

GD for Logistic Regression

Benefits of Early Stopping

Jingfeng Wu

Berkeley
UNIVERSITY OF CALIFORNIA

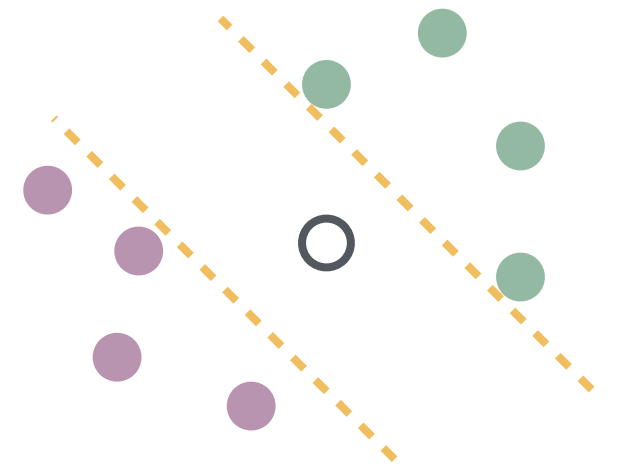
Logistic regression

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

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

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

linear
separability



~~“ERM”~~

~~“uniform convergence”~~

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

Asymptotic implicit bias

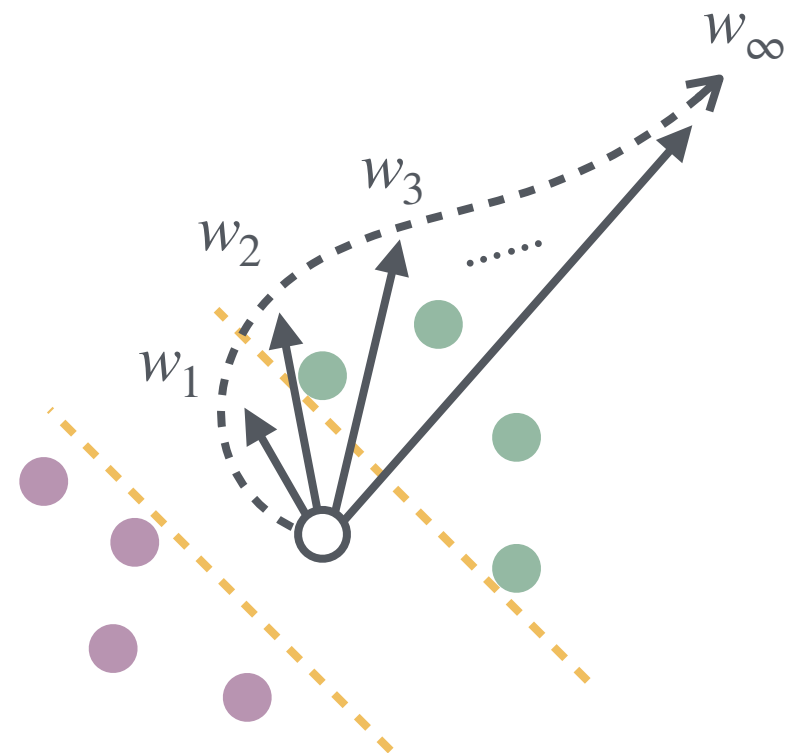
$$\tilde{w} := \arg \max_{\|w\|=1} \min_i y_i x_i^\top w$$

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

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

$$\|w_t\| \rightarrow \infty$$

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



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

old trick
new context



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

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)

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

Calibration $C(w_n) \rightarrow 0$

Data model

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

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

“benign overfitting setup”

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

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

$$\text{C. } \min L \gtrsim 1 \text{ and } \min Z \gtrsim 1$$

Logistic risk bound implies calibration & zero-one

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.

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

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

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

rates improvable

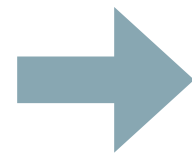
Proof ideas



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

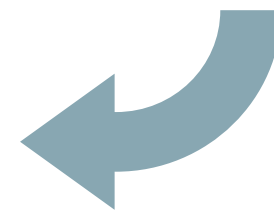
$$\forall u, t, \quad \frac{\|w_t - u\|^2}{2\eta t} + \hat{L}(w_t) \leq \hat{L}(u) + \frac{\|u\|^2}{2\eta t}$$

$$\hat{L}(w_t) \leq \hat{L}(u) \leq \hat{L}(w_{t-1})$$



$$\begin{aligned} \hat{L}(w_t) &\leq \hat{L}(u) \\ \|w_{t-1} - u\| &\leq \|u\| \end{aligned}$$

(local) Rademacher complexity



many loose places; unclear how to improve :(

Rethinking GD

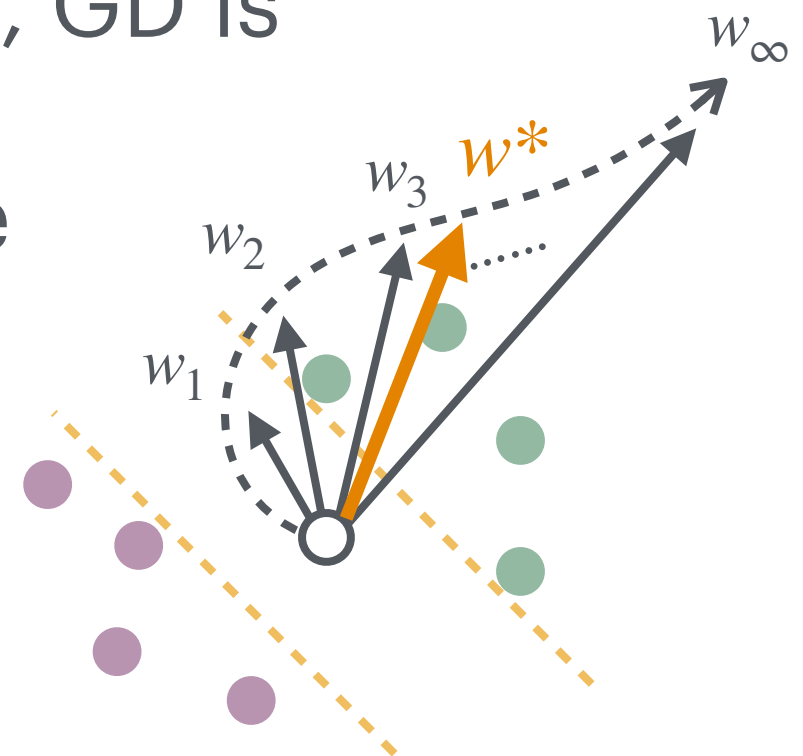
coming next: $t^* \ll \infty$

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 $\text{tr}(\Sigma) \lesssim 1$, $\|w^*\|_\Sigma \lesssim 1$

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

$$n \gtrsim k \ln k, \quad \text{rank}(\Sigma) \approx n \ln n$$

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

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

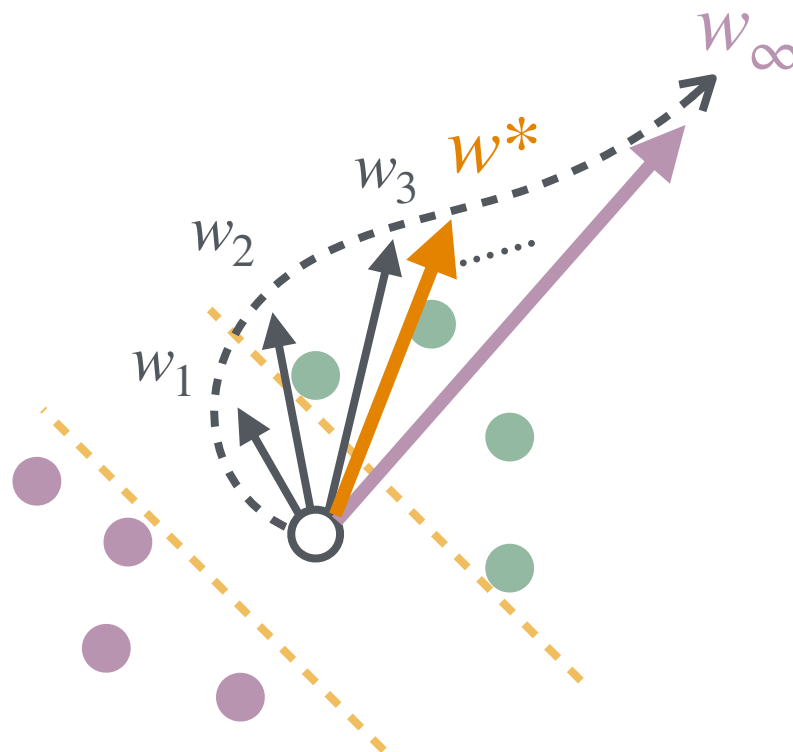
$$Z(w_t) - \min Z \lesssim \text{sqrt} \left(\frac{\|w_{0:k}^*\|}{\sqrt{n}} + \|w_{k:\infty}^*\|_{\Sigma}^2 \right) = \text{poly} \left(\frac{1}{n} \right)$$

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

Benefits of early stopping

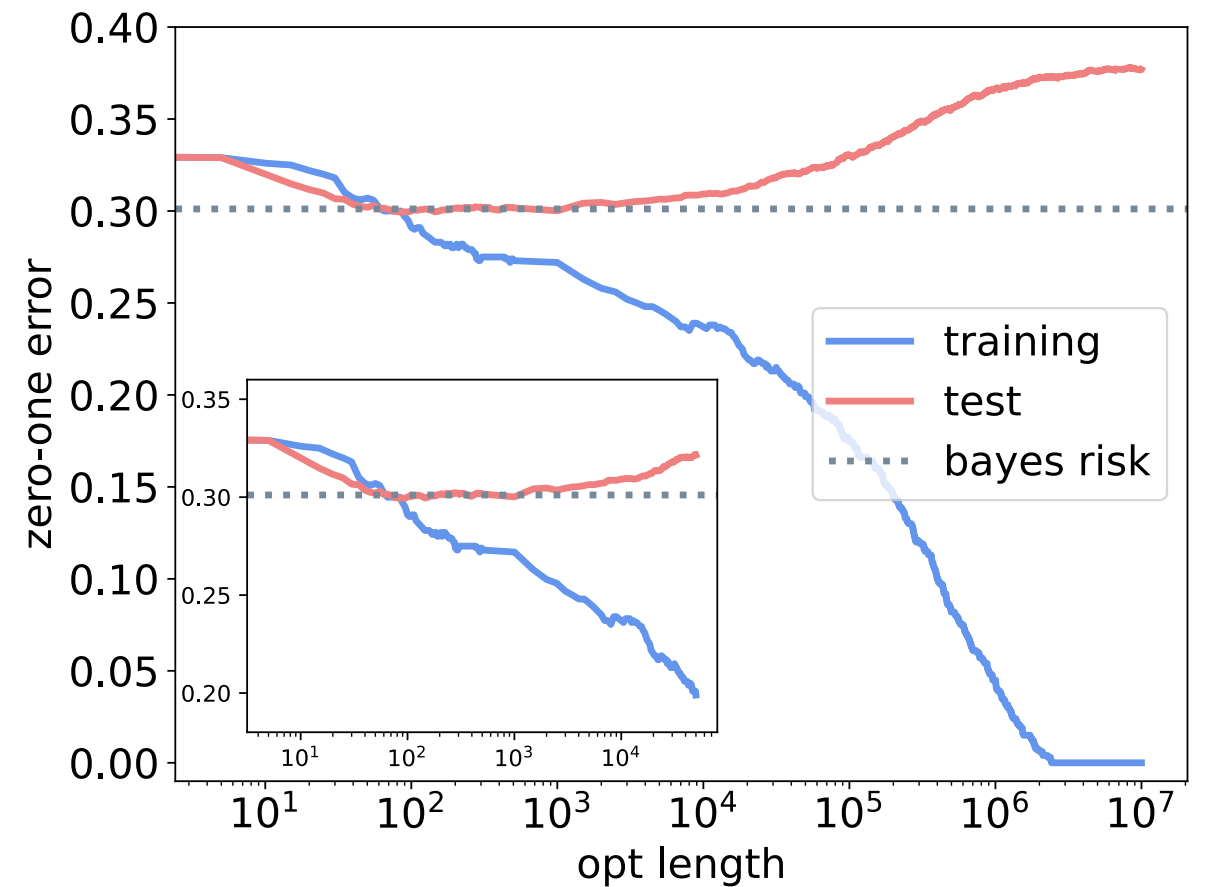
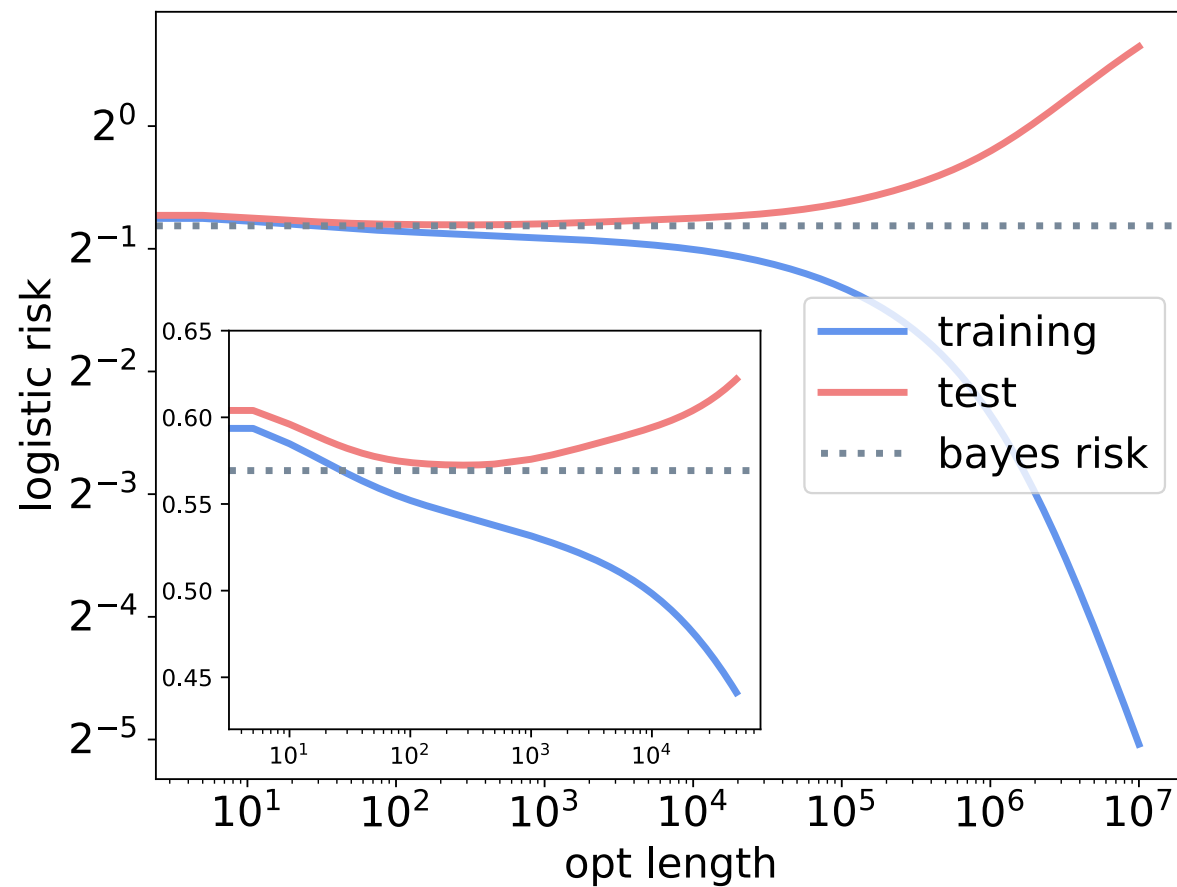
	early-stopped	asymptotic
logistic consistency	always yes	always no
calibration	always yes	always no
zero-one risk	"poly"	"polylog"

GD passes
through w^*



but eventually
diverges from it

Simulations



$$d = 2000, n = 1000, \lambda_i = i^{-2}, w^* = (1, \dots, 1, 0, \dots)$$

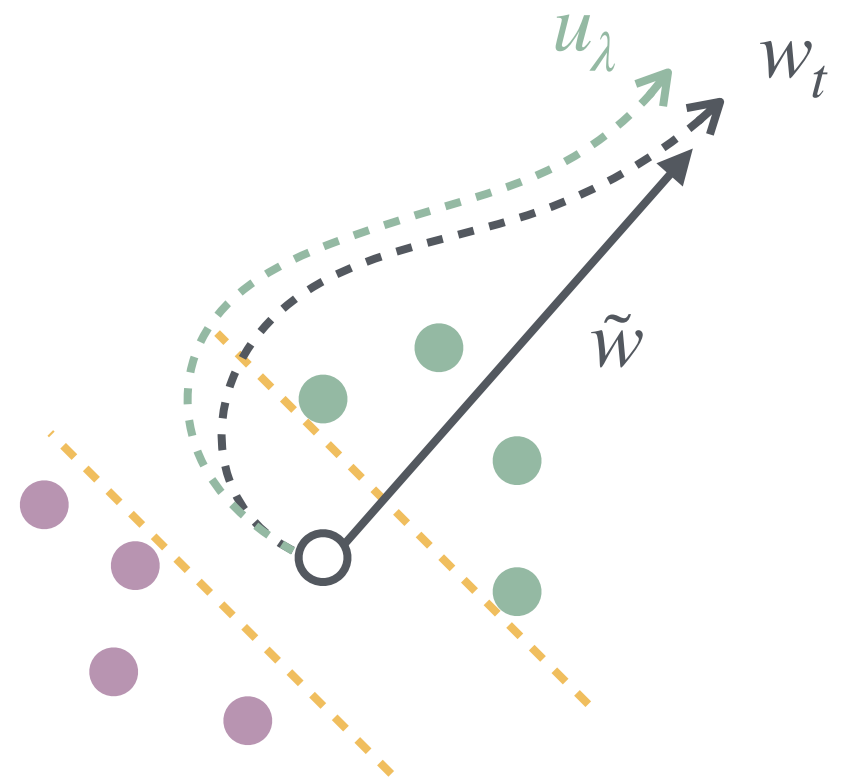
$\underbrace{\hspace{10em}}_{k=100}$

GD and l_2 -regularization

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla \hat{L}(w_t) \\ &= \arg \min \hat{L}(w_t) + \langle \nabla \hat{L}(w_t), u - w_t \rangle + \frac{1}{2\eta} \|u - w_t\|^2 \end{aligned}$$

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

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 $t > 0$,

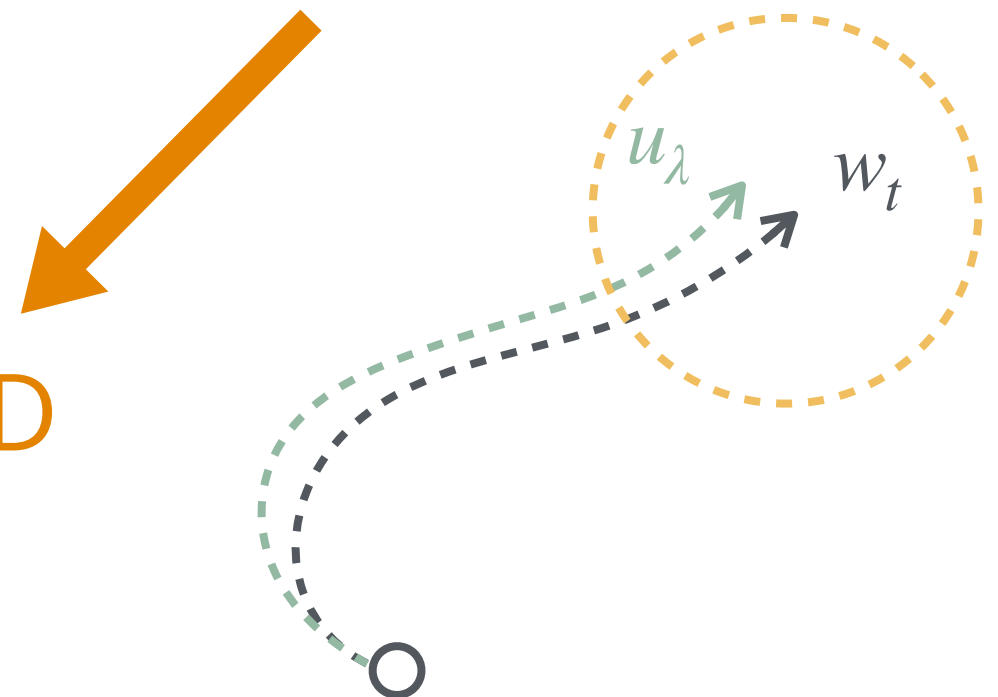
$$\|w_t - u_\lambda\| \leq \frac{1}{\sqrt{2}} \|w_t\| \quad \text{for } \lambda = \eta t$$

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

$\hat{L}$ 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 $\text{rank}\{\text{support vectors}\} = \text{rank}\{\text{data}\}$, then

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

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

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

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

