Benefits of Early Stopping in Gradient Descent for Overparameterized Logistic Regression

Berkeley UNIVERSITY OF CALIFORNIA — NYU

Jingfeng Wu[1] Peter Bartlett[1,3] Matus Telgarsky[2] Bin Yu[1]
[1]UC Berkeley [2]New York University [3]Google DeepMind

Background

Dataset $y_i \in \{\pm 1\},\ x_i \in \mathbb{R}^d,\ i = 1, \dots, n,\ d > n$

overparameterization => linear separability

Empirical risk $\hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n} \ln\big(1 + \exp(-y_i x_i^\top w)\big)$

Gradient descent $w_{t+1} = w_t - \eta \nabla \hat{L}(w_t),\ w_0 = 0$

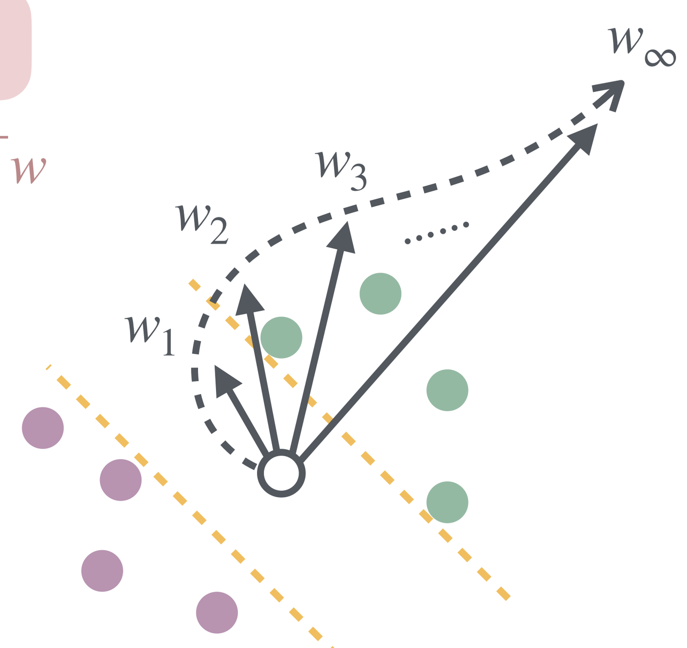
Asymptotic implicit bias

max-margin direction $\tilde{w} = \arg\max_{\|w\|=1} \min_i y_i x_i^\top w$

[Soudry et al, 2018; Ji & Telgarsky, 2018]

If $\eta = \Theta(1)$, then as $t \to \infty$,

$$\|w_t\| \to \infty, \quad \frac{w_t}{\|w_t\|} \to \tilde{w}$$



max-margin is not the full story

Data model

allow $\text{rank}(\Sigma)$, $\|w^*\| = \infty$

[Population distribution] For $\text{tr}(\Sigma) \lesssim 1$ and $\|w^*\|_\Sigma \lesssim 1$,

$$x \sim \mathcal{N}(0, \Sigma) \quad \Pr(y = 1 \mid x) = s(x^\top w^*)$$

sigmoid, $s(t) = \frac{1}{1 + \exp(-t)}$

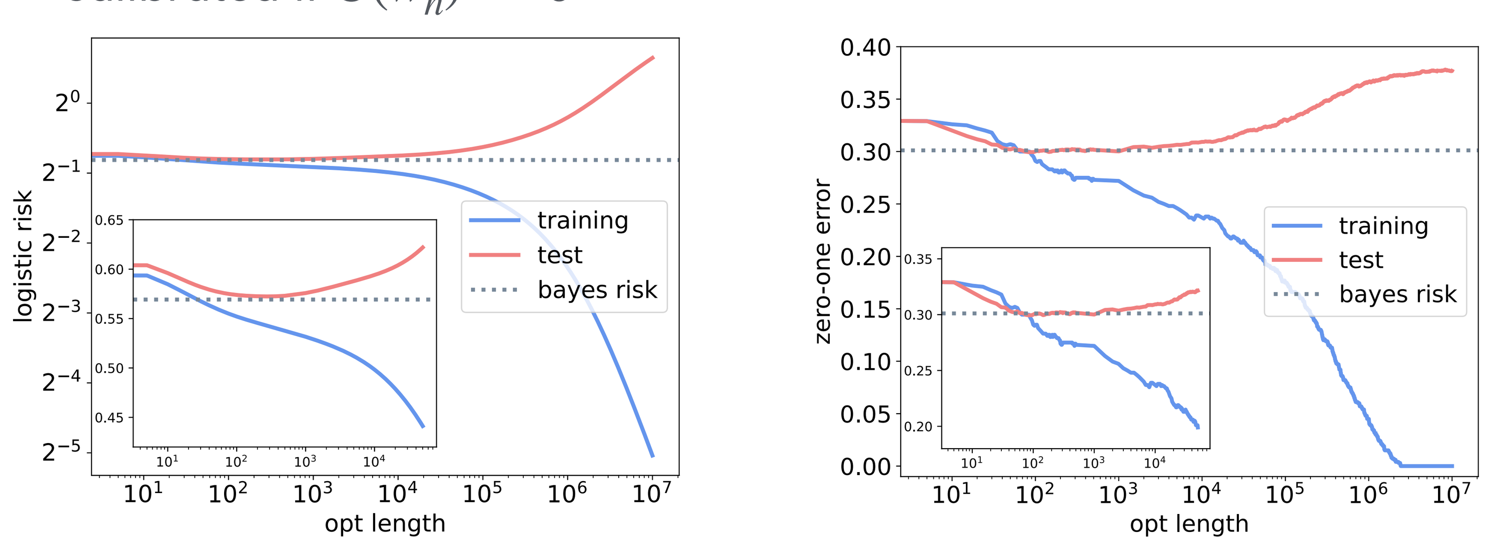
Logistic risk $L(w) = \mathbb{E} \ln\big(1 + \exp(-y x^\top w)\big)$

Zero-one risk $Z(w) = \Pr(y x^\top w \le 0)$

Calibration risk $C(w) = \mathbb{E}\,|s(x^\top w) - \Pr(y = 1 \mid x)|^2$

[Consistency & calibration] An estimator w_n is

- logistic or 0-1 *consistent* if $L(w_n) \to \min L$ or $Z(w_n) \to \min Z$
- *calibrated* if $C(w_n) \to 0$



[Simulations] $d = 2000,\ n = 1000,\ \Sigma_{ii} = i^{-2},\ w^*_{0:100} = 1,\ w^*_{100:d} = 0$

Early-stopped GD

[Basic facts]

logistic consistent => calibration => zero-one consistent

- w^* minimizes L, Z, and C
- $Z(w) - \min Z \le 2\sqrt{C(w)} \le \sqrt{2}\sqrt{L(w) - \min L}$
- $\min L \gtrsim 1$ and $\min Z \gtrsim 1$ — $\Theta(1)$ noise => overfitting

Risk bounds

[Theorem] Let $\eta \lesssim 1$ so GD is stable. Pick a stopping time t

$t(w^*, \Sigma, k_n)$: $\hat{L}(w_t) \le \hat{L}(w^*_{0:k}) \le \hat{L}(w_{t-1})$

Then with high probability

"best" rank-k projection

$$L(w_t) - \min L \lesssim \tilde{O}(1)\sqrt{\frac{\|w^*_{0:k}\|^2}{n} + \|w^*_{k:\infty}\|^2_\Sigma}$$

$o(1)$ for $k_n \uparrow$; $o(1)$ since $k_n \uparrow$ and $\|w^*\|_\Sigma \lesssim 1$

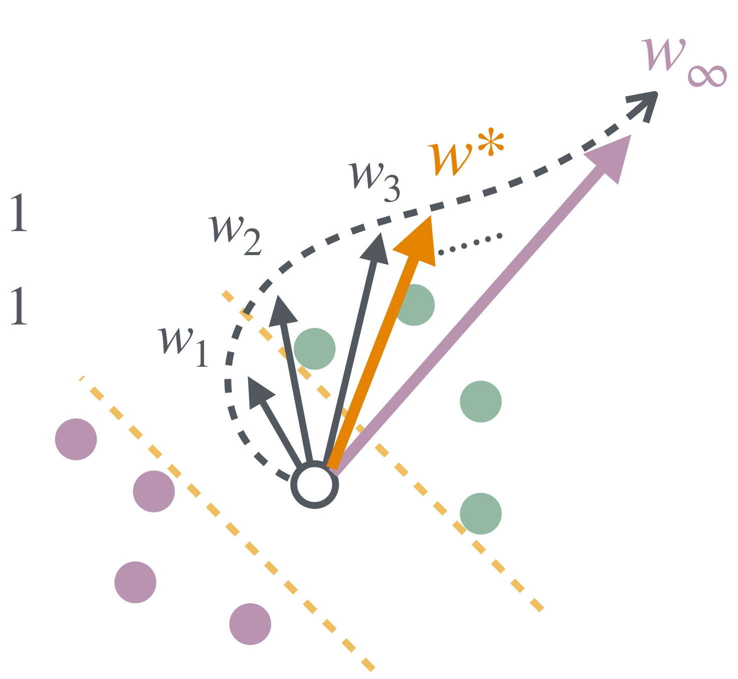
[Examples] (rates are improvable)

- Finite norm: $\|w^*\| \lesssim 1$

$$L(w_t) - \min L \le \tilde{O}(n^{-1/2})$$

- Power laws: $\lambda_i = i^{-a},\ \lambda_i (w_i^*)^2 = i^{-b},\ a, b > 1$

$$L(w_t) - \min L \le \begin{cases} \tilde{O}(n^{-1/2}) & b > a + 1 \\ \tilde{O}(n^{\frac{1-b}{a+b-1}}) & b \le a + 1 \end{cases}$$



GD passes through w^* but eventually diverges from it

Key ideas

[Lemma (known)] For all convex-smooth $\hat{L}$ and small η, we have

$$\forall\, u, t, \quad \frac{\|w_t - u\|^2}{2\eta t} + \hat{L}(w_t) \le \hat{L}(u) + \frac{\|u\|^2}{2\eta t}$$

$\hat{L}(w_t) \le \hat{L}(u) \le \hat{L}(w_{t-1})$ → $\hat{L}(w_t) \le \hat{L}(u)$, $\|w_{t-1} - u\| \le \|u\|$

(local) Rademacher complexity ←

Interpolating estimators

Issue of divergent norm

[Theorem] For all $(w_t)_{t>0}$ such that

apply to GD when overparameterized

$$\lim \|w_t\| = \infty, \quad \lim \frac{w_t}{\|w_t\|} \text{ exists}$$

we have

inconsistent: $L(w_\infty) = \infty$, poorly calibrated: $C(w_\infty) \gtrsim 1$

Issue of interpolation

[Theorem] Assume that $\|w^*\|_\Sigma \approx 1$ and $\Sigma^{1/2} w^*$ is k-sparse. If

$\min_i y_i x_i^\top \hat{w} > 0$, $\quad n \gtrsim k \ln k, \quad \text{rank}(\Sigma) \eqsim n \ln n$

then for every interpolator $\hat{w}$, with high probability

$$Z(\hat{w}) - \min Z \gtrsim \frac{1}{\sqrt{\ln n}}$$

$\text{poly}(1/n)$ for early stopping in "simple problems"

Early stopping and l_2-regularization

$$u_\lambda = \arg\min \hat{L}(u) + \frac{1}{2\lambda}\|u\|^2$$

[Theorem] For all convex-smooth $\hat{L}$, small η, and all $t > 0$,

$$\|w_t - u_\lambda\| \le \frac{1}{\sqrt{2}}\|w_t\| \ \text{ for } \lambda = \eta t$$

global, but relative

As a result: $\angle(w_t, u_\lambda) \le \frac{\pi}{4}, \quad 0.585 < \frac{\|w_t\|}{\|u_\lambda\|} < 3.415$

[Theorem] For logistic regression

- If $\text{rank}\{\text{support vectors}\} = \text{rank}\{\text{data}\}$, then $\|w_t\|, \|u_\lambda\| \to \infty$

  $\lambda \ne \eta t$: $\exists\, \lambda(t) \to \infty, \quad \|w_t - u_\lambda\| \to 0$

- For dataset $x_1 = (\gamma, 0),\ x_2 = (\gamma, \gamma_2),\ y_1 = y_2 = 1$, with $0 < \gamma_2 < \gamma < 1$, which violates the above condition, we have

$$\forall\, \lambda(t), \quad \|w_t - u_\lambda\| \gtrsim \ln\ln \|w_t\| \to \infty$$