GD for Logistic Regression

Benefits of Early Stopping

Jingfeng Wu

Berkeley
UNIVERSITY OF CALIFORNIA

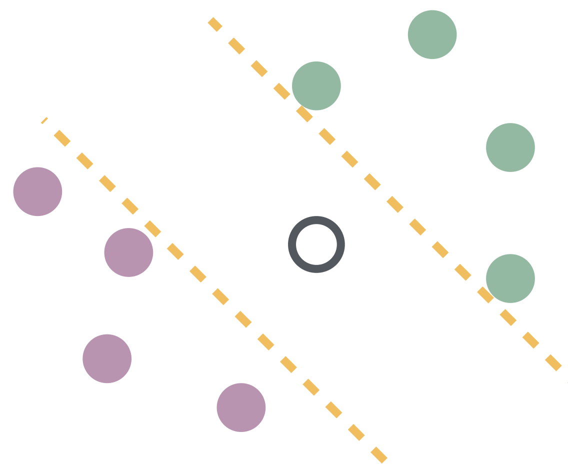
Logistic regression

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

high dim

$d > n$

linear separability



~~"ERM"~~

~~"uniform convergence"~~

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

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

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

Asymptotic implicit bias

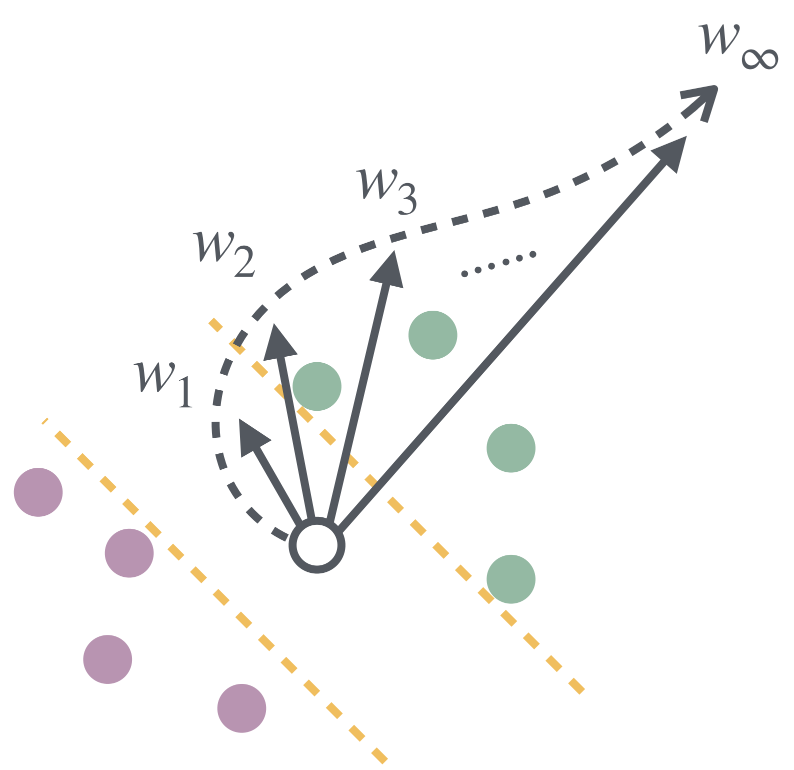
$$\tilde{w} := \arg\max_{\|w\|=1} \min_i y_i x_i^\top w$$

[Soudry et al, 2018; Ji & Telgarsky, 2018; ... **Wu** et al, 2023]

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
- Why logistic not hinge/SVM loss?
- Requiring **exp time**

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

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

Benefits of early stopping

1. Consistency & calibration

2. Advantages over interpolation

3. Connections to l_2-regularization

Peter Bartlett

Matus Telgarsky

Bin Yu

Wu, Bartlett, Telgarsky, and Yu. "Benefits of Early Stopping in Gradient Descent for Overparameterized Logistic Regression" arXiv:2502.13283

Metrics

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

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

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

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

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

"benign overfitting setup"

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

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

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

Logistic risk bound

implies calibration & zero-one

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

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

Then w.h.p.

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

$o(1)$ for some t^*_n as long as "not grow with n"

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

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

Proof ideas

For convex-smooth $\widehat{L}$ and small η, we have

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

$$\widehat{L}(w_t) \le \widehat{L}(u) \le \widehat{L}(w_{t-1})$$

$$\widehat{L}(w_t) \le \widehat{L}(u)$$
$$\|w_{t-1} - u\| \le \|u\|$$

(local) Rademacher complexity

many loose places; unclear how to improve :(

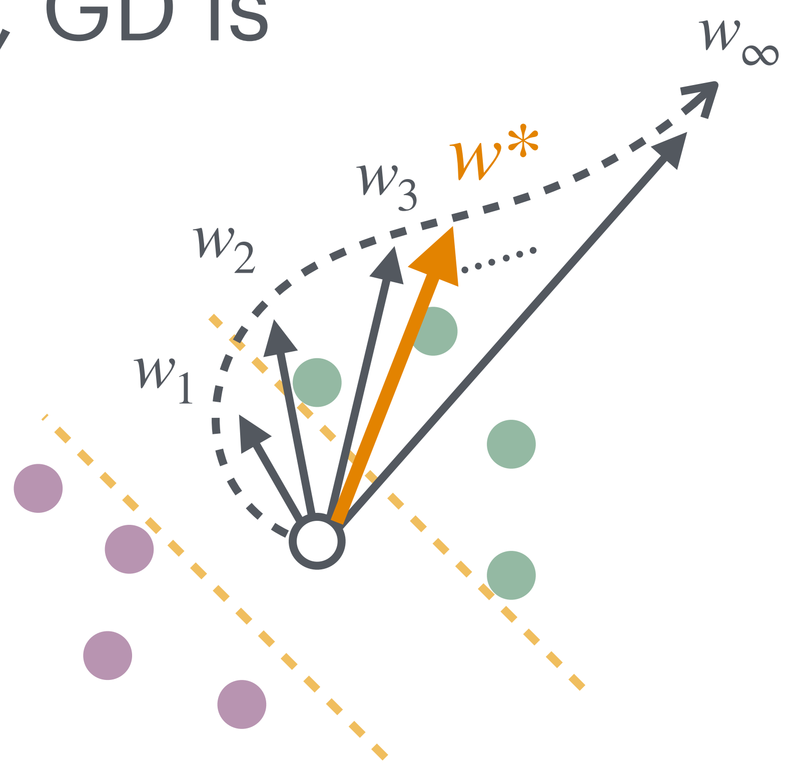
Rethinking GD

coming next: $t^* \ll \infty$

Issues of $t = \infty$:
1. divergent norm
2. interpolation

With an oracle-chosen stopping time, GD is

- consistent in logistic, w/ "poly" rate
- calibrated
- consistent in zero-one



for **every** instance with $\mathrm{tr}(\Sigma) \lesssim 1, \|w^*\|_\Sigma \lesssim 1$

dimension arbitrarily high
l_2-norm arbitrarily large

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

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

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

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

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

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

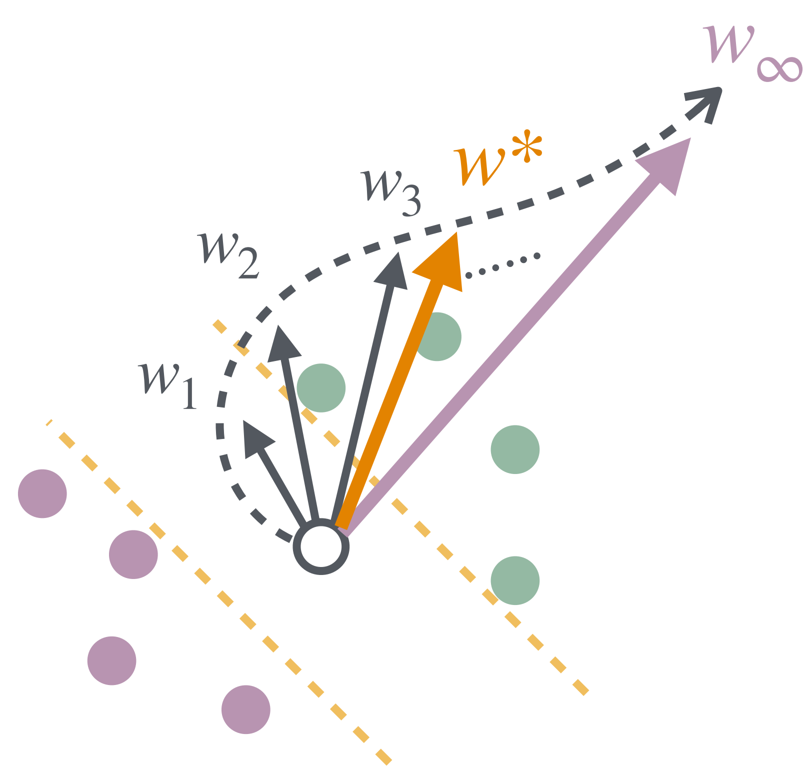
$$Z(w_t) - \min Z \lesssim \mathrm{sqrt}\left(\frac{\|w^*_{0:k}\|}{\sqrt{n}} + \|w^*_{k:\infty}\|_\Sigma^2\right) = \mathrm{poly}\left(\frac{1}{n}\right)$$

for "simple" problems $k = \Theta(1)$ or $\|w^*\| = \Theta(1)$

Benefits of early stopping

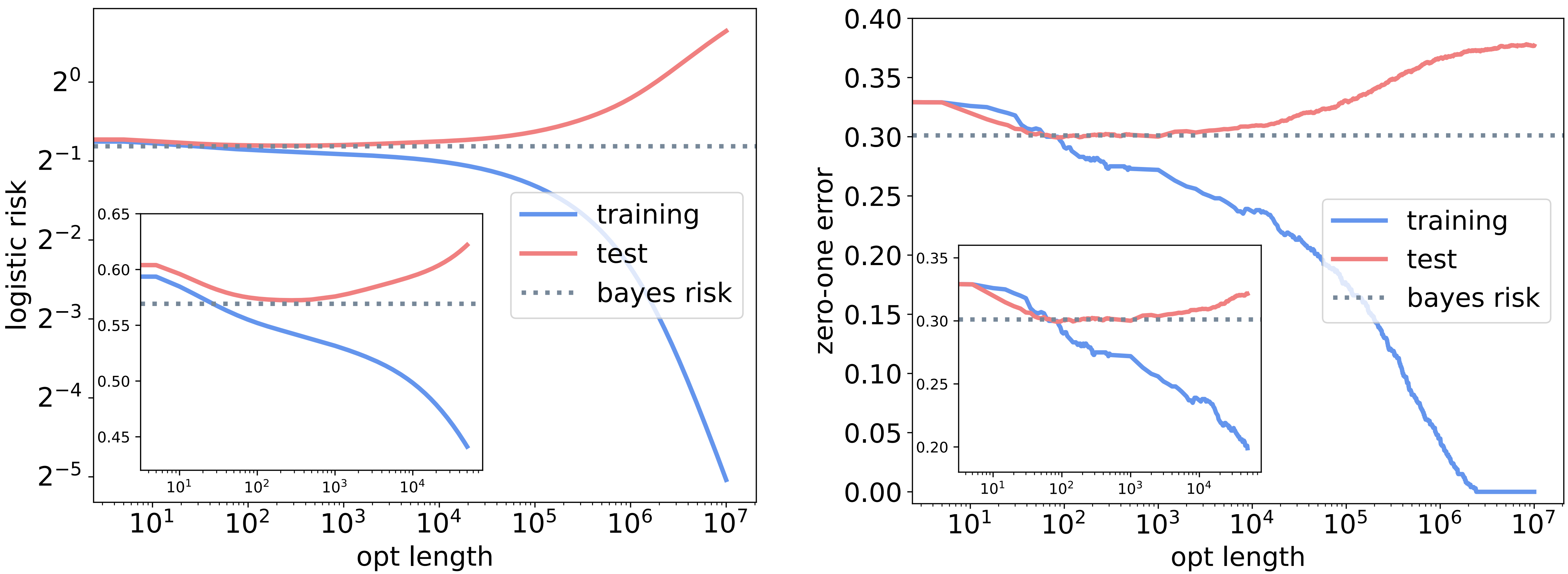
	early-stopped	**asymptotic**
logistic consistency	always yes	always no
calibration	always yes	always no
zero-one risk	“poly”	“polylog”

GD passes through w^*



but eventually diverges from it

Simulations



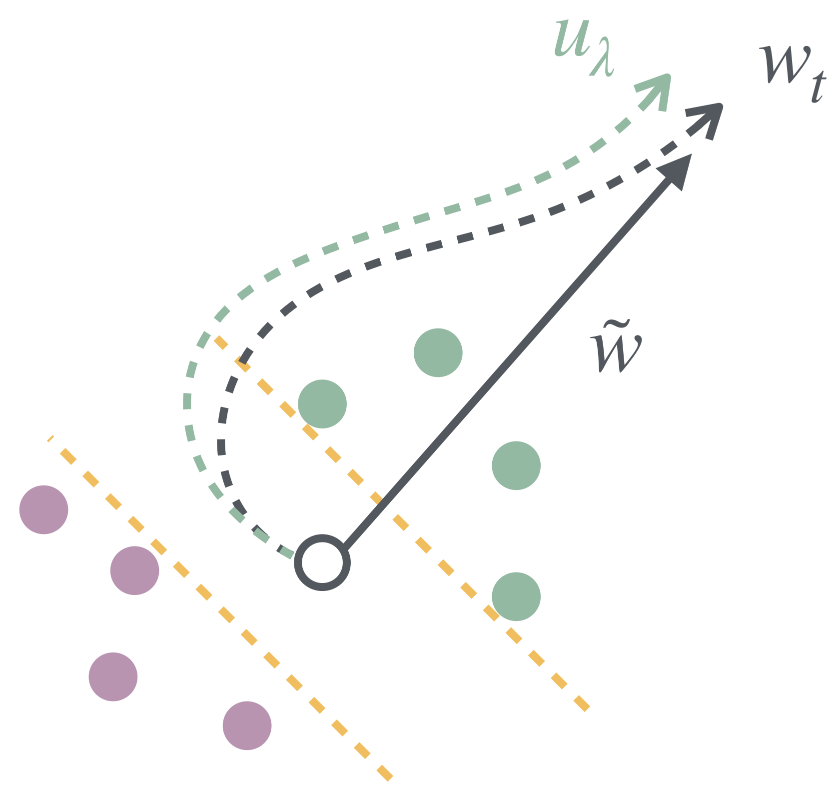
$$d = 2000,\, n = 1000,\, \lambda_i = i^{-2},\, w^* = (\underbrace{1,\ldots,1}_{k=100},0,\ldots)$$

GD and l_2-regularization

$$w_{t+1} = w_t - \eta \nabla \widehat{L}(w_t)$$

$$= \arg\min \widehat{L}(w_t) + \langle \nabla \widehat{L}(w_t), u - w_t \rangle + \frac{1}{2\eta}\|u - w_t\|^2$$

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



preceding GD bounds hold for u_λ
(with similar looseness...)

coming next: rigorous path comparison

Convex function

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

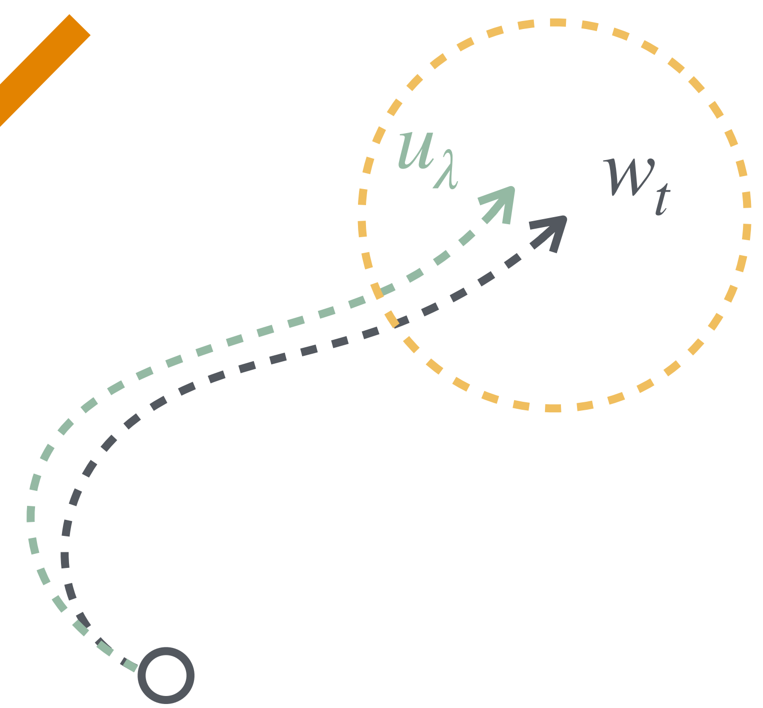
$$\|w_t - u_\lambda\| \leq \frac{1}{\sqrt{2}} \|w_t\| \quad \text{for } \lambda = \eta t$$

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

$\hat{L}$ can be non-strictly convex

$\hat{w}^*$ can be infinite

Theory of l_2-regu applies to GD if it only uses norm



Separable logistic regression

Assume $\mathrm{rank}\{\mathrm{support\ vectors}\} = \mathrm{rank}\{\mathrm{data}\}$, then

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

For dataset $x_1 = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}, x_2 = \begin{pmatrix} \gamma \\ \gamma_2 \end{pmatrix}, y_1 = y_2 = 1$, where $0 < \gamma_2 < \gamma < 1$, we have

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

paths diverge in two directions with different ratios

Contribution

- Implicit regularization via early stopping
- Calibration, consistency, l_2-regu…
- Key diff: logistic vs. linear regression in high-dim