BAYESIAN PARAMETER SHIFT RULE IN VARIATIONAL QUANTUM EIGENSOLVERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter shift rules (PSRs) are key techniques for efficient gradient estimation in variational quantum eigensolvers (VQEs). In this paper, we propose their Bayesian variant, where Gaussian processes with appropriate kernels are used to estimate the gradient of the VQE objective. Our *Bayesian PSR* offers flexible gradient estimation from observations at arbitrary locations with uncertainty information, and reduces to the generalized PSR in special cases. In stochastic gradient descent (SGD), the flexibility of Bayesian PSR allows reuse of observations in previous steps, which accelerates the optimization process. Furthermore, the accessibility to the posterior uncertainty, along with our proposed notion of *gradient confident region* (GradCoRe), enables us to minimize the observation costs in each SGD step. Our numerical experiments show that the VQE optimization with Bayesian PSR and GradCoRe significantly accelerates SGD, and outperforms the state-of-the-art methods, including sequential minimal optimization.

1 INTRODUCTION

The variational quantum eigensolver (VQE) (Peruzzo et al., 2014; McClean et al., 2016) is a hybrid quantum-classical algorithm for approximating the ground state of the Hamiltonian of a given physical system. The quantum part of VQEs uses parameterized quantum circuits to generate trial quantum states and measures the expectation value of the Hamiltonian, i.e., the energy, while the classical part performs energy minimization with noisy observations from the quantum device. Provided that the parameterized quantum circuits can accurately approximate the ground state, the minimized energy gives a tight upper bound of the ground state energy of the Hamiltonian.

The observation noise in the quantum device comes from multiple sources. One source of noise is *measurement shot noise*, which arises from the statistical nature of quantum measurements—outcomes follow the probabilities specified by the quantum state, and finite sampling introduces fluctuations. Since this noise source is random and independent, it can be reduced by increasing the number of measurement shots, to which the variance is inversely proportional. Another source of noise stems from imperfections in the quantum hardware, which have been reduced in recent years by hardware design (Bluvstein et al., 2023), as well as error mitigation (Cai et al., 2023), quantum error correction (Roffe, 2019; Acharya et al., 2024), and machine learning (Liao et al., 2024; Nicoli et al., 2025) techniques. In this paper, we do not consider hardware noise, as is common in papers developing optimization methods (Nakanishi et al., 2020; Nicoli et al., 2023b).

Stochastic gradient descent (SGD), sequential minimal optimization (SMO), and Bayesian optimization (BO) have previously been used to minimize the VQE objective function. Under some mild assumptions (Nakanishi et al., 2020), this objective function is known to have special properties. Based on those properties, SGD methods can use the gradient estimated by so-called *parameter shift rules* (PSRs) (Mitarai et al., 2018), and specifically designed SMO (Platt, 1998) methods, called Nakanishi-Fuji-Todo (NFT) (Nakanishi et al., 2020), perform one-dimensional subspace optimization with only a few observations in each iteration. Iannelli and Jansen (2021) applied BO to solve VQEs as noisy global optimization problems.

Although Gaussian processes (GPs) (Rasmussen and Williams, 2006) have been used in VQEs as common surrogate functions for BO (Frazier, 2018), they have also been used to improve SGD-based and SMO-based methods. Nicoli et al. (2023a) proposed the *VQE kernel*—a physics-informed kernel

that fully reflects the properties of VQEs—and combined SMO and BO with the *expected maximum improvement within confident region* (EMICoRe) acquisition function. This allows for identification of the optimal locations to measure on the quantum computer in each SMO iteration. Tamiya and Yamasaki (2022) combined SGD and BO, and proposed *stochastic gradient line BO* (SGLBO), which uses BO to identify the optimal step size in each SGD iteration. Anders et al. (2024) proposed the *subspace in confident region* (SubsCoRe) approach, where the observation costs are minimized based on the posterior uncertainty estimation in each SMO iteration.

In this paper, we take a different approach to leveraging GPs, and introduce a *Bayesian parameter shift rule* (Bayesian PSR), where the gradient of the VQE objective is estimated using GPs with the VQE kernel. The Bayesian PSR translates into a regularized variant of PSRs if the observations are performed at designated locations. However, our approach offers significant advantages—flexibility and direct access to uncertainty—over existing PSRs (Mitarai et al., 2018; Wierichs et al., 2022). More specifically, the Bayesian PSR can use observations at any set of locations, which allows the reuse of observations performed in previous iterations of SGD. Reusing previous observations along with new observations improves the gradient estimation accuracy, and thus accelerates the optimization process. Furthermore, the uncertainty information can be used to adapt the observation cost in each SGD iteration—in a similar spirit to Anders et al. (2024)—which significantly reduces the cost of obtaining new observations, while maintaining a required level of accuracy. We implement this adaptive observation cost strategy by introducing a novel notion of *gradient confidence region* (GradCoRe)—the region in which the uncertainty of the gradient estimation is below a specified threshold (see Figure 1). Empirical evaluations show that our proposed Bayesian PSR improves the gradient estimator, and SGD equipped with our GradCoRe approach outperforms all previous state-of-the-art methods including NFT and its variants.

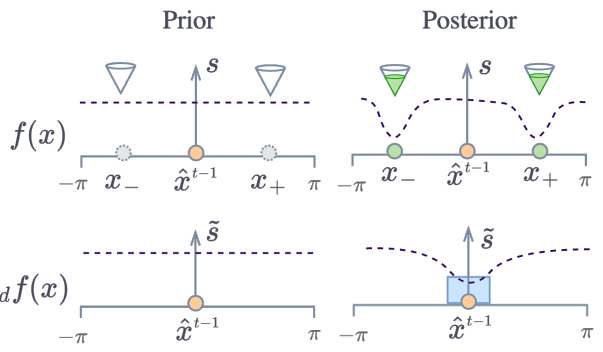

Figure 1: Illustration of our gradient confident region (GradCoRe) approach. Our goal is to minimize the true energy $f^*(\boldsymbol{x})$ over the set of parameters $\boldsymbol{x} \in [0, 2\pi)^D$, where we use a GP surrogate $f(\boldsymbol{x})$ for approximating $f^*(\boldsymbol{x})$. Observing f^* at points $\boldsymbol{x}_-$ and $\boldsymbol{x}_+$ (green circles) along the d-th direction (solid horizontal line) decreases the uncertainty (dashed curves) not only for predicting $f(\boldsymbol{x}_\pm)$, but also for predicting $\partial_d f(\widehat{\boldsymbol{x}}^{t-1})$, so that the current optimal point $\widehat{\boldsymbol{x}}^{t-1}$ falls within the GradCoRe (blue square). Our GradCoRe-based SGD uses the minimum number of measurement shots for achieving required gradient estimation accuracy in each iteration, and thus minimizes the total observation costs over the optimization process.

The main contributions are summarized as follows:

- We propose *Bayesian PSR*, a flexible variant of existing PSRs that provides access to uncertainty information.
- We theoretically establish the relationship between Bayesian PSR and existing PSRs, revealing the optimality of the *shift* parameter in first-order PSRs.
- We introduce the notion of *GradCoRe*, and propose an adaptive observation cost strategy for SGD optimization.
- We numerically validate our theory and empirically demonstrate the effectiveness of the proposed Bayesian PSR and GradCoRe.

Related work: Finding the optimal set of parameters for a variational quantum circuit is a challenging problem, prompting the development of various approaches to improve the optimization in VQEs. Gradient-based methods for VQEs often rely on PSRs (Mitarai et al., 2018; Wierichs et al., 2022), which enable reasonably accurate gradient estimation of the output of quantum circuits with respect to their parameters. Nakanishi et al. (2020) proposed an SMO (Platt, 1998) algorithm, known

as *NFT*, where, at each step of SMO, one parameter is analytically minimized by performing a few observations. Nicoli et al. (2023a) combined NFT with GP and BO by developing a physics-inspired kernel for GP regression and proposing the EMICoRe acquisition function, relying on the concept of confident regions (CoRe). This method improves upon NFT by leveraging the information from observations in previous steps to identify the optimal locations to perform the next observations. Anders et al. (2024) leveraged the same notion of CoRe, and proposed SubsCoRe, where, instead of optimizing the observed locations, the minimal number of measurement shots is identified to achieve the required accuracy defined by the CoRe. The resulting algorithm converges to the same energy as NFT with a smaller quantum computation cost, i.e., the total number of measurement shots on a quantum computer. Tamiya and Yamasaki (2022) combined SGD with BO to tackle the excessive cost of standard SGD approaches and used BO to accelerate the convergence by finding the optimal step size. In a general context of BO, Müller et al. (2021) proposed a gradient information with BO (GIBO) approach, where the uncertainty of the GP-estimated gradient is minimized. Our GradCoRe can be seen as an enhanced version of GIBO, where the theoretically optimal locations are observed with minimum costs based on strong physical information of VQEs.

2 Background

Here we briefly introduce Gaussian process (GP) regression and its derivatives, as well as VQEs with their known properties.

2.1 GP Regression and Derivative GP

Assume that we aim to learn an unknown function $f^*(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ from the training data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathcal{X}^N, \boldsymbol{y} = (y_1, \ldots, y_N)^\top \in \mathbb{R}^N, \boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_N^2) \in \mathbb{R}^N_{++}$ that fulfills

$$y_n = f^*(\boldsymbol{x}_n) + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}_1(y_n; 0, \sigma_n^2), \tag{1}$$

where $\mathcal{N}_D(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the D-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. With the Gaussian process (GP) prior $p(f(\cdot)) = \mathrm{GP}(f(\cdot); 0(\cdot), k(\cdot, \cdot))$, where $0(\cdot)$ and $k(\cdot, \cdot)$ are the prior zero-mean and the kernel (covariance) functions, respectively, the posterior distribution of the function values $\boldsymbol{f}' = (f(\boldsymbol{x}'_1), \ldots, f(\boldsymbol{x}'_M))^\top \in \mathbb{R}^M$ at arbitrary test points $\boldsymbol{X}' = (\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_M) \in \mathcal{X}^M$ is given as

$$p(\boldsymbol{f}'|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}_M(\boldsymbol{f}'; \boldsymbol{\mu}'_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]}, \boldsymbol{S}'_{[\boldsymbol{X},\boldsymbol{\sigma}]}), \quad \text{where} \tag{2}$$

$$\boldsymbol{\mu}'_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]} = \boldsymbol{K}'^\top \left(\boldsymbol{K} + \mathbf{Diag}(\boldsymbol{\sigma})\right)^{-1} \boldsymbol{y} \quad \text{and} \quad \boldsymbol{S}'_{[\boldsymbol{X},\boldsymbol{\sigma}]} = \boldsymbol{K}'' - \boldsymbol{K}'^\top \left(\boldsymbol{K} + \mathbf{Diag}(\boldsymbol{\sigma})\right)^{-1} \boldsymbol{K}' \tag{3}$$

are the posterior mean and covariance, respectively (Rasmussen and Williams, 2006). Here $\mathbf{Diag}(\boldsymbol{v})$ is the diagonal matrix with $\boldsymbol{v}$ specifying the diagonal entries, and $\boldsymbol{K} = k(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{N \times N}, \boldsymbol{K}' = k(\boldsymbol{X}, \boldsymbol{X}') \in \mathbb{R}^{N \times M}$, and $\boldsymbol{K}'' = k(\boldsymbol{X}', \boldsymbol{X}') \in \mathbb{R}^{M \times M}$ are the train, train-test, and test kernel matrices, respectively, where $k(\boldsymbol{X}, \boldsymbol{X}')$ denotes the kernel matrix evaluated at each column of $\boldsymbol{X}$ and $\boldsymbol{X}'$ such that $(k(\boldsymbol{X}, \boldsymbol{X}'))_{n,m} = k(\boldsymbol{x}_n, \boldsymbol{x}'_m)$. We also denote the posterior as $p(f(\cdot)|\boldsymbol{X}, \boldsymbol{y}) = \mathrm{GP}(f(\cdot); \mu_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]}(\cdot), s_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\cdot, \cdot))$ with the posterior mean $\mu_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]}(\cdot)$ and covariance $s_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\cdot, \cdot)$ functions.

Since the derivative operator is linear, the derivative $\boldsymbol{\nabla}_{\boldsymbol{x}} f = (\partial_1 f, \ldots, \partial_D f)^\top \in \mathbb{R}^D$ of GP samples also follows a GP. Here we abbreviate $\partial_d = \frac{\partial}{\partial x_d}$. Therefore, we can straightforwardly handle the derivative outputs at training and test points by modifying the kernel function. Assume that $\boldsymbol{x}$ is a training or test point with non-derivative output $y = f^*(\boldsymbol{x}) + \varepsilon$, and $\boldsymbol{x}'$ and $\boldsymbol{x}''$ are training or test points with derivative outputs, $y' = \partial_{d'} f^*(\boldsymbol{x}') + \varepsilon', y'' = \partial_{d''} f^*(\boldsymbol{x}'') + \varepsilon''$. Then, the kernel functions should be replaced with

$$\widetilde{k}(\boldsymbol{x}, \boldsymbol{x}') = \tfrac{\partial}{\partial x'_{d'}} k(\boldsymbol{x}, \boldsymbol{x}'), \qquad \widetilde{k}(\boldsymbol{x}', \boldsymbol{x}'') = \tfrac{\partial^2}{\partial x'_{d'} \partial x''_{d''}} k(\boldsymbol{x}', \boldsymbol{x}''). \tag{4}$$

The posterior (2) with appropriately replaced kernel matrix entries gives the posterior distribution of derivatives at test points. We denote the GP posterior of a single component of the derivative as

$$p(\partial_d f(\cdot)|\boldsymbol{X}, \boldsymbol{y}) = \mathrm{GP}\left(\partial_d f(\cdot); \widetilde{\mu}^{(d)}_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]}(\cdot), \widetilde{s}^{(d)}_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\cdot, \cdot)\right) \tag{5}$$

with the posterior mean $\widetilde{\mu}^{(d)}(\cdot)$ and covariance $\widetilde{s}^{(d)}(\cdot,\cdot)$ functions for the derivative with respect to x_d. More generally, GP regression can be analytically performed in the case where the training outputs (i.e., observations) and the test outputs (i.e., predictions) contain derivatives with different orders (see Appendix A for more details).

2.2 Variational Quantum Eigensolvers and their Physical Properties

The VQE (Peruzzo et al., 2014; McClean et al., 2016) is a hybrid quantum-classical computing protocol for estimating the ground-state energy of a given quantum Hamiltonian for a Q-qubit system. The quantum computer is used to prepare a parametric quantum state $|\psi_{\boldsymbol{x}}\rangle$, which depends on D angular parameters $\boldsymbol{x} \in \mathcal{X} = [0, 2\pi)^D$. This trial state $|\psi_{\boldsymbol{x}}\rangle$ is generated by applying $D'(\geq D)$ *quantum gate operations*, $G(\boldsymbol{x}) = G_{D'} \circ \cdots \circ G_1$, to an initial quantum state $|\psi_0\rangle$, i.e., $|\psi_{\boldsymbol{x}}\rangle = G(\boldsymbol{x})|\psi_0\rangle$. All gates $\{G_{d'}\}_{d'=1}^{D'}$ are unitary operators, parameterized by at most one variable x_d. Let $d(d') : \{1, \ldots, D'\} \mapsto \{1, \ldots, D\}$ be the mapping specifying which one of the variables $\{x_d\}$ parameterizes the d'-th gate. We consider parametric gates of the form $G_{d'}(x) = U_{d'}(x_{d(d')}) = \exp\left(-ix_{d(d')}P_{d'}/2\right)$, where $P_{d'}$ is an arbitrary sequence of the Pauli operators $\{\mathbf{1}_q, \sigma_q^X, \sigma_q^Y, \sigma_q^Z\}_{q=1}^{Q}$ acting on each qubit at most once. This general structure covers both single-qubit gates, such as $R_X(x) = \exp\left(-i\theta\sigma_q^X\right)$, and entangling gates acting on multiple qubits simultaneously, such as $R_{XX}(x) = \exp\left(-ix\sigma_{q_1}^X \circ \sigma_{q_2}^X\right)$ for $q_1 \neq q_2$, commonly realized in trapped-ion quantum hardware setups (Kielpinski et al., 2002; Debnath et al., 2016).

The quantum computer is used to evaluate the energy of the resulting quantum state $|\psi_{\boldsymbol{x}}\rangle$ by observing

$$y = f^*(\boldsymbol{x}) + \varepsilon, \qquad \text{where} \qquad f^*(\boldsymbol{x}) = \langle\psi_{\boldsymbol{x}}|H|\psi_{\boldsymbol{x}}\rangle = \langle\psi_0|G(\boldsymbol{x})^\dagger H G(\boldsymbol{x})|\psi_0\rangle, \tag{6}$$

and $\dagger$ denotes the Hermitian conjugate. For each observation, repeated measurements, called *shots*, on the quantum computer are performed. Averaging over the number N_{shots} of shots suppresses the variance $\sigma^{*2}(N_{\text{shots}}) \propto N_{\text{shots}}^{-1}$ of the observation noise ε.[1] Since the observation y is the sum of many random variables, it approximately follows the Gaussian distribution, according to the central limit theorem. The Gaussian likelihood (1) therefore approximates the observation y well if $\sigma_n^2 \approx \sigma^{*2}(N_{\text{shots}})$. Using the noisy estimates of $f^*(\boldsymbol{x})$ obtained from the quantum computer, a protocol running on a classical computer is used to solve the following minimization problem:

$$\min_{\boldsymbol{x}\in[0,2\pi)^D} f^*(\boldsymbol{x}), \tag{7}$$

thus finding the minimizer $\widehat{\boldsymbol{x}}$, i.e., the optimal parameters for the (rotational) quantum gates. Given the high expense of quantum computing resources, the computation cost is primarily driven by quantum operations. As a result, the optimization cost in VQE is typically measured by the total number of measurement shots required during the optimization process.[2] We refer to Tilly et al. (2022) for further details about VQEs and their challenges.

Let V_d be the number of gates parameterized by x_d, i.e., $V_d = |\{d' \in \{1, \ldots D'\}; d = d(d')\}|$. Mitarai et al. (2018) proved that the VQE objective (6) for $V_d = 1$ satisfies the parameter shift rule (PSR)

$$\partial_d f^*(\boldsymbol{x}') = \frac{f^*\left(\boldsymbol{x}'+\alpha\boldsymbol{e}_d\right) - f^*\left(\boldsymbol{x}'-\alpha\boldsymbol{e}_d\right)}{2\sin\alpha}, \qquad \forall \boldsymbol{x} \in [0, 2\pi)^D,\ d = 1, \ldots, D,\ \alpha \in [0, 2\pi), \tag{8}$$

where $\{\boldsymbol{e}_d\}_{d=1}^D$ are the standard basis, and the *shift* α is typically set to $\frac{\pi}{2}$. Wierichs et al. (2022) generalized the PSR (8) for arbitrary V_d with equidistant observations $\{\boldsymbol{x}_w = \boldsymbol{x}' + \frac{2w+1}{2V_d}\pi\boldsymbol{e}_d\}_{w=0}^{2V_d-1}$:

$$\partial_d f^*(\boldsymbol{x}') = \frac{1}{2V_d}\sum_{w=0}^{2V_d-1} \frac{(-1)^w f^*(\boldsymbol{x}_w)}{2\sin^2\left(\frac{(2w+1)\pi}{4V_d}\right)}. \tag{9}$$

Most gradient-based approaches rely on those PSRs, which allow reasonably accurate gradient estimation from $\sum_{d=1}^D 2V_d$ observations. Let

$$\boldsymbol{\psi}_\gamma(\theta) = (\gamma, \sqrt{2}\cos\theta, \sqrt{2}\cos 2\theta, \ldots, \sqrt{2}\cos V_d\theta, \sqrt{2}\sin\theta, \sqrt{2}\sin 2\theta, \ldots, \sqrt{2}\sin V_d\theta)^\top \tag{10}$$

[1] We do not consider the hardware noise, and therefore, the observation noise ε consists only of the *measurement shot* noise.

[2] When the Hamiltonian consists of N_{og} groups of non-commuting operators, each of which needs to be measured separately, N_{shots} denotes the number of shots *per operator group*. Therefore, the number of shots *per observation* is $N_{\text{og}} \times N_{\text{shots}}$. In our experiments, we report on the total number of shots per operator group, i.e., the cumulative sum of N_{shots} over all observations, when evaluating the observation cost.

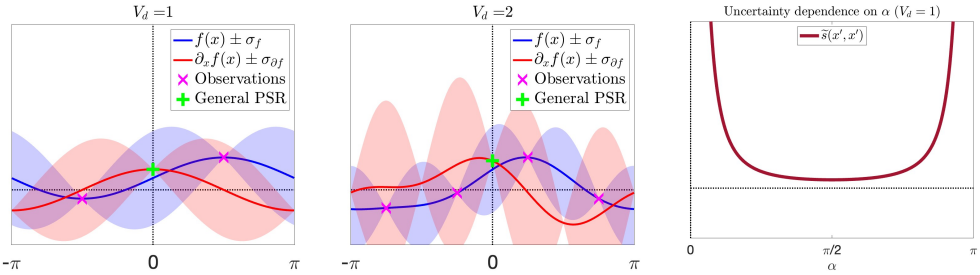

Figure 2: Illustration of the behavior of the Bayesian PSR when $V_d = 1$ (left) and when $V_d = 2$ (middle). Bayesian PSR prediction (red) coincides with general PSR (green cross) for the designed equidistant observations (magenta crosses). The right plot visualizes the variance (the second equation in Eq. (15)) of the derivative GP prediction at $\boldsymbol{x}'$, as a function of the shift α of observations when $V_d = 1$. For all panels, the noise and kernel parameters are set to $\sigma^2 = 0.01, \gamma^2 = 9, \sigma_0^2 = 100$.

be the (1-dimensional) V_d-th order Fourier basis for arbitrary $\gamma > 0$. Nakanishi et al. (2020) found that the VQE objective function $f^*(\cdot)$ in Eq. (6) with any[3] $G(\cdot)$, H, and $|\psi_0\rangle$ can be expressed exactly as

$$f^*(\boldsymbol{x}) = \boldsymbol{b}^\top \mathbf{vec}\left(\otimes_{d=1}^{D} \boldsymbol{\psi}_\gamma(x_d)\right) \tag{11}$$

for some $\boldsymbol{b} \in \mathbb{R}^{\prod_{d=1}^{D}(1+2V_d)}$, where $\otimes$ and $\mathbf{vec}(\cdot)$ denote the tensor product and the vectorization operator for a tensor, respectively. Based on this property, the Nakanishi-Fuji-Todo (NFT) method (Nakanishi et al., 2020) performs SMO (Platt, 1998), where the optimum in a chosen 1D subspace for each iteration is analytically estimated from only $1 + 2V_d$ observations (see Appendix B for the detailed procedure). It was shown that the PSR (8) and the trigonometric polynomial function form (11) are mathematically equivalent (Nicoli et al., 2023a).

Inspired by the function form (11) of the objective, Nicoli et al. (2023a) proposed the VQE kernel

$$k_\gamma(\boldsymbol{x}, \boldsymbol{x}') = \sigma_0^2 \prod_{d=1}^{D} \left(\frac{\gamma^2 + 2\sum_{v=1}^{V_d} \cos\left(v(x_d - x_d')\right)}{\gamma^2 + 2V_d} \right), \tag{12}$$

which is decomposed as $k_\gamma(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}_\gamma(\boldsymbol{x})^\top \boldsymbol{\phi}_\gamma(\boldsymbol{x}')$ with feature maps $\boldsymbol{\phi}_\gamma(\boldsymbol{x}) = \frac{\sigma_0}{(\gamma^2+2V_d)^{D/2}} \mathbf{vec}\left(\otimes_{d=1}^{D} \boldsymbol{\psi}_\gamma(x_d)\right)$, for GP regression. The kernel parameter γ^2 controls the smoothness of the function, i.e., suppressing the interaction terms when $\gamma^2 > 1$. When $\gamma^2 = 1$, the Fourier basis (10) is orthonormal, and the VQE kernel (12) is proportional to the product of Dirichlet kernels (Rudin, 1964). The VQE kernel reflects the physical knowledge (11) of VQE, and thus allows us to perform a Bayesian variant of NFT—*Bayesian NFT* or *Bayesian SMO*—where the 1D subspace optimzation in each SMO step is performed with GP (see Appendix B for more details and the performance comparison between the original NFT and Bayesian NFT). Nicoli et al. (2023a) furthermore enhanced Bayesian NFT with BO, using the notion of confident region (CoRe),

$$\mathcal{Z}_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\kappa^2) = \left\{ \boldsymbol{x} \in \mathcal{X}; s_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa^2 \right\}, \tag{13}$$

i.e., the region in which the uncertainty of the GP prediction is lower than a threshold κ. More specifically, they introduced the EMICoRe acquisition function to find the best observation points in each SMO iteration, such that the maximum expected improvement within the CoRe is maximized.

3 Bayesian Parameter Shift Rules

We propose *Bayesian PSR*, which estimates the gradient of the VQE objective (6) by the GP posterior (5) with the VQE kernel (12) along with its derivatives (4). The advantages of Bayesian PSR include: 1) The gradient estimator has an analytic-form, 2) Estimation can be performed using observations at any set of points, 3) Estimation is optimal for heteroschedastically noisy observations (from the Bayesian perspective), as long as the prior with the kernel parameters, γ and σ_0^2, is appropriately set,

[3] Any circuit consisting of parametrized rotation gates and non-parametric unitary gates.

and 4) The posterior uncertainty can be analytically computed *before* performing the observations. In Section 4, we propose novel SGD solvers for VQEs that leverage the advantages of Bayesian PSR.

As naturally expected, our Bayesian PSR is a generalization of exisiting PSRs, and reduces to the general PSR (9) for noiseless and equidistant observations. Let $\mathbf{1}_D \in \mathbb{R}^D$ be the vector with all entries equal to one.

Theorem 3.1. *For any $x' \in [0, 2\pi)^D$ and $d = 1, \ldots, D$, the mean and variance of the derivative GP prediction, given observations $\boldsymbol{y} = (y_0, \ldots, y_{2V_d-1})^\top \in \mathbb{R}^{2V_d}$ at $2V_d$ equidistant training points $\boldsymbol{X} = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{2V_d-1}) \in \mathbb{R}^{D\times 2V_d}$ for $\boldsymbol{x}_w = \boldsymbol{x}' + \frac{2w+1}{2V_d}\pi \boldsymbol{e}_d$ with homoschedastic noise $\boldsymbol{\sigma} = \sigma^2 \cdot \mathbf{1}_{2V_d}$ for $\sigma^2 \ll \sigma_0$, are*

$$\widetilde{\mu}^{(d)}_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]}(\boldsymbol{x}') = \frac{\sum_{w=0}^{2V_d-1} \frac{(-1)^w y_w}{2\sin^2\left(\frac{(2w+1)\pi}{4V_d}\right)}}{(\gamma^2+2V_d)\frac{\sigma^2}{\sigma_0^2}+2V_d} + O(\tfrac{\sigma^4}{\sigma_0^4}), \quad \widetilde{s}^{(d)}_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\boldsymbol{x}', \boldsymbol{x}') = \sigma^2\left(\frac{2V_d^2+1}{6}\right) + O(\tfrac{\sigma^4}{\sigma_0^2}). \quad (14)$$

The proof, the non-asymptotic form of the mean and the variance, and the numerical validation of the theorem are given in Appendix C. Apparently, the mean prediction (the first equation in Eq. (14)) by Bayesian PSR converges to the general PSR (9) with the uncertainty (the second equation in Eq. (14)) converging to zero in the noiseless limit, i.e., $\sigma^2 \to +0$ and hence $y_w = f^*(\boldsymbol{x}_w)$. In noisy cases, the prior variance $\sigma_0^2 \sim O(\sigma^2)$ suppresses the amplitude of the gradient estimator as a regularizer through the first term in the denominator in the first equation of Eq. (14).

Figure 2 illustrates the behavior of Bayesian PSR when $V_d = 1$ (left panel) and when $V_d = 2$ (middle panel). In each panel, given $2V_d$ equidistant observations (magenta crosses), the blue curve shows the (non-derivative) GP prediction with uncertainty (blue shades), while the red curve shows the derivative GP prediction with uncertainty (red shades). Note the $\frac{\pi}{2V_d}$ shift of the low uncertainty locations between the GP prediction (blue) and the derivative GP prediction (red). The green cross shows the output of the general PSR (9) at $\boldsymbol{x}' = 0$, which almost coincides with the Bayesian PSR prediction (red curve) under this setting. Other examples, including cases where the Bayesian regularization is visible, are given in Appendix C.

In the simplest first-order case, i.e., where $V_d = 1, \forall d = 1, \ldots, D$, we can theoretically investigate the optimality of the choice of the shift α in Eq. (8) (the proof is also given in Appendix C).

Theorem 3.2. *Assume that $V_d = 1, \forall d = 1, \ldots, D$. For any $x' \in [0, 2\pi)^D$ and $d = 1, \ldots, D$, the mean and variance of the derivative GP prediction, given observations $\boldsymbol{y} = (y_1, y_2)^\top \in \mathbb{R}^2$ at two training points $\boldsymbol{X} = (\boldsymbol{x}' - \alpha\boldsymbol{e}_d, \boldsymbol{x}' + \alpha\boldsymbol{e}_d) \in \mathbb{R}^{D\times 2}$ with homoschedastic noise $\boldsymbol{\sigma} = (\sigma^2, \sigma^2)^\top$, are*

$$\widetilde{\mu}^{(d)}_{[\boldsymbol{X},\boldsymbol{y},\boldsymbol{\sigma}]}(\boldsymbol{x}') = \frac{(y_2-y_1)\sin\alpha}{(\gamma^2/2+1)\sigma^2/\sigma_0^2+2\sin^2\alpha}, \qquad \widetilde{s}^{(d)}_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\boldsymbol{x}', \boldsymbol{x}') = \frac{\sigma^2}{(\gamma^2/2+1)\sigma^2/\sigma_0^2+2\sin^2\alpha}. \quad (15)$$

Again, the mean prediction (the first equation in Eq. (15)) is a regularized version of the PSR (8). The uncertainty prediction (the second equation in Eq. (15)) implies that $\alpha = \pi/2$ minimizes the uncertainty in the noisy case, regardless of σ^2, σ_0^2 and γ (see the right panel in Figure 2, where the variance of the derivative GP prediction at $\boldsymbol{x}'$ is visualized as a function of the shift α of observations for $V_d = 1$). This supports most of the use cases of the PSR in the literature (Mitarai et al., 2018), and matches the intuition that the maximum span minimizes the uncertainty.

4 SGD WITH BAYESIAN PSR

In this section, we equip SGD with Bayesian PSR. In the standard implementation of SGD for VQEs, $2V_d$ equidistant points along each direction $d = 1, \ldots, D$ are observed for gradient estimation by the general PSR (9) (or by the PSR (8) if $V_d = 1, \forall d$) in each SGD iteration.

Bayesian SGD (Bayes-SGD): A straightforward application of Bayesian PSR is to replace existing PSRs with Bayesian PSR for gradient estimation, allowing for the reuse of previous observations. We retain $R \cdot 2V_d \cdot D$ latest observations for a predetermined R in our experiments. Reusing previous observations accumulates the gradient information, and thus improves the gradient estimation accuracy, as shown in Section 5.2.

4.1 GRADIENT CONFIDENT REGION (GRADCORE)

We propose an adaptive observation cost control strategy that leverages the uncertainty information provided by the Bayesian PSR. This strategy adjusts the number of measurement shots for gradient estimation in each SGD iteration so that the variances of the derivative GP prediction at the current optimal point $\widehat{\boldsymbol{x}}$ are below certain thresholds. In a similar fashion to the CoRe (13), we define the *gradient confident region* (GradCoRe)

$$\widetilde{\mathcal{Z}}_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\boldsymbol{\kappa}) = \left\{ \boldsymbol{x} \in \mathcal{X}; \widetilde{s}^{(d)}_{[\boldsymbol{X},\boldsymbol{\sigma}]}(\boldsymbol{x},\boldsymbol{x}) \leq \kappa_d^2, \forall d \right\}, \tag{16}$$

where $\boldsymbol{\kappa} = (\kappa_1^2, \ldots, \kappa_D^2)^\top \in \mathbb{R}^D$ are the required accuracy thresholds. Our proposed SGD-based optimizer, named *SGD-GradCoRe*, measures new equidistant points $\breve{\boldsymbol{X}} = \{\{\boldsymbol{x}_w^{(d)} = \widehat{\boldsymbol{x}} + \frac{2w+1}{2V_d}\pi \boldsymbol{e}_d\}_{w=0}^{2V_d}\}_{d=1}^{D}$ for all directions with the minimum total number of shots such that the current optimal point $\widehat{\boldsymbol{x}}$ is in the GradCoRe (see Figure 1).

Before starting optimization, we evaluate the single-shot observation noise variance $\sigma^{*2}(1) = \overline{\sigma}^{*2}$ by collecting measurements at random locations, following Anders et al. (2024). We use this information to estimate the observation noise variance as a function of the number of shots as $\sigma^{*2}(N_{\text{shots}}) = \frac{\overline{\sigma}^{*2}}{N_{\text{shots}}}$. Let $(\boldsymbol{X}^t, \boldsymbol{y}^t, \boldsymbol{\sigma}^t)$ be the training data (all previous observations) at the t-th SGD iteration step, and let $\breve{\boldsymbol{\nu}} \in \mathbb{R}^{2V_d D}$ be the vector of the numbers of measurement shots at the new equidistant measurement points $\breve{\boldsymbol{X}}$ for all directions. Before measuring at $\breve{\boldsymbol{X}}$ in the $(t+1)$-th SGD iteration, we solve the following problem:

$$\min_{\widetilde{\boldsymbol{\nu}}} \|\widetilde{\boldsymbol{\nu}}\|_1 \text{ s.t. } \widehat{\boldsymbol{x}} \in \widetilde{\mathcal{Z}}_{[(\boldsymbol{X}^t, \breve{\boldsymbol{X}}), (\boldsymbol{\sigma}^t, \breve{\boldsymbol{\sigma}}(\widetilde{\boldsymbol{n}}))]}(\boldsymbol{\kappa}(t)), \tag{17}$$

where $\breve{\boldsymbol{\sigma}}(\widetilde{\boldsymbol{\nu}}) = \overline{\sigma}^{*2} \cdot (\widetilde{\nu}_1^{-1}, \ldots, \widetilde{\nu}_{2V_d D}^{-1})^\top$, and $\boldsymbol{\kappa}(t)$ is the required accuracy dependent on the iteration step t. Informally, we minimize the total measurement budget under the constraint that the posterior gradient variance along each direction d is smaller than the required accuracy threshold. For simplicity, we solve the GradCoRe problem (17) by grid search under the additional constraint that all $2V_d D$ points are measured with an equal number of shots.

We set the required accuracy thresholds to $\boldsymbol{\kappa}(t) = \kappa^2(t)\mathbf{1}_D$, where

$$\kappa^2(t) = \max\left(c_0, \tfrac{c_1}{D}\textstyle\sum_{d=1}^{D}\left(\widetilde{\mu}^{(d)}_{[\boldsymbol{X}^t, \boldsymbol{y}^t, \boldsymbol{\sigma}^t]}(\widehat{\boldsymbol{x}}^t)\right)^2\right). \tag{18}$$

Namely, $\kappa(t)$ is set proportional to the L2-norm of the estimated gradient at the current optimal point at the t-th SGD iteration, as long as it is larger than a lower bound. The lower bound c_0 and the slope c_1 are hyperparameters to be tuned. This strategy for setting the required accuracy based on the estimated gradient norm was proposed by Tamiya and Yamasaki (2022).

In the experiment plots in Section 5, we will refer to SGD-GradCoRe as *GradCoRe*. Further algorithmic details, including pseudo-code and used hyperparameter values, are given in Appendix D.

5 EXPERIMENTS

5.1 SETUP

We demonstrate the performance of our Bayesian PSR and GradCoRe approaches in the same setup used by Nicoli et al. (2023a). For all experiments, we prepared 100 different random initial points, from which all optimization methods start. Our Python implementation uses Qiskit (Abraham et al., 2019) for the classical simulation of quantum hardware. The implementation for reproducing our results is attached as supplemental material.

Hamiltonian and Quantum Circuit: We focus on the quantum Heisenberg Hamiltonian with open boundary conditions,

$$H = -\textstyle\sum_{i\in\{X,Y,Z\}}\left[\sum_{j=1}^{Q-1}(J_i\sigma_j^i\sigma_{j+1}^i) + \sum_{j=1}^{Q} h_i\sigma_j^i\right], \tag{19}$$

where $\{\sigma_j^i\}_{i\in\{X,Y,Z\}}$ are the Pauli operators acting on the j-th qubit. For the quantum circuit, we use a common ansatz, called the L-layered `Efficient SU(2)` circuit with open boundary conditions, where $V_d = 1, \forall d$ (see Nicoli et al. (2023a) for more details).

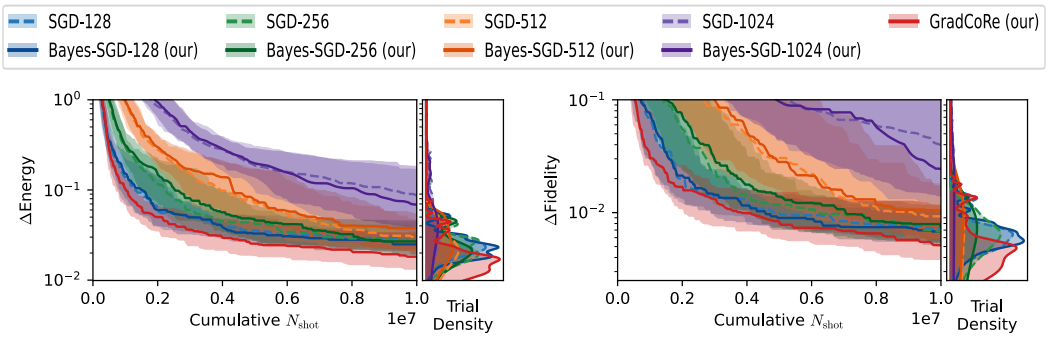

Figure 3: Comparison between SGD with PSR (dashed curves) and SGD with Bayesian PSR (solid curves), as well as GradCoRe (red solid curve), on the Ising Hamiltonian with a $(Q = 5)$-qubits $(L = 3)$-layers quantum circuit. The energy (left) and fidelity (right) are plotted as functions of the cumulative N_{shots}, i.e., the total number of measurement shots. Except GradCoRe equipped with the adaptive shots strategy, the number of shots per observation is set to $N_{\text{shots}} = 128$ (blue), 256 (green), 512 (orange), and 1024 (purple).

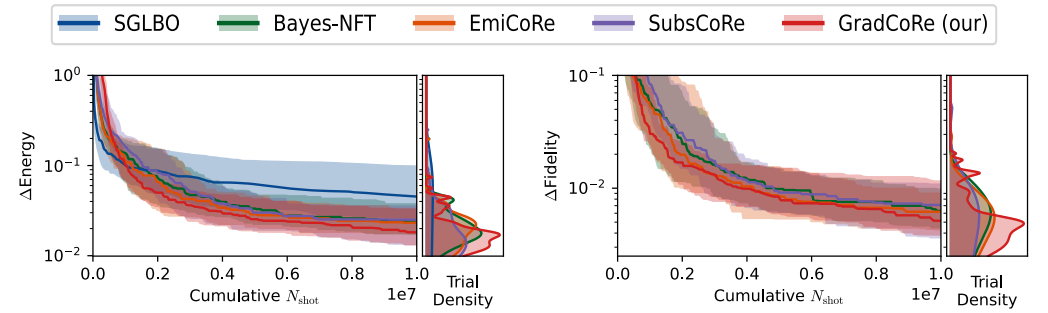

Figure 4: Energy (left) and fidelity (right) achieved within the cumulative number of measurement shots for the Ising Hamiltonian with a $(Q = 5)$-qubits $(L = 3)$-layers quantum circuit. The curves correspond to SGLBO (blue), Bayes-NFT (green), EMICoRe (orange), SubsCoRe (purple), and our proposed GradCoRe (red).

Evaluation Metrics: We compare all methods using two metrics: the best achieved *true energy* $f^*(\widehat{\boldsymbol{x}})$, for $f^*(\cdot)$ defined in Eq. (6), and *fidelity* $\langle\psi_{\text{GS}}|\psi_{\widehat{\boldsymbol{x}}}\rangle \in [0, 1]$. The latter is the inner product between the true ground-state wave function $|\psi_{\text{GS}}\rangle$, computed by exact diagonalization of the target Hamiltonian H, and the trial wave function, $|\psi_{\widehat{\boldsymbol{x}}}\rangle$, corresponding to the quantum state generated by the circuit using the optimized parameters $\widehat{\boldsymbol{x}}$. For both metrics, we plot the difference (smaller is better) to the respective target, i.e.,

$$\Delta\text{Energy} = \langle\psi_{\widehat{\boldsymbol{x}}}|H|\psi_{\widehat{\boldsymbol{x}}}\rangle - \langle\psi_{\text{GS}}|H|\psi_{\text{GS}}\rangle = f^*(\widehat{\boldsymbol{x}}) - \langle\psi_{\text{GS}}|H|\psi_{\text{GS}}\rangle, \tag{20}$$

$$\Delta\text{Fidelity} = \langle\psi_{\text{GS}}|\psi_{\text{GS}}\rangle - \langle\psi_{\text{GS}}|\psi_{\widehat{\boldsymbol{x}}}\rangle = 1 - \langle\psi_{\text{GS}}|\psi_{\widehat{\boldsymbol{x}}}\rangle, \tag{21}$$

in log scale. Here, $|\psi_{\text{GS}}\rangle$ and $\langle\psi_{\text{GS}}|H|\psi_{\text{GS}}\rangle$ are the wave function and true energy at the ground-state, respectively, both of which are computed analytically. As a measure of the quantum computation cost, we consider the total number of measurement shots *per operator group* (see Footnote 2) for all observations over the whole optimization process.

Baseline Methods: We compare our Bayesian SGD and GradCoRe approaches to the baselines, including SGD with the PSR (8), Bayesian NFT, SGLBO (Tamiya and Yamasaki, 2022), EMICoRe (Nicoli et al., 2023a), and SubsCoRe (Anders et al., 2024). We exclude the original NFT (Nakanishi et al., 2020) because it is outperformed by Bayesian NFT (see Figure 5 in Appendix B). We also exclude GIBO (Müller et al., 2021), which is an even weaker baseline than the original NFT (see Appendix G).

Algorithm Setting: All SGD-based methods use the ADAM optimizer with $l_r = 0.05$, $\beta s = (0.9, 0.999)$. For the methods not equipped with adaptive cost control (i.e., all methods except SGLBO,

SubsCoRe and GradCoRe), we set $N_{\text{shots}} = 1024$ for each observation—the same setting as in Nicoli et al. (2023a)—unless specified explicitly. To avoid error accumulation, all SMO-based methods measure the "center", i.e., the current optimal point without shift, every $D + 1$ iterations (Nakanishi et al., 2020). Bayes-SGD and GradCoRe estimate the gradient from the $R \cdot 2V_d \cdot D$ latest observations for $R = 5$, and GradCoRe initially uses the fixed threshold $\kappa^2(t) = \overline{\sigma}^{*2}/256$ before starting the cost adaption after D SGD iterations. Further details on the algorithmic and experimental settings are described in Appendix D and Appendix E, respectively.

5.2 Improvement over SGD with Bayesian PSR and GradCoRe

First, we investigate how our Bayesian PSR and GradCoRe improve SGD. Figure 3 compares SGD with the standard PSR (SGD) and SGD with Bayesian PSR (Bayes-SGD) on the Ising Hamiltonian, i.e., Eq. (19) for $J_{i\in\{X,Y,Z\}} = (-1, 0, 0)$ and $h_{i\in\{X,Y,Z\}} = (0, 0, -1)$, with a $(Q = 5)$-qubits $(L = 3)$-layers quantum circuit. Both for SGD and Bayes-SGD, the optimization performance with $N_{\text{shots}} = 128, 256, 512, 1024$ measurement shots are shown. The left and right panels plot the difference to the ground-state in true energy (20) and fidelity (21) achieved by each method as functions of the cumulative N_{shots}, i.e., the total number of measurement shots. To the right of each panel, the *trial density*, i.e., the distribution over the trials computed by kernel-density estimation, after the use of 1×10^7 total measurement shots is depicted. The median, the 25-th and the 75-th percentiles are shown as a solid curve and shades, respectively. We observe that Bayesian PSR, with a more accurate gradient estimator as shown in Figure 7 in Appendix F, is comparable or compares favorably to the original SGD. More importantly, we observe that GradCoRe automatically selects the optimal number of measurement shots in each optimization phase, thus outperforming SGD and Bayes-SGD with different fixed number N_{shots} of shots through the entire optimization process. The adaptively selected number of shots and the accuracy threshold $\kappa(t)$ for GradCoRe are shown in Appendix F.

5.3 Comparison with State-of-the-art Methods

Figure 4 compares GradCoRe to the baseline methods, SGLBO, Bayes-NFT, EMICoRe, and SubsCoRe. Our GradCoRe, which significantly improves upon SGD as shown in Figure 3, establishes itself as the new state-of-the-art, exhibiting faster convergence and achieving lower overall energy (see Table 3 in Appendix F.1 for statistical significance test results. We also conducted experiments with different Q and L, as well as for the Heisenberg Hamiltonian, on which the results are reported in Appendix F.1.

6 Conclusion

The physical properties of variational quantum eigensolvers (VQEs) allow us to use specialized optimization methods, i.e., stochastic gradient descent (SGD) with parameter shift rules (PSRs) and a specialized sequential minimal optimization (SMO), called NFT (Nakanishi et al., 2020). Recent research has shown that those properties can be appropriately captured by the physics-informed VQE kernel, with which NFT has been successfully improved through Bayesian machine learning techniques. For instance, observations in previous SMO iterations are used to determine the optimal measurement points (Nicoli et al., 2023a), and observation costs are minimized based on the uncertainty prediction (Anders et al., 2024). In this paper, we have shown that a similar approach can also improve SGD-based methods. Specifically, we proposed Bayesian PSR, where the gradient is estimated by derivative Gaussian processes (GPs). Bayesian PSR generalizes existing PSRs to allow for flexible estimation from observations at an arbitrary set of locations. Furthermore, it provides uncertainty information, which enables observation cost adaptation through the novel notion of gradient confident region (GradCoRe). Our theoretical analysis revealed the relation between Bayesian PSR and existing PSRs, while our numerical investigation empirically demonstrated the utility of our approaches. We envisage that Bayesian approaches will facilitate further development of more efficient algorithms for VQEs and, more generally, quantum computing. In future work, we aim to explore the optimal combination of existing methods and strategies for selecting the most suitable approaches for specific tasks, i.e., specific Hamiltonians.

REFERENCES

H. Abraham et al. Qiskit: An open-source framework for quantum computing. *Zenodo*, 2019. doi: 10.5281/zenodo.2562111.

Rajeev Acharya, Dmitry A. Abanin, et al. Quantum error correction below the surface code threshold. *Nature*, 2024. doi: 10.1038/s41586-024-08449-y.

C. J. Anders, K. Nicoli, B. Wu, N. Elosegui, S. Pedrielli, L. Funcke, K. Jansen, S. Kuhn, and S. Nakajima. Adaptive observation cost control for variational quantum eigensolvers. In *Proceedings of 41st International Conference on Machine Learning (ICML2024)*, 2024. doi: 10.5555/3692070.3692133.

Dolev Bluvstein, Simon J Evered, Alexandra A Geim, Sophie H Li, Hengyun Zhou, Tom Manovitz, Sepehr Ebadi, Madelyn Cain, Marcin Kalinowski, Dominik Hangleiter, et al. Logical quantum processor based on reconfigurable atom arrays. *Nature*, pages 1–3, 2023. doi: 10.1038/s41586-023-06927-3.

Zhenyu Cai, Ryan Babbush, Simon C. Benjamin, Suguru Endo, William J. Huggins, Ying Li, Jarrod R. McClean, and Thomas E. O'Brien. Quantum error mitigation. *Rev. Mod. Phys.*, 95:045005, Dec 2023. doi: 10.1103/RevModPhys.95.045005.

S. Debnath, N. M. Linke, C. Figgatt, K. A. Landsman, K. Wright, and C. Monroe. Demonstration of a small programmable quantum computer with atomic qubits. *Nature*, 536(7614):63–66, 2016. doi: 10.1038/nature18648.

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2000.

P. Frazier. A tutorial on Bayesian optimization. *ArXiv e-prints*, 2018. doi: 10.48550/arXiv.1807.02811.

Giovanni Iannelli and Karl Jansen. Noisy Bayesian optimization for variational quantum eigensolvers. *ArXiv e-prints*, 2021. doi: 10.48550/arXiv.2112.00426.

D. Kielpinski, C. Monroe, and D. J. Wineland. Architecture for a large-scale ion-trap quantum computer. *Nature*, 417(6890):709–711, 2002. doi: 10.1038/nature00784.

Haoran Liao, Derek S. Wang, Iskandar Sitdikov, Ciro Salcedo, Alireza Seif, and Zlatko K. Minev. Machine learning for practical quantum error mitigation. *Nature Machine Intelligence*, 6(12):1478–1486, November 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00927-2. URL http://dx.doi.org/10.1038/s42256-024-00927-2.

Jarrod R McClean, Jonathan Romero, Ryan Babbush, et al. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016. doi: 10.1088/1367-2630/18/2/023023.

K. Mitarai, M. Negoro, M. Kitagawa, et al. Quantum circuit learning. *Phys. Rev. A*, 98:032309, 2018. doi: 10.1103/PhysRevA.98.032309.

Sarah Müller, Alexander von Rohr, and Sebastian Trimpe. Local policy search with bayesian optimization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Ken M. Nakanishi, Keisuke Fujii, and Synge Todo. Sequential minimal optimization for quantum-classical hybrid algorithms. *Phys. Rev. Res.*, 2:043158, 2020. doi: 10.1103/PhysRevResearch.2.043158.

K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, S. Kuhn, K.-R. Müller, P. Stornati, P. Kessel, and S. Nakajima. Physics-informed Bayesian optimization of variational quantum circuits. In *Advances in Neural Information Processing Systems (NeurIPS2023)*, 2023a.

K. A. Nicoli, C. J. Anders, et al. EMICoRe: Expected maximum improvement over confident regions. https://github.com/emicore/emicore, 2023b.

Kim A. Nicoli, Luca Wagner, and Lena Funcke. Machine-learning-enhanced optimization of noise-resilient variational quantum eigensolvers. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2501.17689.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, et al. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1):4213, 2014. doi: 10.1038/ncomms5213.

John Platt. Sequential minimal optimization : A fast algorithm for training support vector machines. *Microsoft Research Technical Report*, 1998.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2006. doi: 10.7551/mitpress/3206.001.0001.

Joschka Roffe. Quantum error correction: An introductory guide. *Contemporary Physics*, 60(3): 226–245, 2019. doi: 10.1080/00107514.2019.1667078.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1964. doi: 10.1017/S0013091500008889.

Shiro Tamiya and Hayata Yamasaki. Stochastic gradient line Bayesian optimization for efficient noise-robust optimization of parameterized quantum circuits. *npj Quantum Information*, 8(1):90, 2022. doi: 10.1038/s41534-022-00592-6.

Jules Tilly, Hongxiang Chen, Shuxiang Cao, et al. The variational quantum eigensolver: A review of methods and best practices. *Physics Reports*, 986:1–128, 2022. doi: https://doi.org/10.1016/j.physrep.2022.08.003.

David Wierichs, Josh Izaac, Cody Wang, and Cedric Yen-Yu Lin. General parameter-shift rules for quantum gradients. *Quantum*, 6:677, March 2022. ISSN 2521-327X. doi: 10.22331/q-2022-03-30-677.