

Low Overhead Precise Gate Importance Estimation for Pruning Parametrized Quantum Circuits

Rodrigo M. Sanz^{*}, Rupayan Bhattacharjee[†], Rohit Sarma Sarkar[†],
Andreu Anglès-Castillo^{*,‡}, Eduard Alarcón[†], and Carmen G. Almudéver^{*}

Abstract. Pruning parametrized quantum circuits requires quantifying the local contribution of each gate to the cost function, denoted as the *importance* in the classical domain, a quantity that existing methods approximate through first- or second-order Taylor expansions. We show that this approximation can be avoided by deriving a closed-form expression for the exact importance at zero additional cost. Through VQE simulations, we demonstrate that our criterion outperforms prior pruning methods, and we further analyze how the trainability and implementability of the circuit improve as increasing the number of pruned gates.

Key words. Parameterized Quantum Circuits, Variational QML, Circuit Pruning, Parameter-Shift Rule

1 Introduction and prior work

Parametrized Quantum Circuits (PQCs) underpin a broad class of hybrid quantum-classical algorithms, including Variational Quantum Eigensolvers (VQEs) [16], Quantum Machine Learning (QML) models [1, 19], and combinatorial optimization via QAOA [3]. A PQC ansatz takes the form $U(\boldsymbol{\theta}) = V(\boldsymbol{\theta}_L) W_L \cdots V(\boldsymbol{\theta}_1) W_1$, with $V(\boldsymbol{\theta}_i)$ single-qubit parametrized rotations and W_i fixed entangling blocks, and is optimized against a cost function $E(\boldsymbol{\theta}) = \text{Tr}\{M U(\boldsymbol{\theta}) \rho U^\dagger(\boldsymbol{\theta})\}$, where M is an observable encoding the problem of interest. The parameters $\boldsymbol{\theta}$ are updated with gradient-based optimizers [15, 6] relying on the parameter-shift rule (PSR) [12, 18, 21].

Just as in classical neural networks, one could expect that models with more parameters generally yield better performance, at the expense of computational cost, due to the increase in expressiveness. However, it is widely known that the addition of parametrized gates into a PQC can degrade its performance from two different angles. First, and most intuitively, a higher gate count and depth lead to accumulated errors from imperfect gate implementations and decoherence [17]. Secondly, deeper ansätze can drive the gradient variance to vanish exponentially with system size, giving rise to *Barren Plateaus* [11, 2, 9]. This combination of effects motivates techniques to reduce the size of the ansatz while minimizing the degradation of the cost function as much as possible.

In classical machine learning such methods are known as *pruning* techniques [10, 5]: a chosen criterion identifies parameters that are redundant or contribute weakly, which are removed iteratively to yield sparser models with reduced inference and, sometimes, training cost [13, 4]. A widely adopted choice is Molchanov et al.'s *importance* criterion [13], the squared change in the error E induced by setting a parameter w_m to zero,

$$\mathcal{I}_m = (E(\mathcal{D}, \mathbf{W}) - E(\mathcal{D}, \mathbf{W}|w_m = 0))^2, \quad (1)$$

where $\mathcal{D}$ denotes the input dataset. Direct evaluation is prohibitively expensive (one forward pass per parameter per round), motivating truncated Taylor expansions assembled from quantities already computed during backpropagation, at essentially no overhead:

$$\mathcal{I}_m^{(1)} = (g_m w_m)^2, \quad \mathcal{I}_m^{(2)} = (g_m w_m - \frac{1}{2} w_m \mathbf{H}_m \mathbf{W})^2, \quad (2)$$

where $g_m = \partial E / \partial w_m$ is the gradient and $\mathbf{H}_m$ the m -th row of the Hessian.

This approach has been applied to PQCs, where some works [20, 7] prune parameters whose absolute value $|\theta_l|$ falls below a heuristic cutoff. On the other hand, QADAPRUNE [8] accumulates successive gradient

^{*}Universitat Politècnica de València, Spain (rmarsan3@upv.edu.es).

[†]Universitat Politècnica de Catalunya, Barcelona, Spain.

[‡]Departament de Física Teòrica and IFIC, Universitat de València-CSIC, València, Spain.

differences as a proxy for the Hessian diagonal and freezes gates below an adaptive threshold. In all cases, the underlying score can ultimately be traced back to a first- or second-order approximation of the importance, as in Eq. (2). In this work, we demonstrate that such approximations are unnecessary in the quantum setting: exploiting the sinusoidal nature of parametrized gates [15, 19, 14], we derive an analytic gate-importance expression that depends only on quantities already computed by the PSR during gradient descent. The resulting criterion is zero-overhead and can replace the approximate score in any of the state-of-the-art approaches above. Finally, we study how pruning under this criterion affects trainability and circuit implementability, providing a finer understanding of these techniques for PQCs.

2 Analytical derivation of gate importance without overhead

Consider a variational ansatz $U(\boldsymbol{\theta}) = \prod_{m=1}^N \exp(-i \frac{\theta_m}{2} P_m) W_m$, with the cost function $E(\boldsymbol{\theta}) = \text{Tr}\{M U(\boldsymbol{\theta}) \rho U^\dagger(\boldsymbol{\theta})\}$. Since each Pauli P_m has eigenvalues ± 1 , holding every parameter but θ_m fixed makes E an *exact* single-frequency trigonometric polynomial [19],

$$E(\theta_m) = a_m + b_m \cos \theta_m + c_m \sin \theta_m, \quad (3)$$

for all $\theta_m \in [-\pi, \pi]$. Pruning the gate m sets $\theta_m = 0$, so from the original expression of Eq. (1), its exact importance at the current iterate $\theta_m^{(t)}$ is

$$\mathcal{I}_m \equiv (E(0) - E(\theta_m^{(t)}))^2 = (b_m(1 - \cos \theta_m^{(t)}) - c_m \sin \theta_m^{(t)})^2, \quad (4)$$

in terms of the Fourier coefficients (b_m, c_m) . Differentiating (3) once and twice at $\theta_m^{(t)}$ gives $E' = -b_m \sin \theta_m^{(t)} + c_m \cos \theta_m^{(t)}$ and $E'' = -b_m \cos \theta_m^{(t)} - c_m \sin \theta_m^{(t)}$, with $E' \equiv \partial E / \partial \theta_m$ and $E'' \equiv \partial^2 E / \partial \theta_m^2$. Both are known in closed form from the parameter-shift rule [12, 18],

$$E' = \frac{E_+ - E_-}{2}, \quad E'' = \frac{E_+ + E_-}{2} - E_0, \quad (5)$$

with $E_\pm \equiv E(\theta_m^{(t)} \pm \frac{\pi}{2})$ and $E_0 \equiv E(\theta_m^{(t)})$. The shifts $E_\pm$ are already computed by the PSR for the gradient and E_0 is the cost tracked at every step, so both derivatives come at no additional circuit execution. Inverting this 2×2 system expresses the coefficients as $b_m = -E' \sin \theta_m^{(t)} - E'' \cos \theta_m^{(t)}$ and $c_m = E' \cos \theta_m^{(t)} - E'' \sin \theta_m^{(t)}$; substituting into (4), the importance reduces to

$$\boxed{\mathcal{I}_m = (E''(1 - \cos \theta_m^{(t)}) - E' \sin \theta_m^{(t)})^2} \quad (6)$$

depending only on the PSR gradient and diagonal curvature at the current iterate. Equation (6) is thus available for free under any PSR-based optimizer and remains *exact* for any $\theta_m^{(t)}$, unlike the truncated Taylor estimates of Eq. (2)[§].

3 Results

We benchmark the exact criterion against the two state-of-the-art approximations on a VQE routine, taking the transverse-field Ising Hamiltonian $M = -J \sum_i Z_i Z_{i+1} - h \sum_i X_i$ and the hardware-efficient ansatz with 12 qubits and 8 layers ($N = 192$ single-qubit gates). The circuit is trained with Adam until convergence and then pruned one gate at a time, at each step removing the least important active gate under each criterion and recomputing importances on the surviving parameters, without any re-training. We compare the exact Fourier importance of Eq. (6) with its first-order approximation $(E' \theta_m)^2$ and with the parameter absolute value $|\theta_m|$, and track how pruning under the exact criterion trades circuit accuracy for hardware implementability and trainability.

[§] As $\theta_m^{(t)} \rightarrow 0$, Eq. (6) reduces to $(E' \theta_m^{(t)})^2$, recovering the first-order estimate $\mathcal{I}_m^{(1)}$ of Eq. (2).

As shown in Fig. 1, the exact criterion attains a normalized error up to two orders of magnitude ($\times 10^{-2}$) lower than both baselines in the intermediate pruning regime, at around 30 suppressed gates, while reducing the estimated hardware infidelity along the pruning process. Figure 2 further shows that trainability grows by over an order of magnitude as gates are removed, with the earliest prunes incurring negligible accuracy loss.

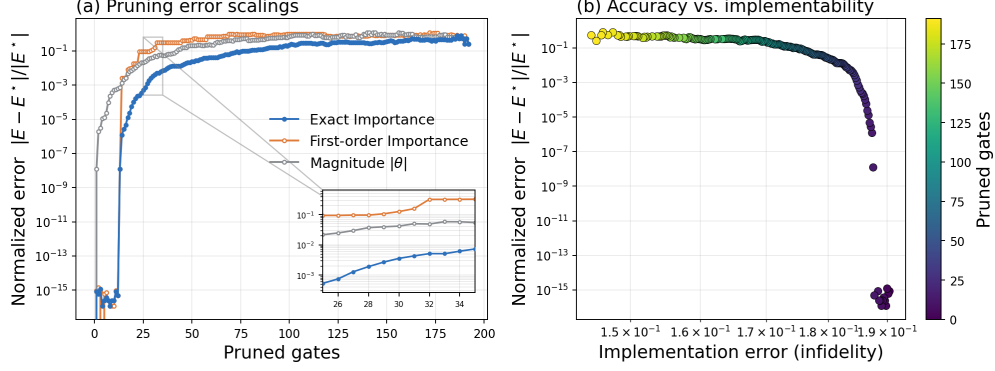


Figure 1: (a) Normalized cost-function error $|E - E^*|/|E^*|$, with E^* the converged value, as gates are pruned from the trained 12-qubit, 8-layer circuit. Removing gates in the order set by the exact importance degrades the cost function up to two orders of magnitude less than the first-order and magnitude criteria. (b) Along the exact-pruning trajectory, normalized error against the estimated implementation error (circuit infidelity under an IBM Heron r2 noise model), colored by number of pruned gates.

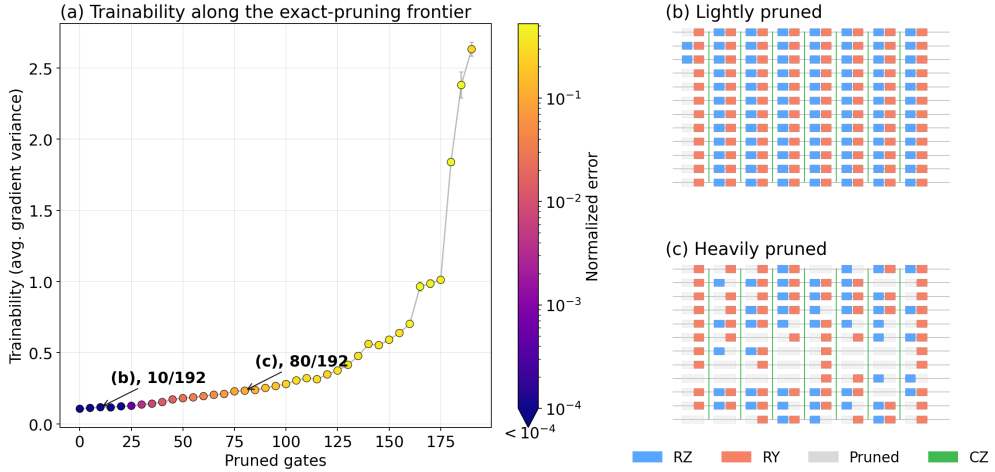


Figure 2: (a) Average gradient variance versus pruned gates, colored by normalized error; trainability grows by over an order of magnitude while the earliest prunes cost little accuracy. (b,c) Circuit structures for ansätze at the two marked operating points: lightly (5%) and heavily (42%) pruned.

Conclusion. In this work, an analytical expression for the gate importance has been derived, depending only on quantities already available from the PSR evaluations and thus incurring zero marginal cost. This criterion is benchmarked against first-order estimates from the literature, which it outperforms by up to two orders of magnitude in final normalized error. Beyond ranking accuracy, our experiments show that pruning under this criterion increases the trainability by over an order of magnitude while reducing the estimated hardware infidelity by up to 25%. Extending the exact criterion to the correlated importance of *filters* of gates, and delimiting when this remains tractable, are natural next steps.

References

- [1] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3:625–644, 2021.
- [2] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12:1791, 2021.
- [3] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm, 2014. arXiv:1411.4028.
- [4] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1803.03635.
- [5] Babak Hassibi, David G. Stork, and Gregory J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299, 1993.
- [6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.
- [7] Michael Kölle, Leonhard Klingert, Julian Schönberger, Philipp Altmann, Tobias Rohe, and Claudia Linnhoff-Popien. Investigating the lottery ticket hypothesis for variational quantum circuits, 2025. arXiv:2509.11190.
- [8] Ankit Kulshrestha, Xiaoyuan Liu, Hayato Ushijima-Mwesigwa, Bao Bach, and Ilya Safro. QAdaPrune: Adaptive parameter pruning for training variational quantum circuits. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, volume 2, pages 120–125, 2024. arXiv:2408.13352.
- [9] Martin Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoe Holmes, and M. Cerezo. Barren plateaus in variational quantum computing. *Nature Reviews Physics*, 2025.
- [10] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pages 598–605, 1990.
- [11] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9:4812, 2018.
- [12] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Physical Review A*, 98:032309, 2018.
- [13] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11264–11272, 2019.
- [14] Ken M. Nakanishi, Keisuke Fujii, and Synge Todo. Sequential minimal optimization for quantum-classical hybrid algorithms. *Physical Review Research*, 2:043158, 2020.
- [15] Mateusz Ostaszewski, Edward Grant, and Marcello Benedetti. Structure optimization for parameterized quantum circuits. *Quantum*, 5:391, 2021.
- [16] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O’Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5:4213, 2014.
- [17] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2:79, 2018.
- [18] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99:032331, 2019.
- [19] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103:032430, 2021.

- [20] Sukin Sim, Jonathan Romero, Jerome F. Gonthier, and Alexander A. Kunitsa. Adaptive pruning-based optimization of parameterized quantum circuits. *Quantum Science and Technology*, 6(2):025019, 2021. arXiv:2010.00629.
- [21] David Wierichs, Josh Izaac, Cody Wang, and Cedric Yen-Yu Lin. General parameter-shift rules for quantum gradients. *Quantum*, 6:677, 2022. arXiv:2107.12390.