Analysis of Data-encoding Induced Barren Plateau in Quantum Machine Learning

Kensuke Kamisoyama

2024/01/24

Contents

Contents

Quantum Computers

Quantum Computers

Quantum circuits consist of the following elements:

- Quantum Bits (Qubits): Two-level quantum systems
- Quantum Gates: Unitaries that change the state of qubits
- Measurements: Extract information from the quantum state

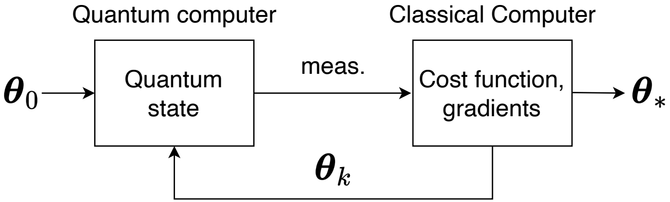
NISQ (Noisy Intermediate-Scale Quantum device) [Preskill2018]

- Quantum computers with a few hundred qubits
- Unable to perform error correction, so noise cannot be ignored
- Limited depth of quantum circuits

Variational Quantum Algorithms (VQA)

Variational Quantum Algorithms (VQA) [review:Cerezo2021]

- Quantum and classical hybrid algorithms feasible on NISQ devices
- $U(\boldsymbol{\theta})$: Parameterized quantum circuit (variational quantum circuit)
- $\ell_{i,k}(\boldsymbol{\theta}) = \mathrm{Tr}\left[U(\boldsymbol{\theta})\rho_i U^\dagger(\boldsymbol{\theta})O_k\right]$, where ρ_i is the initial state, O_k is the observable
- Optimization of cost function $\mathcal{L}(\boldsymbol{\theta}) = f(\{\ell_{i,k}(\boldsymbol{\theta})\}_{i,k})$
- Applications in quantum chemistry, combinatorial optimization, machine learning.



Quantum Machine Learning

Here, we refer to machine learning using variational quantum circuits as Quantum Machine Learning. In supervised quantum machine learning, an encoding circuit for the dataset $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_i$ is necessary.

Supervised Learning (Quantum Circuit Learning Model [Mitarai2018])

Trial function

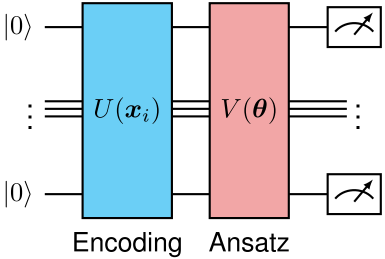
$$|\phi(\boldsymbol{x}_i, \boldsymbol{\theta})\rangle := V(\boldsymbol{\theta})U(\boldsymbol{x}_i)\,|0\rangle^{\otimes n}$$

Predicted label

$$\ell_i(\boldsymbol{\theta}) := \langle\phi(\boldsymbol{x}_i, \boldsymbol{\theta})|O|\phi(\boldsymbol{x}_i, \boldsymbol{\theta})\rangle$$

Cost function

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f(y_i, \ell_i(\boldsymbol{\theta}))$$



Barren Plateau (BP)

Barren Plateau [Mcclean2018]

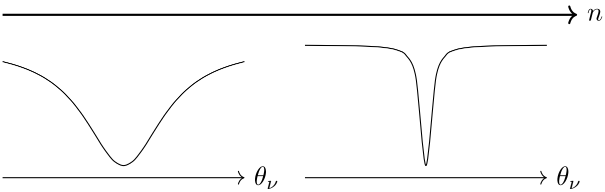
Definition

$\mathcal{L}(\boldsymbol{\theta})$: Cost function, $V(\boldsymbol{\theta})$: Ansatz, n: Number of qubits

$$\mathrm{E}_{V(\boldsymbol{\theta})}\left[\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_\nu}\right] = 0,\ \ \mathrm{Var}_{V(\boldsymbol{\theta})}\left[\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_\nu}\right] \in \mathcal{O}(2^{-\alpha n}),\ \ \alpha > 0$$

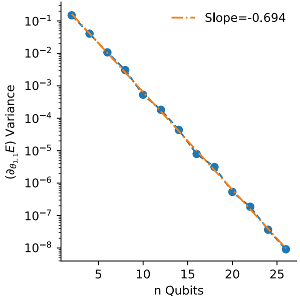
Gradient vanishing → Exponential complexity

Scaling of the variance of gradient is important



Causes

- Depth of the ansatz
- Locality of observables
- Noise
- Data encoding



Variance of cost function gradient

Contents

Research Overview

Research Background

- make machine learning efficient using variational quantum algorithms
- However, Barren Plateau (gradient vanishing) may arise
- Additionally, the effect of data encoding has not been investigated well

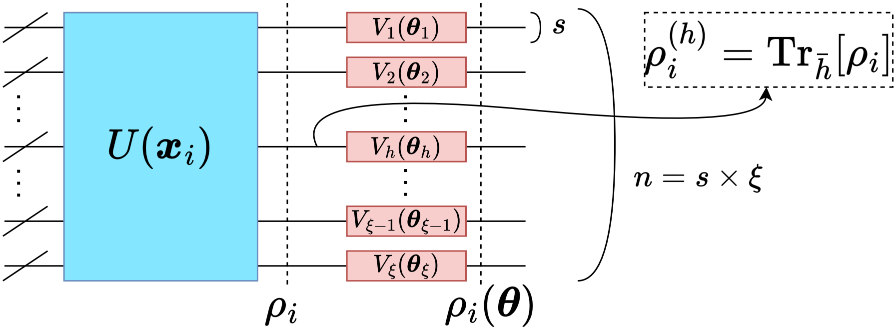
Research Goal and Approaches

Goal: To prevent the Barren Plateau due to data encoding
Approach:

1. Analyze the effect of data encoding on the variance of the cost function gradient. Specifically, we derive upper and lower bounds on the variance of gradient.
2. Numerically verify that the scaling of the variance of gradient is independent of the forms of the cost function.

Contents

Upper Bound on the Variance of the Gradient: Setup 11/27



$$\ell_i(\boldsymbol{\theta}) = \mathrm{Tr}[\rho_i(\boldsymbol{\theta})\, O_L] \in [0,1], \quad O_L = \frac{1}{n}\sum_{j=1}^{n} |0\rangle\langle 0|_j \otimes \mathbb{1}_{\bar{j}}$$

Set the ansatz and O_L so that Barren Plateau does not occur due to factors other than data encoding.

Upper Bound on the Variance of the Gradient: Results 12/27

$$y_i \in \{0,1\}, \quad \ell_i(\boldsymbol{\theta}) := \mathrm{Tr}[\rho_i(\boldsymbol{\theta})\, O_L] \in [0,1], \quad \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f(y_i, \ell_i(\boldsymbol{\theta}))$$

Based on prior research[Thanasilp2021], we derived a new upper bound.

Theorem

Upper bound on the variance of the cost function gradient is given as follows, where $\mathbb{U} := \{U(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{X}\}$:

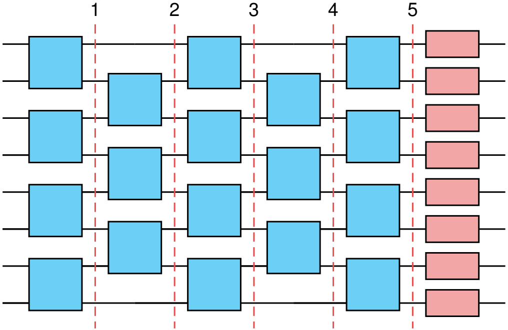
$$\mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu} \mathcal{L}(\boldsymbol{\theta})] \leq A_f \times r_{n,s} \times \overline{D}^s_{\mathrm{HS}} \leq A_f \times r_{n,s} \times \left(\frac{2^s - 2^{-s}}{2^n + 1} + \epsilon_{\mathbb{U}}\right)$$

- A_f is a term depending on the function f such as squared error.
- $r_{n,s}$ is a term depending on the observable being measured.
- $\overline{D}^s_{\mathrm{HS}} := \int_{\mathbb{U}} dU\, D_{\mathrm{HS}}(\rho^{(h)}, \mathbb{1}/2^s)$ is a term depending on the data encoding.
- $\epsilon_{\mathbb{U}}$ is a measure of the expressive power of the encoding circuit, and the higher the expressive power, the smaller this value is.

Assuming $\epsilon_{\mathbb{U}}$ does not go to 0, we analyzed the scaling of $\overline{D}^s_{\mathrm{HS}}$

Upper Bound on the Variance of the Gradient: Circuit 13/27

Assuming the following Alternating Layered Ansatz (ALT) for the encoding circuit $U(x)$, we exactly calculated $\overline{D}_{\mathrm{HS}}^{s=1}$ $\ast$

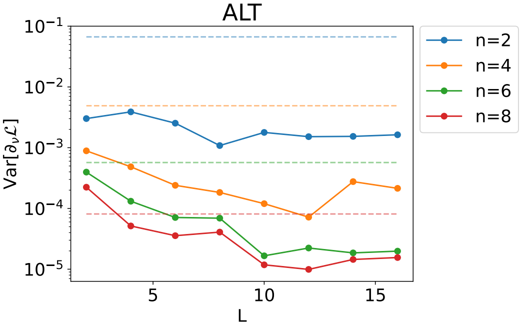


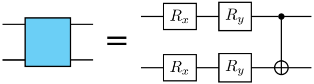
Blue represents the encoding circuit, and red represents the ansatz.

($\ast$ Each blue block is treated as unitary 2-design, and calculations were performed using the Random Tensor Network Integrator (RTNI) [Fukuda2019])

Variance of gradient and Its Upper Bound

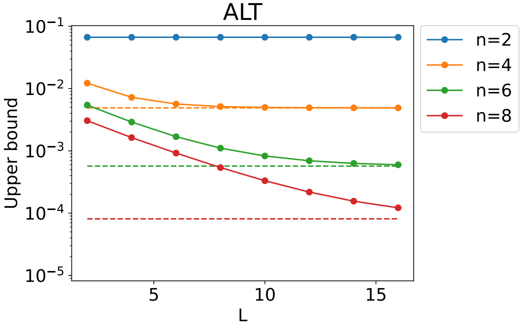
$$\mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu}\mathcal{L}(\boldsymbol{\theta})]$$

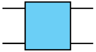




(Define the cost function using the Iris dataset)

Upper Bound $(A_f \times r_{n,s} \times \overline{D}_{\mathrm{HS}}^{s=1})$

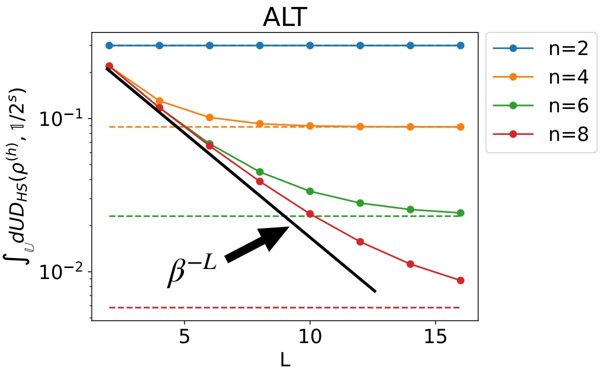



= Unitary 2-Design

Indeed, it is an upper bound.

Necessary Condition for Avoiding Barren Plateau

The number of encoding layers for the upper bound not to decay exponentially



($0 < \alpha$, n: number of qubits)

$\overline{D}_{\mathrm{HS}}^{s} \propto e^{-\alpha n}$

$\rightarrow$ Barren Plateau

Necessary condition for avoiding Barren Plateau:

$(0 < \gamma,\ 1 < \beta)$

$n^{-\gamma} \leq \beta^{-L} \leq \overline{D}_{\mathrm{HS}}^{s}$

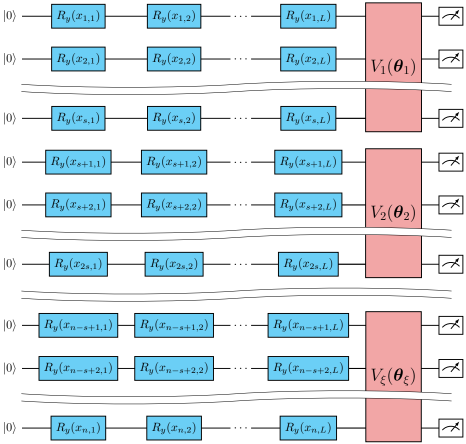
$\Longrightarrow\ L \leq \frac{\gamma}{\log \beta} \log n$

Contents

Lower Bound on the Variance of Gradient: Setup

(Simple encoding circuit for analysis)

- Quantum circuit:
  $n = s \times \xi$ qubits
  Encoding consists of R_y
- Cost function ($y_i \in \{0,1\}$):
  $\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} |\ell_i(\boldsymbol{\theta}) - y_i|$
- Input data:
  $\mathcal{X} = \{\boldsymbol{x}\}$ (label 0)
  $\mathcal{Z} = \{\boldsymbol{z}\}$ (label 1)
  $|\mathcal{X}| : |\mathcal{Z}| = p : q \ (p + q = 1)$
- Gaussian distribution:
  $x_{j,d} \sim \mathcal{N}(\mu_{x|j,d}, \sigma^2_{x|j,d})$
  $z_{j,d} \sim \mathcal{N}(\mu_{z|j,d}, \sigma^2_{z|j,d})$
- Variance: $\sigma_{x|j,d},\ \sigma_{z|j,d} \le \sigma_{\max}$



L layers of R_y gates encoding circuit and ξ s-qubit unitaries $V_\xi(\boldsymbol{\theta}_\xi)$ for the ansatz

Lower Bound on the Variance of Gradient: Results

Theorem

Lower bound on the variance of the cost function gradient is given as follows,

$$\frac{r_{n,s}}{2^s}\sum_{j=1}^{s}\left(p\,e^{-\Sigma_{x|j}/2}-q\,e^{-\Sigma_{z|j}/2}\right)^2 \le \mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_\nu \mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]$$

where $r_{n,s} := \frac{s\,2^{3(s-1)}}{n^2(2^{2s}-1)^2}$, $\Sigma_{x|j} = \sum_{d=1}^{L}\sigma^2_{x|j,d}$, $\Sigma_{z|j} = \sum_{d=1}^{L}\sigma^2_{z|j,d}$

For $|\mathcal{X}| : |\mathcal{Z}| = 1:1 \iff p = q = 1/2$, the lower bound is

$$\frac{r_{n,s}}{2^{s+2}}\sum_{j=1}^{s}\left(e^{-\Sigma_{x|j}/2}-e^{-\Sigma_{z|j}/2}\right)^2 \le \mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_\nu \mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]$$

- The difference between $e^{-\Sigma_{x|j}/2}$ and $e^{-\Sigma_{z|j}/2}$ is crucial in the lower bound
- However, $e^{-\Sigma_{x|j}/2}$ and $e^{-\Sigma_{z|j}/2}$ decay exponentially with #(encoding layers)
- If $e^{-\Sigma_{x|j}/2} = e^{-\Sigma_{z|j}/2}$ for all j, the lower bound becomes 0

Lower Bound on the Variance of Gradient: Results

Theorem

Lower bound on the variance of the cost function gradient is given as follows,

$$\frac{r_{n,s}}{2^s}\sum_{j=1}^{s}\left(p\,e^{-\Sigma_{x|j}/2}-q\,e^{-\Sigma_{z|j}/2}\right)^2\leq \mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_\nu\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]$$

where $r_{n,s}:=\frac{s\,2^{3(s-1)}}{n^2(2^{2s}-1)^2}$, $\Sigma_{x|j}=\sum_{d=1}^{L}\sigma_{x|j,d}^2$, $\Sigma_{z|j}=\sum_{d=1}^{L}\sigma_{z|j,d}^2$

For $|\mathcal{X}|:|\mathcal{Z}|=1:0 \iff p=1,\ q=0$, the lower bound is

$$\frac{r_{n,s}s}{2^s}\,e^{-L\sigma_{\max}^2}\leq\frac{r_{n,s}}{2^s}\sum_{j=1}^{s}e^{-\Sigma_{x|j}}\leq \mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]$$

- Large lower bound when $L\sigma_{\max}^2$ is small
- If $s, L\sigma_{\max}^2\in\mathcal{O}(\log n)$, the lower bound becomes $\mathcal{O}(1/\mathrm{poly}(n))$
  → Sufficient condition for avoiding barren plateaus

Contents

Form of Function f and Variance of Gradient

How does the function f affect the scaling of variance of gradient?

$$\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} |\ell_i(\boldsymbol{\theta}) - y_i|$$

$$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} (\ell_i(\boldsymbol{\theta}) - y_i)^2$$

$$\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}[-y_i \log \ell_i(\boldsymbol{\theta}) - (1-y_i)\log(1-\ell_i(\boldsymbol{\theta}))]$$

$$\Longrightarrow \partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \operatorname{sgn}(\ell_i(\boldsymbol{\theta}) - y_i)\cdot\partial_{\theta_\nu}\ell_i(\boldsymbol{\theta})$$

$$\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} 2|\ell_i(\boldsymbol{\theta}) - y_i| \operatorname{sgn}(\ell_i(\boldsymbol{\theta}) - y_i)\cdot\partial_{\theta_\nu}\ell_i(\boldsymbol{\theta})$$

$$\partial_{\theta_\nu}\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1-|\ell_i(\boldsymbol{\theta}) - y_i|}\operatorname{sgn}(\ell_i(\boldsymbol{\theta}) - y_i)\cdot\partial_{\theta_\nu}\ell_i(\boldsymbol{\theta})$$

where $y_i \in \{0,1\}, \quad \ell_i(\boldsymbol{\theta}) = \mathrm{Tr}[\rho_i(\boldsymbol{\theta})\, O_L] \in [0,1], \quad (O_L = \frac{1}{n}\sum_{j=1}^{n} |0\rangle\langle 0|_j \otimes \mathbb{1}_{\bar{j}})$

Form of Function f and Variance of Gradient

Assuming the ansatz is a unitary 2-design, the mean and variance of $\ell_i(\boldsymbol{\theta})$ are:

$$\mathrm{E}_{\mathcal{U}(d)}[\ell_i(\boldsymbol{\theta})] = \frac{1}{2}$$

$$\mathrm{Var}_{\mathcal{U}(d)}[\ell_i(\boldsymbol{\theta})] = \frac{1}{4n(2^n+1)}$$

Therefore, assuming $\ell_i(\boldsymbol{\theta}) \sim \frac{1}{2}$ for all $i \iff |\ell_i(\boldsymbol{\theta}) - y_i| \sim \frac{1}{2}$ we get:

$$\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta}) \sim \partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta}), \quad \partial_{\theta_\nu}\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta}) \sim 2\,\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})$$

Thus, the ratio of the variance of the gradients for mean squared error to mean absolute error, and cross-entropy error to mean absolute error, is approximately:

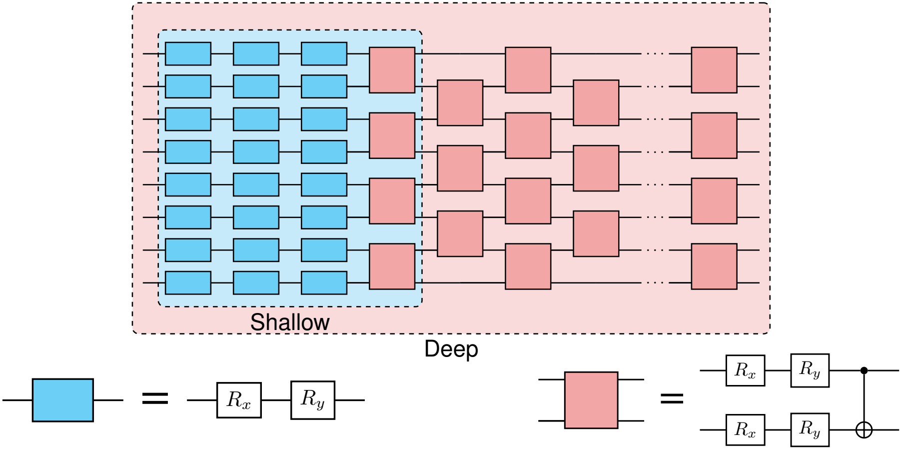
$$\Longrightarrow \quad \frac{\mathrm{Var}_{\mathcal{U}(d)}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta})]}{\mathrm{Var}_{\mathcal{U}(d)}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]} \sim 1$$

$$\frac{\mathrm{Var}_{\mathcal{U}(d)}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta})]}{\mathrm{Var}_{\mathcal{U}(d)}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]} \sim 4$$

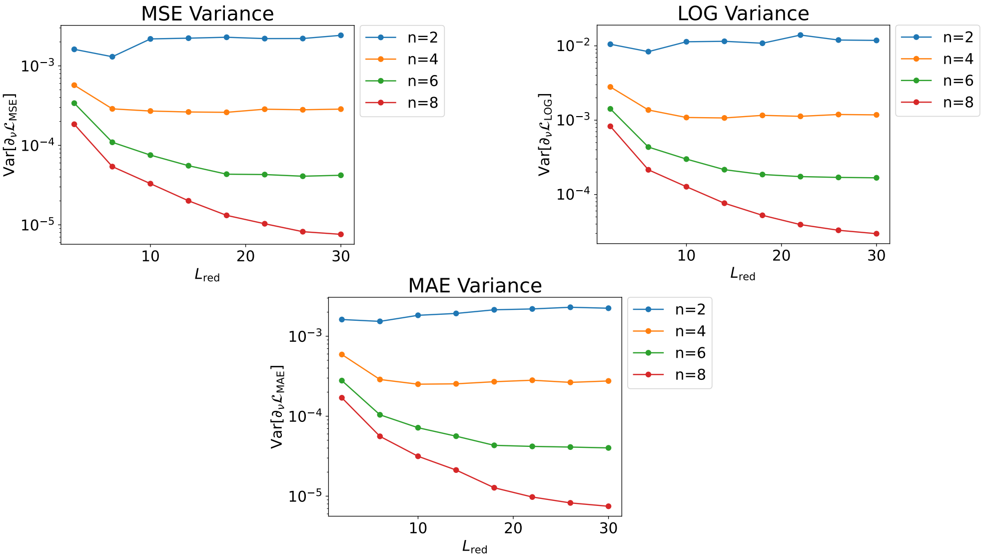
When the ansatz is deep, the scaling of the variance of gradient seems to be similar regardless of the function f.

Form of Function f and Variance of Gradient

Assuming ALT for the ansatz, we examine the ratio of variance of the cost function gradient when the number of layers L_{red} is varied. (Blue represents the encoding circuit, red represents the ansatz.)

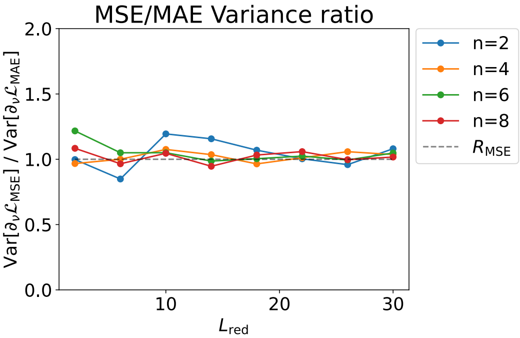


Form of Function f and Variance of Gradient



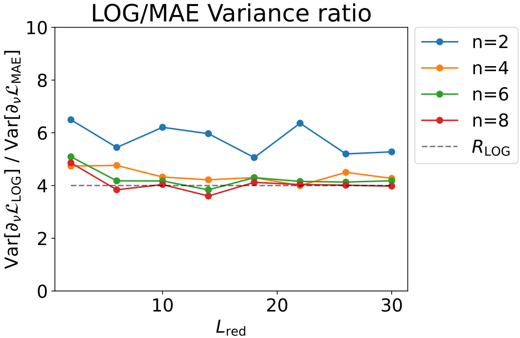
Form of Function f and Variance of Gradient

$\mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta})]\ /\ \mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]$



Dashed line is the approximation ratio $R_{\mathrm{MSE}} = 1$

$\mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta})]\ /\ \mathrm{Var}_{V(\boldsymbol{\theta})}[\partial_{\theta_\nu}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta})]$



Dashed line is the approximation ratio $R_{\mathrm{LOG}} = 4$

In numerical calculations, the scaling of the variance of gradient seems to be similar regardless of the function f, even for shallow circuits.

(Cost functions were defined using the Iris dataset.)

Contents

Summary and Future Work

Study on the Influence of Data Encoding on Barren Plateaus

- Derived and numerically validated an upper bound on variance of the gradient from the perspective of data encoding
- Increased entanglement in the state after data encoding and expressive power of the encoding circuit lead to a smaller upper bound, resulting in barren plateau.
- If #layers for data encoding is $O(\log n)$, the upper bound won't decay exponentially.
- For input data following a Gaussian distribution, the data variance is crucial for the lower bound on the variance of gradient
- Numerically confirmed that the scaling of the variance of gradient is almost independent of the form of the function f

Future Work

- Investigate the lower bound for more general encoding circuits.
- Consider encoding eircuits from the perspective of generalization performance.
- Examine whether the encoding circuit is classically hard to simulate.

Local observable

- $Z_1 := Z \otimes \mathbb{1} \otimes \mathbb{1} \otimes \cdots \otimes \mathbb{1}$
- $|0\rangle\langle 0|_1 = (Z_1 + \mathbb{1}_1)/2$
- $O_L := \frac{1}{n} \sum_{j=1}^{n} |0\rangle\langle 0|_j \otimes \mathbb{1}_{\bar{j}}$ (linear combination of local observables)
- If the depth of ALT ansatz is $\mathcal{O}(\log n)$, no barren plateau (without data encoding)

Global observable

- $Z^{\otimes n} := Z \otimes Z \otimes \cdots \otimes Z$
- The barren plateau occurs regardless of the depth of the ansatz.

Definition

Let $P_{t,t}(U)$ be a homogeneous polynomial of maximum degree t in the components of the unitary U and $U^\dagger$. A set of K unitaries $\{U_k\}$ is said to be a unitary t–design if it satisfies the following condition. (ex. $P_{2,2}(U) = U^\dagger AUBU^\dagger CU$)

$$\frac{1}{K}\sum_{k=1}^{K} P_t(U_k) = \int_{\mathcal{U}(d)} P_t(U) d\mu(U)$$

- Pauli group: $\mathcal{P}(n) = \{e^{i\frac{k\pi}{2}} P_{j_1} \otimes \cdots \otimes P_{j_n} | k, j_l = 0, 1, 2, 3\}$ is a unitary 1–design
- Clifford group: $\mathcal{C}(n) = \{U \in \mathcal{U}(2^n) | UPU^\dagger = \mathcal{P}(n)\}$ is a unitary 3–design
- If $\{U_k\}$ is a unitary t–design, it is also a unitary $(t-1)$–design.

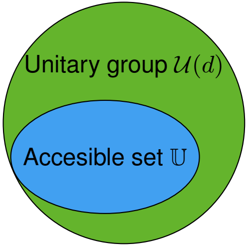
Avoiding Barren Plateaus

- (Without data encoding) It has been shown that the barren plateau does not occur if the depth of the ansatz is $\mathcal{O}(\log n)$ when using local observables.
- Optimize the parameters for each layer of the ansatz. This reduces the effective depth of the ansatz.
- Set the initial values of some of the parameters to cancel out the rest of the circuit. This reduces the effective depth of the ansatz.
- Introduce correlations between parameters. This reduces the expressibility of the ansatz.
- Sample the initial values of the parameters from a normal distribution instead of a uniform distribution.

The expressibility of a quantum circuit is defined as follows.

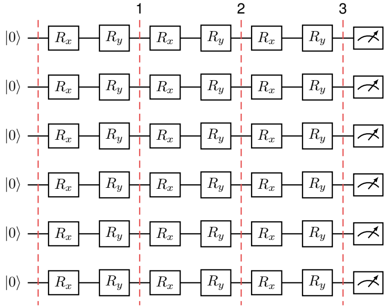
$$\epsilon_{\mathbb{U}}^{(t,p)}(X) := \left\| A_{\mathbb{U}}^{(t)}(X) \right\|_p,$$
$$\mathcal{A}_{\mathbb{U}}^{(t)}(X) := \int_{\mathcal{U}(d)} d\mu_{\mathrm{Haar}}(V)\, V^{\otimes t} X^{\otimes t} (V^{\otimes t})^{\dagger} - \int_{\mathbb{U}} dU\, U^{\otimes t} X^{\otimes t} (U^{\otimes t})^{\dagger}.$$

- $\|\cdot\|_p$: Schatten p–norm
- X : Initial state $|0\rangle\langle 0|^{\otimes n}$
- $\mathcal{U}(d)$: Unitary group of dimension d
- Consider the case of $t = 2$

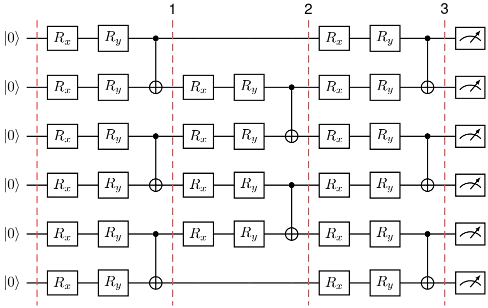


Expressible region of a quantum circuit

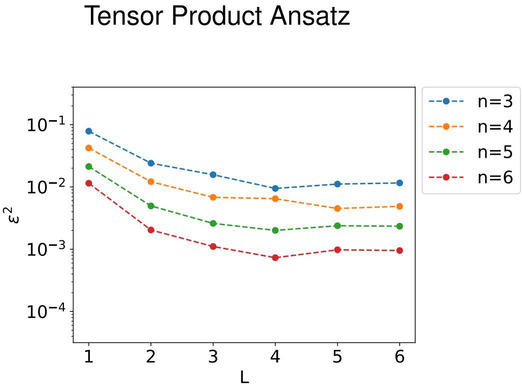
Tensor Product Ansatz

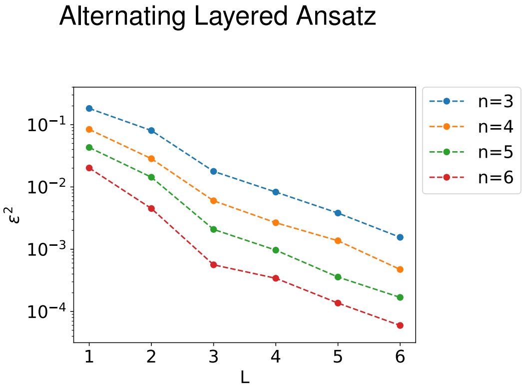


Alternating Layered Ansatz



Expressibility: Plot





$$y_i \in \{0,1\}, \quad \ell_i(\boldsymbol{\theta}) := \mathrm{Tr}[\rho_i(\boldsymbol{\theta})\, O_{\mathrm{L}}] \in [0,1], \quad \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} f(y_i, \ell_i(\boldsymbol{\theta}))$$

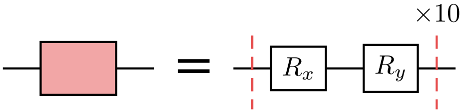
Theorem

Upper bound on the variance of the gradient

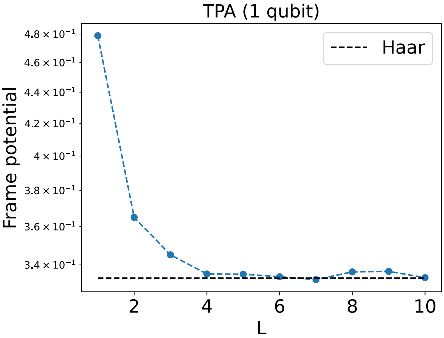
$$\begin{aligned}
&\mathrm{Var}_{\boldsymbol{\theta}}[\partial_{\theta_\nu} \mathcal{L}(\boldsymbol{\theta})] \\
&\leq A_f \times r_{n,s} \times \overline{D}_{\mathrm{HS}} \\
&= 2\max_{i,\boldsymbol{\theta}}[(\partial_{\ell_i(\boldsymbol{\theta})} f)^2] \times \frac{2^{2s-1}}{(2^{2s}-1)^2}\left(\mathrm{Tr}\left[(O_{\mathrm{L}}^h)^2\right] - \frac{\mathrm{Tr}\left[O_{\mathrm{L}}^h\right]^2}{2^s}\right) \times \int_{\mathbb{U}} dU\, D_{\mathrm{HS}}(\rho^{(h)}, \mathbb{I}/2^s)
\end{aligned}$$

, where $O_{\mathrm{L}}^h = \mathrm{Tr}_{\overline{h}}[O_{\mathrm{L}}]$

Unitary 2–design of 1-qubit



Structure used as the gate block of TPA



Frame potential of the Tensor Product Ansatz for 1-qubit. The black dashed line represents the frame potential $(= 1/3)$ when the 1-qubit forms a unitary 2–design.

As an extreme example, consider the case where input data belonging to label 0, $\mathcal{X} = \{\boldsymbol{x}\}$, and input data belonging to label 1, $\mathcal{Z} = \{\boldsymbol{z}\}$, are such that $\mathcal{X} = \mathcal{Z}$. In this case,

$$\begin{aligned}\mathcal{L}_{\mathrm{MAE}}(\boldsymbol{\theta}) &= \frac{1}{N}\Big(\sum_{x\in\mathcal{X}} |\ell_i(\boldsymbol{\theta}) - 0| + \sum_{z\in\mathcal{Z}} |\ell_i(\boldsymbol{\theta}) - 1|\Big) \\ &= \frac{1}{N}\Big(\sum_{x\in\mathcal{X}} \ell_i(\boldsymbol{\theta}) \qquad + \sum_{x\in\mathcal{X}} (1 - \ell_i(\boldsymbol{\theta}))\Big) = \frac{1}{2}\end{aligned}$$

Therefore, the gradient of the cost function becomes 0, and so does the variance.
→ Even if the structure of the encoding circuit is extremely simple, the closer the input data between different labels are, the smaller the variance of the gradient becomes.

As an extreme example, consider the case where input data belonging to label 0, $\mathcal{X} = \{\boldsymbol{x}\}$, and input data belonging to label 1, $\mathcal{Z} = \{\boldsymbol{z}\}$, are such that $\mathcal{X} = \mathcal{Z}$. In this case,

$$\begin{aligned}
\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta}) &= \frac{1}{N}\Big(\sum_{x\in\mathcal{X}}(\ell_i - 0)^2 \qquad + \sum_{x\in\mathcal{X}}(\ell_i - 1)^2\Big) \\
&= \frac{1}{N}\Big(\sum_{x\in\mathcal{X}}\ell_i^2 \qquad + \sum_{x\in\mathcal{X}}(\ell_i - 1)^2\Big) \\
&= \frac{1}{2} + \frac{1}{N}\sum_{x\in\mathcal{X}} 2\left(\ell_i - \frac{1}{2}\right)^2
\end{aligned}$$

In this case, since the cost function $\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta})$ is minimized when ℓ_i is $\frac{1}{2}$, it never approaches the ground truth labels $y_i = \{0, 1\}$.

Lower Bound: Discussion on $\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta})$

As an extreme example, consider the case where input data belonging to label 0, $\mathcal{X} = \{\boldsymbol{x}\}$, and input data belonging to label 1, $\mathcal{Z} = \{\boldsymbol{z}\}$, are such that $\mathcal{X} = \mathcal{Z}$. In this case,

$$
\begin{aligned}
\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta}) &= \frac{1}{N}\left(\sum_{x\in\mathcal{X}} -\log(1-\ell_i) + \sum_{x\in\mathcal{X}} -\log(\ell_i)\right) \\
&= \frac{1}{N}\left(\sum_{x\in\mathcal{X}} -\log \ell_i(1-\ell_i)\right) \\
&= \frac{1}{N}\left(\sum_{x\in\mathcal{X}} -\log\left[-\left(\ell_i - \frac{1}{2}\right)^2 + \frac{1}{4}\right]\right)
\end{aligned}
$$

In this case, since the cost function $\mathcal{L}_{\mathrm{LOG}}(\boldsymbol{\theta})$ is minimized when ℓ_i is $\frac{1}{2}$, it never approaches the ground truth labels $y_i = \{0, 1\}$.