2^{\kappa_d}$, and so the lemma follows. ■

4.2 Integers j yielding α_r

In this section we identify and count all integers j that yield α_r.

Definition 4.3. *Let κ_r denote the greatest integer such that 2^{κ_r} divides r.*

Definition 4.4. *An argument α_r is said to be admissible if there exists an integer j on $0 \leq j < 2^{m+\ell}$ such that $\alpha_r = \{rj\}_{2^{m+\ell}}$.*

Claim 4.2. *All admissible arguments $\alpha_r = \{rj\}_{2^{m+\ell}}$ are multiples of 2^{κ_r}.*

Proof. As $2^{\kappa_r} \mid r$ and the modulus is a power of two the claim follows. ■

Lemma 4.2. *The set of integers j on $0 \leq j < 2^{m+\ell}$ that yield the admissible argument α_r is given by*

$$j = \left(\frac{\alpha_r}{2^{\kappa_r}} \left(\frac{r}{2^{\kappa_r}} \right)^{-1} + 2^{m+\ell-\kappa_r} t_r \right) \mod 2^{m+\ell}$$

as t_r runs trough all integers on $0 \leq t_r < 2^{\kappa_r}$. Each admissible argument α_r hence occurs with multiplicity 2^{κ_r}.

Proof. As $\alpha_r \equiv rj \pmod{2^{m+\ell}}$, it follows by solving for j that

$$j = (\alpha_r r^{-1} + 2^{m+\ell-\kappa_r}\, t_r) \mod 2^{m+\ell}$$

for t_r an integer $0 \leq t_r < 2^{\kappa_r}$, and so the lemma follows. ■

4.3 The distribution of pairs (α_d, α_r)

In this section we analyze the distribution of pairs (α_d, α_r) in the plane.

Definition 4.5. *An argument pair (α_d, α_r) is said to be admissible if there exists an integer pair (j, k) on $0 \leq j < 2^{m+\ell}$ and $0 \leq k < 2^{\ell}$ such that*

$$\alpha_d = \{dj + 2^m k\}_{2^{m+\ell}} \quad \text{and} \quad \alpha_r = \{rj\}_{2^{m+\ell}}.$$

Claim 4.3. *The number of admissible argument pairs (α_d, α_r) is $2^{m+2\ell}$ when accounting for multiplicity as not all pairs need be distinct.*

Proof. The claim follows from Definition 4.5 as each pair (j, k) yields an argument pair for j and k integers on $0 \leq j < 2^{m+\ell}$ and $0 \leq k < 2^{\ell}$. ■

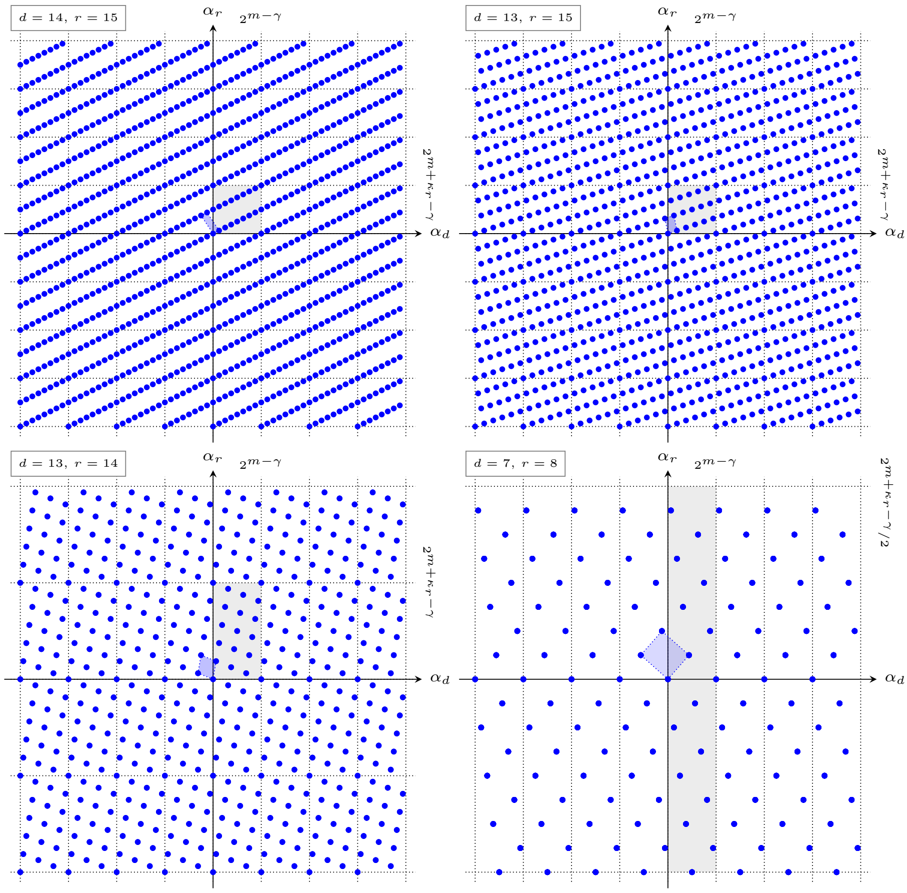

Fig. 2: The distribution of admissible arguments (α_d, α_r) in the region where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ for $m = 4$ and $\ell = 3$, and example combinations of d and r, as indicated. The lattice may be constructed by replicating the fundamental parallelogram (blue) or a rectangle (gray) of size $2^{m-\gamma} \times 2^{m+\kappa_r-\gamma}$.

Definition 4.6. *Let L^α be the lattice generated by the row span of*

$$\begin{bmatrix} \delta_r & 2^{\kappa_r} \\ 2^{m-\gamma} & 0 \end{bmatrix} \qquad \textit{where} \qquad \delta_r = d \left(\frac{r}{2^{\kappa_r}} \right)^{-1} \bmod 2^{m-\gamma}$$

and $\gamma = \max(0,\, \kappa_r - (\ell + \kappa_d))$.

Lemma 4.3. *The admissible argument pairs (α_d, α_r) are vectors in the region $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ in L^α. There are $2^{m+2\ell-\kappa_r+\gamma}$ distinct admissible argument pairs. Each admissible argument pair occurs with multiplicity $2^{\kappa_r-\gamma}$.*

Proof. As $\alpha_r \equiv rj \pmod{2^{m+\ell}}$ it follows that

$$j \equiv \alpha_r r^{-1} + 2^{m+\ell-\kappa_r} t_r \pmod{2^{m+\ell}}$$

for t_r an integer on $0 \leq t_r < 2^{\kappa_r}$. As $\alpha_d \equiv dj + 2^m k \pmod{2^{m+\ell}}$ this implies

$$\alpha_d \equiv d\,(\alpha_r r^{-1} + 2^{m+\ell-\kappa_r} t_r) + 2^m k$$

$$\equiv \alpha_r\, dr^{-1} + \underbrace{2^{m+\ell-\kappa_r+\kappa_d}\, \frac{dt_r}{2^{\kappa_d}}}_{\text{A}} + \underbrace{2^m k}_{\text{B}} \pmod{2^{m+\ell}} \tag{8}$$

for k an integer on $0 \leq k < 2^\ell$. As $2^{m-\gamma}$ is the smallest power of two to divide both 2^m and $2^{m+\ell-\kappa_r+\kappa_d}$, the congruence $\alpha_d \equiv \alpha_r\, dr^{-1} \pmod{2^{m-\gamma}}$ holds.

As t_r and k run through all pairwise combinations, the set of $2^{\ell+\kappa_r}$ arguments α_d generated by (8) is equal to that generated by

$$\begin{aligned} \alpha_d &\equiv \alpha_r\, dr^{-1} + 2^{m-\gamma} t_\gamma \qquad (9) \\ &\equiv \frac{\alpha_r}{2^{\kappa_r}} \left(d \left(\frac{r}{2^{\kappa_r}} \right)^{-1} \bmod 2^{m-\gamma} \right) + 2^{m-\gamma} t'_\gamma \pmod{2^{m+\ell}} \qquad (10) \end{aligned}$$

as t_γ, or equivalently t'_γ, runs through all integers on $0 \leq t_\gamma,\, t'_\gamma < 2^{\ell+\kappa_r}$.

To go from (8) to (9), first note that B runs through all values in $[2^m,\, 2^{m+\ell})$. If $\gamma = 0$, term A introduces multiplicity by repeating the sequence generated by B with various offsets. Such offsets are of no significance to this analysis as we only account for which values occur in the set and with what multiplicity.

If $\gamma > 0$, term A runs through all values in $[2^{m-\gamma},\, 2^{m-\gamma+\kappa_r})$. As $\kappa_r \geq \gamma$ when $\gamma > 0$, term A runs all values in the subrange $[2^{m-\gamma}, 2^m)$. When A assumes values greater than or equal to 2^m, it introduces multiplicity by repeating the sequence of all values on $[2^{m-\gamma}, 2^{m+\ell})$ generated by A and B with various offsets.

This implies that A + B modulo $2^{m+\ell}$ runs through all $2^{m+\ell}/2^{m-\gamma} = 2^{\ell+\gamma}$ values on $[2^{m-\gamma}, 2^{m+\ell})$ with multiplicity $2^{\ell+\kappa_r}/2^{\ell+\gamma} = 2^{\kappa_r-\gamma}$ and this is exactly what is stated in (9). To go from (9) to (10) is trivial.

As there are $2^{m+2\ell}$ admissible argument pairs, and as each pair occurs with multiplicity $2^{\kappa_r-\gamma}$, there are $2^{m+2\ell-\kappa_r+\gamma}$ distinct admissible argument pairs. The lattice L^α is constructed from (10), as the admissible α_r are multiples of 2^{κ_r}, and as the admissible $\alpha_d \equiv (\alpha_r \,/\, 2^{\kappa_r})\, \delta_r + 2^{m+\gamma} t'_\gamma \pmod{2^{m+\ell}}$, in the region of the plane where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$, and so the lemma follows. ■

In Fig. 2 the distribution of arguments in the region of the plane where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ is depicted for various combinations of parameters.

4.4 Pairs (j, k) yielding (α_d, α_r)

In this section we identify all pairs (j, k) that yield (α_d, α_r).

Lemma 4.4. *The set of integer pairs (j, k) for $0 \leq j < 2^{m+\ell}$ and $0 \leq k < 2^\ell$ that yield the admissible argument pair (α_d, α_r) is given by*

$$j = \left(\frac{\alpha_r}{2^{\kappa_r}} \left(\frac{r}{2^{\kappa_r}} \right)^{-1} + 2^{m+\ell-\kappa_r}\, t_r \right) \bmod 2^{m+\ell} \quad \textit{and} \quad k = \frac{\alpha_d - dj}{2^m} \bmod 2^\ell$$

as t_r runs through all integer multiples of 2^γ on $0 \leq t_r < 2^{\kappa_r}$.

Proof. As $\alpha_r \equiv rj \pmod{2^{m+\ell}}$ it follows by solving for j that

$$j = (\alpha_r r^{-1} + 2^{m+\ell-\kappa_r}\, t_r) \bmod 2^{m+\ell}$$

for t_r an integer $0 \leq t_r < 2^{\kappa_r}$.

As $\alpha_d \equiv dj + 2^m k \pmod{2^{m+\ell}}$, for compatibility 2^m must divide $2^{m+\ell-\kappa_r}\, dt_r$ for all $t_r \neq 0$. As $2^{m+(\ell+\kappa_d)-\kappa_r}$ is the greatest power of two to divide $2^{m+\ell-\kappa_r} d$, it follows that t_r must be a multiple of 2^γ, and so the lemma follows. ■

4.5 The density of pairs (α_d, α_r)

In this section we analyze the density of pairs (α_d, α_r) in the argument plane.

Claim 4.4. *The density of admissible argument pairs in the region of the plane where* $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ *is* 2^{-m} *when accounting for multiplicity.*

Proof. There are $2^{m+2\ell}$ admissible (α_d, α_r) when accounting for multiplicity in the region where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ of area $2^{2(m+\ell)}$. The density is hence $2^{m+2\ell}/2^{2(m+\ell)} = 2^{-m}$ and so the claim follows. ■

To construct the histogram for the probability distribution, the argument plane is divided into small rectangular subregions. The below lemma bounds the error when approximating the density in such subregions by 2^{-m}.

Lemma 4.5. *Let* D *be the density of admissible arguments pairs* (α_d, α_r)*, when accounting for multiplicity, in a rectangle* R *of area* A *and circumference* C *in the region where* $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ *of the plane. Then*

$$\left| D - \frac{1}{2^m} \right| \leq 2^{\kappa_r - \gamma} \, \frac{2C\lambda_2 + 4\,(2\lambda_2)^2}{A \, \det L^\alpha}$$

for λ_1 *the norm of the shortest vector* $\boldsymbol{w}_1 \in L^\alpha$*, and* λ_2 *the norm of the shortest vector* $\boldsymbol{w}_2 \in L^\alpha$ *that is linearly independent to* $\boldsymbol{w}_1$*.*

Proof. By Lemma 4.3, the admissible argument pairs (α_d, α_r) are vectors in L^α in the region of the argument plane where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$. Each admissible argument pair occurs with multiplicity $2^{\kappa_r - \gamma}$.

The fundamental parallelogram in L^α contains a single lattice vector. It is spanned by $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, and has area $\det L^\alpha = \lambda_2 \,|\, \boldsymbol{w}_\perp \,| = 2^{m+\kappa_r-\gamma}$, where $\boldsymbol{w}_\perp$ is the component in $\boldsymbol{w}_1$ perpendicular to $\boldsymbol{w}_2$. This implies $\lambda_2 \geq \lambda_1 \geq |\, \boldsymbol{w}_\perp \,|$.

To bound the number of argument pairs $(\alpha_d, \alpha_r) \in R$, we lower- and upper-bound the number of fundamental parallelograms that can at most fit into R, as described below, paying particular attention to the border areas:

To upper-bound the number of vectors in R, we extend each side of R by $2\,\lambda_2$ length units, to ensure that any parallelogram that is only partly in R is included in the count, and divide the area of the resulting rectangle by the area of the fundamental parallelogram. This yields $(A + 2C\lambda_2 + 4\,(2\,\lambda_2)^2) \,/\, \det L^\alpha$.

Conversely, to lower-bound the number of vectors in R, we retract each side of R by $2\lambda_2$ length units, to ensure that all parallelograms that are only partly in the rectangle are excluded from the count, and divide the area of the resulting rectangle by $\det L^\alpha$. This yields $(A - 2C\lambda_2 + 4\,(2\lambda_2)^2) \,/\, \det L^\alpha$.

By combining the upper and lower bounds, dividing by the area A of R, and multiplying by $2^{\kappa_r - \gamma}$ to account for multiplicity, the lemma follows. ■

For known d and r, Lemma 4.5 above provides a bound on the error when approximating the density in a rectangle in L^α by 2^{-m} as λ_2 may then be computed. To bound the error for general problem instances, and when d and r are unknown, we introduce the following less tight lemma:

Lemma 4.6. *Let* D *be the density of admissible argument pairs* (α_d, α_r)*, when accounting for multiplicity, in a rectangle of side lengths* l_d *and* l_r *in the* α_d *and*

α_r directions, respectively, in the region where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ of the argument plane. Then

$$\left| D - \frac{1}{2^m} \right| \leq \frac{2^{\kappa_r}}{2^m l_r} + \frac{1}{2^\gamma l_d} + \frac{1}{l_d l_r}.$$

Proof. By Lemma 4.3, the admissible argument pairs are vectors in L^α.

The vectors in L^α are on horizontal lines (for fixed α_r) evenly separated by a vertical distance of 2^{κ_r}. The number of such lines that intersect the rectangle is upper-bounded by $\lfloor l_r/2^{\kappa_r} \rfloor + 1 \leq l_r/2^{\kappa_r} + 1$ and lower-bounded by $\lfloor l_r/2^{\kappa_r} \rfloor \geq l_r/2^{\kappa_r} - 1$ as may be seen by positioning the rectangle to maximize or minimize the number of lines that intersect the rectangle.

On each line, the vectors in L^α are evenly spaced by a distance of $2^{m-\gamma}$ with varying starting positions. The number of vectors in L^α that fall within the rectangle on each line is upper-bounded by $\lfloor l_d/2^{m-\gamma} \rfloor + 1 \leq l_d/2^{m-\gamma} + 1$ and lower-bounded by $\lfloor l_d/2^{m-\gamma} \rfloor \geq l_d/2^{m-\gamma} - 1$, when not accounting for multiplicity, as may be seen by positioning the line to maximize or minimize the number of vectors that fall within the rectangle.

Hence the number of lattice vectors in the rectangle is upper-bounded by

$$2^{\kappa_r - \gamma}(l_r/2^{\kappa_r} + 1)(l_d/2^{m-\gamma} + 1) = l_d l_r/2^m + l_d 2^{\kappa_r}/2^m + l_r/2^\gamma + 1$$

and lower-bounded by

$$2^{\kappa_r - \gamma}(l_r/2^{\kappa_r} - 1)(l_d/2^{m-\gamma} - 1) = l_d l_r/2^m - l_d 2^{\kappa_r}/2^m - l_r/2^\gamma + 1$$

as each pair occurs with multiplicity $2^{\kappa_r - \gamma}$. By combining these bounds, and dividing by the area $l_d l_r$ of the rectangle, the lemma follows. ■

For unknown d and r, the above lemma provides an error bound, assuming only some bounds on the parameters κ_r and γ. Asymptotically, the error in the approximation tends to zero as the side lengths of the rectangle tend to infinity.

For rectangular subregions of specific dimensions, it may furthermore be shown that the error is zero, as is demonstrated in the following lemma:

Lemma 4.7. *The density of admissible argument pairs in a rectangle of side lengths positive integer multiples of $2^{m-\gamma}$ and $2^{m-\gamma+\kappa_r}$ in α_d and α_r, respectively, in the region where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$ of the argument plane, is 2^{-m} when accounting for multiplicity.*

Proof. By Lemma 4.3, the admissible arguments are vectors in L^α in the region of the argument plane where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$.

From the definition of L^α in Lemma 4.3, it follows that the lattice is cyclic with period $2^{m-\gamma}$ in α_d and $2^{m-\gamma+\kappa_r}$ in α_r. This is illustrated in Fig. 2 where rectangular regions of these dimensions are highlighted in gray. The highlighted regions all extend from the origin in Fig. 2 but the starting point may of course be arbitrarily selected. This implies that the lattice L^α may be generated by replicating and translating any rectangle of side lengths positive multiples of $2^{m-\gamma}$ and $2^{m-\gamma+\kappa_r}$ in α_d and α_r, respectively, see Fig. 2, throughout the plane. The same holds if the rectangle is replicated and translated cyclically throughout the region of the plane where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$.

The number of rectangles that fit in the region when replicated and translated cyclically is $2^{2(m+\ell)}/2^{2(m-\gamma)+\kappa_r} = 2^{2(\ell+\gamma)-\kappa_r}$ as the area of the region is

$2^{2(m+\ell)}$ and the area of the rectangle is $2^{2(m-\gamma)+\kappa_r}$. The total number of lattice vectors in the region is $2^{2m+\ell}$, so each rectangle contains $2^{2^{m+2\ell}/2(\ell+\gamma)-\kappa_r} = 2^{m-2\gamma+\kappa_r}$ vectors when accounting for multiplicity.

By dividing by the area of the rectangle, we see that the density of points in each rectangle is $2^{m-2\gamma+\kappa_r}/2^{2(m-\gamma)+\kappa_r} = 2^{-m}$, and so the lemma follows. ■

5 Soundness of the closed form approximation

In this section, we demonstrate the fundamental soundness of the closed form approximation to $P(\theta_d, \theta_r)$ that we derived in section 3.

5.1 Introduction and recapitulation

Recall that by Theorem 3.1 in section 3, the probability $P(\theta_d, \theta_r)$ of the quantum algorithm in section 2 yielding (j, k), with associated angle pair (θ_d, θ_r), summed over all r group elements $y = [e]\, g \in \mathbb{G}$, may be approximated by

$$\widetilde{P}(\theta_d, \theta_r) = \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} f(\theta_r)\, g(\theta_d, \theta_r)$$

where we have introduced some new notation in the form of the two functions

$$f(\theta_r) = \left| \sum_{n_r = 0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 \qquad g(\theta_d, \theta_r) = \left| \sum_{t = 0}^{2^{\ell-\sigma} - 1} \mathrm{e}^{i(2^\sigma \theta_d + \lceil -2^\sigma d/r \rfloor \theta_r) t} \right|^2$$

that we shall use throughout this section, and that may both be placed on closed form. The error when approximating $P(\theta_d, \theta_r)$ by $\widetilde{P}(\theta_d, \theta_r)$ is bounded by

$$\left| \, \widetilde{P}(\theta_d, \theta_r) - P(\theta_d, \theta_r) \, \right| \leq \tilde{e}(\theta_d, \theta_r),$$

again by Theorem 3.1, where the function $\tilde{e}(\theta_d, \theta_r)$ is specified.

5.1.1 Overview of the soundness argument

In what follows, we demonstrate the fundamental soundness of the above closed form approximation, by summing $\widetilde{P}(\theta_d, \theta_r)$ analytically to show that virtually all probability mass is within a specific region of the plane, and by summing $\tilde{e}(\theta_d, \theta_r)$ analytically to show that the total approximation error in this region is negligible. This implies that $\widetilde{P}(\theta_d, \theta_r)$ is a sound approximation.

Asymptotically, in the limit as m tends to infinity for fixed s, we furthermore show that the fraction of the probability mass captured tends to one whilst the error tends to zero. This implies that $\widetilde{P}(\theta_d, \theta_r)$ asymptotically captures the probability distribution completely and exactly.

5.2 Preliminaries

Before we proceed in accordance with the above outline, we first introduce some preliminary claims, lemmas and definitions in this section.

Lemma 5.1. *Let $\varphi \in \mathbb{R}$ and $\theta(u) = 2\pi u/2^\omega$ for $\omega > 0$ an integer. Then*

$$\sum_{u\,=\,0}^{2^{\omega+c-\varsigma}-1} \left| \sum_{t\,=\,0}^{N-1} \mathrm{e}^{i(2^\varsigma \theta(u)+\varphi)t} \right|^2 = 2^{\omega+c-\varsigma} N$$

for integers c, ς and N such that $c \geq 0$, $0 \leq \varsigma < \omega$ and $0 < N \leq 2^{\omega-\varsigma}$.

Proof. For any $\phi \in \mathbb{R}$ it holds that

$$\begin{aligned}
\left| \sum_{t\,=\,0}^{N-1} \mathrm{e}^{i(2^\varsigma \phi+\varphi)t} \right|^2 &= \left(\sum_{t\,=\,0}^{N-1} \mathrm{e}^{i(2^\varsigma \phi+\varphi)t} \right) \left(\sum_{t\,=\,0}^{N-1} \mathrm{e}^{-i(2^\varsigma \phi+\varphi)t} \right) \\
&= \sum_{t\,=\,-N+1}^{N-1} (N - |\,t\,|)\, \mathrm{e}^{i(2^\varsigma \phi+\varphi)t} \\
&= N + \sum_{t\,=\,1}^{N-1} (N-t)\, (\mathrm{e}^{i(2^\varsigma \phi+\varphi)t} + \mathrm{e}^{-i(2^\varsigma \phi+\varphi)t}).
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{u\,=\,0}^{2^{\omega+c-\varsigma}-1} \left| \sum_{t\,=\,0}^{N-1} \mathrm{e}^{i(2^\varsigma \theta(u)+\varphi)t} \right|^2 \\
&= \sum_{u\,=\,0}^{2^{\omega+c-\varsigma}\,-\,1} \left(N + \sum_{t\,=\,1}^{N-1} (N-t) \left(\mathrm{e}^{i(2^\varsigma \theta(u)+\varphi)t} + \mathrm{e}^{-i(2^\varsigma \theta(u)+\varphi)t} \right) \right) \\
&= 2^{\omega+c-\varsigma} N + \sum_{t\,=\,1}^{N-1} (N-t) \underbrace{\sum_{u\,=\,0}^{2^{\omega+c-\varsigma}\,-\,1} \left(\mathrm{e}^{i(2^\varsigma \theta(u)+\varphi)t} + \mathrm{e}^{-i(2^\varsigma \theta(u)+\varphi)t} \right)}_{=\,0}
\end{aligned}$$

as for any integer t on $0 < |\,t\,| < N \leq 2^{\omega-\varsigma}$ and $\varsigma < \omega$, the series

$$\sum_{u\,=\,0}^{2^{\omega+c-\varsigma}\,-\,1} \mathrm{e}^{i(2^\varsigma \theta(u)+\varphi)t} = \mathrm{e}^{i\varphi t} \frac{\mathrm{e}^{i 2^\varsigma (2\pi/2^\omega)\, 2^{\omega+c-\varsigma} t} - 1}{\mathrm{e}^{i 2^\varsigma (2\pi/2^\omega) t} - 1} = \mathrm{e}^{i\varphi t} \frac{\mathrm{e}^{2\pi i\, 2^c t} - 1}{\mathrm{e}^{2\pi i\, 2^{\varsigma-\omega} t} - 1} = 0$$

as the denominator is non-zero, and so the lemma follows. ■

5.2.1 Bounding tail regions

Claim 5.1. *For Δ and N integers such that $1 < \Delta < N$ it holds that*

$$\int_\Delta^N \frac{\mathrm{d}u}{u^2} < \sum_{z\,=\,\Delta}^{N-1} \frac{1}{z^2} < \int_{\Delta-1}^{N-1} \frac{\mathrm{d}u}{u^2} < \frac{1}{\Delta - 1} \leq \frac{2}{\Delta}.$$

Proof. As

$$\int_z^{z+1} \frac{\mathrm{d}u}{u^2} = \frac{1}{z+z^2} < \frac{1}{z^2} < \frac{1}{z^2 - z} = \int_{z-1}^{z} \frac{\mathrm{d}u}{u^2}$$

for z any integer such that $z > 1$, it follows that

$$\int_{\Delta}^{N} \frac{\mathrm{d}u}{u^2} = \sum_{z\,=\,\Delta}^{N-1} \left(\int_{z}^{z+1} \frac{\mathrm{d}u}{u^2} \right) < \sum_{z\,=\,\Delta}^{N-1} \frac{1}{z^2} < \sum_{z\,=\,\Delta}^{N-1} \left(\int_{z-1}^{z} \frac{\mathrm{d}u}{u^2} \right) = \int_{\Delta-1}^{N-1} \frac{\mathrm{d}u}{u^2}$$

where, for Δ and N integers on $1 < \Delta < N$, it holds that

$$\int_{\Delta-1}^{N-1} \frac{\mathrm{d}u}{u^2} = \frac{1}{\Delta - 1} - \frac{1}{N-1} \leq \frac{1}{\Delta-1} \leq \frac{2}{\Delta}$$

and so the claim follows. ■

Claim 5.2. *For $0 < |\,\phi\,| < \pi$ it holds that*

$$\left| \sum_{t\,=\,0}^{N-1} \mathrm{e}^{i\phi t} \right|^2 \leq \frac{2^4}{\phi^2}.$$

Proof. As $\phi \neq 0$, we have by Claim 5.3 below that

$$\left| \sum_{t\,=\,0}^{N-1} \mathrm{e}^{i\phi} \right|^2 = \left| \frac{\mathrm{e}^{iN\phi} - 1}{\mathrm{e}^{i\phi} - 1} \right|^2 \leq \frac{2^2}{|\,\mathrm{e}^{i\phi} - 1\,|^2} \leq \frac{2^4}{\phi^2}$$

and so the claim follows. ■

Claim 5.3. $|\,\mathrm{e}^{i\phi} - 1\,| \geq |\,\phi\,|/2$ *for any $\phi \in \mathbb{R}$ such that $|\,\phi\,| \leq \pi$.*

Proof. It suffices to show that $|\,\mathrm{e}^{i\phi} - 1\,|^2 = 2(1 - \cos\phi) \geq \phi^2/4$ from which the claim follows as $\cos\phi \leq 1 - \phi^2/8$ for any $\phi \in \mathbb{R}$ such that $|\,\phi\,| \leq \pi$. ■

5.2.2 Intervals of admissible arguments and angles

To facilitate the analysis, we furthermore introduce notation to handle intervals of admissible angles in the two below definitions.

Definition 5.1. *Let $\Theta_r(I)$ be the set of distinct admissible θ_r on the interval I.*

Definition 5.2. *For a fixed admissible θ_r, let $\Theta_d(I, \theta_r)$ be the set of distinct admissible θ_d on the interval I.*

5.2.3 Parameterizing the admissible arguments and angles

Furthermore, we introduce a convenient method for parameterizing the distinct admissible argument pairs (α_d, α_r), or angle pairs (θ_d, θ_r), in the below claim.

Claim 5.4. *The admissible argument α_d and α_r may be parameterized by*

$$\alpha_d(u_d, u_r) = (\delta_r u_r \mod 2^{m-\gamma}) + 2^{m-\gamma} u_d \qquad\qquad \alpha_r(u_r) = 2^{\kappa_r} u_r$$

and the corresponding admissible angles θ_d and θ_r may be parameterized by

$$\theta_d(u_d, u_r) = \frac{2\pi}{2^{m+\ell}}\, \alpha_d(u_d, u_r) \qquad\qquad \theta_r(u_r) = \frac{2\pi}{2^{m+\ell}}\, \alpha_r(u_r)$$

for u_d and u_r integers on

$$-2^{\ell+\gamma-1} \leq u_d < 2^{\ell+\gamma-1} \qquad\qquad -2^{m+\ell-\kappa_r-1} \leq u_r < 2^{m+\ell-\kappa_r-1}$$

when not accounting for multiplicity.

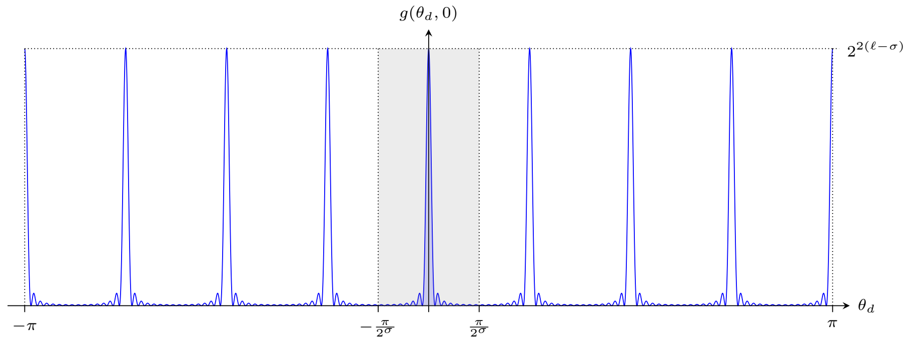

Fig. 3: The function $g(\theta_d, 0)$ plotted continuously in θ_d on the interval $|\,\theta_d\,| \leq \pi$ for $\sigma = 3$ and sample parameters selected to make the figure readable.

Proof. By Lemma 4.3 the admissible arguments (α_d, α_r) are in the region of the lattice L^α introduced in Definition 4.6 where $-2^{m+\ell-1} \leq \alpha_d, \alpha_r < 2^{m+\ell-1}$.

The parameterization takes u_r times the first row and u_d times second row of the basis matrix for L^α. It furthermore uses the second row to reduce the starting point $\delta_r u_r$ modulo $2^{m-\gamma}$. The claim follows from this analysis. ■

5.3 Establishing a baseline

In this section, we show that the sum of $\widetilde{P}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the region where $-\pi \leq \theta_r < \pi$ and $-\pi/2^\sigma \leq \theta_d < \pi/2^\sigma$, tends to one asymptotically in the limit as m tends to infinity for fixed s.

5.3.1 The inner sum over $g(\theta_d, \theta_r)$

Lemma 5.2. *For $\theta_r \in \Theta_r([-\pi, \pi))$ the inner sum*

$$\sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} g(\theta_d, \theta_r) = 2^{2(\ell-\sigma)+\gamma}$$

for $g(\theta_d, \theta_r)$ as defined in section 5.

Proof. The function $g(\theta_d, \theta_r)$ is non-negative and periodic in θ_d for fixed θ_r. It cycles exactly 2^σ times on the interval $-\pi \leq \theta_d < \pi$, as may be seen in Fig. 3 where $g(\theta_d, \theta_r)$ is plotted continuously in θ_d for θ_r fixed to zero. Fixing a different value of θ_r shifts the graph cyclically along the θ_d axis.

This implies that we may parameterize θ_d in u_d and u_r using Claim 5.4 and sum $\theta_d(u_d, u_r)$ over any consecutive sequence of $2^{\ell+\gamma-\sigma}$ values of u_d for the fixed u_r given by θ_r to sum over all $\theta_d \in \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)$.

By using this approach and Lemma 5.1 we thus obtain

$$\sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} g(\theta_d, \theta_r)$$

$$= \sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} \left| \sum_{t\,=\,0}^{2^{\ell-\sigma}-1} \mathrm{e}^{i(2^\sigma \theta_d + \lceil -2^\sigma d/r \rceil \theta_r)t} \right|^2$$

$$= \sum_{u_d\,=\,0}^{2^{\ell+\gamma-\sigma}-1} \left| \sum_{t\,=\,0}^{2^{\ell-\sigma}-1} \mathrm{e}^{i(2^\sigma \theta_d(u_d,u_r) + \lceil -2^\sigma d/r \rceil \theta_r(u_r))t} \right|^2$$

$$= \sum_{u_d\,=\,0}^{2^{\ell+\gamma-\sigma}-1} \left| \sum_{t\,=\,0}^{2^{\ell-\sigma}-1} \mathrm{e}^{i(2^\sigma(2\pi\, 2^{m-\gamma}\, u_d/2^{m+\ell}) + \varphi)t} \right|^2$$

$$= \sum_{u_d\,=\,0}^{2^{\ell+\gamma-\sigma}-1} \left| \sum_{t\,=\,0}^{2^{\ell-\sigma}-1} \mathrm{e}^{i(2\pi\,(2^\sigma u_d)/2^{\ell+\gamma} + \varphi)t} \right|^2 = 2^{\ell+\gamma-\sigma} \cdot 2^{\ell-\sigma}$$

where we have introduced the constant phase

$$\varphi = 2^\sigma(2\pi(\delta_r u_r \bmod 2^{m-\gamma})/2^{m+\ell}) + \lceil -2^\sigma d/r \rceil\, \theta_r(u_r) \in \mathbb{R}$$

and so the lemma follows. ■

5.3.2 The outer sum over $f(\theta_r)$

Lemma 5.3. *The outer sum*

$$\sum_{\theta_r\,\in\,\Theta_r([-\pi,\,\pi))} f(\theta_r) = 2^{m+\ell-\kappa_r} \left\lceil 2^{m+\ell}/r \right\rceil .$$

for $f(\theta_r)$ as defined in section 5.

Proof. The function $f(\theta_r)$ is non-negative and periodic in θ_r. It cycles exactly once on the interval $-\pi \leq \theta_r < \pi$.

This implies that we may parameterize θ_r in u_r using Claim 5.4, and sum over all $2^{m+\ell-\kappa_r}$ values of u_r to sum over all $\theta_r \in \Theta_r([-\pi, \pi))$. By using this approach and Lemma 5.1, we thus obtain

$$\sum_{\theta_r\,\in\,\Theta_r([-\pi,\,\pi))} f(\theta_r) = \sum_{u_r\,=\,0}^{2^{m+\ell-\kappa_r}-1} \left| \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r(u_r)n_r} \right|^2$$

$$= \sum_{u_r\,=\,0}^{2^{m+\ell-\kappa_r}-1} \left| \sum_{n_r\,=\,0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i(2\pi\, 2^{\kappa_r} u_r/2^{m+\ell})n_r} \right|^2$$

$$= 2^{m+\ell-\kappa_r} \left\lceil 2^{m+\ell}/r \right\rceil$$

and so the lemma follows. ■

5.3.3 Combined result

Lemma 5.4. *The combined sum over all distinct admissible (θ_d, θ_r), in the region where $-\pi/2^\sigma \leq \theta_d < \pi/2^\sigma$ and $-\pi \leq \theta_r < \pi$, is*

$$\sum_{\substack{\theta_r\,\in\,\Theta_r([-\pi,\,\pi)) \\ \theta_d\,\in\,\Theta_d([-\pi/2^\sigma,\,\pi/2^\sigma),\,\theta_r)}} \widetilde{P}(\theta_d, \theta_r) = 2^{\gamma-\kappa_r}\, \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil .$$

Proof. By combining Lemmas 5.2 and 5.3, we obtain

$$\begin{aligned}\sum_{\substack{\theta_r \,\in\, \Theta_r([-\pi,\pi)) \\ \theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\,\pi/2^\sigma),\,\theta_r)}} \widetilde{P}(\theta_d,\theta_r) &= \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \sum_{\theta_r \,\in\, \Theta_r([-\pi,\pi))} f(\theta_r) \sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\,\pi/2^\sigma),\,\theta_r)} g(\theta_d,\theta_r) \\ &= \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \cdot 2^{2(\ell-\sigma)+\gamma} \cdot 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \\ &= 2^{\gamma-\kappa_r} \cdot \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \end{aligned}$$

as the inner sum reduces to a constant, and so the lemma follows. ■

It follows from the above lemma that the sum of $\widetilde{P}(\theta_d, \theta_r)$ over all admissible pairs (θ_d, θ_r) in the region where $-\pi/2^\sigma \leq \theta_d < \pi/2^\sigma$ and $-\pi \leq \theta_r < \pi$ tends to one as m tends to infinity for fixed s, when accounting for the fact that each distinct admissible pair (θ_d, θ_r) occurs with multiplicity $2^{\kappa_r - \gamma}$ by Lemma 4.3.

The total approximation error, as upper-bounded by summing $\tilde{e}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the region, is non-negligible however. In the next section we address this problem by reducing the size of the region.

5.4 Adapting the region to reduce the error

In this section, we consider the central part of the limited region where $|\,\theta_d\,| \leq B_d$ and $|\,\theta_r\,| \leq B_r$, for B_d and B_r parameterized in τ. We show that the sum of $\widetilde{P}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r) with multiplicity in this central region captures a fraction of the probability mass in τ.

In the next section, we describe how the approximation error, as upper-bounded by summing $\tilde{e}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the central region, depends on τ. For appropriately selected σ and τ, we show that virtually all probability mass is in the central region, and that the total approximation error in the central region is negligible.

5.4.1 Defining the central region

Definition 5.3. *The central region is the region in the* (θ_d, θ_r)*—plane where* $|\,\theta_d\,| \leq B_d$ *and* $|\,\theta_r\,| \leq B_r$*, for* $B_d = 2^{\tau-\ell+1}\pi$ *and* $B_r = B_d/2$*, where* τ *is an integer constant such that* $1 < \tau < \ell - \sigma - 1$*.*

Note that by the above definition of B_d and B_r, all argument pairs (α_d, α_r) such that $|\,\alpha_d\,| \leq 2^{m+\tau}$ and $|\,\alpha_r\,| \leq 2^{m+\tau-1}$ are in the central region, and $B_r < B_d = 2^{\tau-\ell+1}\pi \leq 2^{-\sigma-1}\pi$, so the central region is a subregion of the limited region we considered in the previous section.

5.4.2 The inner sum over $g(\theta_d, \theta_r)$

Lemma 5.5. *For* $\theta_r \in \Theta_r([-B_r, B_r])$ *the inner sum*

$$\sum_{\theta_d \in \Theta_d([-B_d,\,B_d],\,\theta_r)} g(\theta_d, \theta_r) \geq 2^{2(\ell-\sigma)+\gamma} \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)$$

for $g(\theta_d, \theta_r)$ *as defined in section 5.*

Proof. First observe that for $I_d = [-\pi/2^\sigma,\, -B_d] \cup [B_d,\, \pi/2^\sigma]$ we have

$$\sum_{\theta_d \in \Theta_d([-B_d,\, B_d],\, \theta_r)} g(\theta_d,\, \theta_r) \geq 2^{2(\ell-\sigma)+\gamma} - \sum_{\theta_d \in \Theta_d(I_d,\, \theta_r)} g(\theta_d,\, \theta_r)$$

as $g(\theta_d, \theta_r)$ is non-negative and

$$\begin{aligned}\sum_{\theta_d \in \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} g(\theta_d,\, \theta_r) = & \sum_{\theta_d \in \Theta_d([-\pi/2^\sigma, -B_d),\, \theta_r)} g(\theta_d,\, \theta_r) + \\ & \sum_{\theta_d \in \Theta_d([-B_d,\, B_d],\, \theta_r)} g(\theta_d,\, \theta_r) + \\ & \sum_{\theta_d \in \Theta_d((B_d,\, \pi/2^\sigma),\, \theta_r)} g(\theta_d,\, \theta_r)\end{aligned}$$

where, by Lemma 5.2, the left hand sum

$$\sum_{\theta_d \in \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} g(\theta_d,\, \theta_r) = 2^{2(\ell-\sigma)+\gamma}.$$

To prove the lemma, we seek an upper bound to

$$\begin{aligned}\sum_{\theta_d \in \Theta_d(I_d,\, \theta_r)} g(\theta_d,\, \theta_r) &= \sum_{\theta_d \in \Theta_d(I_d,\, \theta_r)} g(\theta_d + \lceil -2^\sigma d/r \rceil\, \theta_r/2^\sigma, 0) \\ &\leq \sum_{\theta_d \in \Theta_d(I_d,\, \theta_r)} h(\theta_d + \lceil -2^\sigma d/r \rceil\, \theta_r/2^\sigma) \qquad (11)\end{aligned}$$

that is independent of θ_r, where we have used Claim 5.2 to obtain (11), and where we have introduced $h(\theta_d) = 2^4/(2^\sigma\theta_d)^2$ that is strictly decreasing in $|\, \theta_d \,|$.

The situation that arises is illustrated in Fig. 4 where $g(\theta_d, \theta_r)$ for θ_r fixed to zero is plotted continuously in θ_d, for θ_d on $|\, \theta_d \,| \leq \pi$ in the top graph, and on $|\, \theta_d \,| \leq \pi/2^\sigma$ in the middle graph.

Fixing a non-zero $\theta_r \in \Theta_r([-B_r, B_r))$ shifts the top and middle graphs in Fig. 4 cyclically by $\lceil -2^\sigma d/r \rceil\, \theta_r$. As $|\, \lceil -2^\sigma d/r \rceil\, \theta_r/2^\sigma \,| \leq B_r$, the maximum cyclic shift in θ_d is upper bounded by B_r, see the bottom graph in Fig. 4 where $g(\theta_d + B_r, 0)$ is plotted in yellow and $g(\theta_d - B_r, 0)$ in green.

To upper-bound (11) it therefore suffices to sum over all distinct admissible θ_d on $I_r = [-\pi/2^\sigma, -B_r] \cup [B_r, \pi/2^\sigma]$, as this captures all distinct admissible θ_d in the left and right tail regions under any cyclic shift. Hence

$$\begin{aligned}(11) &= \sum_{\theta_d \in \Theta_d(I_d,\, \theta_r)} h(\theta_d + \lceil -2^\sigma d/r \rceil\, \theta_r/2^\sigma) \\ &\leq \max_{\theta_r \,\in\, \Theta_r([-B_r,\, B_r])} \sum_{\theta_d \,\in\, \Theta_d(I_r,\, \theta_r)} h(\theta_d) \qquad (12) \\ &= \sum_{\theta_d \in \Theta_d(I_r,\, 0)} h(\theta_d) = \sum_{\theta_d \in \Theta_d([B_r,\, \pi/2^\sigma],\, 0)} 2h(\theta_d) \qquad (13)\end{aligned}$$

due to symmetry, where we have maximized the set of admissible θ_d over θ_r.

Recall that by Lemma 4.3 there is one distinct admissible α_d on the interval $[0, 2^{m-\gamma})$ for a given fixed α_r. Hence there is one distinct admissible θ_d on

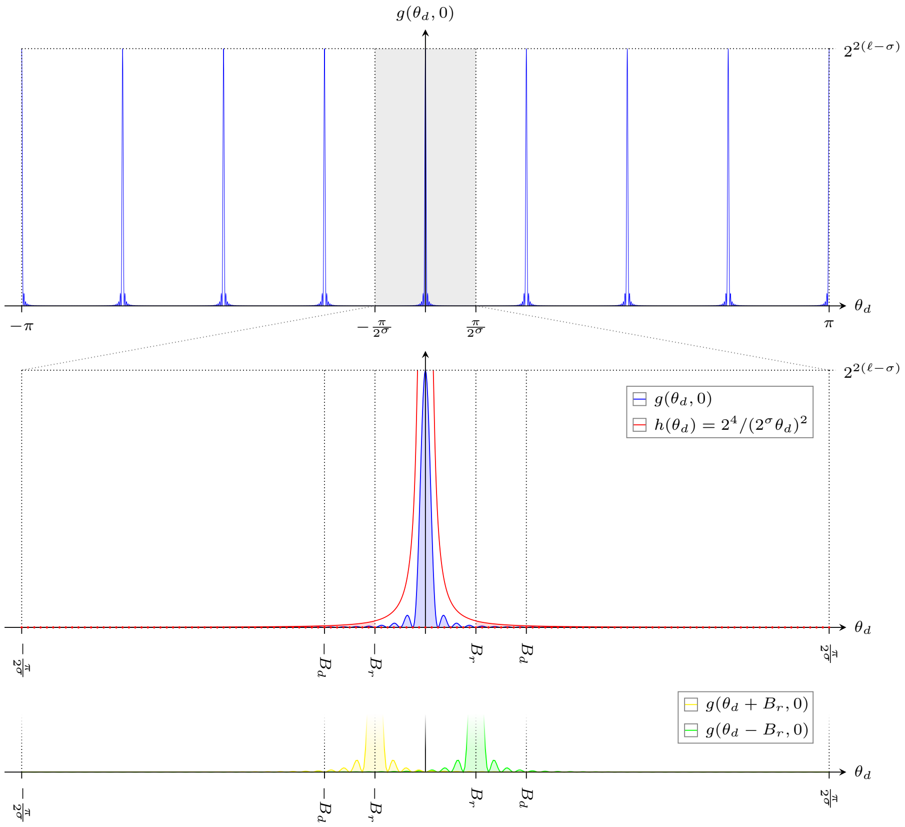

Fig. 4: The functions $g(\theta_d, 0)$ and $h(\theta_d) = 2^4/(2^\sigma\theta_d)^2$ plotted for $\sigma = 3$, $\ell = 9$ and $\tau = 3$. The maximum cyclic shift is bounded by $B_r = B_d/2$.

the interval $[0, 2^{-\ell-\gamma+1}\pi)$ for a given fixed θ_r. All other distinct admissible θ_d spread out from the starting point, equidistantly separated by a distance of $2^{-\ell-\gamma+1}\pi$. The distinct admissible θ_d may occur with multiplicity; however all distinct admissible θ_d occur with the same multiplicity, again see Lemma 4.3.

This implies that the sum in (12) is maximized for θ_r equal to zero, as both endpoints of the interval $B_r \leq |\,\theta_d\,| \leq \pi/2^\sigma$ are then admissible, maximizing both the number of distinct admissible θ_d on the interval, and the contribution from each distinct admissible θ_d as $h(\theta_d)$ is strictly decreasing in $|\,\theta_d\,|$.

By Claim 5.4, the distinct admissible θ_d may be parameterized in u_d and u_r where $\theta_d(u_d, u_r) = 2\pi\,((\delta_r u_r \bmod 2^{m-\gamma}) + 2^{m-\gamma}u_d)/2^{m+\ell}$. Now $\theta_r = 0$ implies $u_r = 0$, which in turn implies $2^{\tau-\ell}\pi = B_d/2 = B_r \leq 2\pi\,u_d/2^{\ell+\gamma} \leq \pi/2^\sigma$, or more succinctly $2^{\tau+\gamma-1} \leq u_d \leq 2^{\ell+\gamma-\sigma-1}$, which yields

$$(13) = \sum_{u_d\,=\,2^{\tau+\gamma-1}}^{2^{\ell+\gamma-\sigma-1}} 2h(\theta_d(u_d, u_r)) = \sum_{u_d\,=\,2^{\tau+\gamma-1}}^{2^{\ell+\gamma-\sigma-1}} \frac{2^5}{(2^\sigma \cdot 2\pi\,u_d/2^{\ell+\gamma})^2}$$

$$= 2^{2(\ell-\sigma+\gamma)}\,\frac{2^3}{\pi^2} \sum_{u_d\,=\,2^{\tau+\gamma-1}}^{2^{\ell+\gamma-\sigma-1}} \frac{1}{u_d^2} \leq 2^{2(\ell-\sigma+\gamma)}\,\frac{2^3}{\pi^2}\,\frac{2}{2^{\tau+\gamma-1}} = 2^{2(\ell-\sigma)+\gamma}\,\frac{2^5}{\pi^2}\,\frac{1}{2^\tau}$$

where we have used Claim 5.1 and that $\gamma \geq 0$ and $\tau > 1$. This implies

$$\sum_{\theta_d \in \Theta_d([-B_d, B_d],\, \theta_r)} g(\theta_d, \theta_r) \geq 2^{2(\ell-\sigma)+\gamma} - 2^{2(\ell-\sigma)+\gamma} \frac{2^5}{\pi^2} \frac{1}{2^\tau} = 2^{2(\ell-\sigma)+\gamma} \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)$$

and so the lemma follows. ■

5.4.3 The outer sum over $f(\theta_r)$

Lemma 5.6. *The outer sum*

$$\sum_{\theta_r \,\in\, \Theta_r([-B_r,\, B_r])} f(\theta_r) \geq 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)$$

for $f(\theta_r)$ as defined in section 5.

Proof. First observe that for $I_r = [-\pi, -B_r] \cup [B_r, \pi]$ it holds that

$$\sum_{\theta_r \,\in\, \Theta_r([-B_r,\, B_r])} f(\theta_r) \geq 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil - \sum_{\theta_r \,\in\, \Theta_r(I_r)} f(\theta_r)$$

as $f(\theta_r)$ is non-negative and

$$\sum_{\theta_r \,\in\, \Theta_r([-\pi,\, \pi))} f(\theta_r) = \sum_{\theta_r \,\in\, \Theta_r([-\pi,\, -B_r))} f(\theta_r) + \sum_{\theta_r \,\in\, \Theta_r([-B_r,\, B_r])} f(\theta_r) + \sum_{\theta_r \,\in\, \Theta_r((B_r,\, \pi))} f(\theta_r)$$

where, by Lemma 5.3, the left hand sum

$$\sum_{\theta_r \,\in\, \Theta_r([-\pi,\, \pi))} f(\theta_r) = 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil .$$

To prove the lemma, we seek an upper bound to

$$\sum_{\theta_r \in \Theta_r(I_r)} f(\theta_r) \leq \sum_{\theta_r \in \Theta_r(I_r)} \frac{2^4}{\theta_r^2} \leq \sum_{\theta_r \in \Theta_r(B_r \leq \theta_r \leq \pi)} \frac{2^5}{\theta_r^2} \tag{14}$$

where we have used Claim 5.2, that $f(\theta_r)$ is symmetric around the origin, and that the distinct admissible θ_r are equidistantly separated by a distance of 2^{κ_r} around the origin by Lemma 4.3. The distinct admissible θ_r may occur with multiplicity; however all distinct admissible θ_r occur with the same multiplicity.

By Claim 5.4, the distinct admissible θ_r may be parameterized in u_r where $\theta_r(u_r) = 2\pi\,(2^{\kappa_r} u_r)/2^{m+\ell}$, which implies $2^{\tau-\ell}\pi = B_r \leq 2\pi\,(2^{\kappa_r} u_r)/2^{m+\ell} \leq \pi$, or more succinctly $2^{m+\tau-\kappa_r-1} \leq u_r \leq 2^{m+\ell-\kappa_r-1}$, which yields

$$\begin{aligned}(14) = \sum_{u_r \,=\, 2^{m+\tau-\kappa_r-1}}^{2^{m+\ell-\kappa_r-1}} \frac{2^5}{(2\pi\, 2^{\kappa_r} u_r/2^{m+\ell})^2} &= 2^{2(m+\ell-\kappa_r)} \frac{2^3}{\pi^2} \sum_{u_r \,=\, 2^{m+\tau-\kappa_r-1}}^{2^{m+\ell-\kappa_r-1}} \frac{1}{u_r^2} \\ \leq 2^{2(m+\ell-\kappa_r)} \frac{2^3}{\pi^2} \frac{2}{2^{m+\tau-\kappa_r-1}} = 2^{m+2\ell-\kappa_r} \frac{2^5}{\pi^2} \frac{1}{2^\tau} &\leq 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \frac{2^5}{\pi^2} \frac{1}{2^\tau}\end{aligned}$$

where we have used Claim 5.1 and that $\gamma \geq 0$ and $\tau > 1$. This implies

$$\begin{aligned} \sum_{\theta_r \in \Theta_r([-B_r,\, B_r])} f(\theta_r) &\geq 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil - 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \frac{2^5}{\pi^2} \frac{1}{2^\tau} \\ &= 2^{m+\ell-\kappa_r} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right) \end{aligned}$$

and so the lemma follows. ■

5.4.4 Combined result

Lemma 5.7. *The combined sum over all distinct admissible (θ_d, θ_r), in the central region where $|\,\theta_d\,| \leq B_d$ and $|\,\theta_r\,| \leq B_r$, is*

$$\sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} \widetilde{P}(\theta_d, \theta_r) \geq 2^{\gamma-\kappa_r} \, \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)^2 .$$

Proof. From Lemmas 5.5 and 5.6 it follows that

$$\begin{aligned} \sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} \widetilde{P}(\theta_d, \theta_r) &= \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \sum_{\theta_r \,\in\, \Theta_r([-B_r,\, B_r])} f(\theta_r) \sum_{\theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)} g(\theta_d, \theta_r) \\ &\geq \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \, 2^{m+\ell-\kappa_r} \cdot 2^{2(\ell-\sigma)+\gamma} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)^2 \\ &= 2^{\gamma-\kappa_r} \, \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)^2 \end{aligned}$$

and so the lemma follows. ■

5.5 Main soundness result

In this section, we combine the above results into our main soundness result.

5.5.1 Bounding the probability mass in the central region

Theorem 5.1. *The sum of $\widetilde{P}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the central region where $|\,\theta_d\,| \leq B_d$ and $|\,\theta_r\,| \leq B_r$, is bounded by*

$$\frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil \left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)^2 \leq \sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} 2^{\kappa_r-\gamma} \, \widetilde{P}(\theta_d, \theta_r) \leq \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil .$$

Proof. The theorem follows immediately by combining Lemmas 5.4 and 5.7. ■

The above theorem states that a constant fraction of the probability mass is located within the central region for fixed τ. The fraction of the probability mass that falls outside the central region decreases exponentially in τ.

5.5.2 Bounding the total error in the central region

Theorem 5.2. *The total error when approximating $P(\theta_d, \theta_r)$ by $\widetilde{P}(\theta_d, \theta_r)$, as upper-bounded by summing $\tilde{e}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the central region where $|\,\theta_d\,| \leq B_d$ and $|\,\theta_r\,| \leq B_r$, is bounded by*

$$\sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} 2^{\kappa_r - \gamma}\, \tilde{e}(\theta_d, \theta_r) \leq 2^{m+2\tau} D \left(\frac{2^6}{2^\sigma} + \frac{2^5}{2^\ell} \right) +$$

$$\frac{2^{\tau+\sigma+2}}{2^\ell} \pi \left(1 + \frac{2^{\tau+\sigma}}{2^\ell} \pi \right) \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil$$

where D is the density of admissible pairs (θ_d, θ_r) in the region.

Proof. The error when approximating $P(\theta_d, \theta_r)$ by $\widetilde{P}(\theta_d, \theta_r)$ is bounded by

$$\tilde{e}(\theta_d, \theta_r) \leq \frac{2^4}{2^{m+\sigma}} + \frac{2^3}{2^{m+\ell}} + \frac{2^\sigma}{2}(|\,\theta_d\,| + |\,\theta_r\,|) \left(2 + \frac{2^\sigma}{2}(|\,\theta_d\,| + |\,\theta_r\,|) \right) \widetilde{P}(\theta_d, \theta_r)$$

by Theorem 3.1. We sum $\tilde{e}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r) with multiplicity in the region where $|\,\theta_d\,| \leq B_d$ and $|\,\theta_r\,| \leq B_r$, where $B_d = 2^{\tau-\ell+1}\pi$ and $B_r = B_d/2$ by Definition 5.3. This is equivalent to summing over all admissible (α_d, α_r) with multiplicity in the region where $|\,\alpha_d\,| \leq 2^{m+\tau}$ and $|\,\alpha_r\,| < 2^{m+\tau-1}$.

As $m > 0$, and as $\tau > 1$ by Definition 5.3, the total area of this region is

$$(2 \cdot 2^{m+\tau} + 1)(2 \cdot 2^{m+\tau-1} + 1) = 2^{2(m+\tau)+1} + 2^{m+\tau+1} + 2^{m+\tau} + 1 \leq 2^{2(m+\tau+1)}$$

from which it follows that the region contains at most $2^{2(m+\tau+1)}D$ admissible pairs (θ_d, θ_r), where D is the density of admissible pairs with multiplicity.

If we furthermore use that $|\,\theta_d\,| + |\,\theta_r\,| \leq 2^{\tau-\ell+2}\pi$, this implies that

$$\sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r)) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d),\, \theta_r)}} 2^{\kappa_r - \gamma}\, \tilde{e}(\theta_d, \theta_r)$$

$$\leq 2^{2(m+\tau+1)} D \left(\frac{2^4}{2^{m+\sigma}} + \frac{2^3}{2^{m+\ell}} \right) + 2^{\tau-\ell+\sigma+2}\pi\,(1 + 2^{\tau-\ell+\sigma}\pi)\, \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil$$

$$\leq 2^{m+2\tau} D \left(\frac{2^6}{2^\sigma} + \frac{2^5}{2^\ell} \right) + \frac{2^{\tau+\sigma+2}}{2^\ell}\pi \left(1 + \frac{2^{\tau+\sigma}}{2^\ell}\pi \right) \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil$$

where we have used that

$$\sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} 2^{\kappa_r - \gamma}\, \widetilde{P}(\theta_d, \theta_r) \leq \frac{r}{2^{m+\ell}} \left\lceil \frac{2^{m+\ell}}{r} \right\rceil$$

by Theorem 5.1, and so the theorem follows. ■

By Lemmas 4.5 and 4.6, the density D of admissible argument pairs in the region is approximately 2^{-m} for random problem instances. Asymptotically, the density tends to 2^{-m} as m tends to infinity for fixed s by Lemma 4.6.

Furthermore, the density is exactly 2^m in rectangular regions of the plane of side length multiples of $2^{m-\gamma}$ and $2^{m-\gamma+\kappa_r}$ by Lemma 4.7. The region in Theorem 5.2 above may be adapted to meet these requirements.

To understand the implications of the above theorem for the bound on the total error in the central region, it remains to select σ to minimize the bound.

5.5.3 Selecting σ to minimize the total error in the central region

To select the integer parameter σ on $0 < \sigma < \ell$ so as to minimize the bound on the total error given in Theorem 5.2, we first approximate the error bound by

$$\underbrace{\frac{2^{2\tau+6}}{2^\sigma}}_{\epsilon_1} + \underbrace{\frac{2^{2\tau+5}}{2^\ell}}_{\epsilon_2} + \underbrace{\frac{2^{\tau+\sigma+2}}{2^\ell}\pi}_{\epsilon_3} + \underbrace{\left(\frac{1}{2} \frac{2^{\tau+\sigma+2}}{2^\ell}\pi \right)^2}_{\epsilon_4}$$

where we have used that $D \approx 2^{-m}$ and $\left(r/2^{m+\ell}\right) \left\lceil 2^{m+\ell}/r \right\rceil \approx 1$, with equality in the limit as m tends to infinity for fixed s.

For the approximation to be good, all error terms are required to be much smaller than one. This implies that ϵ_3 is much greater than ϵ_4 as $\epsilon_4 = (\epsilon_3/2)^2$. As ϵ_2 does not depend on σ, we seek to equate ϵ_1 and ϵ_3 using σ, which yields

$$\frac{2^{2\tau+6}}{2^\sigma} = \frac{2^{\tau+\sigma+2}}{2^\ell}\pi \quad \Rightarrow \quad \sigma = \left\lfloor \frac{1}{2} \left(\ell + \tau + 4 - \log_2 \pi \right) \right\rceil .$$

If σ is fixed accordingly, the bound on the total error as given by summing $\tilde{e}(\theta_d, \theta_r)$ analytically over all admissible (θ_d, θ_r) with multiplicity in the region where $|\, \theta_d \,| \le B_d$ and $|\, \theta_r \,| \le B_r$, is minimized. For this σ, the main error terms

$$\epsilon_1 \approx \epsilon_3 \approx \frac{2^{3\tau/2+4}}{2^{\ell/2}} \sqrt{\pi}. \tag{15}$$

For as long as $2^{3\tau/2+4}\sqrt{\pi}$ is much smaller than $2^{\ell/2}$, we hence expect the upper bound on the total error given in Theorem 5.2 to be negligible.

5.5.4 Asymptotic soundness results

Theorem 5.3. *For fixed s and τ, and $\sigma = \lfloor (\ell + \tau + 4 - \log_2 \pi)/2 \rceil$, the sum of $\widetilde{P}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the central region where $|\, \theta_d \,| \le B_d$ and $|\, \theta_r \,| \le B_r$, is bounded by*

$$\left(1 - \frac{2^5}{\pi^2} \frac{1}{2^\tau} \right)^2 \le \lim_{m \to \infty} \sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} 2^{\kappa_r - \gamma}\, \widetilde{P}(\theta_d, \theta_r) \le 1 \tag{16}$$

in the limit as m tends to infinity. The error $|\, P(\theta_d, \theta_r) - \widetilde{P}(\theta_d, \theta_r) \,| \le \tilde{e}(\theta_d, \theta_r)$ and the sum of $\tilde{e}(\theta_d, \theta_r)$ over the admissible (θ_d, θ_r) with multiplicity tends to

$$\lim_{m \to \infty} \sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r)) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d),\, \theta_r)}} 2^{\kappa_r - \gamma}\, \tilde{e}(\theta_d, \theta_r) = 0. \tag{17}$$

Proof. The bound in (16) follows immediately by taking the limit as m tends to infinity for fixed s and τ of the bound given in Theorem 5.1.

Analogously (17) follows by taking the limit, as m tends to infinity for fixed s and τ, and for σ as in the formulation of this theorem, of Theorem 5.2, where D tends to 2^{-m} in the limit by Lemma 4.6, and so the theorem follows. ■

The above theorem states that an arbitrarily great constant fraction of the probability mass may be captured asymptotically by expanding the region in τ.

As the bound on the error when approximating $P(\theta_d, \theta_r)$ by $\widetilde{P}(\theta_d, \theta_r)$ in the region tends to zero asymptotically, $\widetilde{P}(\theta_d, \theta_r)$ equals $P(\theta_d, \theta_r)$ asymptotically in the region, and all probability mass is in the region asymptotically when τ tends to infinity with m at a moderated rate. This implies that $\widetilde{P}(\theta_d, \theta_r)$ asymptotically captures the probability distribution completely and exactly. The below corollary formalizes these observations:

Corollary 5.1. *For fixed s, for $\tau = \lfloor \ell/6 \rceil$ and $\sigma = \lfloor (\ell + \tau + 4 - \log_2 \pi)/2 \rceil$, the sum of $\widetilde{P}(\theta_d, \theta_r)$ over all admissible (θ_d, θ_r), with multiplicity, in the central region where $|\,\theta_d\,| \le B_d$ and $|\,\theta_r\,| \le B_r$, tends to*

$$\lim_{m \to \infty} \sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r]) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d],\, \theta_r)}} 2^{\kappa_r - \gamma} \, \widetilde{P}(\theta_d, \theta_r) = 1 \tag{18}$$

in the limit as m tends to infinity. The error $|\,P(\theta_d, \theta_r) - \widetilde{P}(\theta_d, \theta_r)\,| \le \tilde{e}(\theta_d, \theta_r)$ and the sum of $\tilde{e}(\theta_d, \theta_r)$ over the admissible (θ_d, θ_r) with multiplicity tends to

$$\lim_{m \to \infty} \sum_{\substack{\theta_r \,\in\, \Theta_r([-B_r,\, B_r)) \\ \theta_d \,\in\, \Theta_d([-B_d,\, B_d),\, \theta_r)}} 2^{\kappa_r - \gamma} \, \tilde{e}(\theta_d, \theta_r) = 0. \tag{19}$$

Proof. The bound in (18) follows immediately by taking the limit as m tends to infinity for fixed s and τ of the bound given in Theorem 5.1.

Analogously (19) follows by taking the limit, as m tends to infinity for fixed s, and for σ and τ as in the formulation of this corollary, of Theorem 5.2, where D tends to 2^{-m} in the limit by Lemma 4.6. This is easy to see as the main error terms ϵ_1 and ϵ_3 in (15) tend to approximately

$$\lim_{m \to \infty} \frac{2^{3(\ell/6)/2+4}}{2^{\ell/2}} \sqrt{\pi} = \lim_{m \to \infty} \frac{2^{\ell/4+4}}{2^{\ell/2}} \sqrt{\pi} = \lim_{m \to \infty} \frac{2^4}{2^{\ell/4}} \sqrt{\pi} = 0,$$

and in the limit the requirement that $1 < \tau < \ell - \sigma - 1$ in Definition 5.3 is respected. The corollary follows from this analysis. ■

6 Simulating the quantum algorithm

In this section, we combine results from the previous sections to construct a high-resolution histogram for the probability distribution for given d and r.

Furthermore, we describe how the histogram may be sampled to simulate output from the quantum algorithm when executed on a quantum computer.

6.1 Constructing the histogram

To construct the high-resolution histogram, we subdivide the argument plane into regions and subregions, and integrate the closed form probability approximation and the associated error bound numerically in each subregion.

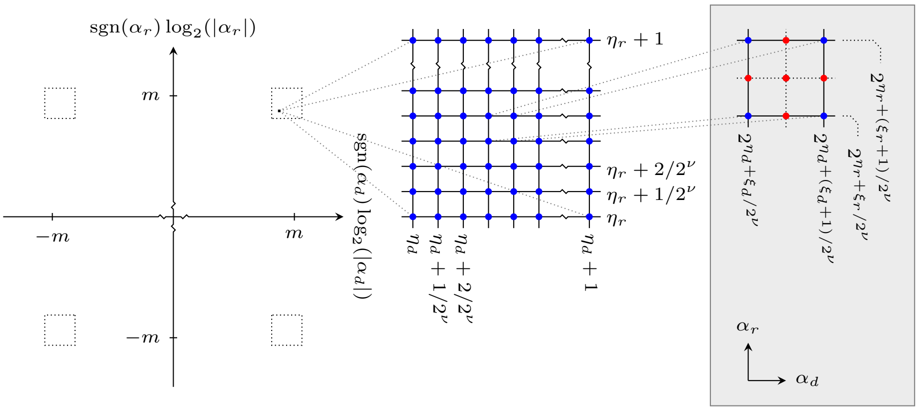

Fig. 5: The subdivision of the plane into regions and subregions. The gray box illustrates Simpson's rule applied to a subregion. The probability is computed in the blue corner points, the four red border midpoints and the red centerpoint.

First, we subdivide each quadrant of the argument plane into $(30+\mu)^2$ rectangular regions where $\mu = \min(\ell - 2, 11)$. Each region thus formed is uniquely identified by $(\eta_d, \eta_r) \in \mathbb{Z}^2$ by requiring that for all (α_d, α_r) in the region

$$2^{|\eta_d|} \le |\,\alpha_d\,| \le 2^{|\eta_d|+1} \quad \text{and} \quad 2^{|\eta_r|} \le |\,\alpha_r\,| \le 2^{|\eta_r|+1},$$

and furthermore $\mathrm{sgn}(\alpha_d) = \mathrm{sgn}(\eta_d)$ and $\mathrm{sgn}(\alpha_r) = \mathrm{sgn}(\eta_r)$, where η_d and η_r are such that $m - 30 \le |\,\eta_d\,|, |\,\eta_r\,| \le m + \mu - 1$, see the illustration in Fig. 5.

Then, we subdivide each region into rectangular subregions identified by an integer pair (ξ_d, ξ_r) by requiring that for all (α_d, α_r) in the subregion

$$2^{|\eta_d|+\xi_d/2^\nu} \le |\,\alpha_d\,| \le 2^{|\eta_d|+(\xi_d+1)/2^\nu} \quad \text{and} \quad 2^{|\eta_r|+\xi_r/2^\nu} \le |\,\alpha_r\,| < 2^{|\eta_r|+(\xi_r+1)/2^\nu}$$

where $0 \le \xi_d, \xi_r < 2^\nu$ for $\nu \in \{6, 7, 8, 9\}$ a resolution parameter adaptively selected as a function of the probability mass and variance in each region.

For each subregion, we compute the approximate probability mass contained within the subregion, and an associated error bound, by applying Simpson's rule in two dimensions, followed by Richardson extrapolation to cancel the linear error term, and division by 2^m to account for the density of pairs, see Lemma 4.5.

Simpson's rule is hence applied $2^{2\nu}(1 + 2^{2\nu})$ times in each region. Each application requires the approximate probability and associated error bound to be computed in up to nine points, for which purpose we use the closed form expressions in Theorem 3.1, with σ adaptively selected to minimize the bounded error. When tracking the error in each region, we take into account the error that arises when approximating the density in the region by 2^{-m} using Lemma 4.5.

6.2 Experiments and results

To illustrate the distribution that arises, a histogram is plotted in the signed logarithmic argument plane in Fig. 6 for $m = 2048$ and $s = 30$, and for d and r selected as explained in section 8.3. As expected, the histogram in Fig. 6 covers approximately 99.99% of the probability mass. The error that arises

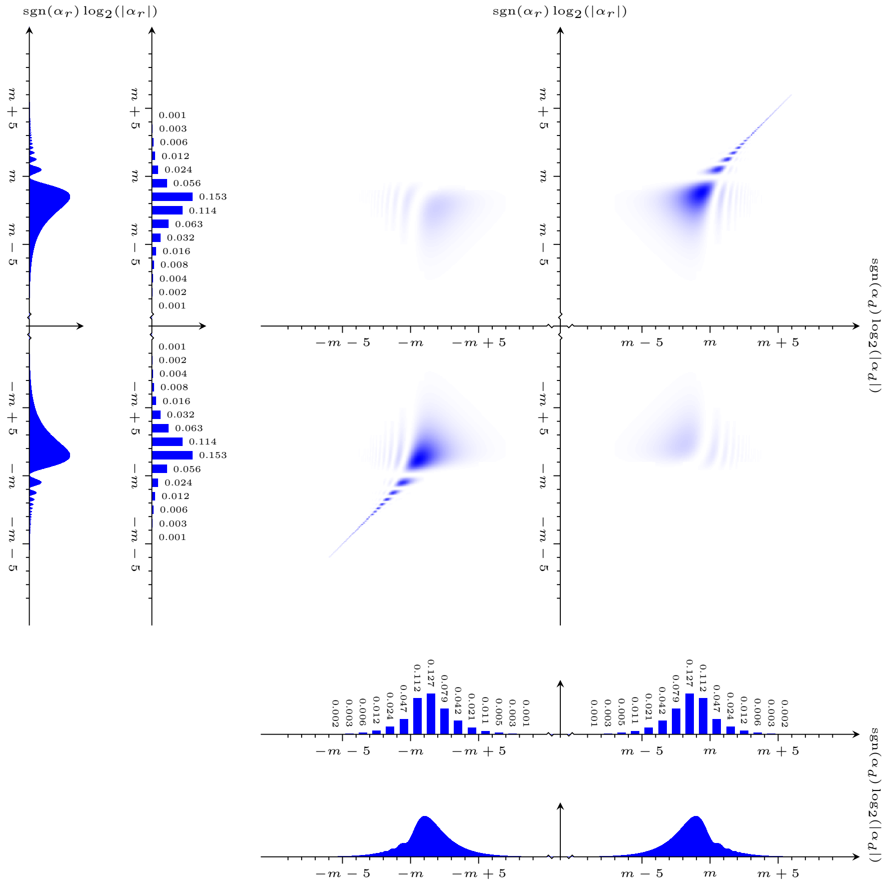

Fig. 6: The probability distribution for general discrete logarithms computed as in section 6.1 for $m = 2048$, $s = 30$, and d and r selected as in section 8.3. To facilitate printing, the resolution has been reduced in this figure.

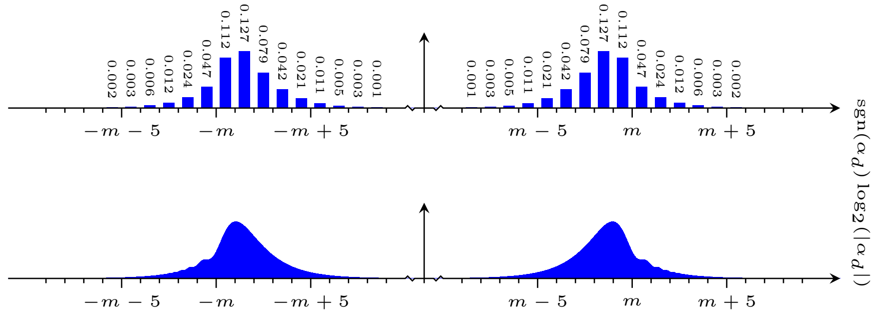

Fig. 7: The probability distribution for short discrete logarithms computed as in appendix B from the closed form expression in [5], for $m = 2048$ and $s = 30$, and d selected as in section 8.3. The resolution has been reduced in this figure.

when $P(\theta_d, \theta_r)$ is approximated by $\widetilde{P}(\theta_d, \theta_r)$ is negligible, as is the error that arises when approximating the density in the various subregions by 2^{-m}.

The histogram plotted in Fig. 6 captures the general characteristics of the distribution. Varying d and r on the interval $2^{m-1} < d < r < 2^m$, for d and r not divisible by large powers of two, in general only slightly affects the distribution. Scaling m and s has virtually no effect on the distribution.

The distribution is symmetric, in that the top right and lower left quadrants are mirrored, as are the top left and lower right quadrants. It hence suffices to compute only two quadrants to construct the histogram.

To see why this is, note that flipping the sign of both arguments in the expression for $\widetilde{P}(\theta_d, \theta_r)$ in Theorem 3.1 has no effect. Flipping the sign of only one argument, on the other hand, may lead to cancellation or lack of cancellation in the angle $\theta_d 2^\sigma + \theta_r \left\lceil -2^\sigma d/r \right\rceil$. This explains the concentration of probability mass in the top right and lower left quadrants, and in the tail that extends along the diagonal in Fig. 6 where $\theta_d 2^\sigma + \theta_r \left\lceil -2^\sigma d/r \right\rceil$ is small.

The marginal distribution along the horizontal α_d axis is virtually identical to the probability distribution induced by d when regarded as a short discrete logarithm, see [5] and Fig. 7 for comparison.

Analogously, the marginal distribution along the vertical α_r axis in Fig. 6 is virtually identical to the probability distribution induced by r when performing order finding, see appendix A and Fig. 8 for comparison. In section 6.3 below we show this analytically by summing $\widetilde{P}(\theta_d, \theta_r)$ over all admissible θ_d.

This implies that the lattice-based post-processing algorithm introduced in [5] may be used to solve sets of pairs (j, k) for both short and general d, with minor modifications, see section 7.1. An analogous lattice-based algorithm may be developed to solve sets of pairs (j, k) for r, see section 7.2.

6.3 Marginal distributions

By using results from the soundness analysis in section 5, we may immediately derive a closed form expression for the marginal distribution that arises when summing $\widetilde{P}(\theta_d, \theta_r)$ over all admissible θ_d with multiplicity.

Lemma 6.1. *For $\theta_r \in \Theta_r([-\pi, \pi])$, the marginal probability distribution that arises when summing $\widetilde{P}(\theta_d, \theta_r)$ over all $\theta_d \in \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)$ is*

$$\sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} \frac{2^{\kappa_r - \gamma}}{2^{\kappa_r}} \, \widetilde{P}(\theta_d, \theta_r) = \frac{r}{2^{2(m+l)}} \left| \sum_{n_r \,=\, 0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2$$

when accounting for multiplicity.

Proof. By Lemma 5.2 we have that

$$\begin{aligned} \sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} \widetilde{P}(\theta_d, \theta_r) &= \frac{2^{2\sigma} r}{2^{2(m+2\ell)}} \, f(\theta_r) \sum_{\theta_d \,\in\, \Theta_d([-\pi/2^\sigma,\, \pi/2^\sigma),\, \theta_r)} g(\theta_d, \theta_r) \\ &= \frac{2^\gamma r}{2^{2(m+\ell)}} \, f(\theta_r) = \frac{2^\gamma r}{2^{2(m+\ell)}} \left| \sum_{n_r \,=\, 0}^{\lceil 2^{m+\ell}/r \rceil - 1} \mathrm{e}^{i\theta_r n_r} \right|^2 \end{aligned}$$

from which the lemma follows, as the pairs (θ_d, θ_r) occur with multiplicity $2^{\kappa_r - \gamma}$ by Lemma 4.3, and the angles θ_r with multiplicity 2^{κ_r} by Lemma 4.2. ■

The above expression for the marginal probability distribution is derived from the approximation $\widetilde{P}(\theta_d, \theta_r)$. It corresponds to the exact expression derived in appendix A for the order finding algorithm with tradeoffs. The difference between the two expressions is explained by $\widetilde{P}(\theta_d, \theta_r)$ being an approximation to $P(\theta_d, \theta_r)$, whilst the expression in appendix A is exact.

A closed form analytical expression for the marginal distribution that arises when summing over all admissible θ_r is less straightforward to derive. Numerically, the marginal distribution may however be seen to correspond to that for short logarithms when $2^{m-1} < d < r < 2^m$ as pointed out in section 6.2.

6.4 Sampling the probability distribution

In this section, we describe how the histogram previously constructed is sampled so as to simulate the quantum algorithm.

6.4.1 Sampling argument pairs (α_d, α_r)

To sample an argument pair (α_d, α_r) from the histogram for the distribution, we first sample a subregion and then sample (α_d, α_r) from this subregion.

To sample the subregion, we first order all subregions in the histogram by probability, and compute the cumulative probability up to and including each subregion in the resulting ordered sequence. Then, we sample a pivot uniformly at random from $[0, 1)$, and return the first subregion in the ordered sequence for which the cumulative probability is greater than or equal to the pivot. Sampling fails if the pivot is greater than the total cumulative probability.

To sample an argument pair (α_d, α_r) from the subregion, we first sample a point $(\alpha'_d, \alpha'_r) \in \mathbb{Z}^2$ uniformly at random from the subregion. Then, we map (α'_d, α'_r) to the closest admissible argument pair $(\alpha_d, \alpha_r) \in L^\alpha$ by reducing the basis for L^α given in Definition 4.6 and applying Babai's algorithm [1].

6.4.2 Sampling integer pairs (j, k)

To sample an integer (j, k) from the distribution, we first sample an argument pair as described above, and then sample (j, k) uniformly at random from the set of all integer pairs (j, k) yielding (α_d, α_r) using Lemma 4.4.

More specifically, we first sample an integer t_r uniformly at random from the set of all admissible values for t_r and then compute (j, k) from (α_d, α_r) and t_r as described in Lemma 4.4

7 The classical post-processing algorithms

In this section, we describe classical post-processing algorithms for recovering the logarithm d and group order r from a set $\{(j_1, k_1), \ldots, (j_n, k_n)\}$ of n pairs produced by executing the quantum algorithm n times.

7.1 Recovering d from a set of n pairs

To recover d, the set of n pairs is used to form a vector

$$\mathbf{v}_d^k = (\,\{-2^m k_1\}_{2^{m+\ell}},\,\ldots,\,\{-2^m k_n\}_{2^{m+\ell}},\,0) \in \mathbb{Z}^D$$

and a D-dimensional integer lattice L^j with basis matrix

$$\begin{bmatrix} j_1 & j_2 & \cdots & j_n & 1 \\ 2^{m+\ell} & 0 & \cdots & 0 & 0 \\ 0 & 2^{m+\ell} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2^{m+\ell} & 0 \end{bmatrix}$$

where $D = n + 1$. For some constants $m_1, \ldots, m_n \in \mathbb{Z}$, the vector

$$\mathbf{u}_d^j = (\{dj_1\}_{2^{m+\ell}} + m_1 2^{m+\ell},\,\ldots,\,\{dj_n\}_{2^{m+\ell}} + m_n 2^{m+\ell},\,d) \in L^j$$

is such that the distance

$$R_d = |\,\mathbf{u}_d^j - \mathbf{v}_d^k\,| = \sqrt{\sum_{i=1}^{n} \big(\{dj_i\}_{2^{m+\ell}} + m_i 2^{m+\ell} - \{-2^m k_i\}_{2^{m+\ell}}\big)^2 + d^2}$$

$$= \sqrt{\sum_{i=1}^{n} \underbrace{\{dj_i + 2^m k_i\}_{2^{m+\ell}}^2}_{\alpha_{d,i}^2} + d^2} = \sqrt{\sum_{i=1}^{n} \alpha_{d,i}^2 + d^2}\,.$$

This implies that $\mathbf{u}_d^j$ and hence d may be found by enumerating all vectors in L^j within a D-dimensional hypersphere of radius R_d centered on $\mathbf{v}_d^k$.

The volume of such a hypersphere is

$$V_D(R_d) = \frac{\pi^{D/2}}{\Gamma\left(\frac{D}{2}+1\right)} R_d^D$$

where Γ is the gamma function.

For comparison, the fundamental parallelepiped in L^j contains a single lattice vector and is of volume $\det L^j = 2^{(m+\ell)n}$. Heuristically, we therefore expect the hypersphere to contain approximately $v_d = V_D(R_d) \,/\, \det L^j$ lattice vectors. The exact number depends on the placement of the hypersphere in $\mathbb{Z}^D$ and on the shape of the fundamental parallelepiped in L^j.

Assuming v_d to be sufficiently small for it to be computationally feasible to enumerate all lattice vectors in the hypersphere in practice, the above algorithm may be used to recover d. As the volume quotient v_d decreases in n, the number of vectors that need be enumerated may be reduced by running the quantum algorithm more times and including the resulting pairs in L^j.

However, there are limits to the number of pairs that may be included in L^j, as a reduced basis must be computed to enumerate L^j, and the complexity of computing such a basis grows rapidly in the dimension of L^j.

In this section, we show how to heuristically estimate the minimum number of runs n required to solve a specific known problem instance for d with minimum success probability q_d, for a given tradeoff factor s, and for a given bound on the number of vectors v_d that we at most accept to enumerate in L^j. Furthermore, we discuss different strategies for solving for d.

7.1.1 Estimating the minimum n required to solve for d

The radius R_d depends on the pairs (j_i, k_i) via the arguments $\alpha_{d,i}$. For fixed n and fixed probability q_d, we may estimate the minimum radius $\widetilde{R}_d$ such that

$$\Pr\left[R_d = \sqrt{\sum_{i=1}^{n} \alpha_{d,i}^2 + d^2} \leq \widetilde{R}_d \right] \geq q_d. \tag{20}$$

by sampling $\alpha_{d,i}$ from the probability distribution. This requires a histogram to be constructed for the probability distribution, and this in turn requires d and r to be known, and the tradeoff factor s to be fixed. For further details on how the estimate is computed, see section 7.3.

Equation (20) now implies that

$$\Pr\left[v_d = \frac{V_D(R_d)}{\det L^j} \leq \frac{V_D(\widetilde{R}_d)}{2^{(m+\ell)n}} \right] \geq q_d \tag{21}$$

providing an heuristic bound on the number of lattice vectors v_d that at most have to be enumerated, that holds with probability at least q_d.

Given an upper limit on the number of lattice vectors that we accept to enumerate, equation (21) may be used as an heuristic to estimate the minimum value of n such that v_d is below this limit with probability at least q_d.

To compute the estimate in practice, we use the heuristic to compute an upper bound on v_d for $n = 1, 2, \ldots$ and return the minimum n for which the bound is below the limit on the number of vectors that we accept to enumerate.

As the volume quotient v_d decreases by approximately a constant factor for every increment in n, the minimum n may be found efficiently via interpolation once the heuristic bound on v_d has been computed for a few values of n.

7.1.2 Selecting n and solving for d

A simple strategy when solving for d is to select n as described in section 7.1.1 such that v_d is below a bound equal to the maximum number of vectors that it is computationally feasible to enumerate with probability q_d.

This strategy seeks to minimize n, for a given problem instance and tradeoff factor, at the expense of potentially computationally expensive post-processing.

7.1.3 Selecting n and solving for d without enumerating

Another strategy is to select n such that $v_d < 2$ with probability q_d, so as to first minimize the number of vectors to enumerate and then minimize n.

By our heuristic, there is then only one vector in the hypersphere. In theory, this enables us to find $\mathbf{u}_d^j$ with probability q_d by mapping $\mathbf{v}_d^k$ to the closest vector in L^j without enumerating vectors in L^j. In practice, however, the situation is more complicated as $\mathbf{u}_r^j = (\{rj_1\}_{2^{m+\ell}}, \ldots, \{rj_n\}_{2^{m+\ell}}, r) \in L^j$ and this vector is short in L^j by construction. This is because $d + tr$ is a solution to the general discrete logarithm problem for t an integer. To recover $\mathbf{u}_d^j$, we therefore first map $\mathbf{v}_d^k$ to the closest vector in L^j, and then add or subtract small integer multiples of the shortest vector in the reduced basis to find $\mathbf{u}_d^j$.

In essence, this amounts to reducing the last component of the vector closest to $\mathbf{v}_d^k$ in L^j by r. However, as the last component of the shortest vector in L^j may be a factor in r, see section 7.2.3, we need to add and subtract multiples.

This complication occurs because we consider general discrete logarithms in groups of prime or composite order r. It does not occur when post-processing short discrete logarithms, see [5] and appendix B.

7.1.4 Selecting n and solving for d by exhausting subsets

Yet another plausible strategy is to independently post-process subsets of the pairs output by the quantum computer.

The greatest argument $\alpha_{d,i}$ essentially determines the bound on the radius R_d, see equation (20), and hence the bound on the number of vectors v_d, see equation (21), that need be enumerated. It may therefore be advantageous to run the quantum algorithm n times and to post-process all subsets of $n-t$ pairs from the resulting n pairs, for t a positive integer constant.

To select n when using this strategy, we specify a bound B on the number of vectors v_d that we at most accept to enumerate in each lattice of reduced dimension $n-t$, and proceed as described in section 7.1.1 to select the minimum n respecting this bound with minimum probability q_d. However, we only include the smallest $n-t$ arguments $\alpha_{d,i}$ when bounding the radius in section 7.3 and make all other necessary modifications to accommodate the reduced dimension.

The post-processing then requires at most B lattice vectors to be enumerated in at most $\binom{n}{t}$ lattices of dimension $n-t+1$. In general, t is limited to small values as the binomial coefficient grows rapidly in t, unless r is known in which case we may heuristically sort the subsets as is explained in the next section.

7.1.5 Sorting the subsets when r is known

If r is known, the argument $\alpha_{r,i} = \{rj_i\}_{2^{m+\ell}}$ is known for $1 \leq i \leq n$, and $\alpha_{r,i}$ provides information on $\alpha_{d,i}$ as the two arguments are correlated.

To see this, note that if $\alpha_{r,i}$ is in a certain position in the tail extending along the diagonal in Fig. 6 on page 36, then it is probable for $\alpha_{d,i}$ to also be in approximately this position in the tail, and vice versa. This may also be seen in (7) and in the expression for $\widetilde{P}(\theta_d, \theta_r)$ in Theorem 3.1 where constructive interference arises when α_r and $\alpha_d - d/r\,\alpha_r$ are both small arguments.

This implies that when constructing subsets of $n-t$ pairs from the n pairs, the pairs (j_i, k_i) should be included in ascending order as induced by $|\,\alpha_{r,i}\,|$. Pairs such that $|\,\alpha_{r,i}\,|$ exceeds some fixed bound may be rejected altogether.

The use of this sorting technique may significantly reduce the post-processing complexity as only the subsets most likely to yield d are solved for d. This technique furthermore enables t to assume large values, in turn enabling the number of runs n to be increased beyond what is possible when using either of the two previous strategies.

7.1.6 Identifying erroneous runs

The correlation between $\alpha_{d,i}$ and $\alpha_{r,i}$ may be used to detect erroneous runs. In general, we assume the quantum computer to execute the quantum algorithm correctly. In a setting where the quantum computer sometimes fails to correctly execute the algorithm, however, and where r is known, some erroneous runs may be identified by testing if $|\,\alpha_{r,i}\,| \leq B$ for B a bound selected so that the probability of observing $|\,\alpha_{r,i}\,| > B$ in a correct run is negligible.

This test is capable of detecting some errors that occur before or in the part of the quantum algorithm where j_i is read out. If j_i is read out before k_i, and an error is detected, the remainder of the computation may be aborted.

7.2 Recovering r from a set of n pairs

To recover r, we use that

$$\mathbf{u}_r^j = (\{rj_1\}_{2^{m+\ell}},\ \ldots,\ \{rj_n\}_{2^{m+\ell}},\ r) \in L^j$$

is a short vector by construction. More specifically, $\mathbf{u}_r^j$ is within a D-dimensional hypersphere in L^j of radius

$$R_r = |\,\mathbf{u}_r^j\,| = \sqrt{\sum_{i=1}^{n} \underbrace{\{rj_i\}_{2^{m+\ell}}^2}_{\alpha_{r,i}^2} + r^2} = \sqrt{\sum_{i=1}^{n} \alpha_{r,i}^2 + r^2}$$

centered at the origin. In analogy with the previous section, we may recover $\mathbf{u}_r^j$ and hence r by enumerating all vectors in this hypersphere. Heuristically, we expect the hypersphere to contain $v_r = V_D(R_r)\,/\det L^j$ lattice vectors.

7.2.1 Estimating the minimum n required to solve for r

The radius R_r depends on the integers j_i via the arguments $\alpha_{r,i}$. For fixed n and a fixed probability q_r, we may estimate the minimum radius $\widetilde{R}_r$ such that

$$\Pr\left[R_r = \sqrt{\sum_{i=1}^{n} \alpha_{r,i}^2 + r^2} \leq \widetilde{R}_r \right] \geq q_r \qquad (22)$$

by sampling $\alpha_{r,i}$ from the probability distribution. This requires a histogram to be constructed for the probability distribution, and this in turn requires d and r to be known, and the tradeoff factor s to be fixed. For further details on how the estimate is computed, see section 7.3.

Equation (22) now implies that

$$\Pr\left[v_r = \frac{V_D(R_r)}{\det L^j} \leq \frac{V_D(\widetilde{R}_r)}{2^{(m+\ell)n}} \right] \geq q_r. \qquad (23)$$

providing an heuristic bound on the number of lattice vectors v_r that we at most have to enumerate. This bound holds with probability at least q_r.

Given an upper limit on the number of lattice vectors that we accept to enumerate, equation (23) may be used as an heuristic to estimate the minimum value of n such that v_r is below this limit with probability at least q_r by proceeding in analogy with the procedure described in section 7.1.1.

7.2.2 Selecting n and solving for r

A simple strategy when solving for r is to select n as described in section 7.2.1 such that v_r is below a bound equal to the maximum number of vectors that it is computationally feasible to enumerate with probability q_r.

This strategy seeks to minimize n, for a given problem instance and tradeoff factor, at the expense of potentially computationally expensive post-processing.

7.2.3 Selecting n and solving for r without enumerating

Another strategy is to select n such that $v_r < 2$ with probability q_r, so as to first minimize the number of vectors to enumerate and then minimize n.

By our heuristic, there is then only one lattice vector in the hypersphere. In theory, this enables us to find $\mathbf{u}_r^j$ with probability q_r by computing the shortest vector in L^j. In practice, this is true in general when r is prime.

Assume the converse that r is composite. Let t be a non-trivial divisor of both r and $\alpha_{r,i}$ for $1 \leq i \leq n$. Then $\mathbf{u}_r^j/t \in L^j$ and $|\,\mathbf{u}_r^j/t\,| < |\,\mathbf{u}_r^j\,|$, so $\mathbf{u}_r^j/t$ and r/t are likely to be recovered by the algorithm instead of $\mathbf{u}_r^j$ and r.

For t a non-trivial divisor of r, the probability of t dividing $\alpha_{r,i}$ for $1 \leq i \leq n$ is approximately $(2^{\kappa_t}/t)^n$, for 2^{κ_t} the greatest power of two to divide t. This implies that r may in practice be recovered from r/t by exhausting t, as the search space in κ_t and $t/2^{\kappa_t}$ is small with great probability.

7.2.4 Selecting n and solving for r by exhausting subsets

A third strategy is to independently post-process subsets of the pairs output by the quantum computer, in analogy with the procedure described in section 7.1.5.

7.3 Estimating $\widetilde{R}_d$ and $\widetilde{R}_r$

To estimate $\widetilde{R}_d$ and $\widetilde{R}_r$ for m, s and n, explicit values of d and r, and a given target success probability q_d or q_r, we take the straightforward approach of sampling N sets of n argument pairs

$$\{(\alpha_{d,1}, \alpha_{r,1}),\, \ldots,\, (\alpha_{d,n}, \alpha_{r,n})\}$$

from the probability distribution as described in section 6.4.

For each set, we compute R_d, sort the resulting list of values in increasing order, and select the value at index $\lfloor (N-1)\, q_d \rceil$ to arrive at our estimate for $\widetilde{R}_d$. The estimate of $\widetilde{R}_r$ is then computed analogously.

The constant N controls the accuracy of the estimate. Assuming N to be sufficiently large in relation to q_d and q_r, and to the variance of the arguments, the above approach yields sufficiently good estimates.

7.3.1 On sampling failures

As is explained in section 6.4 the sampling of argument pairs may fail. This occurs when the pair being sampled is not in any of the regions of the argument plane covered by the histogram constructed for the probability distribution.

If the sampling of at least one pair in a set fails, we err on the side of caution and over-estimate $\widetilde{R}_d$ and $\widetilde{R}_r$ by letting $R_d = R_r = \infty$ for the set. The entries for the failed sets will then all be sorted to the end of the lists. If the values of $\widetilde{R}_d$ and $\widetilde{R}_r$ selected from the sorted lists are ∞ no estimate is produced.

Let p be the total probability mass covered by the histogram. Then the probability of all n argument pairs in a set being in regions covered by the histogram is p^n. When sampling N sets, the expected number of sets with finite R_d and R_r is Np^n. As Nq_d and Nq_r entries, respectively, in the two lists must be finite for the algorithm to produce an estimate, it follows that it is required that $q_d, q_r > p^n$, with some margin to account for the sampling variance, for estimates to be produced.

8 Simulating the complete algorithm

In this section, we estimate the minimum number of runs n required to attain a given minimum success probability q when recovering both d and r for specific problem instances and tradeoff factors s without enumerating the lattice.

8.1 Estimating n

To estimate n for a problem instance given by d, r and s, we proceed as follows:

For $n = s + 1,\ s + 2,\ \ldots$ we first estimate $\widetilde{R}_d$ and $\widetilde{R}_r$ by sampling $N = 10^6$ sets of n argument pairs (α_d, α_r), as explained in section 7.3. We stop and record the smallest n for which the volume quotients $v_d < 2$ and $v_r < 2$ with probability $q = q_d = q_r = 0.99$. As the volume quotients each decrease by approximately a constant factor for every increment in n, the minimum n may be found by interpolation.

For selected problem instances, we verify the above initial estimate of n by simulating the quantum algorithm and post-processing the simulated output. More specifically, we execute the following procedure:

With the initial estimate of n as our starting point, we sample $M = 10^3$ sets of n pairs (j, k), as explained in section 6.4, and test whether recovery of both d and r is successful for at least $\lceil Mq \rceil$ sets when executing the post-processing algorithms in sections 7.1 and 7.2 without enumerating L^j.

Depending on the outcome of the test, we either increment or decrement n, and repeat the process, recursively, until the smallest n such that the test passes has been identified. We record this n alongside the initial estimate of n.

8.2 Selecting m and s

The computational cost of estimating n for a given problem instance is non-negligible, as a high-resolution histogram for the probability distribution must be constructed for each instance. For some problem instances, we furthermore perform simulations to verify the initial estimate of n, and the computational cost of the lattice basis reductions that need be computed to conduct these simulations is non-negligible for large n.

To minimize the overall computational expense, we aim to minimize the number of problem instances considered, whilst at the same time seeking to capture most currently deployed instantiations of cryptologic schemes based on the intractability of the general discrete logarithm problem.

	group and logarithm size m							
tradeoff factor s	128	256	384	512	1024	2048	4096	8192
1	2	2	2	2	2	2	2	2
2	* 3	3	3	3	3	3	3	3
3	–	4	4	4	4	4	4	4
4	–	* 5	5	5	5	5	5	5
5	–	** 6	6	6	6	6	6	6
6	–	–	* 7	7	7	7	7	7
7	–	–	** 9	8	8	8	8	8
8	–	–	–	* 10	9	9	9	9
10	–	–	–	** 12	11	11	11	11
20	–	–	–	–	** 24	22	21	21
30	–	–	–	–	–	35	33 / 32	31
40	–	–	–	–	–	** 49 / 48	44	42
50	–	–	–	–	–	–	57	54 / 53
80	–	–	–	–	–	–	–	– / 88

Tab. 1: The estimated number of runs n required to solve for both a general discrete logarithm d and group order r, selected as described in section 8.3, with $\geq 99\%$ success probability and without enumerating the lattice. For details, see section 8.4. For A the initial and B the simulated estimate, we print B / A, unless B = A; we then only print A. Dash indicates no estimate. For ϵ the total error in the region, one asterisk indicates that $10^{-4} \leq \epsilon < 10^{-3}$ and two asterisks that $10^{-3} \leq \epsilon < 10^{-2}$. For all other entries with estimates $\epsilon < 10^{-4}$.

To this end, for $m \in \{128, 256, \ldots, 8192\}$, we select a single combination of d and r using the method described in section 8.3, and estimate n for a subset of tradeoff factors $s \in \{1, 2, \ldots, 8, 10, 20, \ldots, 50, 80\}$, such that the bounded error in the regions included in the histogram is negligible. Furthermore, we consider $m = 384$ as this is a common elliptic curve group size.

In terms of group size, the above choices of m capture most currently widely deployed elliptic curve groups, Schnorr groups and safe-prime groups.

8.3 Selecting d and r given m

For each value of m, we explicitly select d and r such that $2^{m-1} \leq d < r < 2^m$.

As long as d and r do not have very special properties, such as being divisible by large powers of two or otherwise being smooth, the exact values of d and r are of no great significance. To avoid having to tabulate d and r for m, we read d and r from the decimal expansion of Catalan's constant

$$G = \sum_{i\,=\,0}^{\infty} \frac{(-1)^i}{(2i+1)^2} = \frac{1}{1^2} - \frac{1}{3^2} + \frac{1}{5^2} - \frac{1}{7^2} + \cdots .$$

For a given m on $256 \leq m \leq 8192$, we define $c_{m,i}$ to be the integer formed from the first $m-1$ bits in the i^{th} chunk of 8191 bits from the decimal expansion of G and select $r = 2^{m-1} + c_{m,0}$ and $d = 2^{m-1} + (c_{m,1} \bmod c_{m,0})$.

8.4 Experiments and results

Executing the above experiments yielded the estimates of n in Tab. 1.

In general, the initial estimates of n are verified by the simulations. The volume quotient v_d is, in general, greater than v_r, so v_d determines the initial estimate for n. When the initial estimate of n is such that v_d is close to two, minor discrepancies between the initial estimates and the simulations tend to arise. This is because the volume quotient is only an heuristic metric for the number of lattice vectors that need be enumerated. In particular, the heuristic metric neglects to account for the shape of the fundamental parallelepiped in L^j when comparing its volume to that of a D-dimensional hypersphere. This is not a good heuristic if the quotient between the two volumes is close to two.

For fixed m and s, the volume quotient v_d decreases by an approximately constant factor in n. As s grows for a fixed m, the factor whereby v_d decreases in n becomes increasingly smaller, giving rise to further discrepancies, as v_d is often close to two, and as increasing or decreasing n still keeps v_d close to two. The discrepancies are further amplified by the difference in the sample size N used to compute the initial estimates and M used to perform the simulations.

For a fixed m, the general pattern in Tab. 1 is that for $s = 1, 2, \ldots,$ the estimate $n = s + 1$ up to a cutoff point in s. For greater s, the estimate of n begins to diverge from this pattern. The greater the value of m, the greater the value of s for which divergence occurs.

For a fixed m, the upper bound on the error given in Theorems 3.1 and 5.2 grows in s. We only compute estimates and perform simulations for combinations of m and s such that the error is negligible.

8.4.1 Related results

The marginal distributions plotted along the axes in Fig. 6 on page 36 agree with the distributions induced by the quantum algorithm for computing short discrete logarithms, see Fig. 7 on page 36, and by the quantum algorithm for computing orders with tradeoffs, see Fig. 8 on page 53 in appendix A.

As v_d is greater than v_r in general, we expect the estimates of n for the algorithm for computing general discrete logarithms to agree with the estimates of n for the algorithm for computing short discrete logarithms. This is indeed the case, see Tab. 4 on page 56 in appendix B, where n has been estimated for short discrete logarithms d selected as in section 8.3.

It is reasonable to assume that this is would be the case also for combinations of m and s that are outside the range of our approximation.

8.4.2 Implementation remarks

In practice, we compute the closest vector in the lattice L^j by reducing the basis for the lattice and applying Babai's [1] nearest plane algorithm. The shortest vector in L^j is taken to be the shortest vector in the reduced basis for L^j.

To reduce the lattice basis, we employ the block Korkin-Zolotarev (BKZ) algorithm [7, 11], as it is implemented in fpLLL v5.0, with default parameters and a block size of $\min(n + 1, 10)$ for all combinations of m, s and n. We first compute a Lenstra-Lenstra-Lovàsz (LLL) [9] reduction. Only if the LLL reduction proves insufficient do we proceed to compute a BKZ reduction.

9 Other applications and algorithms

In this section, we discuss some other applications of quantum algorithms for computing general discrete logarithms and orders. Furthermore, we relate our algorithms to other recently proposed quantum algorithms.

9.1 Order finding

The algorithm for computing discrete logarithms introduced in this paper does not require the group order to be known, as neither the quantum algorithm nor the classical post-processing algorithm makes explicit use of the order. If the order of the group is unknown, it may be computed from the same set of pairs (j, k) output by the quantum computer as is used to compute the logarithm.

This implies that the algorithm may be used as an order finding algorithm. When only the order is of interest, only j need be computed, as k is not used by the post-processing algorithm that recovers the order. The second stage of the quantum algorithm where k is computed need therefore not be executed when the goal is to perform order-finding. If the second stage is removed, the quantum algorithm reduces to the algorithm proposed by Seifert [12]. For s equal to one, this algorithm in turn reduces to Shor's order finding algorithm.

This provides a link between our works on computing discrete logarithms, Seifert's work on order finding, and Shor's original work. For post-processing, Seifert generalize Shor's continued fractions-based post-processing algorithm to higher dimensions. We instead use lattice-based post-processing.

In appendix A, we provide a description of the quantum algorithm for order finding with tradeoffs, a complete analysis of the probability distribution that it induces, and estimates for the number of runs n required to solve various problem instances for r when using our lattice-based post-processing algorithm.

9.2 Factoring integers

Quantum algorithms for computing the order of finite cyclic groups may be used to factor integers, as was originally demonstrated by Shor [13].

To factor a composite integer N, we first pick an integer g on $1 < g < N$ such that $\gcd(g, N) = 1$. If the greatest common divisor is not one, we have found a non-trivial factor of N. In general, we would remove small and moderate size factors of N before attempting to factor N via order finding, so this is an unlikely event to occur in practice. Next, we regard g as a generator of the cyclic subgroup $\langle g \rangle \subset \mathbb{Z}_N^*$ and compute its order r using the quantum algorithm.

As $g^r \equiv 1 \pmod N$ we know that $g^r - 1 \equiv 0 \pmod N$ and we may use this fact to find factors of N. When r is even, Shor uses the fact that

$$g^r - 1 \equiv (g^{r/2} - 1)(g^{r/2} + 1) \equiv 0 \pmod N$$

to find non-trivial factors of N by computing $\gcd((g^{r/2} \bmod N) \pm 1, N)$.

In general, if t divides r, it is likely to be the case that $g^{r/t} \equiv 1 \pmod B$ for some non-trivial factor B of N. Non-trivial factors of N may be found by computing $\gcd((g^{r/t} \bmod N) - 1, N)$ as proposed in [6]. Odd orders are also discussed in for example [8].

If no non-trivial factors of N may be derived from the order, the order finding algorithm would typically have to be re-run for a new g. If N is to be completely factored, and composite factors are found, additional runs may also be required.

9.2.1 Factoring RSA integers

If N is an RSA integer, i.e. the product of two prime factors of similar size, factoring by computing a short discrete logarithm as proposed in [4, 5] imposes less demands on the quantum computer than factoring by order finding [12, 13]. The two factors may then be immediately recovered from the logarithm.

9.3 Other quantum algorithms

It was recently proposed by Bernstein, Biasse and Mosca [2] to factor integers by adapting the general number field sieve (GNFS) to run on a quantum computer. As is the case for the classical GNFS, the GNFS as adapted in [2] may also be modified to compute discrete logarithms in subgroups to $\mathbb{F}_p^*$.

Asymptotically, the GNFS as adapted in [2] uses fewer qubits than all of the above algorithms. However, its asymptotic time complexity is subexponential, unlike the above algorithms that are all polynomial time, and as the authors of [2] point out, it is a challenging open question to determine for what size of numbers fewer qubits would actually be used in practice by their algorithm.

Hence, it remains an open question whether the GNFS as adapted in [2] has an advantage over the above algorithms, and in particular whether it has an advantage for numbers of the size currently used in cryptographic applications.

10 Summary and conclusion

In this work, we generalize and bridge Shor's groundbreaking works [13, 14] on order finding and general discrete logarithms, our earlier works [3, 4, 5] on short discrete logarithms with tradeoffs and Seifert's [12] work on order finding with tradeoffs. Our principal contribution is to show how the idea of enabling tradeoffs may be extended to the case of computing general discrete logarithms.

This yields a reduction by up to a factor of two in the number of group operations that need be performed in each run of the quantum algorithm at the expense of having to run the quantum algorithm multiple times.

Combined with our earlier works [4, 5] this implies that the number of group operations that need to be computed in each run of the quantum algorithm equals the number of bits in the logarithm times a small constant factor that depends on the tradeoff factor. The reduction in the number of group operations translates into a reduction of the required circuit depth and coherence time.

We comprehensively analyze the probability distribution induced by our quantum algorithm, describe how the algorithm may be simulated classically, and estimate the number of runs required to compute logarithms with a given minimum success probability for different tradeoff factors s. Only slightly more than s runs are in general required to reach a success probability of at least 99%. This is significantly better than the original bound given by Shor [13].

The group order need not be known when using our algorithm for computing general discrete logarithms. If the order is unknown, it may optionally be computed from the same set of quantum algorithm outputs as is used to compute the general discrete logarithm at no additional quantum cost.

Our algorithms for computing general and short discrete logarithms and group orders use lattice-based post-processing algorithms. If only the group order is sought, as is the case when factoring integers, our lattice-based post-processing algorithm for computing group orders may be combined with the quantum algorithm for order finding originally introduced by Seifert [12]. Only slightly more than s runs of the quantum algorithm are then in general required to reach a probability of at least 99% of recovering the order.

Acknowledgments

Many thanks to Johan Håstad for valuable comments and guidance. Thanks to Lennart Brynielsson and other colleagues for useful comments and discussions.

Funding and support for this work was provided by the Swedish NCSA that is a part of the Swedish Armed Forces. Computations were performed on the Beskow Cray XC40 supercomputer and Tegner post-processing cluster provided by the Swedish National Infrastructure for Computing (SNIC) at PDC at KTH.

References

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

[2] Bernstein, D. J., Biasse, J.-F. and Mosca, M.: A low-resource quantum factoring algorithm. In: Lange T., Takagi T. (Eds) Post-Quantum Cryptography. PQCrypto 2017. LNCS, vol. 10346, pp. 330-346. Springer, Cham (2017)

[3] Ekerå, M.: Modifying Shor's algorithm to compute short discrete logarithms. Cryptology ePrint Archive, Report 2016/1128 (2016)

[4] Ekerå, M., Håstad, J.: Quantum algorithms for computing short discrete logarithms and factoring RSA integers. In: Lange T., Takagi T. (Eds) Post-Quantum Cryptography. PQCrypto 2017. LNCS, vol. 10346, pp. 347–363. Springer, Cham (2017)

[5] Ekerå, M.: On post-processing in the quantum algorithm for computing short discrete logarithms. IACR ePrint Archive Report 2017/1122 (2017).

[6] Johnston, A. M.: Shor's Algorithm and Factoring: Don't Throw Away the Odd Orders. IACR ePrint Archive Report 2017/083 (2017).

[7] Korkine, A., Zolotareff, G.: Sur les formes quadratiques. Math. Ann. **6**(3), pp. 366–389 (1873)

[8] Lawson, T., Odd orders in Shor's factoring algorithm, Quantum Information Processing 14(3), pp. 831–838 (2015).

[9] Lenstra, H.W., Lenstra, A.K., Lovász, L.: Factoring Polynomials with Rational Coefficients. Math. Ann. **261**(4) pp. 515–534 (1982)

[10] Mosca, M., Ekert, A.: The Hidden Subgroup Problem and Eigenvalue Estimation on a Quantum Computer. In: Proceeding from the First NASA International Conference, Quantum Computing and Quantum Communications, vol. 1509, pp. 174–188 (1999).

[11] Schnorr, C.P.: A hierarchy of polynomial time lattice basis reduction algorithms. Theoretical Computer Science, **53**(2–3), pp. 201–224 (1987)

[12] Seifert, J.-P.: Using fewer qubits in Shor's factorization algorithm via simultaneous Diophantine approximation.. In: Naccache, D. (ed.) CT-RSA 2001. LNCS, vol. 2020, pp. 319–327. Springer, Heidelberg (2001)

[13] Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings of the 35th Annual Symposium on Foundations of Computer Science, pp. 124–134 (1994)

[14] Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Comput., **26**(5), pp. 1484–1509 (1997)

A Order finding with tradeoffs

In this appendix, we recapitulate how tradeoffs may be enabled in Shor's order finding algorithm, analyze the distribution that this algorithm induces, and estimate the number of runs n required to solve for the order r.

A.1 The quantum algorithm

In this section we describe the quantum algorithm that upon input of a generator g of a finite cyclic group $\mathbb{G}$ of order r outputs an integer j.

As in section 2 the algorithm is parameterized under a tradeoff factor $s \geq 1$. For $s = 1$ this algorithm reverts to Shor's original order finding algorithm [13] and for $s > 1$, it is the algorithm originally proposed by Seifert [12] to save qubits in Shor's algorithm.

1. Let m be the integer such that $2^{m-1} \leq r < 2^m$, let $\ell = \lceil m/s \rceil$, and let

$$| \, \Psi \, \rangle = \frac{1}{\sqrt{2^{m+2\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} | \, a \, \rangle \, | \, 0 \, \rangle \, .$$

2. Compute $[a] \, g$ and store the result in the second register

$$| \, \Psi \, \rangle = \frac{1}{\sqrt{2^{m+\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} | \, a, [a] \, g \, \rangle$$

3. Compute a QFT of size $2^{m+\ell}$ of the first register to obtain

$$| \, \Psi \, \rangle = \frac{1}{\sqrt{2^{m+\ell}}} \sum_{a\,=\,0}^{2^{m+\ell}-1} | \, a, [a] \, g \, \rangle \quad \xrightarrow{\text{QFT}}$$

$$\frac{1}{2^{m+\ell}} \sum_{a\,=\,0}^{2^{m+\ell}-1} \sum_{j\,=\,0}^{2^{m+\ell}-1} \mathrm{e}^{\,2\pi i\, aj/2^{m+\ell}} \,|\, j, [a]\, g\,\rangle\,.$$

4. Observe the system to obtain j and $y = [e]\, g$ where $e = a \bmod r$.

A.2 The probability of observing j and y

Above, the integer j and element $y = [e]\, g$ are obtained with probability

$$\frac{1}{2^{2(m+\ell)}} \left| \sum_{a} \exp\left[\frac{2\pi i}{2^{m+\ell}}\, aj\right] \right|^2 \tag{24}$$

where the sum is over all a on $0 \le a < 2^{m+\ell}$ such that $a \equiv e \pmod r$. In this section, we seek a closed form expression to (24) that is exact.

To this end, we first perform a variable substitution to obtain a contiguous summation interval. As all a that fulfill the condition that $a \equiv e \pmod r$ are on the form $a = e + n_r r$ where $0 \le n_r \le (2^{m+\ell} - 1 - e)/r$, substituting a for $e + n_r r$ in equation (24) and adjusting the phase yields

$$\frac{1}{2^{2(m+\ell)}} \left| \sum_{n_r\,=\,0}^{\lfloor (2^{m+\ell}-1-e)/r \rfloor} \exp\left[\frac{2\pi i}{2^{m+\ell}}\, \alpha_r n_r\right] \right|^2 = \frac{1}{2^{2(m+\ell)}} \left| \sum_{n_r\,=\,0}^{\lfloor (2^{m+\ell}-1-e)/r \rfloor} \mathrm{e}^{i\theta_r n_r} \right|^2$$

where $\alpha_r = \{rj\}_{2^{m+\ell}}$ and $\theta_r = \theta(\alpha_r) = 2\pi\alpha_r/2^{m+\ell}$. Summing over all e yields

$$\frac{1}{2^{2(m+\ell)}} \sum_{e\,=\,0}^{r-1} \left| \sum_{n_r\,=\,0}^{\lfloor (2^{m+\ell}-1-e)/r \rfloor} \mathrm{e}^{i\theta_r\, n_r} \right|^2 = \tag{25}$$

$$\frac{\beta}{2^{2(m+\ell)}} \left| \sum_{n_r\,=\,0}^{\lfloor (2^{m+\ell}-1)/r \rfloor} \mathrm{e}^{i\theta_r\, n_r} \right|^2 + \frac{r-\beta}{2^{2(m+\ell)}} \left| \sum_{n_r\,=\,0}^{\lfloor (2^{m+\ell}-1)/r \rfloor - 1} \mathrm{e}^{i\theta_r\, n_r} \right|^2 \tag{26}$$

for β such that $\beta \equiv 2^{m+\ell} \pmod r$, as for all $0 \le e < \beta$ we then have that

$$\lfloor (2^{m+\ell} - 1)/r \rfloor = \lfloor (2^{m+\ell} - 1 - e)/r \rfloor$$

whereas for all $\beta \le e < r$ we have

$$\lfloor (2^{m+\ell} - 1)/r \rfloor - 1 = \lfloor (2^{m+\ell} - 1 - e)/r \rfloor\,.$$

A.2.1 Closed form expressions

Assuming $\theta_r \neq 0$, we may write equation (26) on closed form as

$$\frac{\beta}{2^{2(m+\ell)}} \left| \frac{\mathrm{e}^{i\theta_r\, \left(\lfloor (2^{m+\ell}-1)/r \rfloor + 1\right)} - 1}{\mathrm{e}^{i\theta_r} - 1} \right|^2 + \frac{r-\beta}{2^{2(m+\ell)}} \left| \frac{\mathrm{e}^{i\theta_r\, \lfloor (2^{m+\ell}-1)/r \rfloor} - 1}{\mathrm{e}^{i\theta_r} - 1} \right|^2 .$$

otherwise, if $\theta_r = 0$, we may write equation (26) on closed form as

$$\frac{\beta}{2^{2(m+\ell)}} \left((2^{m+\ell} - 1)/r - 1 \right)^2 + \frac{r-\beta}{2^{2(m+\ell)}} \left((2^{m+\ell} - 1)/r \right)^2 .$$

A.3 Simulating the quantum algorithm

In this section, we combine results from the previous sections to construct a high-resolution histogram for the probability distribution for known r. Furthermore, we describe how the histogram may be sampled to simulate output generated by the quantum algorithm

A.3.1 Constructing the histogram

To construct the high-resolution histogram, we divide the argument axis into regions and subregions and integrate the closed form probability expression numerically in each subregion in analogy with the procedure in section 6.1.

First, we subdivide the negative and positive sides of the argument axis into $30 + \mu$ regions where $\mu = \min(\ell - 2, 11)$. Each region thus formed may be uniquely identified by an integer η_r by requiring that for all α_r in the region

$$2^{|\eta_r|} \leq |\, \alpha_r \,| \leq 2^{|\eta_r|+1} \qquad \text{and} \qquad \text{sgn}(\alpha_r) = \text{sgn}(\eta_r)$$

where $m{-}30 \leq |\, \eta_r \,| < m{+}\mu{-}1$. Then, we subdivide each region into subregions identified by an integer ξ_r by requiring that for all α_r in the subregion

$$2^{|\eta_r|+\xi_r/2^\nu} \leq |\, \alpha_r \,| \leq 2^{|\eta_r|+(\xi_r+1)/2^\nu}$$

for ξ_r an integer on $0 \leq \xi_r < 2^\nu$ and ν a resolution parameter. The selection of ν is further elaborated on in section A.3.2 below.

For each subregion, we compute the approximate probability mass contained within the subregion by applying Simpson's rule, followed by Richardson extrapolation to cancel the linear error term. Simpson's rule is hence applied $2^\nu(1 + 2^\nu)$ times in each region. Each application requires the probability to be computed in up to three points (the two endpoints and the midpoint), for which purpose we use the closed form expression developed in section A.2.1.

In theory, we should furthermore multiply by the multiplicity of arguments 2^{κ_r}, see Lemma 4.2 in section 4.2, and divide by 2^{κ_r} to account for the density of distinct pairs in the region. However, these operations cancel out each other.

A.3.2 Selecting the resolution parameter

As it is inexpensive to compute the above linear histogram in α_r compared to computing the two-dimensional histogram in (α_d, α_r) in section 6.1, we fix $\nu = 11$ to obtain a high degree of accuracy in the tail of the linear histogram.

This enables us to linear histogram as a reference when adaptively selecting the resolution for the two-dimensional histogram in section 6.1. Recall that by Lemma 6.1, the linear histogram in α_r constructed above should agree with the marginal distribution in α_r in the two-dimensional histogram in section 6.1.

A.3.3 Experiments and results

The probability distribution is plotted on the signed logarithmic argument axis in Fig. 6 for $m = 2048$ and $s = 30$, and for d and r selected as explained in section 8.3. The regions form two contiguous symmetric areas on the argument axis, as is illustrated in Fig. 8. The distribution is virtually identical to the marginal distribution along the vertical α_r axis in Fig. 6.

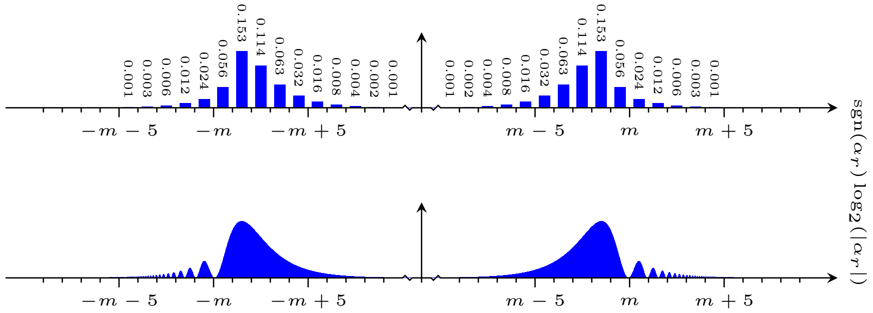

Fig. 8: The probability distribution induced by the order finding algorithm, computed as in section A.3.1, for $m = 2048$ and $s = 30$, and for r selected as in section 8.3. To facilitate printing, the resolution has been reduced in this figure.

A.3.4 Sampling the probability distribution

To sample an argument α_r from the probability distribution, we first sample a subregion from the histogram and then sample α_r uniformly at random this subregion, with the restriction that 2^{κ_r} must divide α_r so that α_r is admissible.

To sample a subregion from the histogram, we order all subregions in the histogram by probability, and compute the cumulative probability up to and including each subregion in the resulting ordered sequence, in analogy with the procedure in section 6.4.

Then, we sample a pivot uniformly at random from $[0, 1)$, and return the first subregion in the ordered sequence for which the cumulative probability is greater than or equal to the pivot. The sampling operation fails if the pivot is greater than the cumulative probability of the last subregion in the sequence.

To sample an integer j from the distribution, we first sample an argument α_r and then select an integer j yielding α_r uniformly at random from the set of all such integers using Lemma 4.2 on page 17. More specifically, we first sample an integer t_r uniformly at random on the admissible interval for t_r and then compute j from α_r and t_r as described in Lemma 4.2.

A.4 Classical post-processing

As the distribution that arises for short discrete logarithms is virtually identical to the marginal distribution along the α_r axis in section 6.1, the classical post-processing algorithm described in section 7.2 may be used to solve sets of n integers j generated by the quantum algorithm in section A.1 for r.

A.5 Simulating the complete algorithm

In analogy with the procedure described in section 8, we may now simulate the complete algorithm to estimate the minimum number of runs n required to attain a given minimum success probability q when recovering both d and r for specific problem instances and tradeoff factors without enumerating L^j.

		group size m						
		128	256	512	1024	2048	4096	8192
tradeoff factor s	1	2	2	2	2	2	2	2
	2	3	3	3	3	3	3	3
	3	4	4	4	4	4	4	4
	4	6	5	5	5	5	5	5
	5	7	6	6	6	6	6	6
	6	9	8	7	7	7	7	7
	7	12 / 11	9	8	8	8	8	8
	8	16 / 15	11	10	9	9	9	9
	10	– / 25	14	12	11	11	11	11
	20	–	– / 54	28 / 29	24	22	21	21
	30	–	–	– / 53	39 / 38	34	32	31
	40	–	–	–	– / 58	48 / 47	44	42
	50	–	–	–	–	– / 63	56	53
	80	–	–	–	–	–	– / 95	– / 87

Tab. 2: The estimated number of runs n required to solve for an order r, selected as in section 8.3, with $\geq 99\%$ success probability, without enumerating the lattice in the post-processing. For details, see appendix A. For A the initial and B the simulated estimate, we print B / A, unless B = A; we then only print A. Dash indicates no estimate.

		group size m						
		128	256	512	1024	2048	4096	8192
tradeoff factor s	1	2	2	2	2	2	2	2
	2	3	3	3	3	3	3	3
	3	4	4	4	4	4	4	4
	4	6	5	5	5	5	5	5
	5	7	6	6	6	6	6	6
	6	9	8	7	7	7	7	7
	7	11 / 12	9	8	8	8	8	8
	8	17 / 16	11	10	9	9	9	9
	10	– / 25	14	12	11	11	11	11
	20	–	– / 55	30 / 29	23 / 24	22	21	21
	30	–	–	– / 53	37 / 39	34	32	31
	40	–	–	–	– / 59	48 / 47	43 / 44	42
	50	–	–	–	–	– / 63	57 / 56	53
	80	–	–	–	–	–	– / 95	– / 87

Tab. 3: The estimated number of runs n required to solve for a maximal order $r = 2^m - 1$ with $\geq 99\%$ success probability, without enumerating the lattice in the post-processing. For details, see appendix A. For A the initial and B the simulated estimate, we print B / A, unless B = A; we then only print A. Dash indicates no estimate.

A.5.1 Estimating n

To estimate n for problem instance given by r, we proceed as follows:

For $n = s + 1, s + 2, \ldots$ we first estimate $\widetilde{R}_r$ by sampling $N = 10^6$ sets of n arguments α_r, as explained in sections A.3.4 and 7.3, and record the smallest n for which the volume quotient $v_r < 2$ with probability $q = q_r = 0.99$.

With this estimate of n as our starting point, we then sample $M = 10^3$ sets of n integers j, as explained in section A.3.4, and test whether recovery of r is successful for at least $\lceil Mq \rceil$ sets when executing the post-processing algorithm in section 7.2 without enumerating L^j. Depending on the outcome of the test, we either increment or decrement n, and repeat the process, recursively, until the smallest n such that the test passes has been identified.

Executing this procedure for m and s selected as described in section 8.2, both for r selected as explained in section 8.3, and for maximal $r = 2^m - 1$, produced the estimates in Tab. 2 and Tab. 3, respectively.

B Short discrete logarithms with tradeoffs

As the method used to sample the distribution in this paper differs somewhat from the method used in [5], we have run experiments analogous to those in appendix A for short discrete logarithms.

A high-resolution histogram for the distribution is constructed in analogy with the procedure in section A.3.1, using the closed form expression in [5]. Arguments α_d are sampled from the histogram in analogy with how arguments α_r are sampled in section A.3.4. If an integer pair (j, k) is to be sampled, an argument α_d is first sampled, and (j, k) is then sampled uniformly at random from the set of all pairs (j, k) yielding α_d using Lemma 4.1.

For classical post-processing, the algorithm in section 7.1 is again used. The number of runs n required to solve different parameterization is estimated in analogy with the procedure described in section A.5.1.

Executing this procedure, for parameters in analogy with section A.5.1, both for maximal $d = 2^m - 1$, and for d selected as described in section 8.2, produced the estimates in Tab. 4 and Tab. 5, respectively.

		logarithm size m						
		128	256	512	1024	2048	4096	8192
tradeoff factor s	1	2	2	2	2	2	2	2
	2	3	3	3	3	3	3	3
	3	4	4	4	4	4	4	4
	4	6	5	5	5	5	5	5
	5	8	6	6	6	6	6	6
	6	10	8	7	7	7	7	7
	7	13	10 / 9	8	8	8	8	8
	8	18	11	10	9	9	9	9
	10	– / 32	15	12	11	11	11	11
	20	–	– / 71	32 / 30	24	22	21	21
	30	–	–	– / 60	40	35	33 / 32	31
	40	–	–	–	– / 62	50 / 48	44	42
	50	–	–	–	–	– / 65	57	54 / 53
	80	–	–	–	–	–	– / 97	– / 88

Tab. 4: The estimated number of runs n required to solve for a short discrete logarithm d, selected as in section 8.3, with $\geq 99\%$ success probability, without enumerating the lattice in the post-processing. For details, see appendix B. For A the initial and B the simulated estimate, we print B / A, unless B = A; we then only print A. Dash indicates no estimate.

		logarithm size m						
		128	256	512	1024	2048	4096	8192
tradeoff factor s	1	2	2	2	2	2	2	2
	2	3	3	3	3	3	3	3
	3	4	4	4	4	4	4	4
	4	6	5	5	5	5	5	5
	5	8	6	6	6	6	6	6
	6	10	8	7	7	7	7	7
	7	13	9	8	8	8	8	8
	8	18	11	10	9	9	9	9
	10	– / 32	16 / 15	12	11	11	11	11
	20	–	– / 71	31	25 / 24	22	21	21
	30	–	–	– / 60	40	35	32	32 / 31
	40	–	–	–	– / 62	49 / 48	45 / 44	42
	50	–	–	–	–	– / 65	57	54 / 53
	80	–	–	–	–	–	– / 97	– / 88

Tab. 5: The estimated number of runs n required to solve for a maximal short discrete logarithm $d = 2^m - 1$ with $\geq 99\%$ success probability, without enumerating the lattice in the post-processing. For details, see appendix B. For A the initial and B the simulated estimate, we print B / A, unless B = A; we then only print A. Dash indicates no estimate.