BP: original paper, local-cost and noise-induced

Ranyiliu Chen

September 26, 2020

This note includes a detailed derivation of the results in [MBS+18], and a brief argument of some conclusions in [CSV+20] and [WFC+20]. Before everything, let me first introduce two useful lemma about unitary Haar measure.

1 Lemmas for uniformly random unitary

Suppose that U forms Haar measure (is uniformly random), we might sometimes want to compute the expectation (average) of some function $f(U)$ of U. Luckily, following lemmas are quite helpful:

Lemma 1: The average of $f(U) = U^\dagger AU$ is

$$\begin{aligned}\mathbb{E}_H(U^\dagger AU) =& \int d\mu(U) U^\dagger AU \\ =& \frac{\mathrm{Tr}A}{d} I\end{aligned}$$

Lemma 2: The average of $f(U) = U^\dagger AUBU^\dagger CU$ is

$$\begin{aligned}\mathbb{E}_H(U^\dagger AUBU^\dagger CU) =& \int d\mu(U) U^\dagger AUBU^\dagger CU \\ =& \left(\frac{\mathrm{Tr}A\mathrm{Tr}C}{d^2-1} - \frac{\mathrm{Tr}(AC)}{d(d^2-1)}\right) B + \mathrm{Tr}B \left(\frac{\mathrm{Tr}(AC)}{d^2-1} - \frac{\mathrm{Tr}A\mathrm{Tr}C}{d(d^2-1)}\right) I\end{aligned}$$

2 Original work in [MBS+18]

Ref. [MBS+18] argues that the gradients of all parameters in the QNN will vanish exponentially in the number of qubits, while their elegant formula uses some approximation. Let us show a concrete evaluation.

The cost function is typically of the form of

$$C = \mathrm{Tr}(H\rho_{\mathrm{out}}) = \mathrm{Tr}(U^\dagger HU\rho)$$

Then the partial derivative of parameter θ_k is

$$\frac{\partial C}{\partial \theta_k} = i\mathrm{Tr}(U_-^\dagger [V, U_+^\dagger HU_+] U_- \rho) = i\mathrm{Tr}([U_-\rho U_-^\dagger, V] U_+^\dagger HU_+) \tag{1}$$

Denote commutator $[V, U_+^\dagger HU_+]$ by T_+ and $[U_-\rho U_-^\dagger, V]$ by T_-. Then compute the square of derivative as follows.

$$\begin{aligned}\left(\frac{\partial C}{\partial \theta_k}\right)^2 &= -\sum_{\alpha,\beta} \mathrm{Tr}\left(U_-^\dagger T_+ U_- \rho|\alpha\rangle\langle\beta|U_-^\dagger T_+ U_-\rho|\beta\rangle\langle\alpha|\right) \\ &= -\sum_{\alpha,\beta} \mathrm{Tr}\left(T_- U_+^\dagger H U_+|\alpha\rangle\langle\beta|T_- U_+^\dagger H U_+|\beta\rangle\langle\alpha|\right)\end{aligned} \tag{2}$$

Now we can compute $\mathbb{E}_{U_-}$ and $\mathbb{E}_{U_+}$ symmetrically, if we assume that when the parameters are randomly chosen, then U_- and U_+ will form a 2-design. Apply Lemma 1 to Eq. 1:

$$\mathbb{E}_{U_-}(\frac{\partial C}{\partial \theta_k}) = i\mathrm{Tr}(\frac{\mathrm{Tr}T_+}{d}\rho) = 0,\ \mathbb{E}_{U_+}(\frac{\partial C}{\partial \theta_k}) = i\mathrm{Tr}(\frac{\mathrm{Tr}H}{d}T_-) = 0$$

Then, apply Lemma 2 to Eq. 2. We compute the average over U_- and U_+, respectively.

$$\begin{aligned}\mathbb{E}_{U_-}(\left(\frac{\partial C}{\partial \theta_k}\right)^2) &= -\sum_{\alpha,\beta}\left(-\frac{\mathrm{Tr}T_+^2}{d(d^2-1)}\langle\alpha|\rho|\alpha\rangle\langle\beta|\rho|\beta\rangle + \frac{\mathrm{Tr}T_+^2}{d^2-1}\langle\beta|\rho|\alpha\rangle\langle\alpha|\rho|\beta\rangle\right) \\ &= -\frac{\mathrm{Tr}T_+^2}{d^2-1}\left(\mathrm{Tr}\rho^2 - \frac{1}{d}\right)\end{aligned} \tag{3}$$

$$\begin{aligned}\mathbb{E}_{U_+}(\left(\frac{\partial C}{\partial \theta_k}\right)^2) &= -\left(\frac{\mathrm{Tr}H^2}{d^2-1}\mathrm{Tr}T_-^2 - \frac{\mathrm{Tr}H^2}{d(d^2-1)}\mathrm{Tr}T_-^2\right) \\ &= -\frac{\mathrm{Tr}T_-^2}{d^2-1}\left(\mathrm{Tr}H^2 - \frac{1}{d}\mathrm{Tr}^2 H\right)\end{aligned} \tag{4}$$

If we continue to compute $\mathbb{E}_{U_-}(\mathrm{Tr}T_-^2)$ and $\mathbb{E}_{U_+}(\mathrm{Tr}T_+^2)$, Eq. 3 and Eq. 4 should yield the same result. Note that

$$\mathrm{Tr}T_+^2 = 2\mathrm{Tr}\left(VU_+^\dagger H U_+ V U_+^\dagger H U_+ - V^2 U_+^\dagger H^2 U_+\right)$$

$$\mathrm{Tr}T_-^2 = 2\mathrm{Tr}\left(U_-\rho U_-^\dagger V U_-\rho U_-^\dagger V - U_-\rho^2 U_-^+ V^2\right)$$

By using Lemma 1 & 2 we have

$$\mathbb{E}_{U_-}(\mathrm{Tr}T_-^2) = \frac{2}{d^2-1}\left(\mathrm{Tr}H^2 - \frac{\mathrm{Tr}^2 H}{d}\right)(\mathrm{Tr}^2 V - d\mathrm{Tr}V^2)$$

$$\mathbb{E}_{U_+}(\mathrm{Tr}T_+^2) = \frac{2}{d^2-1}\left(\mathrm{Tr}\rho^2 - \frac{1}{d}\right)(\mathrm{Tr}^2 V - d\mathrm{Tr}V^2)$$

Then both Eq. 3 and Eq. 4 yield that the average over the whole circuit

$$\mathbb{E}_U(\left(\frac{\partial C}{\partial \theta_k}\right)^2) = -\frac{2}{(d^2-1)^2}\left(\mathrm{Tr}H^2 - \frac{\mathrm{Tr}^2 H}{d}\right)\left(\mathrm{Tr}\rho^2 - \frac{1}{d}\right)(\mathrm{Tr}^2 V - d\mathrm{Tr}V^2) \tag{5}$$

It seems that the conclusion of [MBS+18]:

$$\mathrm{Var}\left[\partial_k E\right] = \begin{cases} -\frac{\mathrm{Tr}(\rho^2)}{d^2} < \mathrm{Tr}\, T_+^2 >_{U_+} \\ -\frac{\mathrm{Tr}(H^2)}{d^2} < \mathrm{Tr}\, T_-^2 >_{U_-}^2 \\ -\frac{2}{d^4}\,\mathrm{Tr}\left(H^2\right)\mathrm{Tr}\left(\rho^2\right)(\mathrm{Tr}^2 V - d\mathrm{Tr}V^2) \end{cases}$$

is an approximation of Eq. 3 4 5 for a large d. Rewrite Eq. 3 4 5 as:

$$\mathrm{Var}\left[\partial_k E\right] = \begin{cases} -\frac{1}{d^2-1}\left(\mathrm{Tr}\rho^2 - \frac{1}{d}\right) < \mathrm{Tr}T_+^2 >_{U_+} \\ -\frac{1}{d^2-1}\left(\mathrm{Tr}H^2 - \frac{1}{d}\mathrm{Tr}^2 H\right) < \mathrm{Tr}T_-^2 >_{U_-} \\ -\frac{2}{(d^2-1)^2}\left(\mathrm{Tr}H^2 - \frac{\mathrm{Tr}^2 H}{d}\right)\left(\mathrm{Tr}\rho^2 - \frac{1}{d}\right)\left(\mathrm{Tr}^2 V - d\mathrm{Tr}V^2\right) \end{cases}$$

In fact, we can make one step further. Typically $V = V_i^j \otimes I_{\bar{j}}$ and V_i^j is a Pauli matrix, then $\mathrm{Tr}V = \mathrm{Tr}(V_i^j)\mathrm{Tr}(I_{\bar{j}}) = 0$ and $\mathrm{Tr}V^2 = \mathrm{Tr}(V_i^j)^2\mathrm{Tr}(I_{\bar{j}}^2) = \mathrm{Tr}(I_d) = d$. In this sense,

$$\mathbb{E}_U(\left(\frac{\partial C}{\partial \theta_k}\right)^2) >= \frac{2d^2}{(d^2-1)^2}\left(\mathrm{Tr}H^2 - \frac{\mathrm{Tr}^2 H}{d}\right)\left(\mathrm{Tr}\rho^2 - \frac{1}{d}\right) \tag{6}$$

3 How is a local cost function different from a global one?

We can quantify the 'locality' by separating the n-qubits target state into m blocks each of which contains n/m qubits. Then the Hamiltonian H of the cost function can be written as:

$$H = \frac{1}{m}\sum_{i=1}^{m} |\phi_i\rangle\langle\phi_i| \otimes I_{2^{n-n/m}}$$

where $|\phi_i$ is the $i-$th block of the target state. Obviously the case $m = 1$ corresponds to the global cost and the case $m = n$ corresponds to the local cost.

Let $F(H) = \mathrm{Tr}H^2 - \frac{\mathrm{Tr}^2 H}{d}$. We can compute $F(H)$ and explore the cases $m = 1$ and $m = n$, respectively:

$$F(H) = \frac{1}{m}(2^{n-n/m} - 2^{n-2n/m}) = \begin{cases} \frac{1}{n}2^{n-2} & m = n \\ 1 - \frac{1}{2^n} & m = 1 \end{cases}$$

Obviously, the local cost performances better than the local cost for a large n, although the variance of cost function is still exponentially small in n.

4 What does Ref [CSV$^+$20] say?

In the *Alternating Layered Ansatz* of Ref [CSV$^+$20], each block W contains m qubits and is assumed to form 2-design. The gradient can be written as

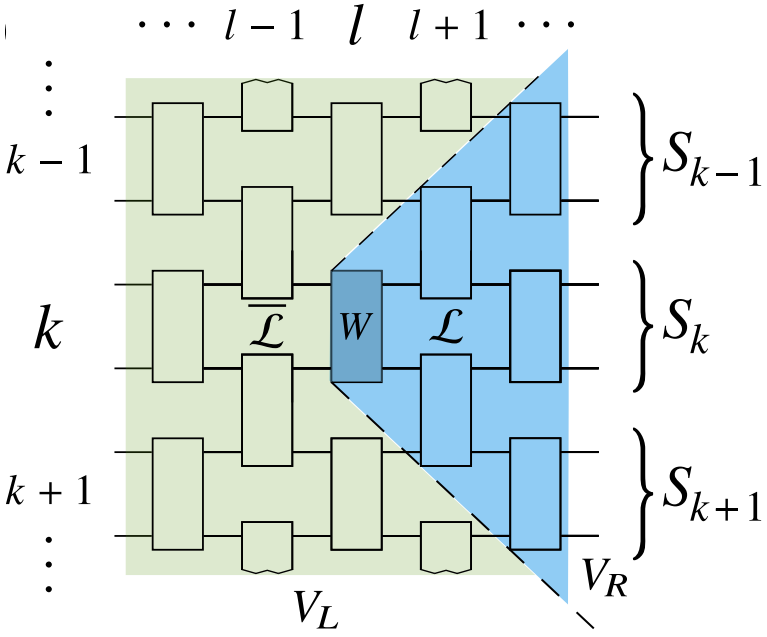
$$V(\boldsymbol{\theta}) = V_L\left(\mathbb{I}_{\bar{w}} \otimes W_A W_B\right) V_R$$

$$\partial_\nu C = \frac{i}{2}\mathrm{Tr}\left[\left(\mathbb{I}_{\bar{w}} \otimes W_B\right) V_L \rho V_L^\dagger \left(\mathbb{I}_{\bar{w}} \otimes W_B^\dagger\right)\left[\mathbb{I}_{\bar{w}} \otimes \sigma_\nu, \left(\mathbb{I}_{\bar{w}} \otimes W_A^\dagger\right) V_R^\dagger O V_R \left(\mathbb{I}_{\bar{w}} \otimes W_A\right)\right]\right]$$

After averaging over W_A and W_B the variance can be written as

$$\left\langle (\partial_\nu C)^2 \right\rangle_V = \frac{2^{m-1}\,\mathrm{Tr}\left[\sigma_\nu^2\right]}{\left(2^{2m}-1\right)^2} \sum_{pqp'q'} \left\langle \Delta\Omega_{pq}^{p'q'} \right\rangle_{V_R} \left\langle \Delta\Psi_{pq}^{p'q'} \right\rangle_{V_L}$$

We note that the case $m = n$ corresponds the model in ref. [MBS$^+$18], and m actually limits the power of the circuit (consider the extreme case when $m = 1$, there is no connection between each qubit).



Ref. [CSV+20] at first yields a polynomially vanishing in the single-layer case, which coincides the model in ref. [MBS+18]. The result of gradient computation is

$$\mathrm{Var}\left[\partial_\nu C_L\right] = \frac{m2^{3(m-1)}}{n^2\left(2^{2m}-1\right)^2}\left(\mathrm{Tr}\left[\rho_h^2\right] - \frac{1}{2^m}\right)$$

This result can yield a polynomially vanishing simply because that, some of the n's in Eq. 6 are replaced by m which is regarded as a constant. If we replace all the m's by n, the above equation perfectly matches Eq. 5.

Then, Ref. [CSV+20] considers the general case where V_L and V_R must be considered as random variables. In this case, they do not assume that V_L and V_R form 2-design. Instead, they introduce a technique called Random Tensor Network Integrator (RTNI) [FKN19] which compute the average over Haar-random tensor networks. This is also the reason that the number of layers L can appear in the lower bound:

$$G_n(L,l) = \frac{2^{m(l+1)-1}}{\left(2^{2m}-1\right)^2\left(2^m+1\right)^{L+l}} \sum_{i\in i_{\mathcal{L}}(k,k')\in k_{\mathcal{L}_B}} c_i^2 \epsilon\left(\rho_{k,k'}\right)\epsilon\left(\widehat{O}_i\right)$$

In this sense, L can be arbitrarily set without violating their assumption that each block is 2-design.

To sum up, the local cost indeed improves the behavior with the requirement of separability. (but the local cost alone is not enough), and parameter m improves the behavior a step further with the price of limitation of the circuit functionality.

5 Barren plateaus in Noise-induced model (outline)

Ref. [WFC+20] introduces how does the Pauli noise cause the gradient vanishment. Unlike the conclusion in previous literature, this one do not assume that the circuit form any type of random distribution, and compute the upper bound of the gradient directly.

There are two critical techniques used in this paper. Note that we can write any quantum state ρ and Hamiltonian H in the Pauli representation:

$$\rho = \boldsymbol{a}_\rho \cdot \boldsymbol{\sigma}_n, H = \boldsymbol{a}_H \cdot \boldsymbol{\sigma}_n$$

where $\boldsymbol{a}_\rho$ and $\boldsymbol{a}_H$ (called Pauli vector) are real vectors of length 4^n.

If we apply this representation in the cost function $C = \mathrm{Tr}(\rho_L O)$, then we get

$$\begin{aligned}|\partial_k C| =& |\mathrm{Tr}(\partial_k \rho_L O)| \\ =& |\mathrm{Tr}(\boldsymbol{g}_L \cdot \boldsymbol{\sigma}_n \ \boldsymbol{a}_O \cdot \boldsymbol{\sigma}_n)| \\ =& |\boldsymbol{g}_L \cdot \boldsymbol{a}_O| \\ \leq& N_O ||\boldsymbol{g}_L||_\infty \cdot ||\boldsymbol{a}_O||_\infty\end{aligned}$$

Now we introduce the second technique which connect the Pauli noise q with $||\boldsymbol{g}_L||_\infty$. Note that The Pauli noise

$$N : \rho \mapsto (1 - P_x - P_y - P_z)\rho + P_x X\rho X + P_y Y\rho Y + P_z Z\rho Z$$

where $P_x + P_y + P_z \leq 1$. And N maps X to $(1 - 2P_y - 2P_z)X = q_x X$. If we define q_y, q_z similarly, and let $q = \max\{|q_x|, |q_y|, |q_z|\}$, then after each noise layer the Pauli vector of the state

$$\left\|\boldsymbol{a}^{(k)}\right\|_2 \leqslant q \left\|\boldsymbol{a}^{(k-1)}\right\|_2 .$$

Recursively using this argument we get

$$\left\|\boldsymbol{a}^{(L)}\right\|_2 \leqslant q^{L-1} \left\|\boldsymbol{a}^{(1)}\right\|_2 .$$

Finally, we can prove that $||\boldsymbol{g}_L||_\infty \leq C \left\|\boldsymbol{a}^{(L)}\right\|_2$ with some constant C irrelevant with ρ. This finishes the proof that the gradient itself is bounded by the noise factor q^L which is exponentially small with the number of layer L.

References

[CSV+20] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost-function-dependent barren plateaus in shallow quantum neural networks, 2020, 2001.00550.

[FKN19] Motohisa Fukuda, Robert König, and Ion Nechita. RTNI—a symbolic integrator for haar-random tensor networks. *Journal of Physics A: Mathematical and Theoretical*, 52(42):425303, sep 2019. doi:10.1088/1751-8121/ab434b.

[MBS+18] 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), Nov 2018. doi:10.1038/s41467-018-07090-4.

[WFC+20] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J. Coles. Noise-induced barren plateaus in variational quantum algorithms, 2020, 2007.14384.