Quantum neural networks as Gaussian processes

Giacomo De Palma
Filippo Girardi

arXiv:2402.08726

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Supervised learning

Goal: classify unlabeled data (e.g., handwritten digits)

Input encoded in vector $x \in \mathbb{R}^a$

For simplicity we assume that the set of the possible inputs is finite, but all results generalize to any compact input set

Classifier: parametric family of functions $\{F_\Theta(x) : \Theta \in \mathbb{R}^b\}$

Binary classification: F_Θ takes real values and label is sign $F_\Theta(x)$

Training data: labeled inputs (X_α, Y_α)

Quality of F_Θ on training data quantified by cost function

We choose square loss $$C(\Theta) = \frac{1}{2}\sum_\alpha \left(F_\Theta(X_\alpha) - Y_\alpha\right)^2$$

Parameters initialized by sampling from iid distribution and trained with (stochastic) gradient descent

Quantum neural networks

Quantum circuits made by **parametric** one- and two-qubit gates (we will assume only one-qubit gates are parametric)

Parameters encode components of Θ and x as evolution times of single-qubit Hamiltonians

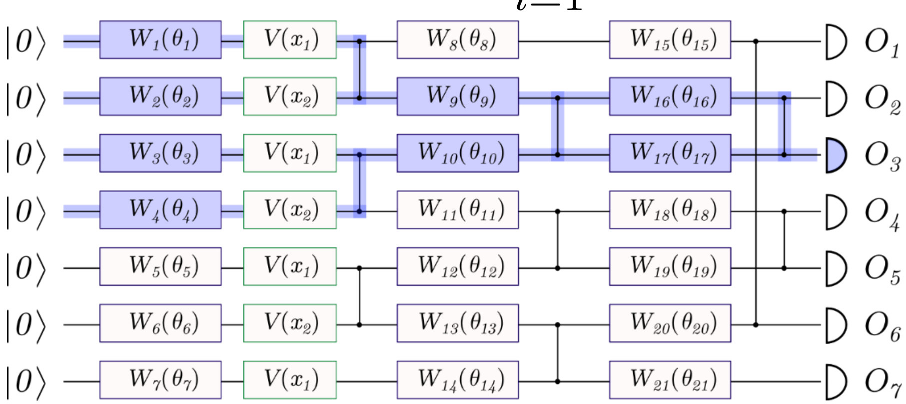
$F_\Theta(x)$ is expectation value of global observable H measured on output state; periodic in each component of x and Θ

Each component of Θ is randomly initialized from uniform distribution

We choose $$H = \frac{1}{N}\sum_{i=1}^{n} Z_i$$

n = #qubits

N normalization factor



Open problems

- Does the empirical risk converge to zero with the training? Possible issues:
  - Limited expressivity
  - Bad local minima

    Anschuetz, Kiani, “Quantum variational algorithms are swamped with traps”, Nat Commun 13, 7760 (2022)
  - Barren plateaus: Gradients of the cost function decay exponentially with # of layers

    Napp, “Quantifying the barren plateau phenomenon for a model of unstructured variational ansätze”, arXiv:2203.06174 (+ many others)
- Does the trained network have good generalization performances, i.e., good performances on inputs that are not part of the training examples? Possible issues:
  - Overfitting (too many parameters)
- Are quantum neural networks better than classical neural networks?

    Cerezo, Larocca, García-Martín, Diaz, Braccia, Fontana, Rudolph, Bermejo, Ijaz, Thanasilp, Anschuetz, Holmes, “Does provable absence of barren plateaus imply classical simulability? Or, why we need to rethink variational quantum computing”, arXiv:2312.09121

The limit of infinite width

Hint from classical deep learning: limit of infinite width is smooth and analytically solvable

Training in the limit **@ constant depth** considered in [Abedi, Beigi, Taghavi, "Quantum Lazy Training", Quantum 7, 989 (2023)]

Key observation: $\langle Z_i \rangle$ depends only on past light-cone of measured qubit i

For constant depth, each $\langle Z_i \rangle$ can be classically computed simulating only $O(1)$ qubits in $O(1)$ time!

We allow polylogarithmic light-cones keeping the depth logarithmic to avoid barren plateaus

A toy model for the infinite-width limit

Assume light-cones are all equal and do not share parameters. Let θ_i be the vector of the parameters in the past light-cone of qubit i

$$F_\Theta(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_{\theta_i}(x) \qquad f_{\theta_i}(x) = \langle Z_i \rangle$$

By central limit theorem, $F_\Theta(x)$ tends to Gaussian process (for any $x_1, \ldots, x_k$, joint probability distribution of $(F_\Theta(x_1), \ldots, F_\Theta(x_k))$ is Gaussian)

Parameters randomly initialized with iid sampling and

$$\mathbb{E}_\theta f_\theta(x) = 0$$

Covariance at initialization

$$\mathbb{E}_\Theta \left(F_\Theta(x)\, F_\Theta(x') \right) = \mathbb{E}_\theta \left(f_\theta(x)\, f_\theta(x') \right) = K_0(x, x')$$

A toy model for the infinite-width limit

Gradient flow: lazy training!

$$\dot{\theta}_i = -\nabla_{\theta_i} C(\Theta) = \frac{1}{\sqrt{n}} \sum_\alpha \left(Y_\alpha - F_\Theta(X_\alpha)\right) \nabla_{\theta_i} f_{\theta_i}(X_\alpha) = O\left(\frac{1}{\sqrt{n}}\right)$$

Finite variation of the generated function

$$\frac{d}{dt} F_\Theta(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\theta}_i \cdot \nabla_{\theta_i} f_{\theta_i}(x) = \sum_\alpha K_\Theta^{\tan}(x, X_\alpha) \left(Y_\alpha - F_\Theta(X_\alpha)\right) = O(1)$$

Empirical neural tangent kernel

$$\begin{aligned} K_\Theta^{\tan}(x, x') &= \nabla_\Theta F_\Theta(x) \cdot \nabla_\Theta F_\Theta(x') \\ &= \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta_i} f_{\theta_i}(x) \cdot \nabla_{\theta_i} f_{\theta_i}(x') = O(1) \end{aligned}$$

A toy model for the infinite-width limit

For $n \to \infty$, the empirical NTK at initialization tends to its expectation

$$K_{\tan}(x, x') = \mathbb{E}_\Theta \left(\nabla_\Theta F_\Theta(x) \cdot \nabla_\Theta F_\Theta(x')\right)$$
$$= \mathbb{E}_\theta \left(\nabla_\theta f_\theta(x) \cdot \nabla_\theta f_\theta(x')\right)$$

The training is lazy and can change the empirical NTK only by $O(1/\sqrt{n})$

The model becomes linear and analytically solvable

$$\frac{d}{dt} F_t^{\mathrm{lin}}(x) = \sum_\alpha K_{\tan}(x, X_\alpha) \left(Y_\alpha - F_t^{\mathrm{lin}}(X_\alpha)\right)$$

$$F_t^{\mathrm{lin}}(x) = F_0(x) - K_{\tan}(x, X)^T K_{\tan}^{-1} \left(I - e^{-tK_{\tan}}\right) (F_0 - Y)$$

$$(K_{\tan})_{\alpha\beta} = K_{\tan}(X_\alpha, X_\beta) \qquad K_{\tan}(x, X)_\alpha = K_{\tan}(x, X_\alpha)$$

$$(F_0)_\alpha = F_0(X_\alpha)$$

A toy model for the infinite-width limit

For $n \to \infty$, $F_t^{\mathrm{lin}}(x)$ converges in distribution to the Gaussian process with mean and covariance

$$\mu_t(x) = \mathbb{E}\, F_t^{\mathrm{lin}}(x) = K_{\mathrm{tan}}(x, X)^T\, K_{\mathrm{tan}}^{-1} \left(I - e^{-tK_{\mathrm{tan}}}\right) Y$$

$$K_t(x, x') = \mathrm{Cov}\left(F_t^{\mathrm{lin}}(x), F_t^{\mathrm{lin}}(x')\right)$$

$$= K_0(x, x') - K_{\mathrm{tan}}(x, X)^T\, K_{\mathrm{tan}}^{-1} \left(I - e^{-tK_{\mathrm{tan}}}\right) K_0(X, x')$$

$$- K_0(x, X)^T \left(I - e^{-tK_{\mathrm{tan}}}\right) K_{\mathrm{tan}}^{-1}\, K_{\mathrm{tan}}(X, x')$$

$$+ K_{\mathrm{tan}}(x, X)^T\, K_{\mathrm{tan}}^{-1} \left(I - e^{-tK_{\mathrm{tan}}}\right) K_0 \left(I - e^{-tK_{\mathrm{tan}}}\right) K_{\mathrm{tan}}^{-1}\, K_{\mathrm{tan}}(X, x')$$

$$(K_0)_{\alpha\beta} = K_0(X_\alpha, X_\beta) \qquad K_0(x, X)_\alpha = K_0(x, X_\alpha)$$

A toy model for the infinite-width limit

Limit $t \to \infty$: Gaussian process perfectly fits the examples.
Mean and covariance

$$\mu_\infty(x) = K_{\tan}(x, X)^T \, K_{\tan}^{-1} \, Y$$

$$\begin{aligned} K_\infty(x, x') = K_0(x, x') - K_{\tan}(x, X)^T \, K_{\tan}^{-1} \, K_0(X, x') \\ - K_0(x, X)^T \, K_{\tan}^{-1} \, K_{\tan}(X, x') \\ + K_{\tan}(x, X)^T \, K_{\tan}^{-1} \, K_0 \, K_{\tan}^{-1} \, K_{\tan}(X, x') \end{aligned}$$

Assumptions

We consider a sequence of QNNs with increasing n trained on a fixed training set with gradient descent

- $K_0(x,x')$ and $K_{\tan}(x,x')$ depend on n. Normalization N chosen such that they have a finite and strictly positive limit

$$\lim_{n\to\infty} K_0^{(n)}(x,x') = K_0(x,x') \succ 0$$

$$\lim_{n\to\infty} K_{\tan}^{(n)}(x,x') = K_{\tan}(x,x') \succ 0$$

Implies no barren plateaus

- Assumptions on the architecture in terms of
  - L = # of layers (needs to be $O(\log n)$ to avoid barren plateaus)
  - Q = maximum # of measured qubits influenced by a single parameter
  - P = maximum # of parameters that influence a single measured qubit

Gaussian process at initialization

Assume that

$$\lim_{n\to\infty} \frac{n\,Q^2\,P^2}{N^3} = 0$$

Then, the random function $F_{\Theta}(x)$ converges in distribution to the Gaussian process with zero mean and covariance $K_0(x,x')$

NTK concentration and lazy training

Assume that $\lim_{n\to\infty} \frac{n\,L\,Q^4\,P^2}{N^4} = 0$

Then, the empirical NTK converges in distribution to $K_{\mathrm{tan}}(x,x')$

Further assume $\lim_{n\to\infty} \frac{n\,Q^2\,P^2}{N^3} = 0$

Then, for any n large enough, with high probability we have

$$\sup_t \|\Theta_t - \Theta_0\|_\infty = O\left(\frac{Q}{\lambda_{\min}\,N}\right)$$

where λ_{min} is the minimum eigenvalue of K_{tan}

Trained QNNs as Gaussian processes

Assume that $$\lim_{n\to\infty} \frac{L^2\, n^2\, Q^6\, P^3 \log N}{N^5} = 0$$

Then, for sufficiently large n, with high probability we have

$$\sup_{x,t} \left| F_{\Theta(t)}(x) - F_t^{\mathrm{lin}}(x) \right| = O\left(\frac{L^2\, n^2\, Q^6\, P^2 \log N}{N^5 \lambda_{\min}^3} \right) = o(1)$$

Moreover, for any t, $F_{\Theta(t)}(x)$ converges in distribution to the Gaussian process with mean $\mu_t(x)$ and covariance $K_t(x,x')$

Noisy gradient descent

We consider training with discrete gradient descent

$$\Theta_{t+1} = \Theta_t - \eta \nabla_{\Theta} C(\Theta_t)$$

Thanks to parameter-shift rule

$$\partial_{\Theta_i} F_{\Theta}(x) = \frac{1}{2}\left(F_{\Theta+\frac{\pi}{2}e_i}(x) - F_{\Theta-\frac{\pi}{2}e_i}(x)\right)$$

gradients can be computed with $O(1)$ evaluations of $F_{\Theta}(x)$

$F_{\Theta}(x)$ estimated by measurements. We assume unbiased estimators for each component of the gradient with variance

$$O\left(\frac{\lambda_{\min}^4 C(\Theta_t)}{Q^2 L^3 n^3 t^2}\right)$$

Can be achieved with $\mathrm{poly}(n)$ measurements for any fixed t

Noisy gradient descent

Assume that $$\lim_{n\to\infty} \frac{L^2\, n^2\, Q^6\, P^4 \log N}{N^5} = 0$$

Then, for any t, $F_{\Theta(t)}(x)$ converges in distribution to the Gaussian process with mean and covariance

$$\mu_t(x) = K_{\tan}(x, X)^T\, K_{\tan}^{-1} \left(I - (I - \eta\, K_{\tan})^t\right) Y$$

$$K_t(x, x') = K_0(x, x')$$

$$- K_{\tan}(x, X)^T\, K_{\tan}^{-1} \left(I - (I - \eta\, K_{\tan})^t\right) K_0(X, x')$$

$$- K_0(x, X)^T \left(I - (I - \eta\, K_{\tan})^t\right) K_{\tan}^{-1}\, K_{\tan}(X, x')$$

$$+ K_{\tan}(x, X)^T\, K_{\tan}^{-1} \left(I - (I - \eta\, K_{\tan})^t\right) K_0 \left(I - (I - \eta\, K_{\tan})^t\right) K_{\tan}^{-1}\, K_{\tan}(X, x')$$

Quantum advantage vs barren plateaus

Effective Hilbert spaces associated to past light-cones of measured qubits have dimension 2^Q

Naïve classical simulation not efficient whenever dimension grows superpolynomially, i.e., $\lim_{n\to\infty} \frac{Q}{\log n} = \infty$

Is this condition compatible with our hypotheses?

Naïve normalization: $N = \sqrt{n}$

Variance decays exponentially with $L \Rightarrow N = \sqrt{\frac{n}{2^{cL}}}$

Choose

$$L = \epsilon \log_2 n \qquad N = n^{\frac{1-c\epsilon}{2}} \qquad 0 < \epsilon < \frac{1}{5\,c}$$

Qubits on 2D square lattice with nearest-neighbor interactions:

$$Q \simeq L^2 = \epsilon^2 \left(\log_2 n\right)^2 = O(\mathrm{polylog}\, n) \qquad P \leq L\,Q = O(\mathrm{polylog}\, n)$$

Our hypotheses are satisfied!

$$\frac{L^2\, n^2\, Q^6\, P^4 \log N}{N^5} = O\left(n^{\frac{5c\epsilon-1}{2}}\, \mathrm{polylog}\, n\right) \to 0$$

Conclusions

- Trained QNNs in the limit of infinite width are Gaussian processes
- Training always converges in poly time and perfectly fits the training examples
- Generated function is smooth despite infinitely many parameters
- Results robust to statistical noise
- QNNs with qubits on 2D square lattice with nearest-neighbor interactions and logarithmic depth satisfy hypotheses and do not allow naïve efficient classical simulations
- Provable advantages??