Benefits of Early Stopping in GD
for Overparameterized Logistic Regression

Jingfeng Wu
Joint work with Peter Bartlett, Matus Telgarsky, and Bin Yu

Logistic regression

$y_i \in \{\pm 1\},\ x_i \in \mathbb{R}^d,\ i \leq n$

$\ell(t) := \ln(1 + e^{-t})$

$$\widehat{L}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w)$$

Logistic regression

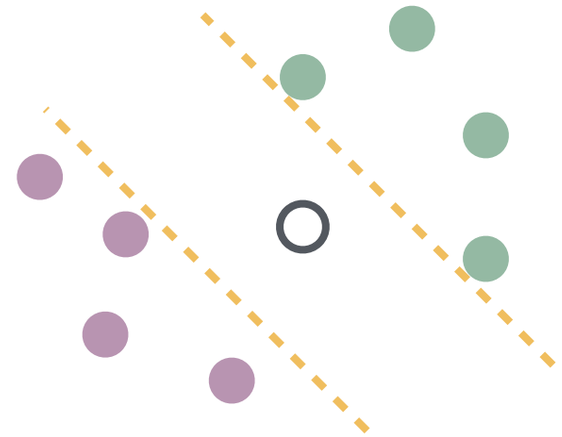
$y_i \in \{\pm 1\},\ x_i \in \mathbb{R}^d,\ i \le n$

high dim

$d > n$

$\ell(t) := \ln(1 + e^{-t})$

linear separability



$$\widehat{L}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w)$$

~~"ERM"~~

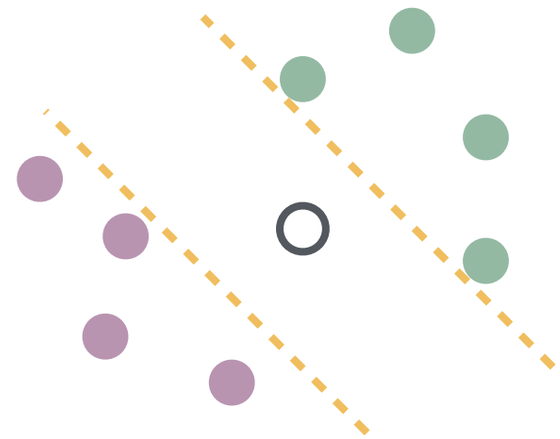
~~"uniform convergence"~~

Logistic regression

$y_i \in \{\pm 1\},\ x_i \in \mathbb{R}^d,\ i \le n$ high dim $d > n$

$\ell(t) := \ln(1 + e^{-t})$

linear separability



$$\widehat{L}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top w)$$

~~"ERM"~~

~~"uniform convergence"~~

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

Asymptotic implicit bias

$$\tilde{w} := \arg\max_{\|w\|=1} \min_i y_i x_i^\top w$$

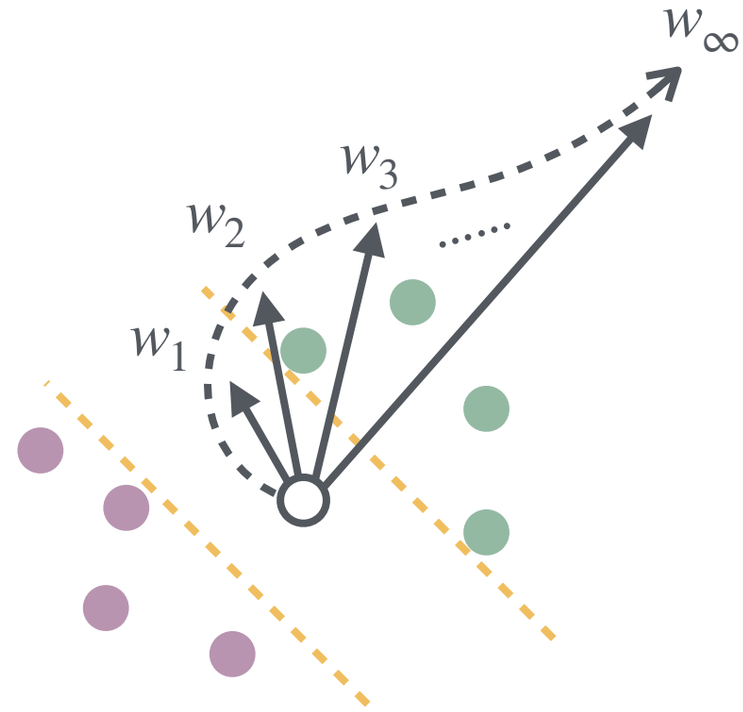
max-margin direction

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

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

$$\|w_t\| \to \infty$$

$$\frac{w_t}{\|w_t\|} \to \tilde{w}$$



Is max-margin the full story?

Missing aspects

- Divergent norm (bad for metrics other than zero-one)
- Max-margin feels unstable

Missing aspects

- Divergent norm (bad for metrics other than zero-one)
- Max-margin feels unstable
- Requiring **exp time**

$$\frac{w_t}{\|w_t\|} = \tilde{w} + O\left(\frac{\ln\ln(t)}{\ln(t)}\right)$$

$$\|w_t\| = \Theta(\ln t)$$

Metrics

Logistic $L(w) := \mathbb{E}\ell(yx^\top w) \qquad \ell(t) := \ln(1 + e^{-t})$

Zero-one $Z(w) := \Pr(yx^\top w \leq 0)$

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

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

Metrics

Logistic $L(w) := \mathbb{E}\ell(yx^\top w) \qquad \ell(t) := \ln(1 + e^{-t})$

Zero-one $Z(w) := \Pr(yx^\top w \leq 0)$

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

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

Consistency (logistic or zero-one)

$$L(w_n) \to \min L \quad \text{or} \quad Z(w_n) \to \min Z$$

Calibration $C(w_n) \to 0$

Data model

sigmoid

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

for $\mathrm{tr}(\Sigma) \lesssim 1$ and $\|w^*\|_\Sigma \lesssim 1$ "not grow with n"

Data model

allow $\operatorname{rank}(\Sigma) = \infty$

sigmoid

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

for $\operatorname{tr}(\Sigma) \lesssim 1$ and $\|w^*\|_\Sigma \lesssim 1$ "not grow with n"

allow $\|w^*\| = \infty$

"benign overfitting setup"

Data model

allow $\operatorname{rank}(\Sigma) = \infty$

sigmoid

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

for $\operatorname{tr}(\Sigma) \lesssim 1$ and $\|w^*\|_\Sigma \lesssim 1$ "not grow with n"

allow $\|w^*\| = \infty$

"benign overfitting setup"

A. w^* minimizes L, Z, and C

B. $Z(w) - \min Z \le 2\sqrt{C(w)} \le \sqrt{2}\sqrt{L(w) - \min L}$

C. $\min L \gtrsim 1$ and $\min Z \gtrsim 1$

Data model

allow $\operatorname{rank}(\Sigma) = \infty$

sigmoid

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

for $\operatorname{tr}(\Sigma) \lesssim 1$ and $\|w^*\|_\Sigma \lesssim 1$ "not grow with n"

allow $\|w^*\| = \infty$

"benign overfitting setup"

A. w^* minimizes L, Z, and C

B. $Z(w) - \min Z \le 2\sqrt{C(w)} \le \sqrt{2}\sqrt{L(w) - \min L}$

C. $\min L \gtrsim 1$ and $\min Z \gtrsim 1$

logistic consistent => calibration
=> zero-one consistent

Data model

allow $\operatorname{rank}(\Sigma) = \infty$

sigmoid

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

for $\operatorname{tr}(\Sigma) \lesssim 1$ and $\|w^*\|_\Sigma \lesssim 1$ "not grow with n"

allow $\|w^*\| = \infty$

"benign overfitting setup"

A. w^* minimizes L, Z, and C

B. $Z(w) - \min Z \le 2\sqrt{C(w)} \le \sqrt{2}\sqrt{L(w) - \min L}$

C. $\min L \gtrsim 1$ and $\min Z \gtrsim 1$

$\Theta(1)$ noise => overfitting

logistic consistent => calibration
=> zero-one consistent

Logistic risk bound

Let $\eta \lesssim 1$ so GD is stable. Pick stopping time t

$$\widehat{L}(w_t) \leq \widehat{L}(w^*_{0:k}) \leq \widehat{L}(w_{t-1})$$

Then w.h.p.

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

Logistic risk bound

Let $\eta \lesssim 1$ so GD is stable. Pick stopping time t

$$\widehat{L}(w_t) \leq \widehat{L}(w^*_{0:k}) \leq \widehat{L}(w_{t-1})$$

Then w.h.p.

"best" rank-k projection

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

Logistic risk bound

Let $\eta \lesssim 1$ so GD is stable. Pick stopping time t

$$\widehat{L}(w_t) \leq \widehat{L}(w^*_{0:k}) \leq \widehat{L}(w_{t-1})$$

Then w.h.p.

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

"best" rank-k projection

$o(1)$ for $k_n \uparrow$

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

Logistic risk bound

Let $\eta \lesssim 1$ so GD is stable. Pick stopping time t

$$\widehat{L}(w_t) \leq \widehat{L}(w^*_{0:k}) \leq \widehat{L}(w_{t-1})$$

Then w.h.p.

"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 some t^*_n as long as "not grow with n"

$o(1)$ for $k_n \uparrow$

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

Logistic risk bound

Let $\eta \lesssim 1$ so GD is stable. Pick stopping time t

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

Then w.h.p.

$t(w^*, \Sigma, k_n)$

"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 some t^*_n as long as "not grow with n"

$o(1)$ for $k_n \uparrow$

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

Logistic risk bound

implies calibration & zero-one

Let $\eta \lesssim 1$ so GD is stable. Pick stopping time t

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

Then w.h.p.

$t(w^*, \Sigma, k_n)$

"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 some t^*_n as long as "not grow with n"

$o(1)$ for $k_n \uparrow$

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

Examples

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

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

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

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

$\|w^*\| = \infty$

Examples

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

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

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

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

$\|w^*\| = \infty$

rates improvable

Issue of divergent norm

We have

inconsistency

poor calibration

$$L(w_\infty) = \infty, \quad C(w_\infty) \gtrsim 1$$

for all $(w_t)_{t>0}$ such that

applies to GD when overparameterized

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

metrics sensitive to estimator norm

but $\|w_\infty\| = \infty$

inherent in “ERM”

Issue of interpolation

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

$$n \gtrsim k \ln k, \quad \operatorname{rank}(\Sigma) \eqsim n \ln n$$

then for every interpolator $\hat{w}$, w.h.p.

$\min_i y_i x_i^\top \hat{w} > 0$

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

Issue of interpolation

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

$$n \gtrsim k \ln k, \quad \operatorname{rank}(\Sigma) \eqsim n \ln n$$

then for every interpolator $\hat{w}$, w.h.p.

$\min_i y_i x_i^\top \hat{w} > 0$

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

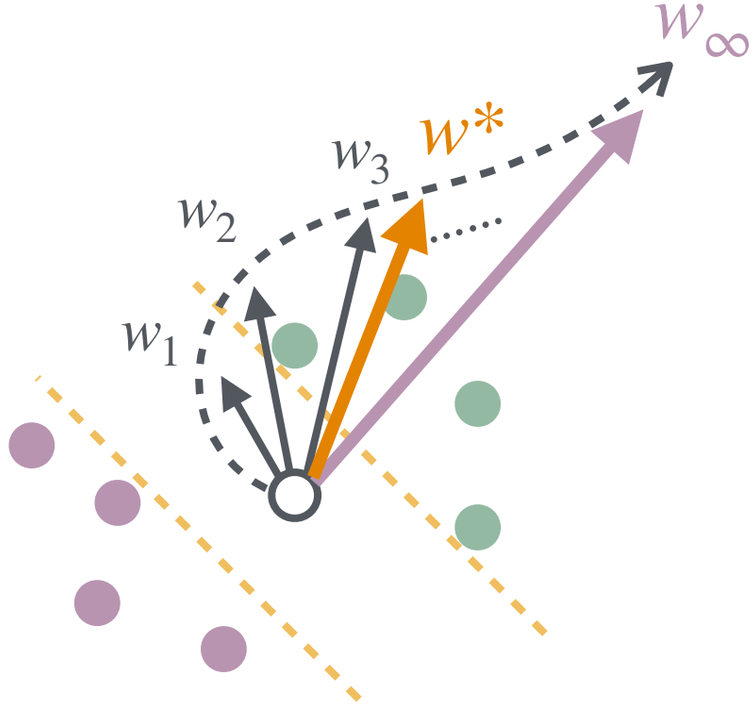
$$Z(w_t) - \min Z \lesssim \operatorname{sqrt}\left(\frac{\|w^*_{0:k}\|}{\sqrt{n}} + \|w^*_{k:\infty}\|_\Sigma^2 \right) = \operatorname{poly}\left(\frac{1}{n}\right)$$

for “simple” problems $\|w^*\| = \Theta(1)$ or power laws

Benefits of early stopping

	early-stopped	**asymptotic**
logistic consistency	always yes	always no
calibration	always yes	always no
zero-one risk	“poly”	“polylog”

GD passes through w^*

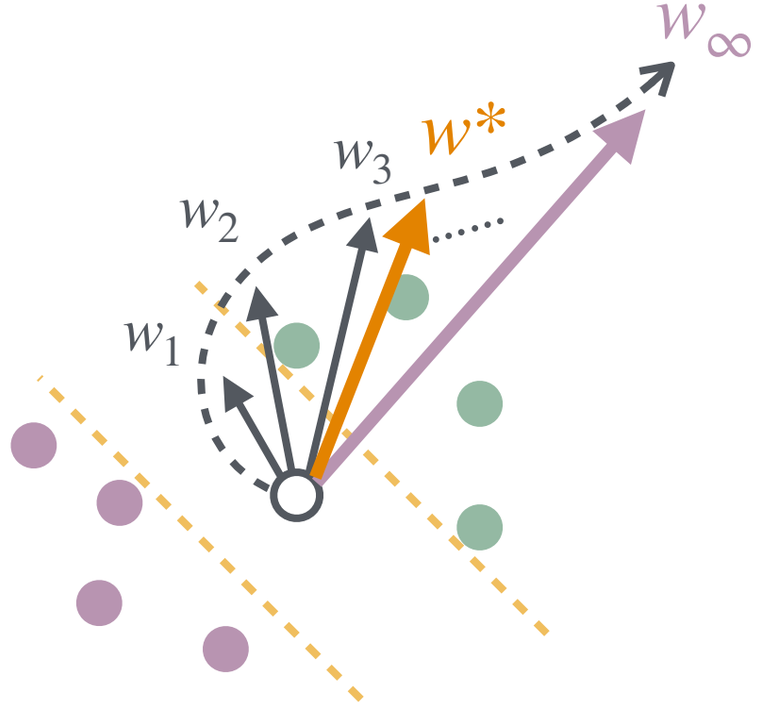


but eventually diverges from it

Benefits of early stopping

	early-stopped	asymptotic
logistic consistency	always yes	always no
calibration	always yes	always no
zero-one risk	“poly”	“polylog”

GD passes through w^*



but eventually diverges from it

More in paper: early stopping vs. ℓ_2-regularization