Backpropagation scaling in parameterised quantum circuits

arXiv 2306.14962

Joseph Bowles, David Wierichs, Chae-Yeun Park

28.07.2023

QML seminar | AG Eisert | FU Berlin

XANADU

Backpropagation scaling in parameterised quantum circuits

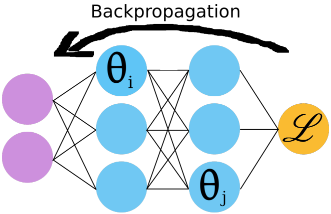
Joseph Bowles

David Wierichs

Chae-Yeun Park

XANADU

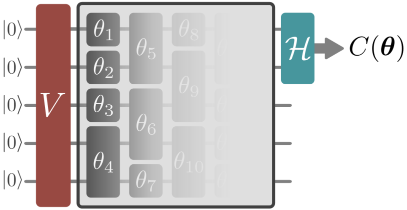
Backpropagation



- Cost function gradient from chain rule
- Gradient about as expensive as function itself
- Made machine learning scalable

Parameterised quantum circuits (PQCs)

$$C(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|\,\mathcal{H}\,|\psi(\boldsymbol{\theta})\rangle$$
$$|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\,V|0\rangle$$



- State preparation circuit V
- Variational circuit $U(\boldsymbol{\theta}) = \prod_j e^{-i\theta_j G_j}$
- Estimate $C(\boldsymbol{\theta})$ with $M \in \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ shots to precision ϵ

Gradients of PQCs

Gradients of PQCs

◦Circuit differentiation through generator decomposition

- Train circuit parameters using $\nabla C(\boldsymbol{\theta})$
- Large number of methods to choose from

◦ Linear combination of unitaries

◦ Pulse-generator parameter-shift rule

◦Generalised circuit differentiation through spectral decomposition

◦ Finite differences

◦ “Proper” parameter-shift rules

◦ Parameter-shift rule

◦ Generalized parameter-shift rules

◦ Stochastic parameter-shift rule

◦ Effective generator parameter-shift rule

◦ Stochastic pulse parameter-shift rule

Gradients of PQCs

- Train circuit parameters using $\nabla C(\boldsymbol{\theta})$
- Large number of methods to choose from
- *All scale linearly with the number of parameters:*

$$M_{\nabla C} \in \mathcal{O}\left(\frac{n}{\epsilon^2}\right)$$

  for n parameters and ℓ_∞-norm precision ϵ

- More precisely:

$$\frac{TIME(\nabla C) MEM(\nabla C)}{TIME(C) MEM(C)} \in \Omega(n)$$

Gradient cost of PQCs - Example

Example: Circuit with n Pauli rotation gates

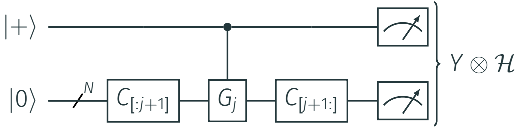
- Parameter-shift rule

$$\partial_j C(\boldsymbol{\theta}) = \frac{1}{2}\left[C\left(\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_j\right) - C\left(\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_j\right)\right]$$

$\Rightarrow$ $2n$ circuits (sequential / parallel)

- Linear combination of unitaries / Hadamard test

$\partial_j C(\boldsymbol{\theta}) = -2\times$



$\Rightarrow$ n circuits with additional qubit and depth

Scaling of gradient cost

Linear scaling sounds fine, what's the fuss about?

- QML: Model with $n = 10000$ for 1000 data points
  $1\mu s$ / circuit

$\Rightarrow$ One gradient to precision $\epsilon = 10^{-3}$: 0.6 years

- VQE: Measure water to chemical accuracy: 2.3d[1]

$\Rightarrow$ *One* gradient for single-layer HEA ($n = 208$): 2.6yrs

[1] Gonthier, Radin et al. Phys. Rev. Res. **4** 033154, 2022.

Scaling of gradient cost

- Backpropagation scaling:

Overhead at most logarithmic in n

$$TIME(\nabla C) \leq c_t TIME(C),\ c_t \in \mathcal{O}(\log(n))$$
$$MEM(\nabla C) \leq c_m MEM(C),\ c_m \in \mathcal{O}(\log(n))$$

- Modern large-scale ML unfeasible without backprop
- $n \approx 10^{11}$ parameters[2] $\Rightarrow$ Speedup of $\frac{n}{\log n} \approx 10^{10}$
- No backpropagation scaling for generic PQCs[3]

[2] GPT-3 has 175 billion parameters
[3] Abbas, King et al. arXiv 2305.13362, 2023.

Commuting-generator circuits

Theorem (Commuting-generator circuits)

Consider an N-qubit circuit $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})V|0\rangle$.
Assume that the generators commute ($[G_j, G_k] = 0$) and that each generator commutes or anticommutes with $\mathcal{H}$ ($[G_j, \mathcal{H}] = 0 \vee \{G_j, \mathcal{H}\} = 0$).
Then $\nabla C(\boldsymbol{\theta})$ can be estimated without bias to precision ϵ using $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ shots of an N-qubit circuit.

Commuting-generator circuits

Why does this work?

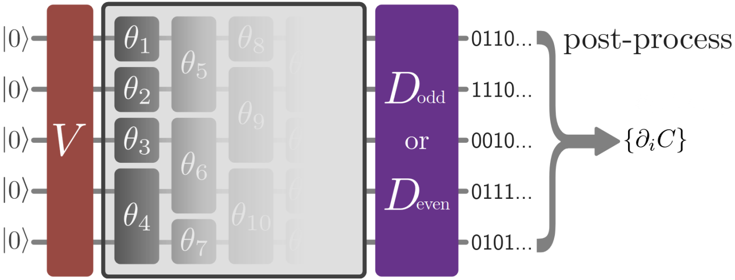
$$\frac{\partial C}{\partial \theta_j}(\boldsymbol{\theta}) = i \langle \psi(\boldsymbol{\theta}) | [G_j, \mathcal{H}] | \psi(\boldsymbol{\theta}) \rangle =: \langle \psi(\boldsymbol{\theta}) | O_j | \psi(\boldsymbol{\theta}) \rangle$$

$$O_j = \begin{cases} 2iG_j\mathcal{H} & \{G_j, \mathcal{H}\} = 0 \\ 0 & [G_j, \mathcal{H}] = 0 \end{cases}$$

$$\Rightarrow \quad [O_j, O_k] = 0$$

We can measure all $\{O_j\}_j$ simultaneously.

Commuting-generator circuits



- $c_m = 1$
- Additional unitary D to diagonalise $\{O_j\}_j$
- No guarantee for backpropagation scaling yet

Higher-order derivatives

- Get all derivatives of given order simultaneously
- Even get all odd(even)-order derivatives at once
- May require more involved diagonalisation D
- Quantum Fisher information becomes constant

$$\mathcal{F}_{jk} = \mathrm{Cov}(G_j, G_k)_{V|0\rangle}$$

- Approximations of advanced optimization schemes

Can we/should we train commuting-generator PQCs?

Are commuting-generator PQCs still powerful?

- PQC function $C(\boldsymbol{\theta})$ may be classically hard due to V
- Sampling is hard even for $V = \mathbb{I}$ (IQP circuits)[4]

Are commuting-generator PQCs trainable?

- Conjecture: DLA captures gradient magnitude[5]
- Commuting-generator PQCs: $\dim(DLA) = n \leq Nd_C$

[4] Bremner, Montanaro and Shepherd. Phys. Rev. Let. **117** 080501, 2016.
[5] Larocca, Czarnik et al. Quantum **6** 824, 2022.

Commuting-Pauli-generator circuits

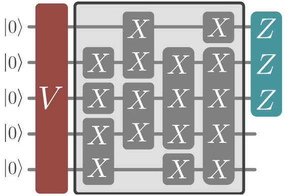
- Choose $\mathcal{H}, G_j \in \{\mathbb{I}, X, Y, Z\}^N$
- (Anti)commutativity guaranteed
- Depth of D bounded[6] $\Rightarrow\ c_t - 1 \in \mathcal{O}\left(\frac{N}{d_C \log N}\right)$

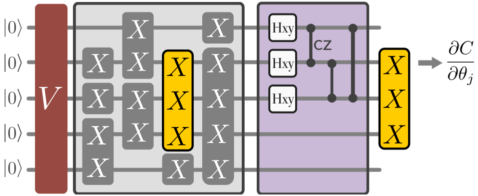
$c_t - 1$	$d_C = \log N$	$d_C = \sqrt{N}$	$\boxed{d_C = N}$
(single layer) $n = N$	$\frac{n}{\log^2 n}$	$\frac{\sqrt{n}}{\log n}$	$\frac{1}{\log n}$
(all gates) $n = Nd_C$	$\frac{n}{\log^3 n}$	$\frac{\sqrt[3]{n}}{\log n}$	$\frac{1}{\log n}$

[6] Jiang, Sun et al. Proc. of 14. Annual ACM-SIAM Symposium, 213-229, 2020.

X-generator ansatz

- Choose generators $G_j \in \{\mathbb{I}, X\}^N$, $\mathcal{H} \in \{\mathbb{I}, Z\}^N$
- D has depth $d_D \approx |\mathcal{H}| + 2$
- Backpropagation scaling for $|\mathcal{H}| \in \mathcal{O}\left(d_C \log n\right)$
- $d_D = 1$ for $\mathcal{H} = Z_r \Rightarrow \; c_t \approx c_m = 1$

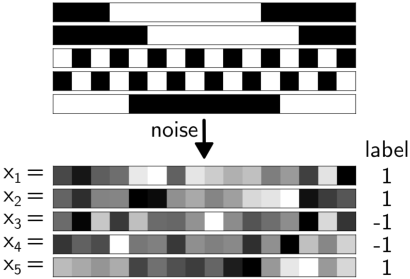




Numerical example

Numerics: Classification problem

- "Bars and dots" dataset



- Data points of length 16
- Task: Classify bars vs dots

Numerics: Models

- Encoding $V(\boldsymbol{x}_i) = \prod_{r=1}^{16} RY_r(x_{i,r}/2)$

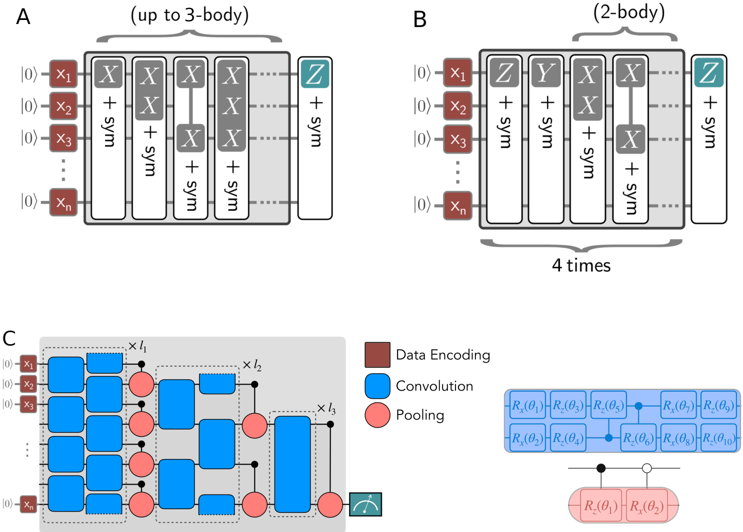
A: Translation-equivariant X-generator circuit
$\mathcal{H} = \sum_r Z_r$; parallel gradients

B: Translation-equivariant non-commuting circuit
$\mathcal{H} = \sum_r Z_r$; parameter-shift

C: Quantum convolutional neural network
$\mathcal{H} = Z_{16}$; parameter-shift

Powered by PENNYLANE

Numerics: Models

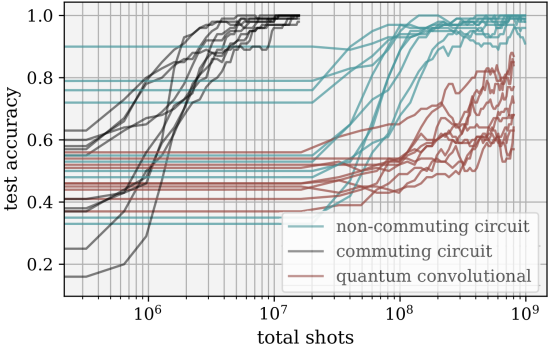


QCNN figures adapted from Hur, Kim and Park. Quant. Mach. Int. **4** 3,

Numerics: Models

Model	n	$\lvert\{G_j\}\rvert$	# circuits/gradient
A	44	696	$N = 16$
B	40	608	$N^2(L-1) + 3N(L+1) - 2 = 1006$
C	48	320	$56N - 80 = 816$

Numerics: Results



- Significantly reduced optimization cost
- Performance maintained for commuting generators

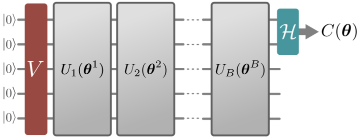
Theorem (Commuting-block circuits)

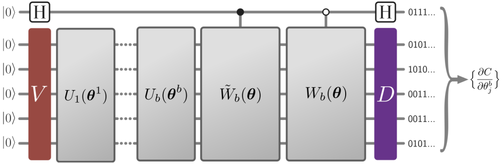
Consider an N-qubit circuit $|\psi(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})V|0\rangle$ with B blocks.

Assume that the generators commute within each block, commute or anticommute between blocks and that each generator commutes or anticommutes with $\mathcal{H}$.

Then $\nabla C(\boldsymbol{\theta})$ can be estimated without bias to precision ϵ using $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ shots of each of $2B-1$ circuits on $N+1$ qubits.

Commuting-block circuits





- $B = 1$ for commuting-generator circuit
- $B = n$ for arbitrary circuit ($\Rightarrow$ LCU/Hadamard test)
- $c_t \geq 2$ generically

Tailored training methods

- Grow circuit ansätze (e.g. ADAPT-VQE, ROTOSELECT)
- Parallelized evaluation of building block quality[7]
- “Greedy” training of newly added (commuting) blocks
- Many circuits can be seen as commuting-block circuits

[7] Anastasiou, Mayhall et al. arXiv 2306.03227, 2023.

Conclusion

- Backpropagation scaling in PQCs
    - excluded for fully general circuits
    - guaranteed for some circuit classes
    - achievable in practice
    - important for QML and NISQ use cases
- Use tailored PQCs instead of generic ansätze!

Thank you for your attention

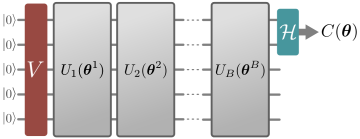
github.com/xanaduAI/backprop_scaling_pqcs

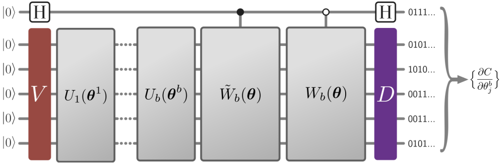
arxiv.org/abs/2306.14962

Gradients of commuting-block circuits

$$
\begin{aligned}
W_b &= U_B(\boldsymbol{\theta}^B)\cdots U_{b+2}(\boldsymbol{\theta}^{b+2})U_{b+1}(\boldsymbol{\theta}^{b+1}) \\
W_b &= G_j\tilde{W}_b \quad \forall\ G_j \quad \text{(block (anti)commutation)} \\
\Rightarrow \quad \frac{\partial C}{\partial \theta_j^b} &= \langle\psi_b|\left(\tilde{W}_b{}^\dagger iG_j\mathcal{H}W_b - (-1)^{g_j}W_b{}^\dagger iG_j\mathcal{H}\tilde{W}_b\right)|\psi_b\rangle \\
&= \frac{1}{2}\left[\left\langle O_j\right\rangle_{L^+_{W_b}|\psi_b\rangle} - \left\langle O_j\right\rangle_{L^-_{W_b}|\psi_b\rangle}\right] \\
&= \left\langle 2Z\otimes O_j\right\rangle_{|\phi_b\rangle} \quad \text{(Hadamard test construction)}
\end{aligned}
$$

Gradients of commuting-block circuits



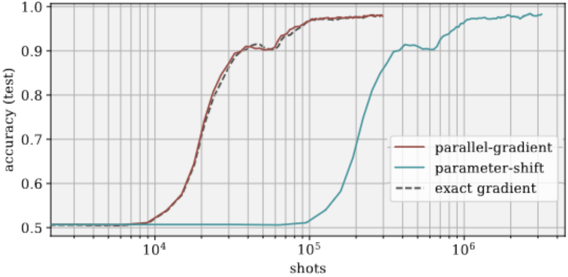


- One additional qubit
- Additional depth via D and $c\tilde{W}_b, \bar{c}W_b$

Numerics: Details

- Train with Adam for 50 epochs
- 1000 training data points, train with batches of 20
- Evaluate test accuracy on 100 test data points
- 10 runs per model

Optimization at finite shots



- Compare finite shots with exact optimization
- 6-qubit circuit
- Sufficiently close in behaviour

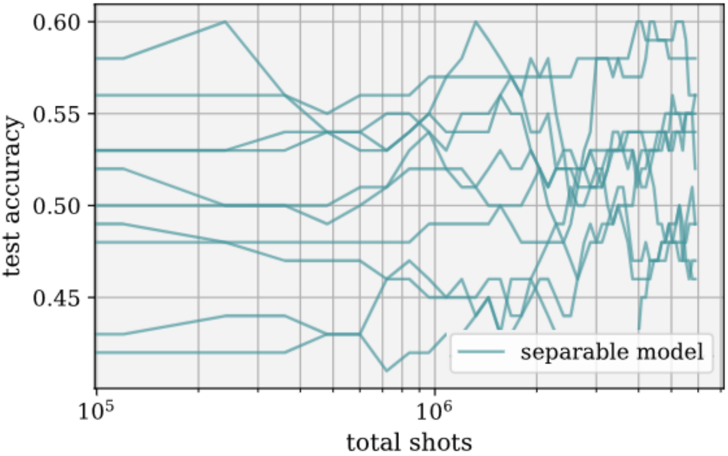
Numerics: Separable model

D: Separable model

Model	n	$\lvert\{G_j\}\rvert$	# circuits/gradient	
A	44	696	N	$= 16$
B	40	608	$N^2(L-1)+3N(L+1)-2$	$= 1006$
C	48	320	$56N-80$	$= 816$
D	48	48	6	

- Arbitrary rotation on each qubit $d = 3,\ n = 3N$
- Sanity check for hardness of classification task
- Trivial to simulate classically
- Gradient computation parallelized over qubits

Numerics: Separable model



- Very poor performance at very low cost