Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks

Ziwei Ji, Matus Telgarsky

Main results.

Setting. One-hidden-layer ReLU network, logistic loss.
The neural tangent kernel (NTK) setting.

Theorem (GD). Test error ϵ with prob $1-\delta$ if
width $m \geq \widetilde{\Omega}\left(\frac{1}{\gamma^8}\right)$, samples $n \geq \widetilde{\Omega}\left(\frac{1}{\epsilon^2\gamma^4}\right)$, steps $t = \widetilde{\Theta}\left(\frac{1}{\epsilon\gamma^2}\right)$,
where γ is the NTK margin on the data distribution,
and poly $\log(n, 1/\delta, 1/\epsilon)$ omitted.

Theorem (SGD). Even better: steps $=$ samples $= \widetilde{\Theta}(1/\epsilon\gamma^2)$.

Main results.

Setting. One-hidden-layer ReLU network, logistic loss.
The neural tangent kernel (NTK) setting.

Theorem (GD). Test error ϵ with prob $1-\delta$ if
width $m \geq \widetilde{\Omega}\left(\frac{1}{\gamma^8}\right)$, samples $n \geq \widetilde{\Omega}\left(\frac{1}{\epsilon^2\gamma^4}\right)$, steps $t = \widetilde{\Theta}\left(\frac{1}{\epsilon\gamma^2}\right)$,
where γ is the NTK margin on the data distribution,
and poly $\log(n, 1/\delta, 1/\epsilon)$ omitted.

Theorem (SGD). Even better: steps $=$ samples $= \widetilde{\Theta}(1/\epsilon\gamma^2)$.

- Prior work requires width polynomial in n, or $1/\delta$, or $1/\epsilon$.
- γ is positive under mild conditions;
  width lower bound and tight sample complexity for NTK.

Setting and proof ideas

Basic setting.

- Shallow ReLU ($\sigma(z) = \max\{0, z\}$) network

$$f(x; W) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(w_s^\top x),$$

  where w_s are trained and a_s are held fixed.

- $a_s \sim \text{uniform}(\{-1, +1\})$ and initially $w_s \sim \mathcal{N}(0, I_d)$.

Basic setting.

- Shallow ReLU ($\sigma(z) = \max\{0, z\}$) network

$$f(x; W) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(w_s^{\mathsf{T}} x),$$

  where w_s are trained and a_s are held fixed.

- $a_s \sim \text{uniform}(\{-1, +1\})$ and initially $w_s \sim \mathcal{N}(0, I_d)$.
- Binary classification: $\|x_i\|_2 = 1$ and $y_i \in \{-1, +1\}$.
- Logistic loss

$$\widehat{\mathcal{R}}(W) = \frac{1}{n} \sum_i \ell(y_i f(x_i; W)) = \frac{1}{n} \sum_i \ln\left(1 + \exp(-y_i f(x_i; W))\right).$$

Basic setting.

- Shallow ReLU ($\sigma(z) = \max\{0, z\}$) network

$$f(x; W) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(w_s^\mathsf{T} x),$$

  where w_s are trained and a_s are held fixed.

- $a_s \sim \text{uniform}(\{-1, +1\})$ and initially $w_s \sim \mathcal{N}(0, I_d)$.
- Binary classification: $\|x_i\|_2 = 1$ and $y_i \in \{-1, +1\}$.
- Logistic loss

$$\widehat{\mathcal{R}}(W) = \frac{1}{n} \sum_i \ell(y_i f(x_i; W)) = \frac{1}{n} \sum_i \ln\left(1 + \exp(-y_i f(x_i; W))\right).$$

- GD and SGD

$$W_{t+1} := W_t - \nabla_W \widehat{\mathcal{R}}(W_t),$$
$$\tilde{W}_{t+1} := \tilde{W}_t - \ell'(\tilde{y}_t f(\tilde{x}_t; \tilde{W}_t)) \tilde{y}_t \nabla_W f(\tilde{x}_t; \tilde{W}_t).$$

The neural tangent kernel (Jacot-Gabriel-Hongler '18, Li-Liang '18, Du-Zhai-Poczós-Singh '18).

Linearize the predictor around the initialization W_0:

$$f(x; W) \approx \hat{f}(x; W) := f(x; W_0) + (W - W_0)^\top \nabla_W f(x; W_0).$$

$\hat{f}$ is basically a linear model on $\{(\nabla_W f(x_i; W_0), y_i)\}_{i=1}^n$.

The neural tangent kernel (Jacot-Gabriel-Hongler '18, Li-Liang '18, Du-Zhai-Poczós-Singh '18).

Linearize the predictor around the initialization W_0:

$$f(x; W) \approx \hat{f}(x; W) := f(x; W_0) + (W - W_0)^\top \nabla_W f(x; W_0).$$

$\hat{f}$ is basically a linear model on $\{(\nabla_W f(x_i; W_0), y_i)\}_{i=1}^n$.

Intuition:

- GD on $\hat{f}$ minimizes the risk to 0.
- With a large width, GD on f roughly follows GD on $\hat{f}$.
- Various norm controlled, giving generalization.

The NTK for classification.

Note that

$$f(x; W) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(w_s^\top x)$$

is 1-positively homogeneous: $f(x; rW) = rf(x; W)$ for any $r > 0$.

It follows that (Euler's homogeneous function theorem)

$$f(x; W_t) = \langle W_t, \nabla_W f(x; W_t) \rangle.$$

The NTK for classification.

Note that

$$f(x; W) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(w_s^\top x)$$

is 1-positively homogeneous: $f(x; rW) = rf(x; W)$ for any $r > 0$.

It follows that (Euler's homogeneous function theorem)

$$f(x; W_t) = \langle W_t, \nabla_W f(x; W_t) \rangle.$$

The proof ideas:

- $\exists U$ with a positive margin on $\{(\nabla_W f(x_i; W_0), y_i)\}_{i=1}^n$.
- $\nabla_W f(x_i; W_t) \approx \nabla_W f(x_i; W_0)$: needs small $\|w_{s,t} - w_{s,0}\|$.
  Small empirical risk: needs large $\|W_t - W_0\|_F$.
- A Rademacher complexity bound using $\|w_{s,t} - w_{s,0}\|$.

Intuition: why is polylog width possible?

$H^{(m)}(t)$, the width-m NTK at step t:

$$H_{ij}^{(m)}(t) := \left\langle \nabla_W f(x_i; W_t), \nabla_W f(x_j; W_t) \right\rangle .$$

$H^{(\infty)}$, the expected / infinite-width NTK:

$$H_{ij}^{(\infty)} := \mathbb{E}\left[\left\langle \nabla_W f(x_i; W_0), \nabla_W f(x_j; W_0) \right\rangle \right] .$$

Intuition: why is polylog width possible?

$H^{(m)}(t)$, the width-m NTK at step t:

$$H_{ij}^{(m)}(t) := \langle \nabla_W f(x_i; W_t), \nabla_W f(x_j; W_t) \rangle .$$

$H^{(\infty)}$, the expected / infinite-width NTK:

$$H_{ij}^{(\infty)} := \mathbb{E}\left[\langle \nabla_W f(x_i; W_0), \nabla_W f(x_j; W_0) \rangle \right] .$$

- For regression, some prior work used $\lambda_{\min}\left(H^{(m)}(t)\right) > 0$; which involves controlling $\sum_{1 \le i,j \le n} \left| H_{ij}^{(m)}(t) - H_{ij}^{(\infty)} \right|$.

Intuition: why is polylog width possible?

$H^{(m)}(t)$, the width-m NTK at step t:

$$H_{ij}^{(m)}(t) := \langle \nabla_W f(x_i; W_t), \nabla_W f(x_j; W_t) \rangle .$$

$H^{(\infty)}$, the expected / infinite-width NTK:

$$H_{ij}^{(\infty)} := \mathbb{E}\left[\langle \nabla_W f(x_i; W_0), \nabla_W f(x_j; W_0) \rangle \right].$$

- For regression, some prior work used $\lambda_{\min}\left(H^{(m)}(t)\right) > 0$; which involves controlling $\sum_{1 \le i,j \le n} \left| H_{ij}^{(m)}(t) - H_{ij}^{(\infty)} \right|$.
- For classfication, need a positive margin:

$$\min \left\{ \sum_{1 \le i,j \le n} q_i q_j y_i y_j H_{ij}^{(m)}(t) \,\middle|\, q \ge 0, \sum_{i=1}^{n} q_i = 1 \right\}.$$

Need to bound $\max_{1 \le i,j \le n} \left| H_{ij}^{(m)}(t) - H_{ij}^{(\infty)} \right|$; scales with $\ln(n)$.

The NTK margin γ.

Linear margin with infinite-width NTK features;
in the proof we use a primal definition (Nitanda & Suzuki '19).

The NTK margin γ.

Linear margin with infinite-width NTK features;
in the proof we use a primal definition (Nitanda & Suzuki '19).

- $\gamma > 0$ if there are no parallel inputs,
  or if the labels can be represented by a continuous function
  (Ji-Telgarsky-Xian '19).

The NTK margin γ.

Linear margin with infinite-width NTK features;
in the proof we use a primal definition (Nitanda & Suzuki '19).

- $\gamma > 0$ if there are no parallel inputs,
  or if the labels can be represented by a continuous function
  (Ji-Telgarsky-Xian '19).
- There exists a dataset, if $m < O(1/\sqrt{\gamma})$, then with const prob,
  the finite-width NTK does not have a positive margin.

The NTK margin γ.

Linear margin with infinite-width NTK features;
in the proof we use a primal definition (Nitanda & Suzuki '19).

- $\gamma > 0$ if there are no parallel inputs,
  or if the labels can be represented by a continuous function
  (Ji-Telgarsky-Xian '19).
- There exists a dataset, if $m < O(1/\sqrt{\gamma})$, then with const prob,
  the finite-width NTK does not have a positive margin.
- (Wei-Lee-Liu-Ma '19) introduced the noisy 2-XOR distribution,
  proved the infinite-width NTK needs d^2 samples.

  Our SGD upper bound matches this lower bound:
  We prove $\widetilde{O}(1/\gamma^2)$ samples, and $\gamma = \Omega(1/d)$ for this case.

Thanks!