Sign Gradient Descent for Quantum Optimisation

Thomas Crilly, Amira Abbas

University of Amsterdam, The Netherlands

Abstract

In this paper, we explore the possibility of efficient optimisation for variational quantum circuits through the use of Sign Gradient Descent, or SignGD. We prove that variational models satisfy the properties needed for convergence to a minima using SignGD in combination with a decaying learning rate. We further prove that the scaling efficiency of SignGD is in line with backpropagation for various regimes and support our findings with numerical experiments.

Introduction and motivation

Sign Gradient Descent, or SignGD, is an optimisation algorithm that has become prevalent in classical machine learning tasks due to its simplicity and efficiency in training deep learning models [1]. Unlike traditional gradient descent methods that use the magnitude and sign of the gradient vector, SignGD updates model parameters using only the sign information of the gradient. The result of such a technique can lead to a reduction in the impact of gradient noise, more stable convergence and faster training times when paired with a time-decaying learning rate [2, 3]. Interestingly, studies have shown that SignGD can achieve state-of-the-art performance with significantly less computational overhead, making it a promising alternative for large-scale machine learning tasks.

Currently, the scaling of quantum optimisation methods for variational quantum circuits (VQCs) presents significant challenges. As the number of qubits and circuit depth increase, traditional gradient-based optimisation techniques face issues such as vanishing gradients, commonly referred to as barren plateaus, which severely impede the convergence of these methods [4]. Perhaps even more concerning, is the required resources for computing gradients of VQCs with current methods, which do not achieve backpropagation-scaling, resulting in a slow and inefficient training process [5]. These issues ultimately render it difficult to optimise VQCs with many parameters efficiently and thus, scaling VQCs to parameter regimes akin to neural networks quickly becomes infeasible. We, therefore, aim to alleviate these issues through the use of SignGD, enhancing scalability and efficiency in optimising variational quantum circuits. In this work, we focus on a general subclass of VQCs where the parameterised exponentiated unitary operators are Pauli operators acting on all (or a subset of) qubits. Furthermore, we limit our observables to the Pauli-Z operator for simplicity. We prove that SignGD for VQCs converges at a comparable rate to gradient descent, where the adaptive learning rate controls the speed of convergence. Additionally, there are provable regimes where backpropagation-scaling for the class of VQCs considered, is achieved – which we outline in this work. Supporting these results are numerical simulations which show that SignGD, combined with a decaying learning rate, indeed attains a similar convergence to gradient descent on a collection of randomly generated circuits.

SignGD convergence with variational models

In order to prove that SignGD converges to a local minima, we first require a lemma on the smoothness of convergence.

Lemma 1. *Let $f'(\theta_t)$ denote the gradient of the objective function of our quantum variational model $f(\cdot)$ evaluated at point θ_t where $t \in \mathbb{N}$ denotes the training iteration. Then $\forall\ \theta_{t+1}, \theta_t$, we require that for some non-negative constant $\tilde{L} := [L_1, \ldots, L_d]$,*

$$f(\theta_{t+1}) \leq f(\theta_t) + f'(\theta_t)^T(\theta_{t+1} - \theta_t) + \frac{1}{2}\sum_i L_i(\theta_{t+1} - \theta_t)^2, \tag{1}$$

Proof sketch: This lemma can be proved in the standard way for any quantum variational circuit. First, we expand our function up to second order in terms of powers of $(\theta_{t+1} - \theta_t)$. Using this expansion, one can then bound the Hessian of the function in terms of a vector L by an explicit construction. After some rearrangement, the result follows straightforwardly.

Using this lemma, we can prove the central point of this paper.

Theorem 2 (SignGD convergence for VQCs)**.** *Given a time decaying learning rate η_t combined with a variational quantum circuit, SignGD will converge to a local minima.*

Proof sketch: Using the smoothness result from Lemma 1, we analyse the sign update rule $\theta_{t+1} = \theta_t - \eta_t \cdot \text{sign}(f'(\theta_t))$. We derive that the function value decreases with each iteration. Using a decaying learning rate, we then show that the average gradient norm weighted by the learning rate converges to zero as the number of iterations goes to infinity. Thus, SignGD with a decaying learning rate ensures convergence.

Provably favourable resource scaling

A simple operational definition for backpropagation-scaling using relative complexity bounds derived from classical automatic differentiation literature [6] is outlined in Ref. [5]. The key characteristic can be summarised in one sentence: computational and memory resources employed to compute gradients of a function are bounded multiples of those used to compute the function. We make this explicit in the following definition, and prove regimes for which SignGD achieves this scaling with VQCs.

Definition 3 (Backpropagation scaling)**.** Given a parameterised function $f(\theta)$, $\theta \in \mathbb{R}^M$, let $f'(\theta)$ be an estimate of the gradient vector accurate to within some constant ε in the infinity norm. The total computational cost incurred to obtain $f'(\theta)$ with backpropagation is bounded such that

$$\text{TIME}(f'(\theta)) \leq c_t \cdot \text{TIME}(f(\theta)), \tag{2}$$

and

$$\text{MEMORY}(f'(\theta)) \leq c_m \cdot \text{MEMORY}(f(\theta)), \tag{3}$$

where $c_t, c_m = O(\log(M))$, and TIME$(\cdot)$ and MEMORY$(\cdot)$ capture the time and space complexity respectively, for either computing the function f or its gradient f'.

With this definition at hand, there are several regimes for which we prove backpropagation-scaling for VQAs employing SignGD.

Promise on gradient magnitudes: The first regime for provably favourable resource scaling is where one has a bound on the gradient magnitudes, i.e. each gradient component is $> \varepsilon$. If we have such a promise, this leads to the following theorem.

Theorem 4 (SignGD with gradient bounds)**.** *Let $U(\theta)$ be the parameterised operations applied in the VQC model and define $f(\theta) = \langle\psi(\theta)| O |\psi(\theta)\rangle = \text{Tr}[O\rho(\theta)]$ as the model function. The gradient with respect to the k^{th} parameter component is*

$$[f'(\theta)]_{\theta_k} = -2 \text{ Im}\big[\langle 0| U(\theta)^\dagger O \partial_{\theta_k} U(\theta) |0\rangle \big] := -2 \text{ Im}[\langle\psi|\lambda_k\rangle]$$

Given m copies of a state $\rho = |\psi\rangle\langle\psi|$, and the guarantee that

$$|\text{Tr}[O_k\rho]| > \varepsilon, \tag{4}$$

for all $k = 1, ..., M$, where $O_k = |\lambda_k\rangle\langle\lambda_k|$. Then, one can obtain

$$\text{sign}(\text{Tr}[O_k\rho]) \ \forall \ k = 1, ..., M \tag{5}$$

with probability $1 - \delta$ using $m = \Theta(\log(M/\delta)/\varepsilon^2)$ copies of ρ, which is in line with backpropagation scaling.

Proof. Given a two-outcome majority vote observable E consisting of a tensor product of projectors onto the eigenspace of O_k, the probability of obtaining the correct sign for $|\mathrm{Tr}[O_k\rho]|$ can be bounded by Hoeffding's inequality (as shown in Ref. [7])

$$1 - e^{-\frac{1}{2}|\mathrm{Tr}[O_k\rho]|^2 m} > 1 - e^{-\frac{1}{2}\varepsilon^2 m} = 1 - \frac{\delta^2}{2M^2},$$

by choosing $m = 4\log(2M/\delta)/\varepsilon^2$. By the union bound, the result then follows for estimating $|\mathrm{Tr}[O_k\rho]|$ for all k with the desired probability $1-\delta$. □

The assumption that gradient magnitudes are $> \varepsilon$ is quite reasonable in optimisation, as if a component is close to zero, one may merely stop optimising in this direction.

Promise on mistake bound: Another, perhaps less reasonable but still interesting, regime for which we can prove backpropagation-scaling is when we have a guarantee on the number of mistakes for an estimate of the sign vector.

Theorem 5 (SignGD with mistake bound)**.** *Consider the vector of sign information $g' \in \{-1,1\}^M$. Promised that f' differs from g' on only $O(\mathrm{polylog}(M))$ entries, one can find all the incorrect entries and flip the sign with a cost that is in line with backpropagation-scaling.*

Proof sketch: Using the gentle search procedure from Ref. [8], one can use $O(\mathrm{polylog}(M))$ copies of ρ to find the components of g' that differ from f', provided that the number of discrepencies is $O(\mathrm{polylog}(M))$.

Empirical analysis and outlook: Finally, we present a figure demonstrating an application of our results. In the following experiment, we run two versions of gradient descent simultaneously on ten different circuits. Each circuit undergoes 25 trials, with the initial and true parameters for each trial chosen at random. We then plot the average loss for each method at each iteration step. As expected, SignGD achieves comparable performance with gradient descent when optimising VQCs. Our full manuscript broadens the scope of SignGD for VQCs and analyses more cases for which convergence provably achieves backpropagation-scaling. Overall, we find that SignGD may be a promising candidate for quantum optimization through alleviation of resource and noise scaling issues.

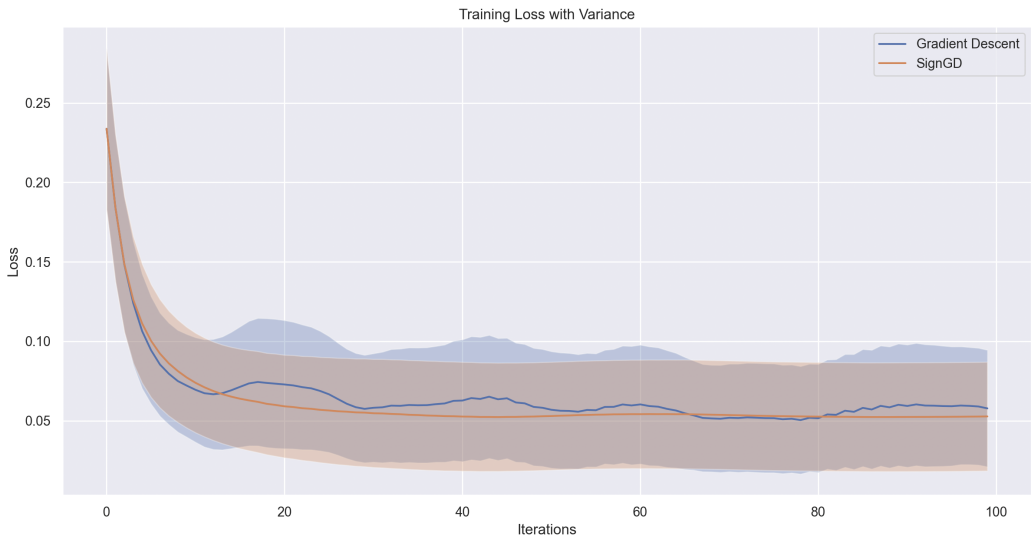

Figure 1: A graph showing the convergence of SignGD and gradient descent, over ten random circuits.

References

[1] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

[2] Emmanuel Moulay, Vincent Léchappé, and Franck Plestan. Properties of the sign gradient descent algorithms. *Information Sciences*, April 2019.

[3] Xiuxian Li, Kuo-Yi Lin, Li Li, Yiguang Hong, and Jie Chen. On faster convergence of scaled sign gradient descent, 2021.

[4] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1), November 2018.

[5] Amira Abbas, Robbie King, Hsin-Yuan Huang, William J. Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod R. McClean. On quantum backpropagation, information reuse, and cheating measurement collapse, 2023.

[6] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

[7] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.

[8] Costin Bădescu and Ryan O'Donnell. Improved quantum data analysis. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1398–1411, 2021.