

Overshifted Parameter-Shift Rules: Optimizing Complex Quantum Systems with Few Measurements

Leonardo Banchi,^{1,2,*} Dominic Branford,¹ and Chetan Waghela¹

¹*Department of Physics and Astronomy, University of Florence,
via G. Sansone 1, I-50019 Sesto Fiorentino (FI), Italy*

²*INFN Sezione di Firenze, via G. Sansone 1, I-50019, Sesto Fiorentino (FI), Italy*

Gradient-based optimization is a key ingredient of variational quantum algorithms, with applications ranging from quantum machine learning to quantum chemistry and simulation. The parameter-shift rule provides a hardware-friendly method for evaluating gradients of expectation values with respect to circuit parameters, but its applicability is limited to circuits whose gate generators have a particular spectral structure. In this work, we present a generalized framework that, with optimal minimum measurement overhead, extends parameter shift rules beyond this restrictive setting to encompass basically arbitrary gate generator, possibly made of complex multi-qubit interactions with unknown spectrum and, in some settings, even infinite dimensional systems such as those describing photonic devices or qubit-oscillator systems. Our generalization enables the use of more expressive quantum circuits in variational quantum optimization and enlarges its scope by harnessing all the available hardware degrees of freedom.

I. INTRODUCTION

The rapid development of quantum computing hardware has spurred a new era of algorithms designed to leverage the unique properties of quantum mechanics. Among these, Variational Quantum Algorithms (VQAs) have shown potential for applications in many areas, such as chemistry, materials science, artificial intelligence and optimization [1], and represent one of the most promising approaches to harnessing near-term quantum devices [2]. VQAs operate by optimizing the parameters of a quantum circuit to minimize a cost function that is estimated directly in hardware from measurable quantities. This optimization process is often the most demanding part of the algorithm and typically relies on gradient-based methods [3]. Central to these methods is the estimation—directly in hardware—of gradients of expectation values with respect to circuit parameters, which then guide classical optimization routines.

Among the techniques to directly estimate gradients in hardware, parameter-shift rules have emerged as particularly elegant and practical methods [4–10]. They enable an unbiased estimation of the gradient with $\mathcal{O}(1)$ variance, thus avoiding the large variance associated to finite difference methods, which typically result in convergence issues due to the excess stochastic noise. While remarkably effective for a wide class of quantum gates, the standard parameter-shift rule [4] is often derived for and applied specifically to single-qubit gates. Further generalizations are still limited to special cases, e.g. gates generated by Hamiltonians with equally spaced frequencies [5].

These limitations restrict the design space of VQAs and may hinder their applicability to certain problems. For instance, in quantum chemistry applications [11], variational circuit ansätze are typically built by exponentiating complex Hamiltonians, which often involve many-qubits and have a complex and possibly unknown spectrum; here conventional parameter shift rules do not apply. Moreover, some novel or established quantum computing architectures go beyond the qubit representation, e.g. those based on qudits [12], on the vibrational modes of ions [13], on hybrid oscillator-qubit systems [14] or on continuous variable systems, such as in photonic quantum com-

* leonardo.banchi@unifi.it

puting [15–18]. The Hamiltonians that describe these architectures may act on infinitely dimensional spaces where conventional methods do not apply. Although for some quantum algorithms there may exist some reasonable mapping—exact or approximate in some limit—from these systems to qubits, for variational quantum algorithms this mapping is not necessary and one can fully exploit the peculiarity of these more complex quantum systems.

This paper addresses the limitations of known methods by presenting a generalization of the parameter-shift rule, applicable to basically *any* arbitrary gate and quantum operation. This generalization significantly expands the practical applicability of VQAs, enabling the use of more complex and expressive ansätze, as well as the use of all available degrees of freedom in different quantum computing architectures. Central to our analysis is the development of “overshifted” rules, where the number of parameter shifts is larger than what would be required by a mere counting argument. Within this extended space there are infinitely many solutions and, among them, we can select the parameter shift rule with minimum variance, which hence requires less measurement shots for estimating derivatives in hardware. The resulting problem for defining new parameter shift rules is convex and, in some important limits, it can be approximated analytically and efficiently even for large dimensional systems.

We demonstrate that known parameter shift rules are special cases of our generalized framework, and we provide the theoretical underpinnings for its optimality in terms of the total number of measurement shots. Our generalization not only broadens the theoretical foundations of variational quantum optimization but also provides a practical toolkit for implementing gradient-based learning in more expressive quantum models, opening pathways to improved algorithmic performance on near- and long-term quantum hardware.

The remainder of this paper is structured as follows: Sec. II provides the necessary background to formalize different parameter shift rules. Sec. III defines overshifted parameter shift rules and the convex optimization problem to find the ones with minimum variance, discussing also the connections with signal processing. Sec. IV introduces analytic approximations that are sometimes based on infinite or continuously many shifts. Numerical simulations and different applications are considered in Secs. V and VI. Conclusions and further research directions are drawn in Sec. VII.

II. PROBLEM DEFINITION

We focus on parametric quantum circuits expressed as a cascade of gates

$$|\psi(\boldsymbol{\theta})\rangle = \hat{W}_L e^{i\theta_L \hat{H}_L} \dots e^{i\theta_2 \hat{H}_2} \hat{W}_1 e^{i\theta_1 \hat{H}_1} |\psi_0\rangle, \quad (1)$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_L)$, $\theta_j \in \mathbb{R}$ with $j = 1, \dots, L$ define the tunable parameters, L is the number of parameters, $\hat{H}_j$ are Hermitian operators (e.g. “Hamiltonians”), and $\hat{W}_j$ are constant gates. We are interested in derivatives of expectation values $f(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | \hat{O} | \psi(\boldsymbol{\theta}) \rangle$, for a certain observable $\hat{O}$. Without loss of generality, we can focus on a single parameter θ_k for a certain k and study $\partial f(\boldsymbol{\theta}) / \partial \theta_k$, with all other parameters constant. Computing the full gradient is then trivial by repeating the same procedure for all possible k . Fixing $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_k$, we may focus on

$$f(\theta) \equiv f(\boldsymbol{\theta}) \Big|_{\substack{\theta_k = \theta, \\ \theta_{j \neq k} = \text{const}}} = \langle \psi | e^{-i\hat{H}_k \theta} \hat{M} e^{i\hat{H}_k \theta} | \psi \rangle, \quad (2)$$

where we have dropped the dependence on k to simplify the notation and set $|\psi\rangle = \prod_{j=1}^{k-1} \hat{U}_j |\psi_0\rangle$, $\hat{M} = W_k^\dagger \left(\prod_{j=k+1}^L \hat{U}_j \right)^\dagger \hat{O} \left(\prod_{j=k+1}^L \hat{U}_j \right) W_k$, where $\hat{U}_j = \hat{W}_j e^{i\theta_j \hat{H}_j}$. Without loss of generality, we

may assume diagonal $\hat{H}_k$ with eigenvalues $E_j^{(k)}$, since we can always reabsorb the diagonalizing unitaries in the constant gates $\hat{W}_k$ and $\hat{W}_{k-1}$, and $|\psi_0\rangle$ for $k = 1$. Therefore, we may express $f(\theta)$ as a Fourier-like series

$$f(\theta) = \sum_{\omega \in \Omega} M_\omega e^{i\omega\theta} \quad (3)$$

where

$$\Omega = \{E_j^{(k)} - E_i^{(k)} \text{ for } i, j = 1, \dots, N_k\}, \quad (4)$$

is the set of “beat” frequencies (energy differences), N_k is the number of distinct eigenvalues of $\hat{H}_k$ and the complex numbers M_ω depend on the sum of operator elements $\langle E_j^{(k)} | \hat{M} | E_i^{(k)} \rangle$ in the energy basis with $E_j^{(k)} - E_i^{(k)} = \omega$. Therefore, the total number of distinct frequencies satisfies $|\Omega| \leq \mathcal{O}(N_k^2)$. Since $f(\theta) \in \mathbb{R}$, the complex coefficients satisfy $M_\omega^* = M_{-\omega}$, and if $\omega \in \Omega$ then also $-\omega \in \Omega$.

While Ω is determined by the eigenvalues of $\hat{H}_k$, as only eigenstates $\{|E_j^{(k)}\rangle\}$ with support on both $|\psi\rangle$ and $\hat{M}$ contribute to the M_ω terms, a valid shift rule can be found for states $|\psi\rangle \in \Psi$ and observables $\hat{M} \in \mathcal{M}$ from only

$$\Omega = \{E_j^{(k)} - E_i^{(k)} | E_i^{(k)}, E_j^{(k)} \in \mathcal{E}_\Psi(\hat{H}_k) \cap \mathcal{E}_\mathcal{M}(\hat{H}_k)\}, \quad (5)$$

where $\mathcal{E}_\Psi(\hat{H}_k) = \{E_j^{(k)} | \langle \psi | E_j^{(k)} \rangle \neq 0 \text{ for any } \psi \in \Psi\}$ is the set of eigenstates which any $|\psi\rangle \in \Psi$ have support on, and $\mathcal{E}_\mathcal{M}(\hat{H}_k) = \{E_j^{(k)} | \langle \hat{M} | E_j^{(k)} \rangle \neq 0 \text{ for any } \hat{M} \in \mathcal{M}\}$ is the set of eigenstates which any $\hat{M} \in \mathcal{M}$ have support on.

A. Computing Gradients in the Quantum Hardware

From the definition (3), it is now trivial to express the derivative as

$$f'(\theta) \equiv \frac{\partial f(\theta)}{\partial \theta_k} = \sum_{\omega \in \Omega} M_\omega e^{i\omega\theta} i\omega. \quad (6)$$

However, for complex quantum circuits the coefficients M_ω may be very hard to compute with classical computers. Therefore, we try to express the derivative (6) as a linear combination of evaluations of $f(\theta)$, since we already know how to estimate that quantity in a quantum computer: we sequentially apply the parametric gates and the constant gates $\hat{W}_j$ to first create the state (1), and then measure the observable $\hat{O}$.

In parameter shift rules [5, 6, 8, 9] we look for expressions like

$$\frac{df(\theta)}{d\theta} = \sum_p c_p f(\theta + \vartheta_p), \quad (7)$$

where the real coefficients c_p and shifts ϑ_p are unknown and must be obtained.

Particular solutions to the above equation have been extensively studied in the literature [4–7, 19]. The most popular one is for when $\hat{H}_k$ is a Pauli operator, with two distinct eigenvalues $E_j^{(k)} = \pm 1$, for which two shifts $\vartheta_\pm = \pm\pi/4$, with weightings $c_\pm = \pm 1$ are used. For more general Hamiltonians, there are no explicit guidelines to define the shifts ϑ_p and the coefficients c_p .

By asking that (7) and (6) are equal for all possible values of M_ω , which are unknown and hard to compute, we get the following expression

$$\sum_p c_p e^{i\omega\vartheta_p} = i\omega, \quad \forall \omega \in \Omega. \quad (8)$$

For general frequencies the above problem can be solved by the Nonequidistant Fast Fourier Transform, for which there are several numerical libraries, e.g. [20].

B. Symmetric parameter shift rules

We can simplify the above linear system, by considering only shifts that are symmetric around the origin, coming in pairs $\pm\theta_p$ with equal and opposite coefficients. In this case we require that

$$\frac{df(\theta)}{d\theta} = \sum_{p=1}^P c_p [f(\theta + \vartheta_p) - f(\theta - \vartheta_p)], \quad (9)$$

where P is the number of positive shifts $\vartheta_p > 0$ and

$$2 \sum_{p=1}^P c_p \sin(\omega\vartheta_p) = \omega, \quad \forall \omega \in \Omega^+, \quad (10)$$

where the real part of Eq. (8) is automatically satisfied for any symmetric parameter-shift rule of the form (9). The set Ω^+ is the subset of Ω with strictly positive frequencies. Clearly $|\Omega| = 2|\Omega^+| + 1$ since Ω contains all negative frequencies and the zero frequency. The linear system Eq. (10) may be solvable in general provided that the number of positive shifts P satisfies

$$P \geq N \quad N := |\Omega^+|. \quad (11)$$

In general though there is no guideline to choose the shifts θ_p . Since the functions $f(\theta \pm \vartheta_p)$ are typically estimated in a quantum hardware, one may naively guess that one should look for solutions with the smallest number of shifts, namely with $P = N$. However, since the optimal shifts θ_p are not known, we will show that it is in general beneficial to work in the overparametrized regime, where the total number of shifts exceeds the number of constraints and the problems (8),(10), have infinitely many solutions. We call the corresponding parameter shift rule “overshifted”, as more shifts than those required to solve Eq. (8) will be used.

As we will show, with the proper regularization, the solutions of many overshifted parameter shift rules will be sparse, namely the number of non-zero coefficients c_p will still be small. Moreover, these solutions also minimize the variance and hence the number of measurements in the quantum hardware. In general, with $P \gg N$ we can find better parameter shift rules with several advantages.

The “infinitely overshifted” limit $P \rightarrow \infty$ will also be quite useful: in that limit parameter shift rules can be defined even without knowing the frequencies, but just by knowing an upper bound on the bandwidth

$$\Lambda = \max_{\omega \in \Omega} |\omega|. \quad (12)$$

This is particularly useful when we are interested in gradients of complex quantum evolutions where the “Hamiltonians” $\hat{H}_\ell$ in Eq. (1) are complex, e.g. many-body Hamiltonians with unknown spectrum. Another interesting application is for dependent parameters, e.g. for parameter sharing, as we define in the next section.

C. Parameter sharing

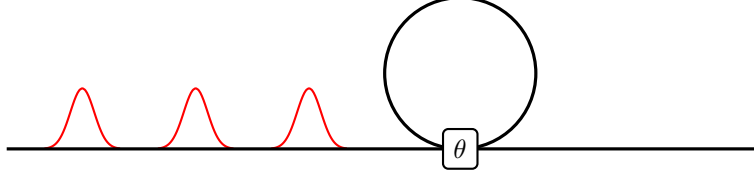


Figure 1. Example circuit with hardware parameter sharing. Three time-bin encoded photonic qubits (red pulses) enter into a beam splitter, with angle θ . The delay line length is tuned according to the time separation between each photons and applies the same beam splitting operation to each neighbouring pair.

In some settings, some parameters in the quantum circuit definition (1) may be equal, either by design choices, e.g. to approximate Floquet dynamics [21], or by hardware limitations. These limitations are quite common, for instance, in photonic quantum computing [15, 22, 23]. An example is a beam splitter between time-bin encoded qubits, shown in Fig. 1, where the same parametric operation is applied to each neighbouring pair of qubits.

The conventional approach would be to use the chain rule of derivatives, but this breaks the time-translational invariance. We will show this with the following example

$$f(\theta_1, \theta_2) = \langle \psi_0 | e^{-i\theta_1 \hat{H}_1} \hat{W} e^{-i\theta_2 \hat{H}_2} \hat{M} e^{i\theta_2 \hat{H}_2} \hat{W} e^{i\theta_1 \hat{H}_1} | \psi_0 \rangle. \quad (13)$$

Suppose that hardware limitations force $\theta_1 = \theta_2 = \theta$, and assume $\hat{H}_1 = \hat{H}_2$, as in the example of Fig. 1. Then we can compute gradients using the chain rule and Eq. (7) as

$$\frac{df}{d\theta} = \frac{\partial f}{\partial \theta_1} \frac{d\theta_1}{d\theta} + \frac{\partial f}{\partial \theta_2} \frac{d\theta_2}{d\theta} = \sum_p c_p [f(\theta + \vartheta_p, \theta) + f(\theta, \theta + \vartheta_p)]. \quad (14)$$

However, to estimate gradients in hardware using the above formula we need to evaluate the circuit for non-symmetric parametrizations, which might not be possible. In the example of Fig. 1, this would require the dynamic modulation of θ at different times, which is more challenging than a static θ .

On the other hand, within the setting of Sec. II, we may express $f(\theta, \theta)$ using the expansion (3) with

$$\Omega = \{E_j + E_\ell - E_i - E_k \text{ for } i, j, k, \ell = 1, \dots, N\}, \quad (15)$$

where E_j are the eigenvalues of $\hat{H}_1$. Thanks to Eq. (7) we can now express all derivatives of $f(\theta, \theta)$ as a linear combination of $f(\theta + \vartheta_p, \theta + \vartheta_p)$, which maintains the parametrization symmetry.

More generally, suppose that the parameter θ is shared among T gates which are located at the layers $\ell_t \in [1, \dots, L]$. Then, we may use the solution of Sec. (II A) with

$$\Omega = \left\{ \sum_{t,s=1}^T E_{j_t}^{(\ell_t)} - E_{i_t}^{(\ell_s)} \forall i_t, j_t = 1, \dots, N \right\}, \quad (16)$$

The evaluation of the set Ω becomes prohibitively expensive for large L , as the number of frequencies increases exponentially with L . Nonetheless, we will show that some parameter shift rules only depend on the bandwidth Eq. (12) which is much simpler to compute. For instance, in (16) we can estimate the bandwidth Λ by summing the largest frequencies in each layer.

III. OVERSHIFTED PARAMETER-SHIFT RULE

Among the infinitely many solutions of (8) or (10) in the overparametrized regime, we need to select the most appropriate given the hardware constraints. The most severe one is typically the measurement overhead, namely the number of measurement shots to be performed in-hardware to have a reliable estimate of the gradient.

Motivated by this, we adapt an error estimation technique from Ref. [5]. Suppose that the variance of $f(\theta)$ for any θ is bounded by σ^2 . Then if we estimate the p th element of the sum in (7) with S_p measurement shots, the variance of the gradient estimator is

$$\text{Var} \left[\frac{df(\theta)}{d\theta} \right] \leq \sum_p \frac{|c_p|^2 \sigma^2}{S_p}. \quad (17)$$

If we use a total of $S = \sum_p S_p$ shots, then the minimum variance is obtained by choosing

$$S_p = S |c_p| / \|\mathbf{c}\|_1, \quad (18)$$

where $\mathbf{c}$ is the vector with components c_p . With such optimal shot allocation, if we can tolerate an error of at most ε , then the total number of measurement shots required satisfies

$$S \geq \frac{\sigma^2}{\varepsilon^2} \|\mathbf{c}\|_1^2. \quad (19)$$

Based on the above estimate, we can try to reduce the number of measurement shots by working with overparametrized problems and—among the infinitely many solutions—choose the one with minimum $\|\mathbf{c}\|_1$. This problem can then be formally defined as a convex optimization problem, which can be easily solved using available libraries [24]:

$$\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{such that} \quad \sum_p c_p e^{i\omega\vartheta_p} = i\omega, \quad \forall \omega \in \Omega, \quad (20)$$

or for symmetric parameter-shift rules

$$\min_{\mathbf{c}} \|\mathbf{c}\|_1 \quad \text{such that} \quad 2 \sum_{p=1}^P c_p \sin(\omega\vartheta_p) = \omega, \quad \forall \omega \in \Omega^+. \quad (21)$$

Optimization problems like the above have been extensively studied in the literature [25, 26], with applications in error correction [27], signal reconstruction and magnetic resonance imaging [28], and compressed sensing [29]. In the quantum setting, they have been used to probabilistically interpolate quantum circuits [30, 31].

The resulting algorithm is then summarized in Algorithm 1.

Algorithm 1 Overshifted Parameter Shift Rule for a given Ω

- 1: Fix a “shift bandwidth”, namely a B such that $|\vartheta| \leq B$. For periodic functions with period 2π we may set $B = \pi$.
 - 2: Fix a suitably large number of shifts. For symmetric shifts we require that $P \geq |\Omega^+|$, with larger P meaning larger overshifting.
 - 3: Define the shifts: e.g. for symmetric parameter-shift rules and generic B , we might set $\vartheta_p = pB/P$. Alternative choices, directly connected to Discrete Fourier Transforms when $B = \pi$, are with $\vartheta_p = 2Bp/(2P+1)$ or $\vartheta_p = B(2p-1)/(2P)$.
 - 4: Check if the linear system (8) [or (10)] has at least a solution, e.g. via Rouché-Capelli’s theorem. If not, repeat the previous steps with different choices.
 - 5: Find the coefficients $\mathbf{c}$ by solving either Eq. (20) [or (21)] to enforce minimum measurement overhead, or (23), if we want to enforce continuity.
 - 6: Optimize the measurement shots as in Eq. (18) and estimate all values of $f(\theta + \vartheta_p)$ with S_p shots. The final estimate is then given by the weighted average Eq. (9).
-

A simpler, yet non optimal, alternative consists in replacing the L_1 norm with the L_2 norm, for which the resulting convex optimization problem has a closed form solution as

$$\mathbf{c}_{L_2} = D^\# i\boldsymbol{\omega}, \quad (22)$$

where $\boldsymbol{\omega}$ is the vectors with components $\omega \in \Omega$, D is the matrix with elements $D_{jp} = e^{i\omega_j \vartheta_p}$ with $j = 1, \dots, |\Omega|$, and $D^\#$ is its Moore-Penrose pseudo-inverse.

A. Towards smooth solutions

One possible issue with the solutions in (20) is that they might be highly oscillatory, i.e. with $|c_p - c_q|$ large even when $|\vartheta_p - \vartheta_q|$ is small. This might be undesirable in experimental settings with a non-negligible calibration error, namely where precisely calibrating the shifts might be challenging. In order to get smoother solutions, we might then change (20) with

$$\min_{\mathbf{c}} \sum_p |c_{p+1} - c_p| \quad \text{such that} \quad \sum_p c_p e^{i\omega \vartheta_p} = i\omega, \quad \forall \omega \in \Omega, \quad (23)$$

and amend Algorithm 1 accordingly. In this way, solutions with highly different neighbouring shifts are penalized. Moreover, from the triangle inequality $\sum_p |c_{p+1} - c_p| \leq 2\|\mathbf{c}\|_1$, and from Eqs. (19) and (27), we note that the solutions to the above optimization problem might require at most four times the number of measurement shots compared with the solutions of (20), which might still be acceptable. The convex problem (23) was already studied in [28], where it was shown to share many of the desired properties of that in Eq. (20).

B. Continuous limit: stochastic parameter shift rule

We have seen that increasing the number of shifts may be beneficial to reduce a bound on the number of measurement shots. In the limit where the number of shifts tends to infinity, we may replace Eqs. (7) and (8) with their continuous versions

$$\frac{df(\theta)}{d\theta} = \int d\vartheta c(\vartheta) f(\theta + \vartheta), \quad \int d\vartheta c(\vartheta) e^{i\omega \vartheta} = i\omega, \quad \forall \omega \in \Omega, \quad (24)$$

with a shift density $c(\vartheta)$. Alternatively, working with positive shifts only we get

$$\frac{df(\theta)}{d\theta} = \int_0^\infty d\vartheta c(\vartheta) [f(\theta + \vartheta) - f(\theta - \vartheta)], \quad \int_0^\infty d\vartheta c(\vartheta) \sin(\omega \vartheta) = \omega, \quad \forall \omega \in \Omega^+, \quad (25)$$

At first the above formulae may seem to be of limited interest, as it is impossible to experimentally measure a continuous number of circuits. However, the above expression is useful to derive a stochastic parameter shift rule, which was originally developed for gate parametrizations that contain a drift Hamiltonian [6].

The main motivation behind the stochastic parameter shift rule is stochastic gradient descent, which is routinely used in machine learning applications. In stochastic gradient descent, the optimizer does not employ the exact gradient but rather an approximation estimated via a finite number of samples. To express (24) in the form of stochastic gradient descent, we define $c_\pm(\vartheta) = \max\{0, \pm c(\vartheta)\}$, such that $c(\vartheta) = c_+(\vartheta) - c_-(\vartheta)$. Notice that this decomposition is necessary also for Eq. (25), where $c(-\vartheta) = -c(\vartheta)$ but $c(\vartheta)$ may be negative even for $\vartheta > 0$. Then, from Eq. (24) with $\omega = 0$, which is always included in Ω thanks to the definitions (4) and (16),

we see that $\int d\vartheta c_+(\vartheta) = \int d\vartheta c_-(\vartheta) = \|c\|_1/2$. Therefore, we can define two normalized probability distributions $p_{\pm}(\vartheta) = c_{\pm}(\vartheta) \frac{2}{\|c\|_1}$ and write

$$\frac{df(\theta)}{d\theta} = \frac{\|c\|_1}{2} \left(\mathbb{E}_{\vartheta_+ \sim p_+, \vartheta_- \sim p_-} [f(\theta + \vartheta_+) - f(\theta + \vartheta_-)] \right). \quad (26)$$

The gradient can then be estimated by sampling $\vartheta_{\pm}$ a certain number of times from the distributions $p_{\pm}$ and then estimating the abstract average in Eq. (26) with the empirical average using the finite number of samples. How to optimally allocate the samples to minimize the variance is discussed in Appendix A. Remarkably, convergence can be proven even when each gradient is estimated with a single sample [32, 33]. Indeed, let G be an upper bound on the gradient estimator, then stochastic gradient descent converges to a local optimum of $f(\theta)$ with an error that is upper bounded by $R \frac{G}{\sqrt{I}}$, where I is the number of iterations, and R is a constant that depends on the function and on the parameter space. Since at each iteration we need to estimate $f(\theta + \vartheta_{\pm})$, the number of iterations is proportional to the number of measurement shots S . Similarly to Eq. (19), in order to be η -close to the optimum after I iterations, each using 2 measurement shots, we get

$$S \gtrsim \frac{R^2 G^2}{\eta^2} \propto \|c\|_1^2, \quad (27)$$

where upperbounds $G \propto \|c\|_1$ exist due to Eq. (26). Therefore, we recover the analysis of the previous section: in order to minimize the overall number of measurement shots we need parameter shift rules that minimize $\|c\|_1$, while also solving Eq. (24).

The above steps are summarized in Algorithm 2. Note that the most computationally demanding parts, steps 1 and 2, must be done only once.

Algorithm 2 Overshifted (Smooth) Stochastic Parameter Shift Rule

- 1: Fix the shifts ϑ_p and find the coefficients $\mathbf{c}$, by repeating the steps 1–5 of Algorithm 1.
 - 2: Define the probability distributions $p_{\pm}(t) = \max\{0, \pm c_t\} / (\sum_s \max\{0, \pm c_s\})$.
 - 3: Sample $t_{\pm}$ from $p_{\pm}(t)$, e.g. using Algorithm 10 from Appendix C.
 - 4: Estimate $f(\theta + \vartheta_{t_{\pm}})$ in a quantum device and call the unbiased estimated result $g_{\pm}$.
 - 5: Define an unbiased estimate of the gradient as $G = (g_+ - g_-)\|\mathbf{c}\|_1/2$.
 - 6: Repeat steps 3–5 $S/2$ times and return an average of the estimated G .
-

As we show in Appendix A, the optimal way to estimate each $f(\theta + \vartheta)$ in a quantum devices is via a single shot. Namely, given a certain number of total shots S , the optimal shot allocation is to sample S different values of ϑ and then use a single-shot estimation of each $f(\theta + \vartheta)$. However, for nowadays quantum computers, this might be expensive. Indeed, for running $f(\theta)$ in a quantum device the abstract circuit must be first compiled into native gates and control pulses. If we need to run $f(\theta)$ for many values of θ this complex procedure must be performed each time. A simple solution to avoid this problem is to first sample S values of ϑ , and count how many times we have sampled the same shift ϑ . If each unique value of ϑ is found $S(\vartheta)$ times, then we can reproduce the same statistics of Algorithm 2 by estimating $f(\theta + \vartheta)$ with $S(\vartheta)$ shots and then performing a weighted average. The resulting procedure is formally described in Algorithm 9 from Appendix A.

C. Uncertainty Principle in Parameter Shifts

For simplicity, suppose that all the frequencies are commensurate, namely that they can be expressed as $\omega_j = \alpha n_j$ for a fixed $\alpha \in \mathbb{R}$ and integers n_j . If we reabsorb the global α into the

definition of θ , then $f(\theta)$ is periodic with period 2π and we can focus on functions $c(\vartheta)$ in Eq. (24) which share the same periodicity. As such, we may expand $c(\vartheta)$ as a discrete Fourier series

$$c(\vartheta) = \sum_{n=-\infty}^{\infty} f_n e^{in\vartheta}, \quad (28)$$

where $f_n^* = f_{-n}$ since $c(\vartheta)$ is real. Plugging this into Eq. (24) we get

$$f_n = 2\pi i n, \quad \forall n \in \Omega, \quad (29)$$

while all the other coefficients f_n with $n \notin \Omega$ can be chosen freely. The choice $f_n = 0$ for $n \notin \Omega$ is not optimal. Indeed, solutions to the optimization problem (20) have been linked [34] to a discrete version of the uncertainty principle [28, 35], which essentially states that such solutions cannot be sparse in both the real and Fourier domain. In other terms—if $c(\vartheta)$ is large only for a few values of ϑ —as we need to minimize $\|c\|_1$, then the number of non-zero Fourier coefficients f_n must be large.

The interplay between sparsity in one domain and spread in the conjugate domain was discovered in [27, 28], where it was considered the problem of reconstructing a signal from highly incomplete frequency information, rather than from sampling at the Nyquist rate. In those settings, when only a small random subset of Fourier coefficients is known, it was shown that for signals that are sparse in the time domain, one can exactly recover the full signal by solving an L_1 -norm minimization problem, which promotes sparsity.

IV. ANALYTIC APPROXIMATIONS

In this section, we derive some analytic approximations that are valid for any finite set of frequencies. In Sec. V we will subsequently study some particular cases, and derive simpler and optimal shift rules. The general principle is to extend the linear system in Eq. (24) with an interpolating function $I_\Omega(\omega)$, such that $I_\Omega(\omega) = \omega$ for all $\omega \in \Omega$. From Eq. (24) we then get

$$c(\vartheta) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{-i\omega\vartheta} i I_\Omega(\omega), \quad (30)$$

which is valid for all interpolating functions $I_\Omega(\omega)$. Eq. (25) is then recovered for odd interpolating functions $I_\Omega(-\omega) = -I_\Omega(\omega)$ with

$$c(\vartheta) = \int_0^\infty \frac{d\omega}{\pi} \sin(\omega\vartheta) I_\Omega(\omega), \quad (31)$$

whence we see that $c(-\vartheta) = -c(\vartheta)$.

Solutions involving Dirac deltas are not optimal, due to the uncertainty principle, as they have unbounded ϑ . Another trivial solution $I_\Omega(\omega) = \omega$ is not optimal as its Fourier transform is $\propto \delta'$. In the following sections we will discuss some approximate solutions that can be tuned in order to minimize $\|c\|_1$.

A. Triangle wave

The first interpolating function is based on the triangle wave, which linearly interpolates all points in Ω , and also extrapolates over the whole infinite domain of Ω in a zigzag way. Since it uses the entire infinite set of frequencies, from the uncertainty principle this solution is expected to be

good. Moreover, as we will see in Sec. V, numerical solutions of the problem (20) with equispaced frequencies resemble a triangular wave, see e.g. Fig. 3.

Triangle waves of period $2T$ and amplitudes in $[-1, 1]$ have the following functional form

$$W_{2T}(x) = \frac{8}{\pi^2} \sum_{t=0}^{\infty} \frac{(-1)^t}{(2t+1)^2} \sin\left(\frac{(2t+1)\pi x}{T}\right). \quad (32)$$

Let $\Lambda \geq |\omega|$ be a number bigger than all beat frequencies in Ω , e.g. the bandwidth Eq. (12). Then $W_{4\Lambda}(\omega) = \omega/\Lambda$ for all $\omega \in \Omega$ and, accordingly, $I_{\Omega}(\omega) = \Lambda W_{4\Lambda}(\omega)$ defines an interpolating function. From Eq. (30) we then get

$$c(\vartheta) = \frac{4\Lambda}{\pi^2} \sum_{t=0}^{\infty} \frac{(-1)^t}{(2t+1)^2} \left[\delta\left(\vartheta - \frac{\pi(2t+1)}{2\Lambda}\right) - \delta\left(\vartheta + \frac{\pi(2t+1)}{2\Lambda}\right) \right], \quad (33)$$

which results in Algorithm 3. Such an algorithm returns a single unbiased estimate of the gradient. Being a stochastic parameter shift rule, we can optimize shot allocations as described in Algorithm 9 from Appendix A. The resulting shift rule is given by Algorithm 4.

Algorithm 3 Triangle Shift Rule: single-shot unbiased estimator of $f'(\theta)$.

- 1: Fix $\Lambda \geq \max_{\omega \in \Omega} |\omega|$.
 - 2: Sample u uniformly from $[0, 1] \subset \mathbb{R}$.
 - 3: Repeat the iteration $q_i = q_{i-1} + 8/\pi^2(2i+1)^{-2}$ with $q_{-1} = 0$ while $q_i \leq u$. Let t be the index such that $q_{t-1} \leq u < q_t$.
 - 4: Sample a fair coin $p \in \{0, 1\}$ and set $\vartheta = (-1)^p \pi(2t+1)/(2\Lambda)$.
 - 5: Estimate $f(\theta + \vartheta)$ in a quantum device and call the outcome g .
 - 6: Return $(-1)^{t+p} \Lambda g$.
-

Algorithm 4 Cost-Efficient Triangle Shift Rule: unbiased estimator of $f'(\theta)$ with S shots.

- 1: Fix $\Lambda \geq \max_{\omega \in \Omega} |\omega|$.
 - 2: **for** $s = 1, \dots, S$ **do**
 - 3: Sample u uniformly from $[0, 1] \subset \mathbb{R}$.
 - 4: Repeat the iteration $q_i = q_{i-1} + 8/\pi^2(2i+1)^{-2}$ with $q_{-1} = 0$ while $q_i \leq u$.
 - 5: Let t be the index such that $q_{t-1} \leq u < q_t$.
 - 6: Sample a fair coin $p \in \{0, 1\}$, set $\vartheta_s = (-1)^p \pi(2t+1)/(2\Lambda)$ and $c_s = \Lambda(-1)^{p+t}$.
 - 7: **end for**
 - 8: Define $u_i = \theta_{v_i}$ as the set of different values of $\{\theta_s\}_{s=1}^S$, where v_i defines the first occurrence of such shift in the set. Let n be the number of u_i , namely the number of distinct θ_s . Let also $S_i = |\{s : u_i = \theta_s\}|$ be the number of occurrences of u_i in the sampled shifts.
 - 9: Estimate $f(\theta + u_i)$ in a quantum device using S_i measurement shots and call the outcomes f_{ij} where $i = 1, \dots, n$ and $j = 1, \dots, S_i$.
 - 10: Return the average $\sum_{i=1}^n c_{v_i} \sum_{j=1}^{S_i} f_{ij} / S$.
-

From Eq. (33) we find $\|c(\vartheta)\|_1 \leq \Lambda$, so the optimal Λ is indeed the bandwidth Eq. (12), $\Lambda = \max_{\omega \in \Omega} |\omega| := \omega_{\max}$, namely the minimum value compatible with the constraints. The downside of this method is that ϑ has infinite support: although values with large t have a low probability $\mathcal{O}(t^{-2})$ to occur, the distribution has long tails. In order to force a more bounded ϑ , it is possible to chose a larger Λ , at the expense though of increasing the number of measurements due to (19) and (27).

B. Single zig-zag

A similar—yet more continuous—solution can be obtained using only the first period of the triangle wave, namely with $I_\Omega(\omega) = |t + \Lambda| - |t - \Lambda| - \frac{1}{2}(|t + 2\Lambda| - |t - 2\Lambda|)$ for which we get

$$c(\vartheta) = \frac{4 \sin^2\left(\frac{\theta\Lambda}{2}\right) \sin(\theta\Lambda)}{\pi\theta^2}, \quad (34)$$

and $\|c(\vartheta)\|_1 \leq 2\Lambda$, so the optimal choice is again $\Lambda = \omega_{\max}$. Due to the factor of 2, this solution has at most twice the L_1 norm of the triangle wave. However, the above solution is smooth while the triangle wave requires very specific values of ϑ . Therefore, this solution might be less affected by imperfect applications of the shifts (24) in real quantum hardware.

The above solution can be expressed as a stochastic parameter shift rule with probability

$$p(\vartheta) = \frac{2 \sin^2\left(\frac{\vartheta\Lambda}{2}\right)}{\pi\Lambda\vartheta^2}, \quad (35)$$

and cumulative distribution

$$F(\vartheta) = \int_{-\infty}^{\vartheta} d\phi p(\phi) = \frac{\pi\Lambda\vartheta + 2 \cos(\Lambda\vartheta) + 2\Lambda\vartheta \text{Si}(\Lambda\vartheta) - 2}{2\pi\Lambda\vartheta}, \quad (36)$$

where $\text{Si}(x) = \int_0^x \sin(t)/t dt$.

Algorithm 5 Zigzag shift rule: unbiased estimator of $f'(\theta)$.

- 1: Fix $\Lambda \geq \max_{\omega \in \Omega} |\omega|$.
 - 2: Use Algorithm 11 from Appendix C to sample ϑ from (35) via (36).
 - 3: Estimate $f(\theta + \vartheta)$ in a quantum device and call the outcome g .
 - 4: Return $2\Lambda \sin(\theta\Lambda)g$.
-

We can then use inverse sampling to sample ϑ from $p(\vartheta)$, the resulting algorithm is summarized in Algorithm 5. This algorithm can also be expressed in the language of Sec. IIIB, but sampling from the probabilities $p_{\pm}(\vartheta)$ is more complicated due to the lack of an explicit expression like (36) for their cumulative distributions.

The zigzag shift rule has a smooth $c(\vartheta)$, so results may be less affected by experimental fluctuations of the parameters. However, sampling from the distribution (35) requires finding the zeros of a non-linear equation, which may be slow. In order to find parameter shift rules with simpler, classical sampling, we turn to kernel interpolation in the next section.

C. Kernel Interpolation

Gaussian process regression is a popular technique with wide applications in machine learning [36]. In the noiseless case, the interpolating function can be expressed as

$$I_\Omega(\omega) = \sum_i y_{\omega_i} k(\omega - \omega_i), \quad y_{\omega_i} = \sum_{j=1}^{|\Omega|} (K_\Omega^{-1})_{ij} \omega_j, \quad (37)$$

where ω_i are the elements of Ω and $(K_\Omega)_{ij} = k(\omega_i - \omega_j)$ is a matrix with $|\Omega| \times |\Omega|$ components. The positive semidefinite function $k(\omega, \omega') = k(\omega - \omega')$ is called the kernel and, due to Bochner's theorem, it can be expressed as

$$k(\omega - \omega') = \int_{-\infty}^{\infty} d\theta e^{i\theta(\omega - \omega')} p(\theta), \quad (38)$$

where $p(\theta)$ is a probability density function. Plugging these definitions into Eq. (30) we get

$$c(\vartheta) = p(\vartheta) \sum_{j=1}^{|\Omega|} i y_{\omega_j} e^{-i \omega_j \vartheta} = p(\vartheta) \sum_{\omega \in \Omega^+} 2 y_{\omega} \sin(\omega \vartheta), \quad (39)$$

where in the second equation we assume that y_{ω} is real with $y_{-\omega} = y_{\omega}$, which holds for even probability densities $\rho(\theta) = \rho(-\theta)$, namely when the distribution is symmetric around $\theta = 0$ and the kernel has real codomain. The resulting numerical procedure is described in Algorithm 6, different kernel choices are summarised in Table I.

Algorithm 6 Kernel-based shift rule: unbiased estimator of $f'(\theta)$.

- 1: Choose a suitable distribution $p(\vartheta)$ with the desirable properties described in the main text.
 - 2: Sample ϑ from $p(\vartheta)$.
 - 3: Compute $\lambda = \sum_{\omega \in \Omega^+} 2 y_{\omega} \sin(\omega \vartheta)$. If $\lambda \approx 0$, discard this sample ϑ and go back to the previous step.
 - 4: Estimate $f(\theta + \vartheta)$ in a quantum device and call the outcome g .
 - 5: Return λg .
-

Distribution	$p(\vartheta)$	$k(\omega)$	$\Delta \vartheta^2$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\vartheta^2}{2B^2}}$	$\exp(-B^2\omega^2/2)$	B^2
Uniform	$\frac{H(\vartheta+B)-H(\vartheta-B)}{2B}$	$\text{sinc}(\omega B/\pi)$	$B^2/3$
Cauchy	$\frac{2}{\pi} [1 + (B\vartheta)^2]^{-1}$	$\exp(- \omega /B)$	∞
Cosine	$\frac{1}{2B} [1 + \cos(\frac{\vartheta}{B}\pi)]$	$\frac{\text{sinc}(B\omega/\pi)}{1-(B\omega/\pi)^2}$	$B^2 (\frac{1}{3} - \frac{2}{\pi^2})$
Wigner	$\frac{2}{\pi B^2} \sqrt{B^2 - \vartheta^2}$	$\frac{2}{B\omega} J_1(B\omega)$	$\frac{B^2}{4}$

Table I. Some distributions with efficient sampling algorithms [37], their kernel, and the variance of the parameter shifts. All distributions depend on a single tunable hyperparameter B . In the table entries, H is the step function, $\text{sinc}(x) = \sin(\pi x)/(\pi x)$ and $J_1(x)$ is a Bessel function.

Following Eq. (39), we can now choose a probability distribution $p(\vartheta)$ that is easy to sample from. However, the hyperparameters of such distribution must be chosen carefully to make the coefficients y_{ω} small, as $\|c\|_1 \leq \|y\|_1$. Although the sampling part in this algorithm can be made very easy with a suitable choice of the distribution $p(\vartheta)$, e.g. a normal or uniform distribution, Algorithm 6 has two bottlenecks. First, depending on Ω , the coefficient $\lambda = \sum_{\omega \in \Omega^+} 2 y_{\omega} \sin(\omega \vartheta)$ might be close to zero for many values of ϑ . Although such values can be *classically discarded* without calling the quantum hardware, and hence without increasing the measurement cost, they still increase the classical computation part. Another possibility is use a higher number of measurement shots for the shifts ϑ with larger λ , as shown in Appendix A – see in particular Eq. (A6).

The second complication that we observe in numerical experiments is that the kernel matrix K_{Ω} can be singular. In order to have a guideline about this possibility and develop countermeasures, we might use Gershgorin's circle theorem, which basically states that the eigenvalues of K_{Ω} are within a radius $R_i = \sum_{j \neq i} |k(\omega_i - \omega_j)|$ of the diagonal element $k(\omega_i - \omega_i) = 1$. Therefore, to avoid any singularities, it is sufficient (but not necessary) to request that $R_i \ll 1$, namely that the off-diagonal

elements of the kernel matrix are small. For a minimum separation $\Delta\omega$ among the frequencies in Ω , this normally implies that the distribution must be broad enough, with $\Delta\vartheta \approx (\Delta\omega)^{-1}$.

D. Constrained solutions

Ideally, stochastic parameter shift rule should have the following properties

- I. The shifts should be constrained within a finite interval, $\vartheta \in [-B, B]$.
- II. It should be simple to sample from their probability distribution.
- III. The value of $\|c\|_1$ should be as small as possible.
- IV. The rule should be easy to compute even when there are many (exponentially) many frequencies.

None of the shift rules that we have introduced have all of these properties. The triangle shift rule (Algorithm 3) have basically all of these, with the exception of point I. Numerical solutions of Eq. (20) satisfy I-III by design, but become challenging when the number of frequencies is very large, namely they don't satisfy IV. Kernel methods can have limited support (e.g. the Uniform distribution in Table I), but they require the numerical inversion of a matrix that depends on the number of frequencies, which can be large.

Since we have unveiled the tight connection between shift rules and interpolation, an analytical approach to define shift rules that satisfy all of the above constraints might be to use band-limited interpolating functions [38], which are often based on Prolate Spheroidal Wave Functions [39, 40], the eigenfunctions of the sinc kernel. As such, these methods are tightly connected to the kernel interpolation with sinc kernels described above—see e.g. [41]. However, no simple analytical construction exists.

On the other hand, numerical solutions of Eq. (21) show lots of flexibility, as we can choose the shifts in the range we want and have guarantees of optimality, within that range. However, they don't satisfy point IV. Motivated by the triangle shift rule, which only depends on the bandwidth Eq. (12), that can easily be computed in many cases, even when $|\Omega|$ is exponentially large, we propose Algorithm 7 which defines an approximate interpolating function in $[-\Lambda, \Lambda]$.

Algorithm 7 Approximate shift rule for a given bandwidth

1: Given the bandwidth Λ , discretize the function $f(\omega) = \omega$ for $\omega \in [-\Lambda, \Lambda]$ and define

$$\Omega_\Lambda = \left\{ \frac{\ell}{L}\Lambda \quad : \quad \text{for } \ell = -L, \dots, L, \text{ and } L \leq P \right\}.$$

2: Use Algorithm 1 with the above Ω_Λ .

V. NUMERICAL SIMULATIONS

We now test the performance of the different Algorithms proposed in the previous section.

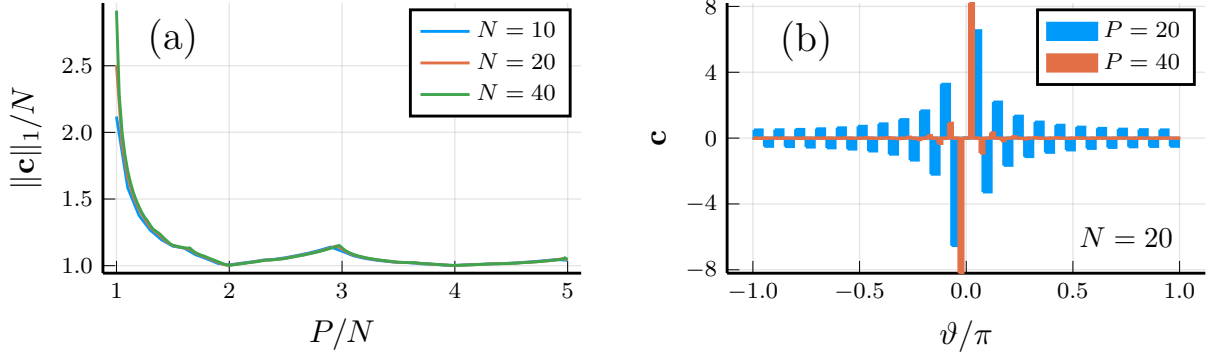


Figure 2. (a) Norm of solutions of (20) vs P/N for different values of N . (b) Solutions of (20) vs ϑ for $N=20$ and two values of P , $P=20$ (not overshifted) and $P=40$ (overshifted).

A. Equispaced frequencies

A common scenario, e.g. when using parameter sharing (Sec. II C) in qubit-based quantum circuits, or when dealing with photonic quantum circuits (Sec. VI A), consists of equispaced frequencies. In that setting,

$$\Omega = \{-N, -N+1, \dots, -1, 0, 1, \dots, N-1, N\}, \quad (40)$$

for a given integer N . This case has been extensively considered in the literature [5, 8, 9].

Since the resulting function is periodic with period 2π and Eq. (8) can be inverted using discrete Fourier transforms, there are basically two main approaches to define the shifts. Fix any integer $P \geq N$. Then, following Pappalardo *et al.* [8], we may set

$$\vartheta_p = \frac{2\pi p}{2P+1}, \quad \text{for } p \in \{-P, -P+1, \dots, P\}. \quad (41)$$

Alternatively, following Wierichs *et al.* [5] we may define the shifts as

$$\vartheta_p = \frac{\pi(2p-1)}{2P}, \quad \vartheta_{-p} = -\vartheta_p, \quad \text{for } p \in \{1, \dots, P\}. \quad (42)$$

In general, there is an odd number of shifts in Eq. (41), due to the extra “zero shift” $\vartheta_0 = 0$, while the number of shifts in Eq. (42) is even, without the “zero shift”. Both choices lead to a linear system with N equations and P variables in Eq. (10), so overshifting occurs when $P > N$. Note that in this particular case $c_0 = 0$ for odd shifts, so the extra shift does not play any role in the expansion Eq. (7).

For equispaced shifts the choice of Eq. (42) should be preferred, as it provides analytic shift rules when $P = N$, as shown by Wierichs *et al.* [5]. Suppose though that we were not aware of this explicit solution and that we decided to focus on the suboptimal choice Eq. (41). This is motivated by the fact that, for general Ω with possibly incommensurable frequencies, there is no explicit guideline to choose a particular set of shifts.

We use the shifts from Eq. (41) and solve Eq. (21) for different values of N and P . The results are shown in Fig. 2. As we see in Fig. 2(a), overshifting reduces $\|\mathbf{c}\|_1$, and hence the number of measurement shots thanks to Eq. (19). Our results show that, with the suboptimal choice of the shifts from Eq. (41), overshifting is always beneficial to this formulation, with the optimal at around $P = 2N$, where $\|\mathbf{c}\|_1 \simeq N$. Moreover, the relative advantage between the overshifted result

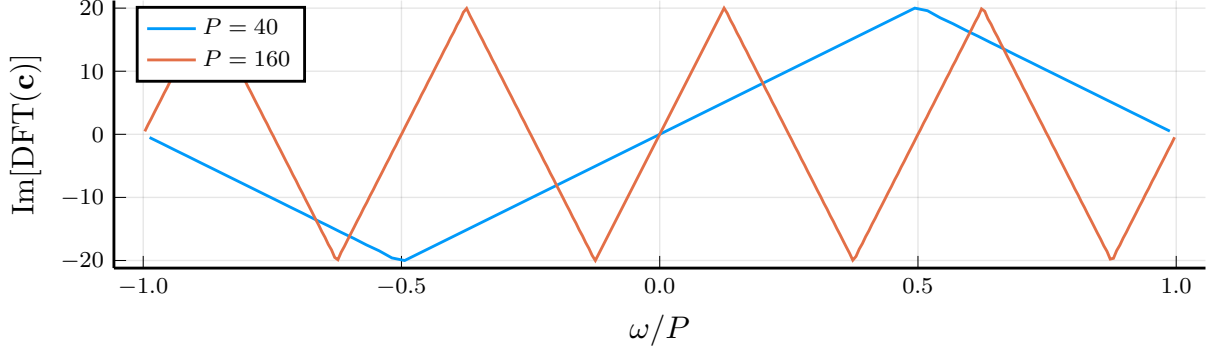


Figure 3. Discrete Fourier Transform of the solutions of (20) for $N = 20$ and $P = 40, 160$. The real parts are always zero.

for $P = 2N$, and the standard result for $P = 2N$ grows for larger N , mostly because the standard solution grows more than linearly. For instance, according to Fig. 2(a), for $N = 40$, the number of measurement shots required when $P = 2N$ is basically one third of that for $P = N$. The reason why overshifting is advantageous is apparent from the result from Fig. 2(b). Indeed, although overshifted rules have more shifts, their coefficients are much more sparse, with only a few shifts θ_p having a non-zero value. On the other hand, when $P = N$ all shifts have a relatively large coefficient c_p .

In Sec. IV we discussed several parameter shift rules with an analytic form. To have a better understanding about whether the optimal coefficients according to problem (20)—shown in Fig. 2(b)—can be linked to any of those analytic strategies, in Fig. 3 we plot the Discrete Fourier Transform (DFT) of such coefficients. In that plot, we clearly see a resemblance with triangle waves whose period depends on P . We recall that in Sec. III C we have shown that the Fourier coefficients must be non-zero even for $n \notin \Omega$, namely for $|n| > N$. From Fig. 3 we see that the solution of (20) basically extends (29) over $|n| > N$ with the zigzag trend of triangular waves. Therefore, we can use the analytic solution of Sec. (IV A) for any for any $\Lambda \geq N$, with optimal choice $\Lambda = N$. Indeed, for this choice $\|c\|_1 \leq N$ and the resulting triangle shift rule (Algorithm 3) has the same performance of numerical solutions of problem (20) shown in Figure 2(a).

The triangle shift rule can be further simplified in this setting. Indeed, exploiting the periodicity of $f(\theta)$ in Eq. (24), in appendix C we then find that the coefficients (33) can be manipulated to get

$$\frac{df(\theta)}{d\theta} = \sum_{t=0}^{N-1} \frac{(-1)^t}{2N(1 - \cos \vartheta_t)} [f(\theta + \vartheta_t) - f(\theta - \vartheta_t)], \quad (43)$$

where $\vartheta_t = \frac{\pi(2t+1)}{2N}$. The final expression is then equivalent to the one already obtained in [5, 9].

In summary, we started from the suboptimal choice Eq. (41), we used the analytic solution Eq. (33) based on the continuous limit $P \rightarrow \infty$, and then by manipulating the resulting expression we get Eq. (43) which uses the shifts from Eq. (42), which are different from our starting point Eq. (41). From our derivation, it is now clear that the choice from [5, 9] is optimal to minimize the number of measurement shots. It also shows how to use the continuous limit $P \rightarrow \infty$ to find the best set of shifts.

Moreover, from Eq. (43) we can define a stochastic parameter shift rule, which is now summarized in Algorithm 8.

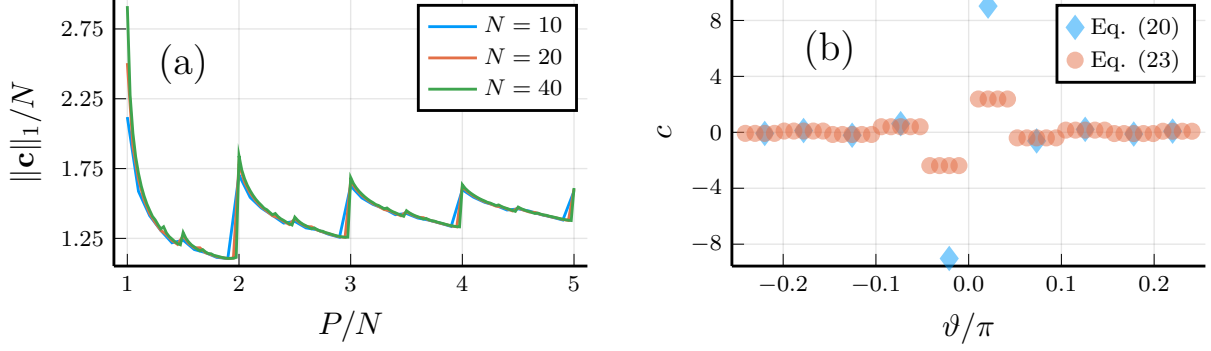


Figure 4. (a) Norm of solutions of (23) vs P/N for different values of N . (b) Solutions of (20) and (23) vs ϑ for $N = 20$ and $P = 95$. For visual clarity, only the values of $c(\vartheta)$ with $|c| > 0.05$ are displayed.

Algorithm 8 Equispaced Stochastic Parameter Shift Rule

- 1: Sample $t \in \{0, \dots, N-1\}$ from the probability distribution $p_N(t) = \frac{1}{N^2(1 - \cos \frac{\pi(2t+1)}{2N})}$, e.g. using Algorithm 10.
 - 2: Compute $\vartheta = \frac{\pi(2t+1)}{2N}$.
 - 3: Estimate $f(\theta \pm \vartheta)$ in a quantum device and call the unbiased estimated result $g_{\pm}$.
 - 4: Return an unbiased estimate of the gradient $(-1)^t(g_+ - g_-)N/2$.
-

Finally, in Fig. 4 we study the solutions of Eq. (23). In Fig. 4(a), we see that, as expected, $\|\mathbf{c}\|_1$ is typically larger than the one obtained from the solution of Eq. 20, shown in Fig. 2(a). In Fig. 4(b) we then show the coefficients obtained by solving either Eq. (20) or (23) for the same value of N and P . We note that the solutions of Eq. (20) display only a non-zero values of $c(\vartheta)$ are a re scattered in $[-\pi, \pi]$. On the other hand, the solutions of Eq. (23) are *clustered*. Therefore, we expect that these solutions are less affected by imperfections in tuning the shifts ϑ .

B. Arbitrary shifts

While equally-spaced shifts may be the most natural setting to consider, Eq. (8) also allows us to identify other solutions, for example we can obtain a qubit parameter shift rule $f'(x) \approx -f(x - \frac{\pi}{4}) + 0.707107f(x) + 0.292893f(x + \frac{\pi}{2})$, although for uniform noise assumptions both the conventional $\pm\pi/2$ and the noisier $\pm\pi/4$ parameter shift rules outperform it. Nonetheless this introduces further control that may be valuable to more limited systems or more convoluted noise budgets.

Randomly selecting n shifts from $[-\pi, \pi]^n$, we see in Fig. 5 that it is generally feasible to find better rules matching Wierichs *et al.* [5]. Moreover, these randomly selected shifts are more likely to lead to the convex solver finding a valid shift rule, and any such shift rule is less likely to be high cost.

VI. APPLICATIONS

A. Photonic Quantum Circuits

In photonic quantum circuits parametric gates are normally implemented via linear optical elements [42, 43]. All linear optical components can be expressed as a fixed component, e.g. a

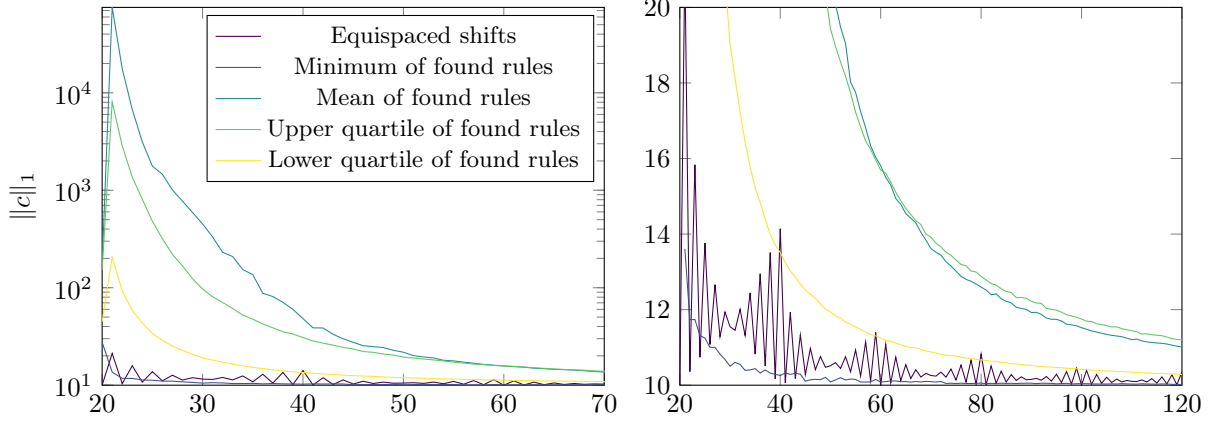


Figure 5. Result of 10000 randomly-generated arbitrary shifts for $N = 10$ with an equispaced spectrum. Equispaced shifts are $\{(-1 + \frac{1}{N_{\text{Shifts}}})\pi, (-1 + \frac{3}{N_{\text{Shifts}}})\pi, \dots, (1 - \frac{1}{N_{\text{Shifts}}})\pi\}$, encompassing the Wierichs *et al.* [5] and Pappalardo *et al.* [8] shifts.

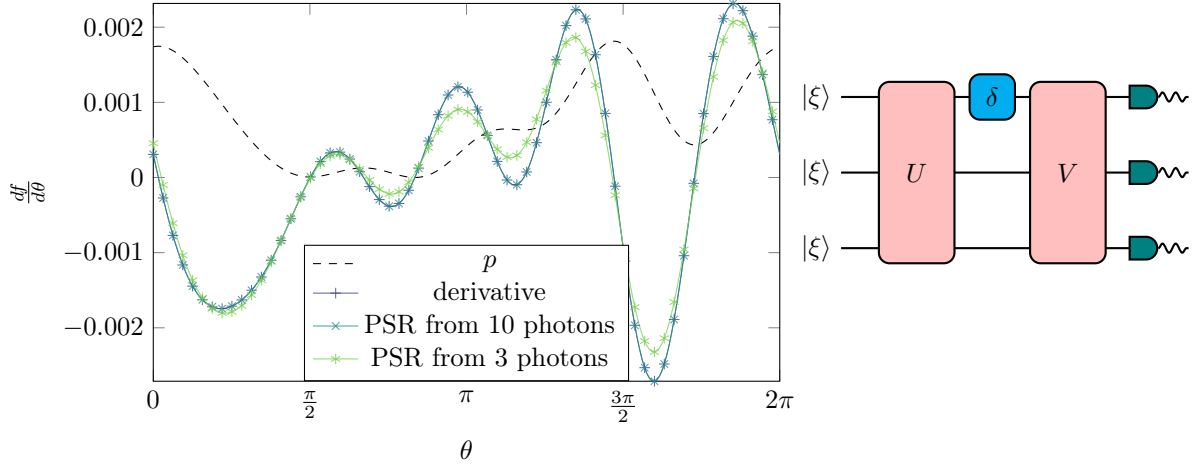


Figure 6. Probability of detecting five photons as (3,1,1) in a three-mode interferometer with squeezed vacuum input as a function of a single phase in the first mode.

50/50 beam splitter, and a phase gate $e^{i\theta\hat{n}}$ where $\hat{n} = \hat{a}^\dagger\hat{a}$ is the photon number operator. When using single-photon sources, let N be the maximum number of photons entering into a mode. We are then within the settings of Sec. V A, where the possible frequencies (40) are due to the allowed eigenvalues of the $\hat{n}$ operator.

B. Gaussian States

While Gaussian states are in principle infinite-dimensional, with an unbounded spectrum, finite energy states can be approximated by a Fock space expansion, while certain events can be evaluated exactly from the relevant fixed-number subspaces following Eq. (5).

When working with photon number resolving detectors, the proper subspace can be identified from the measurement outcomes. For instance, suppose that $|\psi\rangle$ is a Gaussian state and $|\mathbf{n}\rangle\langle\mathbf{n}|$ are projectors modelling photon number resolving measurements with outcomes n_k on mode k . Here $|\mathbf{n}\rangle = |n_1, n_2, \dots\rangle$ where $|n_j\rangle$ are Fock states with n_j photons. We can reinterpret the probability of

getting a particular outcome, $|\langle \mathbf{n} | \psi \rangle|^2$, as the creation of the multi-mode Fock state $|\mathbf{n}\rangle$ followed by the measurement of the Gaussian observable $|\psi\rangle\langle\psi|$. In other terms, if measurement results always satisfy $\sum_k n_k \leq N$ for a certain cutoff N , then the Hilbert space can be approximated as finite dimensional and, for linear optical circuits, the results of Sec. V still apply.

Fig. 6 shows the probability of detecting a given number of photons at the output of a circuit consisting of a phase shift in one arm of a three-mode Mach–Zehnder-like interferometer, with a squeezed vacuum input. As the detection is on a Fock state it can—as the observable is equivalent to the outcome of homodyne detection on a fixed photon number state—be solved exactly with a PSR for that number of photons, shift rules for a lower number of photons can still approximate the derivative.

C. Hamiltonian Dynamics of Many-Body Systems

As an another example application we consider functions obtained by letting a many-body quantum system evolve with some Hamiltonian $\hat{H}$, e.g.

$$f_H(\theta) = \langle \psi | e^{i\hat{H}\theta} \hat{O} e^{-i\hat{H}\theta} | \psi \rangle, \quad (44)$$

where $|\psi\rangle$ is a suitable initial state and $\hat{O}$ is an observable. In order to focus on a non-trivial, yet analytically solvable model we focus on a spin chain with L qubits interacting via the XY Hamiltonian

$$\hat{H} = \frac{1}{4} \sum_{i=1}^{L-1} (\hat{X}_i \hat{X}_{i+1} + \hat{Y}_i \hat{Y}_{i+1}), \quad (45)$$

where $\hat{X}_i$ and $\hat{Y}_i$ are Pauli matrices acting on qubit i . The above Hamiltonian can be exactly diagonalized (see e.g. [44]) with energies $E_k = \cos(\pi k / (L + 1))$ for $k = 1, \dots, L$. We assume that $|\psi\rangle$ and $\hat{O}$ are chosen such that $f_H(\theta)$ can be expanded as

$$f_H(\theta) = \sum_{\omega \in \Omega_H^+} \frac{\cos(\omega\theta)}{\omega}, \quad (46)$$

where Ω_H^+ is the set of positive frequencies $\omega = E_k - E_\ell$, with $\omega > 0$.

In Fig. 7 we plot the derivative, and the standard deviations of different estimators discussed in the previous section for $L = 10$ qubits. In such case $N = |\Omega_H^+| = 25$. All the estimators were capable of almost perfectly reproducing the value of the derivative, but some estimators have larger variances. In particular, those obtained by numerically solving Eq. (20) with different values of P show little differences, and the analytical estimator obtained by sampling from the Cauchy distribution almost matches their performance. Note though that the latter comes at the price of a non-negligible probability of sampling large values of θ , since the Cauchy distribution has long tails. On the other hand, in numerical solutions with $P = N, 2N, 4N$, the range of possible ϑ in Eq. (20) is constrained by design to a fixed interval, here $[-2\pi, 2\pi]$, so their performance can be beaten by other methods with a different interval.

D. Jaynes-Cummings

As another example, we focus on the Jaynes-Cummings Hamiltonian, a popular model of the interactions between an atom and an optical cavity [45]:

$$\hat{H}_{\text{JC}} = \frac{\delta}{2} \hat{Z} + \frac{\lambda}{2} (\hat{a}^\dagger \hat{\sigma}_- + \hat{a} \hat{\sigma}_+), \quad (47)$$

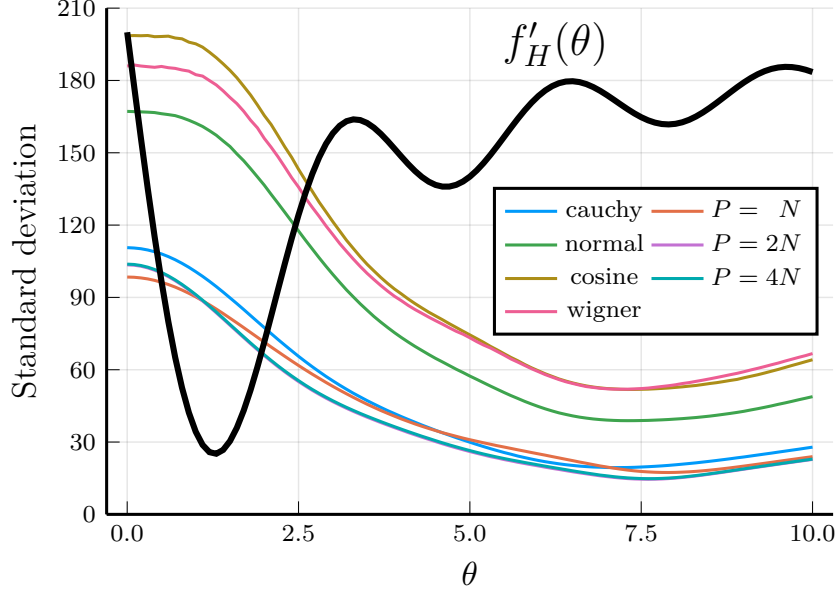


Figure 7. Standard deviation of different estimators of the derivative $f'_H(\theta)$, with the $f_H(\theta)$ defined in Eq. (46). The exact value of $f'_H(\theta)$, rescaled in order to be shown with the same axes of the standard deviation, is shown in black as a reference. The function values at each point, and their variances, are empirically computed using 10^7 samples. For these numbers, the error in the estimation of $f'_H(\theta)$ is $\approx 200/\sqrt{10^7}$ and all the estimators basically reproduce the true derivative without observable errors in the plot. Cauchy, normal, cosine, and Wigner refer to the analytic estimators discussed in Sec. IV and have been numerically computed using Algorithm 6, while $P = nN$ with $n = 1, 2, 4$ refer to numerical solutions of Eq. (20) where N is the number of positive frequencies, and the estimators have been numerically computed using Algorithm 2.

where $\hat{\sigma}_{\pm} = (\hat{X} \pm i\hat{Y})/2$, $\hat{X}, \hat{Y}, \hat{Z}$ are the Pauli matrices, while $\hat{a}^\dagger$ and $\hat{a}$ are, respectively, bosonic creation and annihilation operators. Since $[\hat{Z} + \hat{a}^\dagger\hat{a}, \hat{H}] = 0$, the Hamiltonian can be diagonalized in each subsector where $\hat{Z} + \hat{a}^\dagger\hat{a}$ is diagonal and constant. The resulting eigenvalues are

$$E_n = \sqrt{\delta^2 + \lambda^2(n+1)}, \quad n = 0, 1, 2, \dots, \infty. \quad (48)$$

Although the bandwidth can grow up to infinity, states that are produced in the lab have an energy constraint. We can then put an arbitrary cut-off on the photon number, meaning that we can approximate the infinite operators $\hat{a}$ and $\hat{a}^\dagger$ as $(n+1)$ -dimensional matrices.

As an example we focus on the function

$$f_{\text{JC}}(\theta) = \langle \psi(\alpha) | e^{i\hat{H}_{\text{JC}}\theta} \hat{Z} e^{-i\hat{H}_{\text{JC}}\theta} | \psi(\alpha) \rangle, \quad (49)$$

where $|\psi(\alpha)\rangle \propto (e^{\alpha\hat{a}^\dagger}|0\rangle) \otimes |1\rangle$. Numerical results are shown in Fig. 8, where we compare the performance of the Approximate shift rule and the Triangle shift rule, both of which only require an estimate of the bandwidth Eq. (12). However, since the bandwidth is infinite for this model, we perform two approximations. Firstly, we estimate the bandwidth Λ by fixing an energy truncation to 10 bosons. If we now try to mimic the experimental evaluation of Eq. (7), even when the coefficients c_p are estimated by assuming this energy truncated model, then for each function evaluation $f_{\text{JC}}(\theta + \vartheta_p)$ we should basically sum over an infinite number of frequencies. Since we cannot perform this limit exactly in numerical simulations, as a second approximation we estimate each $f_{\text{JC}}(\theta + \vartheta_p)$ with a much larger cutoff (100 photons) than that used to define the shift rules.

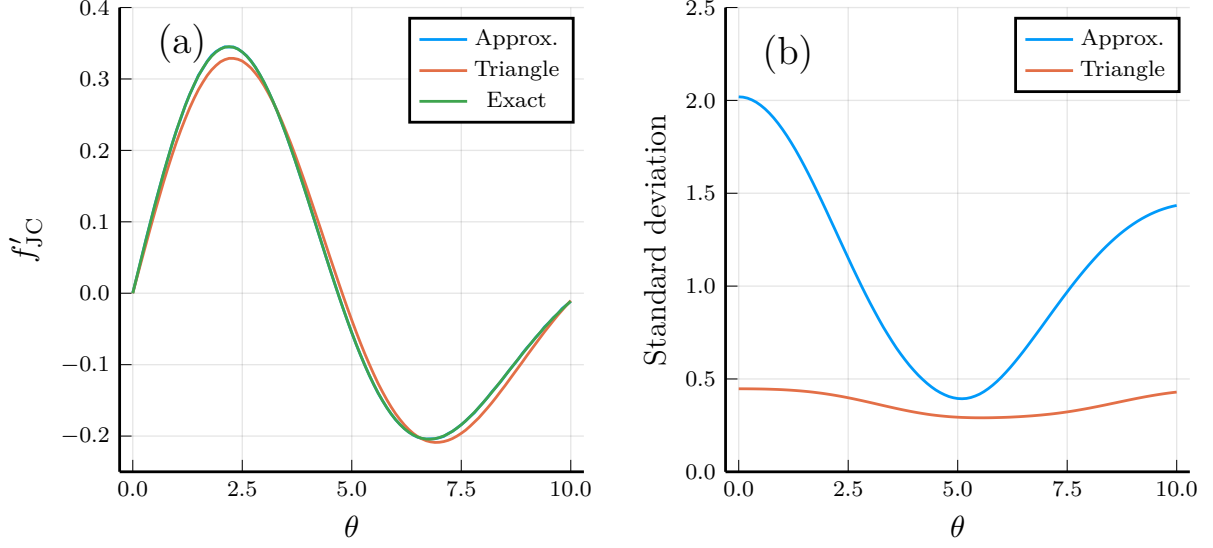


Figure 8. Different estimators of the derivative $f'_{JC}(\theta)$, with the $f_{JC}(\theta)$ defined in Eq. (49) and their standard deviations. We used $\alpha = 1$, $\delta = 0.2$ and $\lambda = 0.5$. We considered two algorithms: the triangle shift rule from Algorithm 3 and the Approximate shift rule Algorithm 7 with $L = 100$, $P = 1000$ and $\vartheta \in [-\pi, \pi]$. Since the model has infinite bandwidth, we define parameter shift rules with a bandwidth Eq. (12) estimated by truncating the model to 10 bosons. Nonetheless, the functions are computed using a larger Hilbert space with a truncation up to 100 bosons, in order to approximate the infinite limit. In (a) the Approximate shift rule is basically indistinguishable from the exact expressions (error at most 10^{-3}), while the triangle shift rule shows some bias. In (b) we see that the triangle shift rule has a lower standard deviation, as we expect from its optimality, though it introduces a bias in (a) due to the wrong estimation of the bandwidth.

In this way we try to mimic the experimental setting where we define our derivatives assuming an energy constraint, but the functions in Eq. (7) are estimated without such restriction.

From our numerical solutions, shown in Fig. 8, we see that the both the Approximate and Triangle shift rules have a $\mathcal{O}(1)$ variance, with the triangle shift rule having a smaller variance since it has access to an unbounded set of shifts. Nonetheless, in Fig. 8(a) we see that the triangle shift rule introduces a bias, while the estimate of the approximate shift rule is basically indistinguishable from the exact one.

The reason behind such bias may be that the triangle shift rule sometimes samples a large shift ϑ , so possibly the errors due to energy truncation get amplified. As a further proof of our intuition, we note that if we use the same truncation to estimate the bandwidth and to evaluate the function $f_{JC}(\theta + \vartheta)$, then both shift rules produce almost exact results, with a $\mathcal{O}(10^{-3})$ error that is compatible with the finite amount of samples (10^7).

Our results show that we can play with energy truncation to get a reliable estimate of derivatives even for models with incommensurable and infinitely many energies.

E. Parameter sharing

Finally, we focus on quantum circuits with dependent parameters.

1. Variational quantum circuits

As a simple case, consider a problem similar to a Variational Quantum Eigensolver (VQE), where the task is to variationally approximate the ground state of a Hamiltonian $\hat{H}$. The variational circuit is constructed as in Eq. (1) with some entangling layers $\hat{W}_\ell$ and fixed rotations. For simplicity, here we assume that the rotations are always around the Z axis, namely $\hat{H}_\ell = \hat{Z}_{q_\ell}$ is a Pauli Z gate on qubit q_ℓ . In this setting

$$f(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | \hat{H} | \psi(\boldsymbol{\theta}) \rangle. \quad (50)$$

We assume a parametrization

$$\theta_i = w_i \theta, \quad (51)$$

with some fixed weights w_i and the goal is to take gradients with respect to the tunable θ . Approaches like this, namely to project the parameter space on a reduced manifold, are routinely used in deep neural networks to improve generalization [46].

Gradients with respect to the parametrization Eq. (51) can be obtained using standard parameter shift rule as

$$\frac{df}{d\theta} = \sum_i \frac{df}{d\theta_i} w_i = \sum_i w_i \left[f(\boldsymbol{\theta} + \frac{\pi}{4} \mathbf{e}_i) - f(\boldsymbol{\theta} - \frac{\pi}{4} \mathbf{e}_i) \right], \quad (52)$$

where $\mathbf{e}_i$ is the basis vector with elements $(\mathbf{e}_i)_j = \delta_{ij}$. The above can be rewritten as in Eq. (7) with $\mathbf{c} = (w_1, -w_1, w_2, -w_2, \dots)$ with $\|\mathbf{c}\|_1 = 2\|\mathbf{w}\|_1$.

On the other hand, finding the optimal shift rule is prohibitively expensive by working with the convex problem (20), as the set of frequencies (16) increases exponentially with the number of layers. However, since each Z rotation has a frequency ± 1 , as discussed in Sec. II C, finding the bandwidth is straightforward and we get.

$$\Lambda = 2\|\mathbf{w}\|_1. \quad (53)$$

Therefore, it is still possible to use Algorithms like the *Triangle shift rule* that only depend on the bandwidth and work even for incommensurable frequencies, without having to solve complicated equations. From the discussion in Appendix A regarding stochastic shift rules and from that in Eq. (17) about the shot allocation with Eq. (52), we find that the error coming from the application of the standard parameter shift rule together with the chain rule of derivatives is comparable to that of Triangle shift rule.

As discussed previously, one of the disadvantages of the triangle shift rule is that it may sample large values of the shift ϑ . Nonetheless, this is not a problem for this example as Z rotations are periodic, so we can always take $\theta_i = w_i \theta \bmod 2\pi$, which never gets larger than 2π .

2. Structured photonic quantum circuits

In order to demonstrate Sec. II C we consider a small temporarily multiplexed cluster state generation scheme, in the manner of Larsen *et al.* [15], where three temporally spaced pairs pass through the same beam splitter—with a transmittivity tunable through a phase—before a time delay in one mode, a second—fixed and balanced—beam splitter, a further time delay in one mode, and a final—fixed and balanced—beam splitter. Fig. 9 then shows the probability of detecting a specific number of counts in each mode, and the PSR attainable without needing to tune the beam splitter transmittivity within an iteration.

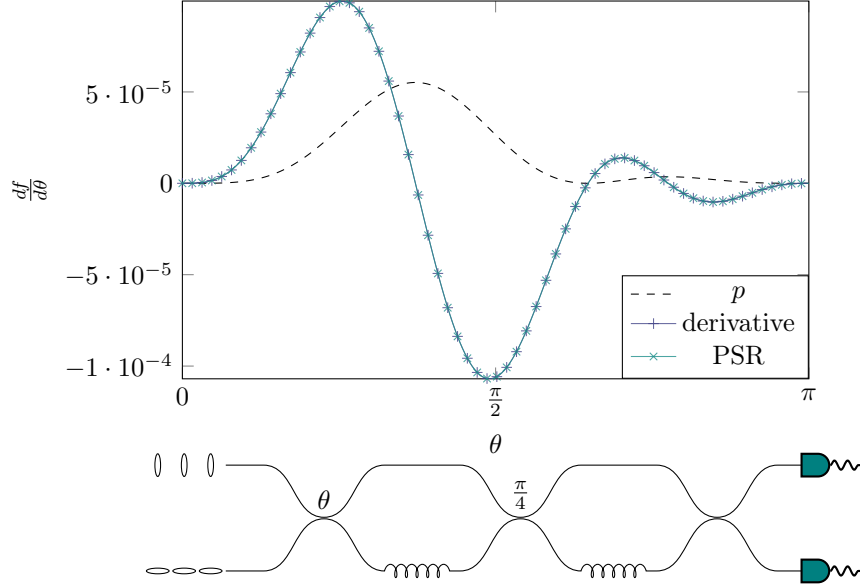


Figure 9. Probability of detecting a given set of counts $\{n_j\}$ in a 2-spatial mode interferometer, where the first physical beam splitter (that acts on each time-binned pair) varies with transmittance $\cos\theta$. The PSR is exact, being evaluated for more energies than $\sum_j n_j$.

VII. CONCLUSIONS

We have generalized the parameter shift rule to circuits generated by general gates $e^{i\theta\hat{H}}$ with arbitrary Hamiltonians $\hat{H}$ and arbitrary dimensions of the Hilbert space. Our results allow for the estimation of gradients of complex quantum circuits and quantum evolutions directly via measurements on the quantum hardware, with guaranteed minimal overhead. We have shown how to define rules that work even when the spectrum of the $\hat{H}$ is unknown or unbounded, e.g. for infinite dimensional systems. We have applied our findings to estimate derivatives of structured quantum circuits, Gaussian circuits and circuits made with spin-boson interactions.

Although not explicitly studied here, our results can be readily generalized to other settings. For instance, following Ref. [5] we can extend our rules to estimate higher order derivatives by simply changing Eq. (20). Moreover, by trivially adapting the stochastic simulation techniques from Ref. [6], our rules can be also extended to gates where the parameter θ is one of the parameters in the system Hamiltonian, e.g. for gates $e^{i(\hat{H}_0 + \theta\hat{V})}$ with arbitrary operators $\hat{H}_0$ and $\hat{V}$.

Upon completion of this work, a paper by Lai *et al.* [47] appeared that proposes to use gradient descent to find new parameter shift rules for general spectra. Their results complement our work rather than competing with it. Indeed, their problem is non-convex, meaning that gradient descent is not guaranteed to find a solution, nor the optimal solution with minimum measurement overhead. On the other hand, our Eq. (20) is convex and optimal, but requires a fixed choice of the shifts. These two methods can be combined together. For instance, one could use Eq. (20) to get a first estimate of the shifts, and then fine tune the results (selecting only the shifts with large enough coefficients) via gradient descent following [47]. Nonetheless, our techniques has another advantage, as it can be applied even when the spectrum of the Hamiltonian that generates the gate is unknown and possibly unbounded, as we have shown with the triangle shift rule and with Algorithm 9.

ACKNOWLEDGMENTS

This work is supported by the European Union's Horizon Europe research and innovation program under the EPIQUE Project (Grant Agreement No. 101135288).

Appendix A: Shot allocation in stochastic parameter shift rules

We now focus on deciding the optimal shot allocation to estimate derivatives via Eq. (26). We first note that we can rewrite Eq. (26) as

$$\frac{df(\theta)}{d\theta} = \|c\|_1 \sum_{s=\pm} p_s \int d\vartheta p_s(\vartheta) f(\theta + \vartheta) s = \|c\|_1 \mathbb{E}_{s,\vartheta} [sf(\theta + \vartheta)], \quad (\text{A1})$$

where $p_{\pm} = 1/2$. In hardware, f is estimated from the measurement of an observable that here we call $\hat{Y}$. Let us assume that the measured observable has eigenvalues y and eigenvectors $|y\rangle$. From Born's rule, the measurement of $\hat{Y}$ results in a probability distribution $p(y|\vartheta) = |\langle y|\psi(\theta + \vartheta)\rangle|^2$ that depends on ϑ (and the fixed θ). Let $x = (s, \vartheta)$ be a tuple and $e(x) = s\|c\|_1$ be the function that extract the first element of such tuple (the sign) and multiplies it by $\|c\|_1$. Then we can rewrite Eq. (A1) as

$$\frac{df(\theta)}{d\theta} = \mathbb{E}_{x \sim p(x)} \mathbb{E}_{y \sim p(y|x)} [e(x)y], \quad (\text{A2})$$

where $p(x) = p_s p_s(\vartheta)$ for $x = (s, \vartheta)$ and $p(y|x) = p(y|\vartheta)$. The problem is then reduced to a standard problem in statistics.

Let $\mu = \mathbb{E}[e(X)Y]$ be the quantity that we want to estimate. Suppose we draw n distinct $x_i \sim p(x)$. For each x_i , we take $m_i \geq 1$ conditionals $y_{ij} \sim p(y|x_i)$ and use the estimator

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{m_i} \sum_{j=1}^{m_i} e(x_i) y_{ij} \right). \quad (\text{A3})$$

By the law of total variance,

$$\text{Var}(\hat{\mu}) = \frac{1}{n} \text{Var}[e(X)Y(X)] + \mathbb{E} \left(\frac{1}{n^2} \sum_{i=1}^n \frac{e(x_i)^2 \text{Var}(Y|x_i)}{m_i} \right), \quad (\text{A4})$$

where $Y(x) = \mathbb{E}_{y \sim p(y|x)}[y] \equiv f(\theta + \vartheta)$. The first term denotes the variance due to the uncertainty on x , while the second term denotes the variance due to the uncertainty coming from quantum measurements, with $\text{Var}(Y|x) = \langle \hat{Y}^2 \rangle_x - \langle \hat{Y} \rangle_x^2$. Clearly the first term does not depend on m_i , while the second does. Assuming that the total number of measurement shots is constrained by $\sum_{i=1}^n m_i = S$, then we can optimize the shot allocation by the following program

$$\text{minimize} \quad \left[\sum_{i=1}^n \frac{w_i^2}{m_i} \right] \quad \text{with} \quad \sum m_i = S, \quad \text{where } w_i^2 := e(x_i)^2 \text{Var}(Y|x_i). \quad (\text{A5})$$

Via a Lagrange multiplier we get

$$m_i \propto w_i = |e(x_i)| \sqrt{\text{Var}(Y|x_i)}, \quad (\text{A6})$$

namely we should allocate more samples where where $|e(x)|$ is large and/or where $Y|x$ is more noisy. For our problem Eq. (A2), $|e(x)| = \|c\|_1$ is constant and the variance of the quantum observable

can always be bounded by a constant quantity – e.g., for Pauli observables $Y^2 = 1$. With that suboptimal solution $S = mn$ and

$$\text{Var}(\hat{\mu}) = \frac{m}{S} \text{Var}[e(X)Y(X)] + \frac{1}{S} \mathbb{E}[e(X)^2 \text{Var}(Y|X)]. \quad (\text{A7})$$

Since the above function is increasing for larger m , the optimal choice is

$$n = S, \quad m = 1, \quad (\text{A8})$$

namely use all the shots to sample θ and s , and use the quantum hardware to estimate $f(\theta + \vartheta)$ with a single shot.

As an example, let us consider the case where $\hat{Y}$ is such that $\hat{Y}^2 = 1$. Then

$$\text{Var}[e(X)Y(X)] = \|c\|_1^2 \mathbb{E}_\theta \langle \hat{Y} \rangle_\theta^2 - \frac{df^2}{d\theta}, \quad (\text{A9})$$

$$\mathbb{E}[e(X)^2 \text{Var}(Y|X)] = \|c\|_1^2 \mathbb{E}_\theta \left(\langle \hat{Y}^2 \rangle_\theta - \langle \hat{Y} \rangle_\theta^2 \right) = \|c\|_1^2 \left(1 - \mathbb{E}_\theta \langle \hat{Y} \rangle_\theta^2 \right). \quad (\text{A10})$$

Inserting that expression in Eq. (A7) we get

$$\text{Var} \left[\frac{df}{d\theta} \right] = \frac{1}{S} \left(\|c\|_1^2 - \frac{df^2}{d\theta} \right), \quad (\text{A11})$$

which is compatible with Eq. (17) with the optimal shot allocation, where $\sigma = 1$.

Finally, we consider how to reshape the above estimator in order to avoid unnecessary circuit recompilations, as mentioned at the end of Sec. III B. The resulting algorithm is shown as Algorithm 9. Other estimators might also be more accurate, e.g. those based on median of means or more recent variations [48], but we found the simple mean to be accurate enough in our numerical simulations.

Algorithm 9 Estimate Eq. (A3) without unnecessary circuit compilations

- 1: Consider the estimator Eq. (A3) with a total number of shots S and shot allocation $m_i = 1$ and $n = S$.
- 2: Sample $\mathcal{X} = \{x_i\}_{i=1}^S$ from the joint distribution of parameter shifts and signs $p(x)$.
- 3: Let $\{\hat{x}_i\}_{i=1}^n$ be the set different elements of $\mathcal{X}$ and let $n \leq S$ be its cardinality.
- 4: Let m_i be the number of occurrences of $\hat{x}_i$ in $\mathcal{X}$, so $\sum_{i=1}^n m_i = S$.
- 5: Then we can rewrite our estimator as

$$\hat{\mu} = \frac{1}{S} \sum_{i=1}^n \sum_{j=1}^{m_i} e(\hat{x}_i) \hat{y}_{ij} = \sum_{i=1}^n \frac{m_i}{S} \frac{1}{m_i} \sum_{j=1}^{m_i} e(\hat{x}_i) \hat{y}_{ij}.$$

where now $\hat{y}_{ij}$ is sampled from $p(y|\hat{x}_j)$, where $\hat{x}_j$ are all different. In other terms, we should call the quantum device only with the different values of the sampled shifts $\hat{x}_j$ and with an adaptive number of shots m_i that depends on the number of occurrences of $\hat{x}_j$ in $\mathcal{X}$. Note the differences with Eq. (A3): now there is only a final mean over the total S shots rather than a “mean of means”. In the second equality we show that the result can also be expressed as a weighted mean of means, with relative weight m_i/S .

Appendix B: Some useful algorithms

In this appendix we simply remind a few useful algorithms to sample from complex distributions, namely Algorithms 10 and 11.

Algorithm 10 Sample $t \in [0, \dots, T-1]$ from a distribution $p_T(t)$ with $\sum_{t=0}^{T-1} p_T(t) = 1$

- 1: Sample x uniformly from $[0, 1] \subset \mathbb{R}$.
- 2: Divide $[0, 1]$ into T intervals

$$[c_0, c_1], [c_1, c_2], \dots, [c_{T-1}, c_T]$$

where $c_0 = 0$, $c_T = 1$ and $c_t = \sum_{k=0}^{t-1} p_T(k)$ is the cumulative distribution.

- 3: Return the index of the interval where x belongs, namely t such that $c_t \leq x < c_{t+1}$.
-

Algorithm 11 Sample from a continuous distribution $p(\theta)$: inverse sampling

- 1: Sample u uniformly from $[0, 1] \subset \mathbb{R}$.
 - 2: Find θ such that $F(\theta) = u$, where $F(\theta) = \int_{-\infty}^{\theta} p(\vartheta) d\vartheta$ is the cumulative distribution.
 - 3: Return θ .
-

Appendix C: Recovering shift rules for equispaced frequencies

Starting from Eq. (33) we note that, after defining $\vartheta_t = \frac{\pi(2t+1)}{2N}$, we have $\vartheta_{t+N} = \pi + \vartheta_t$ and $\vartheta_{t+2N} = 2\pi - \vartheta_t$. Therefore, we can write (33) as

$$c(\vartheta) = \frac{N}{2} \sum_{t=0}^{2N-1} (-1)^t \eta_t^{(2N)} [\delta(\vartheta - \vartheta_t) - \delta(\vartheta + \vartheta_t)] \quad (\text{C1})$$

where

$$\eta_t^{(2N)} = \frac{8}{\pi^2} \sum_{k=0}^{\infty} \frac{(-1)^{2Nk}}{(2t + 4Nk + 1)^2} = \frac{1}{2\pi^2 N^2} \psi^{(1)}\left(\frac{2t+1}{4N}\right), \quad (\text{C2})$$

$\psi^{(1)}(z) = d^2/dz^2 \log \Gamma(z)$ is the “trigamma” function and $\Gamma(z)$ is Gamma function, namely the analytic extension to the factorial function. Applying this function in the parameter shift rule (24) we get

$$\frac{df(\theta)}{d\theta} = \frac{N}{2} \sum_{t=0}^{2N-1} (-1)^t \eta_t^{(2N)} [f(\theta + \vartheta_t) - f(\theta - \vartheta_t)] \quad (\text{C3})$$

$$= \frac{N}{2} \sum_{t=0}^{N-1} (-1)^t \eta_t^{(2N)} [f(\theta + \vartheta_t) - f(\theta - \vartheta_t)] + \quad (\text{C4})$$

$$(-1)^{t+N} \eta_{t+N}^{(2N)} [f(\theta + \vartheta_{t+N}) - f(\theta - \vartheta_{t+N})]$$

$$= \frac{N}{2} \sum_{t=0}^{N-1} (-1)^t \eta_t^{(2N)} [f(\theta + \vartheta_t) - f(\theta - \vartheta_t)] - \quad (\text{C5})$$

$$(-1)^t \eta_{2N-1-t}^{(2N)} [f(\theta + \vartheta_{2N-1-t}) - f(\theta - \vartheta_{2N-1-t})]$$

$$= \frac{N}{2} \sum_{t=0}^{N-1} (-1)^t \left[\eta_t^{(2N)} + \eta_{2N-1-t}^{(2N)} \right] [f(\theta + \vartheta_t) - f(\theta - \vartheta_t)], \quad (\text{C6})$$

where we used the fact that $f(\theta) = f(\theta + 2\pi)$ and $\vartheta_{2N-1-t} = 2\pi - \vartheta_t$. Using known functional forms, we then get

$$\frac{N}{2} \left[\eta_t^{(2N)} + \eta_{2N-1-t}^{(2N)} \right] = \frac{1}{4n \sin^2\left(\frac{2\pi t + \pi}{4n}\right)} \quad (\text{C7})$$

and the solution (43).

-
- [1] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, Variational quantum algorithms, *Nature Reviews Physics* **3**, 625 (2021).
 - [2] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, *et al.*, Noisy intermediate-scale quantum algorithms, *Reviews of Modern Physics* **94**, 015004 (2022).
 - [3] A. W. Harrow and J. C. Napp, Low-depth gradient measurements can improve convergence in variational hybrid quantum-classical algorithms, *Physical Review Letters* **126**, 140502 (2021).
 - [4] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, Evaluating analytic gradients on quantum hardware, *Physical Review A* **99**, 032331 (2019).
 - [5] D. Wierichs, J. Izaac, C. Wang, and C. Y.-Y. Lin, General parameter-shift rules for quantum gradients, *Quantum* **6**, 677 (2022).
 - [6] L. Banchi and G. E. Crooks, Measuring analytic gradients of general quantum evolution with the stochastic parameter shift rule, *Quantum* **5**, 386 (2021).
 - [7] R. Wiersema, D. Lewis, D. Wierichs, J. Carrasquilla, and N. Killoran, Here comes the su (n): multi-variate quantum gates and gradients, *Quantum* **8**, 1275 (2024).
 - [8] A. Pappalardo, P.-E. Emeriau, G. de Felice, B. Ventura, H. Jaunin, R. Yeung, B. Coecke, and S. Mansfield, Photonic parameter-shift rule: Enabling gradient computation for photonic quantum computers, *Physical Review A* **111**, 032429 (2025).
 - [9] F. Hoch, G. Rodari, T. Giordani, P. Perret, N. Spagnolo, G. Carvacho, C. Pentangelo, S. Piacentini, A. Crespi, F. Ceccarelli, R. Osellame, and F. Sciarrino, Variational approach to photonic quantum circuits via the parameter shift rule, *Physical Review Research* **7**, 023227 (2025).
 - [10] G. Facelli, D. D. Roberts, H. Wallner, A. Makarovskiy, Z. Holmes, and W. R. Clements, Exact gradients for linear optics with single photons, *arXiv preprint arXiv:2409.16369* (2024).
 - [11] B. Bauer, S. Bravyi, M. Motta, and G. K.-L. Chan, Quantum algorithms for quantum chemistry and quantum materials science, *Chemical reviews* **120**, 12685 (2020).
 - [12] T. Roy, T. Kim, A. Romanenko, and A. Grassellino, *Qudit-based quantum computing with SRF cavities at Fermilab*, Tech. Rep. (Fermi National Accelerator Laboratory (FNAL), Batavia, IL (United States), 2024).
 - [13] W. Chen, Y. Lu, S. Zhang, K. Zhang, G. Huang, M. Qiao, X. Su, J. Zhang, J.-N. Zhang, L. Banchi, *et al.*, Scalable and programmable phononic network with trapped ions, *Nature Physics* **19**, 877 (2023).
 - [14] E. Crane, K. C. Smith, T. Tomesh, A. Eickbusch, J. M. Martyn, S. Kühn, L. Funcke, M. A. DeMarco, I. L. Chuang, N. Wiebe, *et al.*, Hybrid oscillator-qubit quantum processors: Simulating fermions, bosons, and gauge fields, *arXiv preprint arXiv:2409.03747* (2024).
 - [15] M. V. Larsen, X. Guo, C. R. Breum, J. S. Neergaard-Nielsen, and U. L. Andersen, Deterministic generation of a two-dimensional cluster state, *Science* **366**, 369 (2019).
 - [16] T. Giordani, F. Hoch, G. Carvacho, N. Spagnolo, and F. Sciarrino, Integrated photonics in quantum technologies, *La Rivista del Nuovo Cimento* **46**, 71 (2023).
 - [17] N. Maring, A. Fyrrillas, M. Pont, E. Ivanov, P. Stepanov, N. Margaria, W. Hease, A. Pishchagin, A. Lemaître, I. Sagnes, *et al.*, A versatile single-photon-based quantum computing platform, *Nature Photonics* **18**, 603 (2024).
 - [18] H. Aghaee Rad, T. Ainsworth, R. Alexander, B. Altieri, M. Askarani, R. Baby, L. Banchi, B. Baragiola, J. Bourassa, R. Chadwick, *et al.*, Scaling and networking a modular photonic quantum computer, *Nature* **638**, 912 (2025).
 - [19] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, Quantum circuit learning, *Physical Review A* **98**, 032309 (2018).
 - [20] T. Knopp, M. Boberg, and M. Grosser, NFFT.jl: Generic and fast julia implementation of the nonequidistant fast Fourier transform, *SIAM Journal on Scientific Computing* **45**, C179 (2023).
 - [21] T. N. Ikeda, S. Sugiura, and A. Polkovnikov, Robust effective ground state in a nonintegrable floquet quantum circuit, *Physical Review Letters* **133**, 030401 (2024).

- [22] P. C. Humphreys, B. J. Metcalf, J. B. Spring, M. Moore, X.-M. Jin, M. Barbieri, W. S. Kolthammer, and I. A. Walmsley, Linear Optical Quantum Computing in a Single Spatial Mode, *Physical Review Letters* **111**, 150501 (2013).
- [23] L. S. Madsen, F. Laudenbach, M. F. Askarani, F. Rortais, T. Vincent, J. F. Bulmer, F. M. Miatto, L. Neuhaus, L. G. Helt, M. J. Collins, *et al.*, Quantum computational advantage with a programmable photonic processor, *Nature* **606**, 75 (2022).
- [24] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd, Convex optimization in Julia, in *Proceedings of the 1st First Workshop for High Performance Technical Computing in Dynamic Languages* (IEEE Press, 2014) pp. 18–28.
- [25] S. S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM review* **43**, 129 (2001).
- [26] D. L. Donoho and Y. Tsaig, Fast solution of l_1 -norm minimization problems when the solution may be sparse, *IEEE Transactions on Information theory* **54**, 4789 (2008).
- [27] E. J. Candes and T. Tao, Decoding by linear programming, *IEEE transactions on information theory* **51**, 4203 (2005).
- [28] E. J. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE Transactions on information theory* **52**, 489 (2006).
- [29] D. L. Donoho, Compressed sensing, *IEEE Transactions on information theory* **52**, 1289 (2006).
- [30] B. Koczor, Sparse probabilistic synthesis of quantum operations, *PRX Quantum* **5**, 040352 (2024).
- [31] B. Koczor, J. J. Morton, and S. C. Benjamin, Probabilistic interpolation of quantum rotation angles, *Physical Review Letters* **132**, 130602 (2024).
- [32] S. Bubeck *et al.*, Convex optimization: Algorithms and complexity, *Foundations and Trends® in Machine Learning* **8**, 231 (2015).
- [33] L. Gentini, A. Cuccoli, S. Pirandola, P. Verrucchi, and L. Banchi, Noise-resilient variational hybrid quantum-classical optimization, *Physical Review A* **102**, 052414 (2020).
- [34] D. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Transactions on Information Theory* **47**, 2845 (2006).
- [35] D. L. Donoho and P. B. Stark, Uncertainty principles and signal recovery, *SIAM Journal on Applied Mathematics* **49**, 906 (1989).
- [36] K. P. Murphy, *Machine learning: a probabilistic perspective* (MIT press, 2012).
- [37] M. Besançon, T. Papamarkou, D. Anthoff, A. Arslan, S. Byrne, D. Lin, and J. Pearson, Distributions.jl: Definition and modeling of probability distributions in the juliastats ecosystem, *Journal of Statistical Software* **98**, 1 (2021).
- [38] J. Knab, Interpolation of band-limited functions using the approximate prolate series, *IEEE transactions on Information Theory* **25**, 717 (1979).
- [39] D. J. Thomson, Spectrum estimation and harmonic analysis, *Proceedings of the IEEE* **70**, 1055 (2005).
- [40] F. J. Simons, Slepian functions and their use in signal estimation and spectral analysis, in *Handbook of geomathematics* (Springer, 2010) pp. 891–923.
- [41] D. Slepian, Prolate spheroidal wave functions, fourier analysis, and uncertainty—v: The discrete case, *Bell System Technical Journal* **57**, 1371 (1978).
- [42] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, Experimental realization of any discrete unitary operator, *Physical review letters* **73**, 58 (1994).
- [43] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, Optimal design for universal multiport interferometers, *Optica* **3**, 1460 (2016).
- [44] L. Banchi, Ballistic quantum state transfer in spin chains: General theory for quasi-free models and arbitrary initial states, *The European Physical Journal Plus* **128**, 1 (2013).
- [45] B. W. Shore and P. L. Knight, The jaynes-cummings model, *Journal of Modern Optics* **40**, 1195 (1993).
- [46] S. Lotfi, M. Finzi, S. Kapoor, A. Potapczynski, M. Goldblum, and A. G. Wilson, Pac-bayes compression bounds so tight that they can explain generalization, *Advances in Neural Information Processing Systems* **35**, 31459 (2022).
- [47] Z. Lai, J. Hu, D. An, and Z. Wen, Extended parameter shift rules with minimal derivative variance for parameterized quantum circuits, *arXiv preprint arXiv:2508.08802* (2025).
- [48] J. C. Lee and P. Valiant, Optimal sub-gaussian mean estimation in $\mathbb{R}$, in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022) pp. 672–683.