Trainability Proof Notes

Zoe Holmes

December 2022

1 Local to global state bound

Proposition 1. *Let ρ be an arbitrary quantum state, let $U_{\boldsymbol{\theta}}$ be a parameterized unitary matrix and write $\rho_U = U_{\boldsymbol{\theta}}^\dagger \rho U_{\boldsymbol{\theta}}$ for short. Consider a global cost of the form*

$$C_G(\boldsymbol{\theta}) = 1 - \mathrm{Tr}\left[\rho_U H_{\mathrm{G}}\right] \quad \textit{where} \quad H_G = |0\rangle\langle 0|^{\otimes n} \tag{1}$$

is the all zero projector, and a local cost of the form

$$C_L(\boldsymbol{\theta}) = 1 - \mathrm{Tr}\left[\rho_U H_{\mathrm{L}}\right] \quad \textit{where} \quad H_L = \frac{1}{n}\sum_j^n |0\rangle\langle 0|_j \otimes \mathbb{I}_{\overline{j}} \tag{2}$$

It follows that

$$C_L(\boldsymbol{\theta}) \leqslant C_G(\boldsymbol{\theta}) \leqslant n C_L(\boldsymbol{\theta}) \tag{3}$$

Thus $C_L(\boldsymbol{\theta}) = 0$ iff $C_G(\boldsymbol{\theta}) = 0$.

Proof. We can write $H_{\mathrm{L}} = \frac{1}{n}\sum_j^n H_{\mathrm{L},j}$, where

$$H_{L,j} = |0\rangle\langle 0|_j \otimes \mathbb{I}_{\overline{j}} \tag{4}$$

are projectors that mutually commute. Note that $\prod_{j=1}^n H_{\mathrm{L},j} = H_{\mathrm{G}}$. We can associate events E_j with the projectors $H_{\mathrm{L},j}$ such that $\Pr[E_j] = \mathrm{Tr}\left[\rho_U H_{\mathrm{L},j}\right]$. Then, $\mathrm{Tr}\left[\rho \prod_{j=1}^n H_{\mathrm{L},j}\right] = \Pr \cap_{j=1}^n E_j$. Recall, from basic probability theory, that for any set of events $\mathcal{A} \coloneqq \{A_1, A_2, \ldots, A_n\}$, it holds that

$$\Pr[\cup_{i=1}^n A_i] \geqslant \frac{1}{n}\sum_{i=1}^n \Pr[A_i]\,. \tag{5}$$

Choosing $A_i = \overline{E_j}$, we see

$$\begin{aligned}
&\Pr\big[\cup_{j=1}^n \overline{E_j}\,\big] \geqslant \frac{1}{n}\sum_{i=j}^n \Pr\big[\,\overline{E_j}\,\big] \\
\implies &1 - \Pr\big[\cap_{j=1}^n E_j\big] \geqslant \frac{1}{n}\sum_{j=1}^n \left(1 - \Pr[E_j]\right) \\
\implies &1 - \Pr\big[\cap_{j=1}^n E_j\big] \geqslant 1 - \frac{1}{n}\sum_{j=1}^n \mathrm{Tr}[\rho_U H_{\mathrm{L},j}] \\
\implies &1 - \mathrm{Tr}[\rho_U H_G] \geqslant 1 - \mathrm{Tr}[\rho_U H_{\mathrm{L}}]\,.
\end{aligned} \tag{6}$$

This is precisely the first desired inequality $C_L \leqslant C_G$.

To prove the remaining inequality, observe that, via the union bound,

$$\Pr\big[\cup_{j=1}^n A_j\,\big] \leqslant \sum_{j=1}^n \Pr[A_j] \tag{7}$$

we have

$$\begin{aligned} \Pr\left[\cup_{j=1}^{n} \overline{E_j}\right] &\leqslant \sum_{j=1}^{n} \Pr\left[\overline{E_j}\right] \\ \implies 1 - \Pr\left[\cap_{j=1}^{n} E_j\right] &\leqslant \sum_{j=1}^{n} \left(1 - \Pr[E_j]\right) \\ \implies 1 - \mathrm{Tr}[\rho_U H_G] &\leqslant n\left(1 - \mathrm{Tr}[\rho_U H_L]\right) \end{aligned} \tag{8}$$

Thus, $C_G \leqslant nC_L$ as required. □

2 Barren Plateaus

A parameterized unitary circuit can be represented as

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} \exp(-i\theta_l V_l) W_l = \underbrace{\prod_{l=k+1}^{L} \exp(-i\theta_l V_l) W_l}_{U_+} \underbrace{\prod_{l=1}^{k} \exp(-i\theta_l V_l) W_l}_{U_-} \tag{9}$$

where $\{V_l\}_{l=1}^{L}$ are hermitian, involutionary ($V_l^2 = \mathbb{I}$) generators of rotations $\exp(-i\theta_l V_l)$ around independent, trainable angles $\boldsymbol{\theta} \in [0, 2\pi)^L$ and $\{W_l\}_{l=1}^{L}$ are fixed gates.

Taking the partial derivative of the circuit gives

$$\partial_k U(\boldsymbol{\theta}) = -iU_+ V_k U_- \tag{10}$$

$$\partial_k U^\dagger = iU_-{}^\dagger V_k U_+{}^\dagger \tag{11}$$

We will take our cost function to be the standard VQE cost, i.e.

$$E(\boldsymbol{\theta}) = \langle 0|U(\boldsymbol{\theta})^\dagger H U(\boldsymbol{\theta})|0\rangle = \mathrm{Tr}\Big(U(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger H\Big). \tag{12}$$

This is the expectation value of a Hermitian operator H for a state $U(\boldsymbol{\theta})\rho U(\boldsymbol{\theta})^\dagger$ where $\rho = |0\rangle\langle 0|^{\otimes n}$.

The derivative of the cost function with respect to θ_k is given by

$$\partial_{\theta_k} E(\boldsymbol{\theta}) = i\,\mathrm{Tr}\Big(\rho(U_-{}^\dagger V_k U_+{}^\dagger H U - U^\dagger H U_+ V_k U_-)\Big) \tag{13}$$

$$= i\,\mathrm{Tr}\left(\rho U_-{}^\dagger (V_k \underbrace{U_+{}^\dagger H U_+}_{H_+} - \underbrace{U_+{}^\dagger H U_+}_{H_+} V_k) U_-\right) \tag{14}$$

$$= i\,\mathrm{Tr}\left(\underbrace{U_-\rho U_-{}^\dagger}_{\rho_-}[V_k, H_+]\right) \tag{15}$$

$$= i\,\mathrm{Tr}\Big([\rho_-, V_k] U_+{}^\dagger H U_+\Big) \tag{16}$$

In the final line, we use the cyclicity of the trace operation, i.e. $\mathrm{Tr}[A[B,C]] = \mathrm{Tr}[ABC - ACB] = \mathrm{Tr}[ABC - BAC] = \mathrm{Tr}[[A,B]C]$.

2.1 Average of the cost gradients

The average value of the gradient of the cost function, over a random initialisation of the parameters $\boldsymbol{\theta}$, is given by

$$\langle \partial_{\theta_k} E(\boldsymbol{\theta})\rangle = \int \mathrm{d}U p(U) \partial_{\theta_k} E(\boldsymbol{\theta}) = \int \mathrm{d}U_+ p(U_+) \mathrm{d}U_- p(U_-) \partial_{\theta_k} E(\boldsymbol{\theta}) \tag{17}$$

Thus we have

$$\langle \partial_{\theta_k} E(\boldsymbol{\theta})\rangle = i \int\int \mathrm{d}U_- p(U_-)\,\mathrm{Tr}\left(\rho_- \left[V_k, \int \mathrm{d}U_+ p(U_+) H_+\right]\right) \tag{18}$$

$$\tag{19}$$

Assuming U_- is at least a 1-design using Eq.(39) (from Section 2.3) we have that

$$\langle \partial_{\theta_k} E(\boldsymbol{\theta})\rangle = i \int \mathrm{d}\mu(U_-)\,\mathrm{Tr}\left(\rho_- \left[V_k, \int \mathrm{d}U_+ p(U_+) H_+\right]\right) \tag{20}$$

$$= i\frac{\mathrm{Tr}(\rho)}{d}\,\mathrm{Tr}\left(\left[V_k, \int \mathrm{d}U_+ p(U_+) H_+\right]\right) \tag{21}$$

$$= 0 \tag{22}$$

as the trace of a commutator is zero.

Similarly, assuming U_+ is at least a 1-design, results in:

$$\langle \partial_{\theta_k} E(\boldsymbol{\theta}) \rangle = i \int \mathrm{d}U_- p(U_-) \operatorname{Tr}\Big(\rho_- \Big[V_k, \int \mathrm{d}\mu(U_+) H_+ \Big] \Big) \tag{23}$$

$$= i \int \mathrm{d}U_- p(U_-) \operatorname{Tr}\Big(\rho_- \Big[V_k, \frac{\operatorname{Tr}(H)}{d} \mathbb{I} \Big] \Big) \tag{24}$$

$$= 0 \tag{25}$$

Hence, the gradients are not biased in any single direction.

2.2 Variance of the cost gradients

The probability that the cost function gradient deviates from its average of zero can be bounded using Chebyshev's inequality,

$$P(|\partial_{\theta_k} E| \geqslant \epsilon) \leqslant \frac{\operatorname{Var}(\partial_{\theta_k} E)}{\epsilon^2}\,, \tag{26}$$

where the variance of the gradient is given as:

$$\operatorname{Var}(\partial_{\theta_k} E) = \big\langle (\partial_{\theta_k} E)^2 \big\rangle - \underbrace{\langle \partial_{\theta_k} E \rangle^2}_{0} = \int \mathrm{d}U p(U) (\partial_{\theta_k} E)^2 \tag{27}$$

$$= \int \mathrm{d}U_- p(U_-) \int \mathrm{d}U_+ p(U_+) i^2 \operatorname{Tr}\Big(\rho_-^{\otimes^2} [V_k, H_+]^{\otimes^2} \Big) \tag{28}$$

Assuming U_- is a 2-design, using Eq.(40), results in:

$$\begin{aligned} \operatorname{Var}(\partial_{\theta_k} E) &= -\int \mathrm{d}U_+ p(U_+) \int \mathrm{d}\mu(U_-) \operatorname{Tr}\Big(U_-^{\otimes^2} \rho^{\otimes^2} {U_-}^{\dagger \otimes^2} [V_k, H_+]^{\otimes^2} \Big) \\ &= -\int \mathrm{d}U_+ p(U_+) \Bigg(\frac{\operatorname{Tr}\Big([V_k, H_+]^{\otimes^2}\Big) \operatorname{Tr}\Big(\rho^{\otimes^2}\Big) + \operatorname{Tr}\Big([V_k, H_+]^2\Big) \operatorname{Tr}\Big(\rho^2\Big)}{d^2 - 1} \\ &\quad - \frac{\operatorname{Tr}\Big([V_k, H_+]^2\Big) \operatorname{Tr}\Big(\rho^{\otimes^2}\Big) + \operatorname{Tr}\Big([V_k, H_+]^{\otimes^2}\Big) \operatorname{Tr}(\rho^2)}{d(d^2-1)} \Bigg) \\ &= -\left(1 - \frac{1}{d}\right) \frac{\Big\langle \operatorname{Tr}\Big([V_k, H_u]^2\Big) \Big\rangle_{U_+}}{d^2 - 1} \end{aligned} \tag{29}$$

As $\operatorname{Tr}(\rho) = 1$ and $\operatorname{Tr}\Big([V_k, H_+]^{\otimes^2}\Big) = \operatorname{Tr}([V_k, H_+])^2 = 0$.

To bound $\operatorname{Tr}\Big([V_k, H_+]^2\Big)$ we use the triangle inequality and then Cauchy-Schwarz to show that

$$\big|\operatorname{Tr}\big([V_k, H_+]^2\big)\big| = \Big| 2 \operatorname{Tr}\big((V_k H_+)^2\big) - 2 \operatorname{Tr}\Big(\underbrace{V_k^2}_{\mathbb{I}} H_+^2 \Big) \Big| \tag{30}$$

$$\leqslant 2\big|\operatorname{Tr}\big((V_k H_+)^2\big)\big| + 2 \underbrace{\big|\operatorname{Tr}\big(H_+^2\big)\big|}_{\|H\|_2^2} \tag{31}$$

$$\leqslant 2||V_k||_2 ||V_k H_+ V_k||_2 + 2\|H\|_2^2 \tag{32}$$

$$= 2 \operatorname{Tr}\Big[V_k H_+ V_k V_k^\dagger H_+ V_k^\dagger \Big] + 2\|H\|_2^2 \tag{33}$$

$$= 2 \operatorname{Tr}\big[H_+^2\big] + 2\|H\|_2^2 \tag{34}$$

$$= 4\|H\|_2^2 \tag{35}$$

Assuming that $\|H\|_2^2 \in \mathcal{O}(d)$, it follows that $\mathrm{Var}(\partial_{\theta_k} E) \in \mathcal{O}\left(\frac{1}{d}\right)$.

Assuming U_+ is a 2-design, results in:

$$\mathrm{Var}(\partial_{\theta_k} E) = -\int \mathrm{d}U_- p(U_-) \int \mathrm{d}\mu(U_+) \,\mathrm{Tr}\Big(U_+^{\otimes^2} [\rho_-, V_k]^{\otimes^2} {U_+}^{\dagger \otimes^2} H^{\otimes^2}\Big) \tag{36}$$

$$= -\frac{1}{d^2-1}\left(\mathrm{Tr}\big(H^2\big) - \frac{\mathrm{Tr}(H)^2}{d}\right) \Big\langle \mathrm{Tr}\Big([\rho_u, V_k]^2\Big)\Big\rangle_{U_-} \tag{37}$$

For simplicity let's assume that H is expressed as a sum of Pauli operators and so $\mathrm{Tr}(H) = 0$. Then using Cauchy-Schwarz (41) we get:

$$\Big|\mathrm{Tr}\Big([\rho_-, V_k]^2\Big)\Big| \leqslant 4\,\mathrm{Tr}\big(\rho_-^2\big) = 4 \tag{38}$$

Therefore, $\mathrm{Var}(\partial_{\theta_k} E) \in \mathcal{O}\left(\frac{1}{d}\right)$.

Therefore the probability that the cost gradient deviates from 0 is suppressed exponentially in the number of qubits.

2.3 Useful formulas

Let $\mathcal{U}(d)$ denote the unitary group of degree $d = 2^n$. Let $d\mu(U)$ be the volume element of the Haar measure, where $U \in \mathcal{U}(d)$. The following identities hold.

$$\int \mathrm{d}\mu(U)\,\mathrm{Tr}\big(UAU^\dagger B\big) = \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}{d} \tag{39}$$

$$\int \mathrm{d}\mu(U)\,\mathrm{Tr}\Big(AU^{\otimes^2} BU^{\dagger\otimes^2}\Big) = \frac{\mathrm{Tr}(A)\,\mathrm{Tr}(B) + \mathrm{Tr}(AW)\,\mathrm{Tr}(BW)}{d^2-1} - \frac{\mathrm{Tr}(AW)\,\mathrm{Tr}(B) + \mathrm{Tr}(A)\,\mathrm{Tr}(BW)}{d(d^2-1)} \tag{40}$$

where W is the swap operator $W\,|i\rangle\,|j\rangle = |j\rangle\,|i\rangle$.

Note that $\mathrm{Tr}(A \bigotimes C)W) = \mathrm{Tr}(AC)$ (prove this to yourself!) and therefore $\mathrm{Tr}\Big(A^{\otimes^2} W\Big) = \mathrm{Tr}\big(A^2\big)$.

The Cauchy-Schwarz inequality in trace form is given by:

$$\big|\mathrm{Tr}\big(A^\dagger B\big)\big| \leqslant ||A||_2 ||B||_2 = \sqrt{\mathrm{Tr}(A^\dagger A)}\sqrt{\mathrm{Tr}(B^\dagger B)} \tag{41}$$