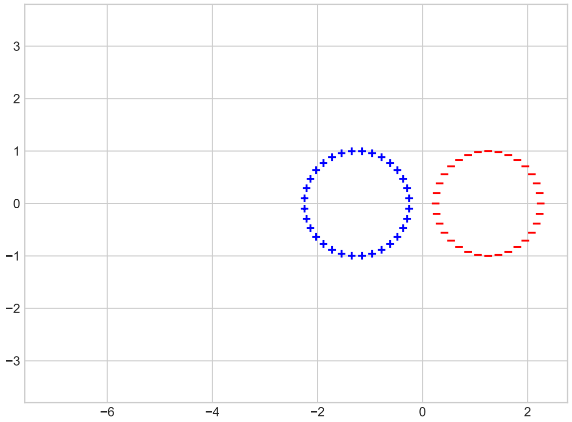
Implicit bias of first-order optimization.

Matus Telgarsky, with special thanks to Ziwei Ji, *Fanny Yang*.

Implicit bias: first-order optimization methods *automatically balance* norm and objective.

- **Old idea:** dates at least to 1962 (Novikoff).
- **Recent interest:** good generalization and other phenomena in deep learning?
- **This talk:**
  - Linear cases: clean proofs and good intuition.
  - Non-linear cases: still a murky mess ☹.



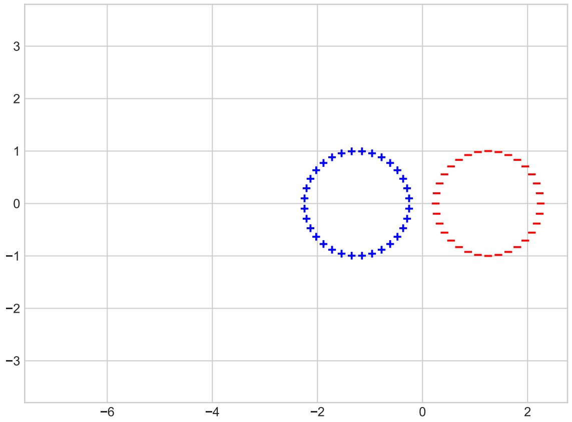
Empirical risk:

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^{\mathsf{T}} w))$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w).$$

Gradient descent:

$$w_{t+1} := w_t - \eta \nabla \widehat{\mathcal{R}}(w_t)$$
$$= \arg\min_{w} \left\{ \left\langle w, \nabla \widehat{\mathcal{R}}(w_t) \right\rangle + \frac{1}{2\eta} \|w - w_t\|^2 \right\}.$$

Separability: $\inf_u \widehat{\mathcal{R}}(u) = 0$.



Empirical risk:

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-y_i x_i^{\mathsf{T}} w))$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w).$$

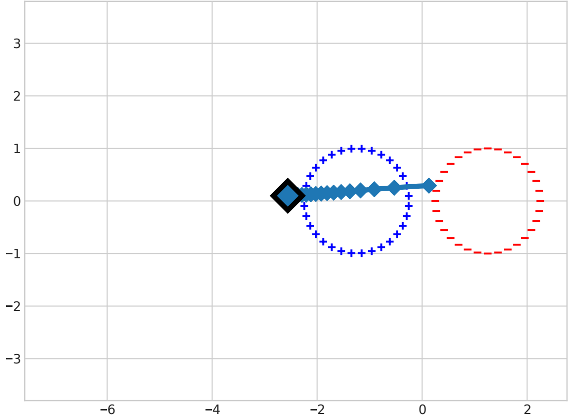
Gradient descent:

$$w_{t+1} := w_t - \eta \nabla \widehat{\mathcal{R}}(w_t)$$
$$= \arg\min_{w} \left\{ \left\langle w, \nabla \widehat{\mathcal{R}}(w_t) \right\rangle + \frac{1}{2\eta}\|w - w_t\|^2 \right\}.$$

Separability: $\inf_u \widehat{\mathcal{R}}(u) = 0.$

Margins:

$$\text{margin}(w_t) := \frac{\min_i y_i x_i^{\mathsf{T}} w_t}{\|w_t\|} \approx \frac{-\ln \sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w_t)}{\|w_t\|}.$$



Empirical risk:

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^{\mathsf{T}} w))$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w).$$

Gradient descent:

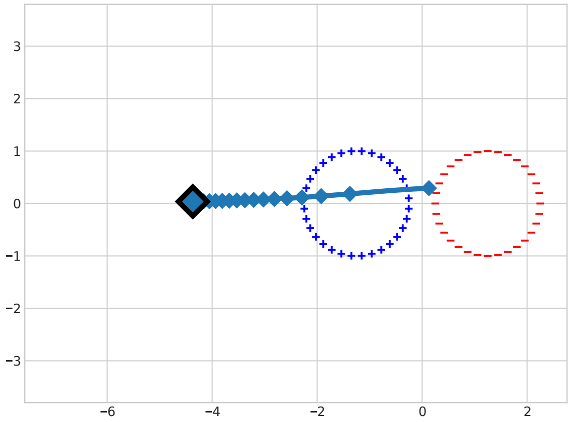
$$w_{t+1} := w_t - \eta \nabla \widehat{\mathcal{R}}(w_t)$$

$$= \arg\min_w \left\{ \left\langle w, \nabla \widehat{\mathcal{R}}(w_t) \right\rangle + \frac{1}{2\eta} \|w - w_t\|^2 \right\}.$$

Separability: $\inf_u \widehat{\mathcal{R}}(u) = 0$.

Margins:

$$\operatorname{margin}(w_t) := \frac{\min_i y_i x_i^{\mathsf{T}} w_t}{\|w_t\|} \approx \frac{-\ln \sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w_t)}{\|w_t\|}.$$



Empirical risk:

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-y_i x_i^{\mathsf{T}} w))$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w).$$

Gradient descent:

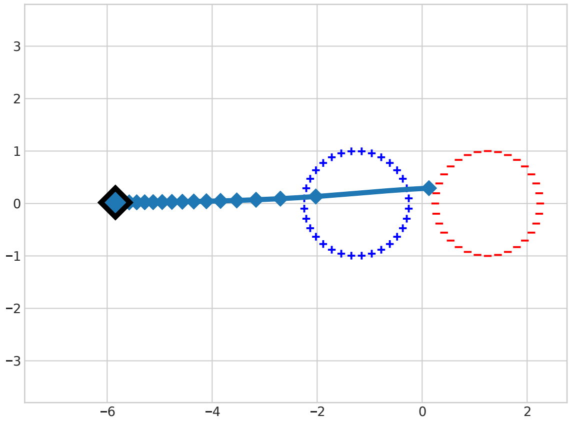
$$w_{t+1} := w_t - \eta \nabla \widehat{\mathcal{R}}(w_t)$$

$$= \arg\min_{w} \left\{ \left\langle w, \nabla \widehat{\mathcal{R}}(w_t) \right\rangle + \frac{1}{2\eta}\|w - w_t\|^2 \right\}.$$

Separability: $\inf_u \widehat{\mathcal{R}}(u) = 0$.

Margins:

$$\operatorname{margin}(w_t) := \frac{\min_i y_i x_i^{\mathsf{T}} w_t}{\|w_t\|} \approx \frac{-\ln \sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w_t)}{\|w_t\|}.$$



Empirical risk:

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^{\mathsf{T}} w))$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w).$$

Gradient descent:

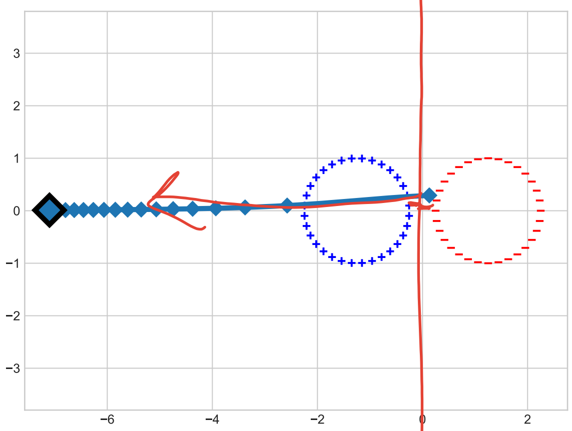
$$w_{t+1} := w_t - \eta \nabla \widehat{\mathcal{R}}(w_t)$$

$$= \arg\min_{w} \left\{ \left\langle w, \nabla \widehat{\mathcal{R}}(w_t) \right\rangle + \frac{1}{2\eta} \|w - w_t\|^2 \right\}.$$

Separability: $\inf_u \widehat{\mathcal{R}}(u) = 0$.

Margins:

$$\operatorname{margin}(w_t) := \frac{\min_i y_i x_i^{\mathsf{T}} w_t}{\|w_t\|} \approx \frac{-\ln \sum_{i=1}^{n} \exp(-y_i x_i^{\mathsf{T}} w_t)}{\|w_t\|}.$$



Empirical risk:

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^{\top} w))$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i x_i^{\top} w).$$

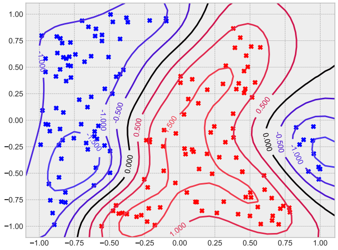
Gradient descent:

$$w_{t+1} := w_t - \eta \nabla \widehat{\mathcal{R}}(w_t)$$
$$= \arg\min_{w} \left\{ \left\langle w, \nabla \widehat{\mathcal{R}}(w_t) \right\rangle + \frac{1}{2\eta} \|w - w_t\|^2 \right\}.$$

Separability: $\inf_u \widehat{\mathcal{R}}(u) = 0.$

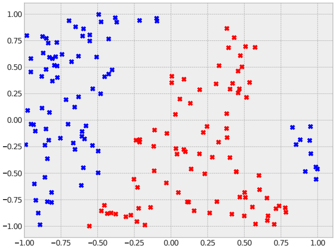
Margins:

$$\text{margin}(w_t) := \frac{\min_i y_i x_i^{\top} w_t}{\|w_t\|} \approx \frac{-\ln \sum_{i=1}^{n} \exp(-y_i x_i^{\top} w_t)}{\|w_t\|}.$$

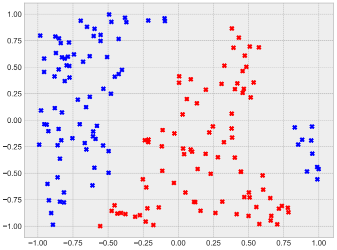


RBF SVM. *Explicitly* solves

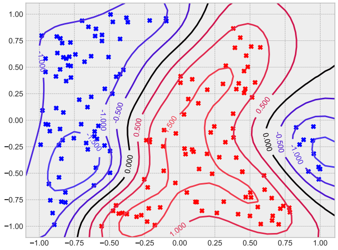
$$
\begin{aligned}
\min \quad & \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\
\text{s.t.} \quad & y_i f(x_i) \geq 1 \quad \forall i.
\end{aligned}
$$



AdaBoost.

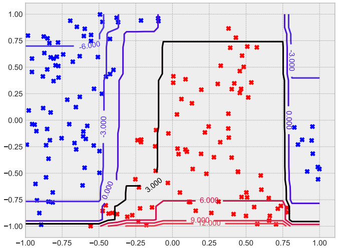


2-layer ReLU.



RBF SVM. *Explicitly* solves

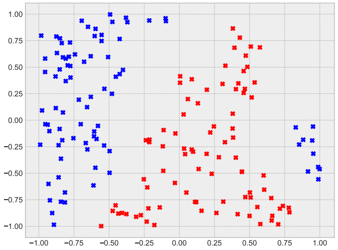
$$\min \quad \frac{1}{2}\|f\|_{\mathcal{H}}^2$$
$$\text{s.t.} \quad y_i f(x_i) \geq 1 \quad \forall i.$$



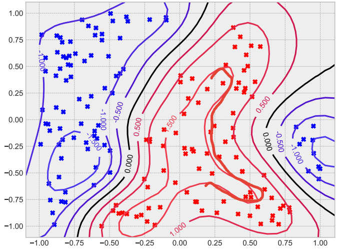
AdaBoost. *Implicitly* solves

$$\min \quad \|w\|_1$$
$$\text{s.t.} \quad y_i \sum_j w_j h_j(x_i) \geq 1 \quad \forall i.$$

[Zhang-Yu '04, T '13].

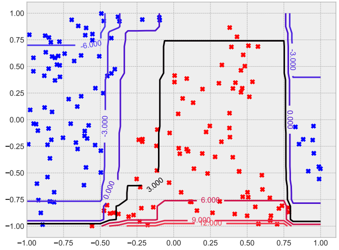


2-layer ReLU.



RBF SVM. *Explicitly* solves

$$\min \quad \frac{1}{2}\|f\|_{\mathcal{H}}^2$$
$$\text{s.t.} \quad y_i f(x_i) \geq 1 \quad \forall i.$$



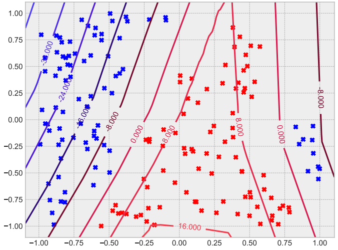
AdaBoost. *Implicitly* solves

$$\min \quad \|w\|_1$$
$$\text{s.t.} \quad y_i \sum_j w_j h_j(x_i) \geq 1 \quad \forall i.$$

[Zhang-Yu '04, T '13].

"decision stump":

$x \mapsto \mathbb{1}[x_i \geq b]$



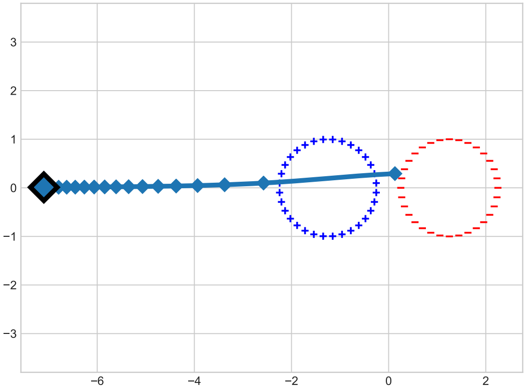
2-layer ReLU.

Situation unclear ☹.

$$\min \quad \frac{1}{2}\|w\|_2^2$$
$$\text{s.t.} \quad y_i F(x_i; w) \geq 1 \quad \forall i$$

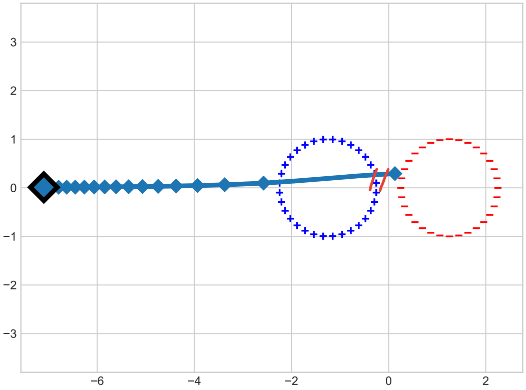
Linear case.

- One family of proof techniques:
  $\ln \sum \exp$ and its dual.
- Two open problems:
  regularization path and logistic.



Theorem *[Ji-T '18]*. For linear predictors,

$$\max_{\|u\|\le 1} \text{margin}(u) - \text{margin}(w_t) = \mathcal{O}(\ln(n)) \cdot \begin{cases} \frac{1}{\ln t} & \text{when } \eta = \mathcal{O}(1), \\ \frac{1}{t} & \text{when } \eta_s = \frac{\mathcal{O}(1)}{\widehat{\mathcal{R}}(w_s)}. \end{cases}$$



separable training set

Theorem *[Ji-T '18]*. For linear predictors, $\|x_i\| \le 1$

$$\max_{\|u\|\le 1} \text{margin}(u) - \text{margin}(w_t)$$

$$= \mathcal{O}(\ln(n)) \cdot \begin{cases} \frac{1}{\ln t} & \text{when } \eta = \mathcal{O}(1), \\ \frac{1}{t} & \text{when } \eta_s = \frac{\mathcal{O}(1)}{\widehat{\mathcal{R}}(w_s)}. \end{cases}$$

γ 0

Jason Altschuler

Remarks.

- **Proof techniques:**
  - Smoothness of $\ln \sum \exp$, rate $1/\sqrt{t}$ *[T '13]*.
  - Rates $\frac{1}{t}$ and $\frac{1}{t^2}$ via duality *[Ji-T '19, Ji-Srebro-T '21]*. (Fastest SVM solvers!)
  - Duality/SVM proof, rate $\frac{1}{\ln(t)}$ *[Soudry-Srebro-etal '17]*.
- Logistic loss presents some difficulties.

$\text{margin}(u) := \min_i \langle y_i x_i, \frac{u}{\|u\|} \rangle$

test error $\lesssim \frac{1}{n\gamma^2}$

T '13 proof technique (for coordinate descent); margin γ with direction $\bar{u}$:

$$\frac{\min_i y_i x_i^\mathsf{T} w_t}{\|w_t\|} \geq \frac{-\ln \sum_i \exp(-y_i x_i^\mathsf{T} w_t)}{\|w_t\|}$$

T '13 proof technique (for coordinate descent); margin γ with direction $\bar{u}$:

$$
\begin{aligned}
\frac{\min_i y_i x_i^{\mathsf T} w_t}{\|w_t\|} &\geq \frac{-\ln\sum_i \exp(-y_i x_i^{\mathsf T} w_t)}{\|w_t\|} \\
&= \frac{\int_0^t \left\langle -\nabla_w \ln \widehat{\mathcal{R}}(w_s), \dot{w}_s \right\rangle ds}{\|w_t\|} - \frac{\ln\sum_i \exp(-y_i x_i^{\mathsf T} w_0)}{\|w_t\|}
\end{aligned}
$$

$$\nabla \ln \sum \exp(-) = \frac{\nabla \exp}{\sum \exp(-)}$$

$$\|w_t\| = \left\| \int_0^t \dot{w}_s \, ds \right\|$$

$$\geq \left\langle \int_0^t -\frac{\sum_i \exp(x_i^{\mathsf T} w_s y_i) x_i y_i}{\sum_i \exp(-y_i x_i^{\mathsf T} w_s)}, \bar{u} \right\rangle \quad (\geq \gamma)$$

$$= t\gamma.$$

T '13 proof technique (for coordinate descent); margin γ with direction $\bar{u}$:

$$
\begin{aligned}
\frac{\min_i y_i x_i^\mathsf{T} w_t}{\|w_t\|} &\geq \frac{-\ln \sum_i \exp(-y_i x_i^\mathsf{T} w_t)}{\|w_t\|} \\
&= \frac{\int_0^t \left\langle -\nabla_w \ln \widehat{\mathcal{R}}(w_s), \dot{w}_s \right\rangle \mathrm{d}s}{\|w_t\|} - \frac{\ln \sum_i \exp(-y_i x_i^\mathsf{T} w_0)}{\|w_t\|} \\
&\geq \frac{\gamma \int_0^t \|\dot{w}_s\| \, \mathrm{d}s}{\int_0^t \|\dot{w}_s\| \mathrm{d}s} - \frac{\ln n}{\left\langle \int_0^t \dot{w}_s \mathrm{d}s, \bar{u} \right\rangle}
\end{aligned}
$$

T '13 proof technique (for coordinate descent); margin γ with direction $\bar{u}$:

$$\begin{aligned}
\frac{\min_i y_i x_i^{\mathsf{T}} w_t}{\|w_t\|} &\geq \frac{-\ln \sum_i \exp(-y_i x_i^{\mathsf{T}} w_t)}{\|w_t\|} \\
&= \frac{\int_0^t \left\langle -\nabla_w \ln \widehat{\mathcal{R}}(w_s), \dot{w}_s \right\rangle \mathrm{d}s}{\|w_t\|} - \frac{\ln \sum_i \exp(-y_i x_i^{\mathsf{T}} w_0)}{\|w_t\|} \\
&\geq \frac{\gamma \int_0^t \|\dot{w}_s\| \, \mathrm{d}s}{\int_0^t \|\dot{w}_s\| \mathrm{d}s} - \frac{\ln n}{\left\langle \int_0^t \dot{w}_s \mathrm{d}s, \bar{u} \right\rangle} \\
&\geq \gamma - \frac{\ln n}{t\gamma}.
\end{aligned}$$

T '13 proof technique (for coordinate descent); margin γ with direction $\bar{u}$:

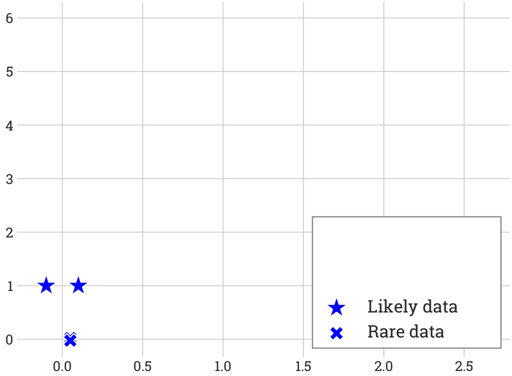
$$
\begin{aligned}
\frac{\min_i y_i x_i^{\top} w_t}{\|w_t\|} &\geq \frac{-\ln \sum_i \exp(-y_i x_i^{\top} w_t)}{\|w_t\|} \\
&= \frac{\int_0^t \left\langle -\nabla_w \ln \widehat{\mathcal{R}}(w_s), \dot{w}_s \right\rangle \mathrm{d}s}{\|w_t\|} - \frac{\ln \sum_i \exp(-y_i x_i^{\top} w_0)}{\|w_t\|} \\
&\geq \frac{\gamma \int_0^t \|\dot{w}_s\| \mathrm{d}s}{\int_0^t \|\dot{w}_s\| \mathrm{d}s} - \frac{\ln n}{\left\langle \int_0^t \dot{w}_s \mathrm{d}s, \bar{u} \right\rangle} \\
&\geq \gamma - \frac{\ln n}{t\gamma}.
\end{aligned}
$$

Remarks.

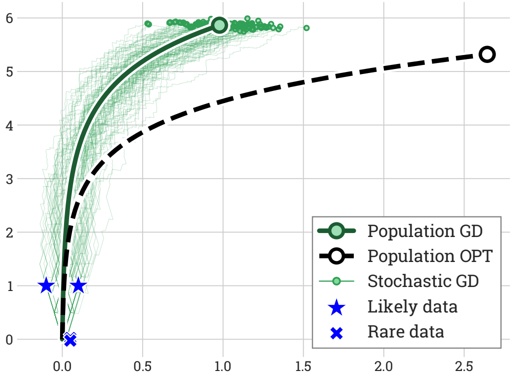
- Stated for GD by *[Gunasekar-Lee-Srebro-etal '18, Ji-T '18]*; different proof slower rate *[Soudry-Srebro-etal '17]*.
- Rate $\frac{1}{t}$ with CD impossible, with GD hard (needs duality?).
- For logistic use $\ell^{-1}(\sum_i \ell_i)$; needs burn-in.

1. get positive margin
2. do this
Jingfeng Wu - Jason Lee '23
$\ln \sum \exp(-)$
ℓ^{-1}, ℓ
Kaifeng Lyu - Li
"smooth margin"
$\ell^{-1}(\sum_i \ell(v_i))$

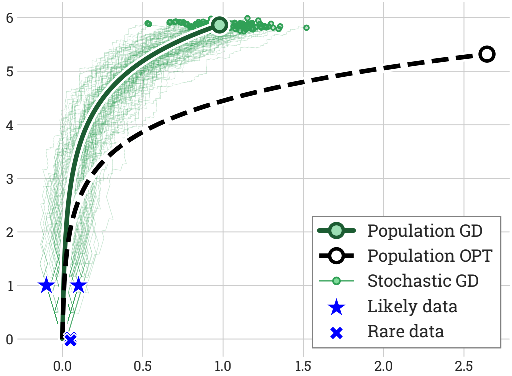
Open #1: intermediate solutions.



Open #1: intermediate solutions.



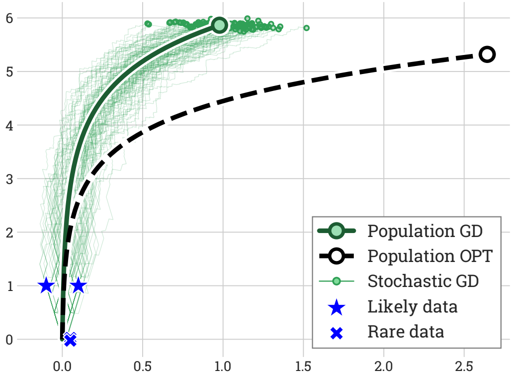
Open #1: intermediate solutions.



GD follows regularization path

For regression *[Efron et al. '04, Rosset-Zhu '07]*; also GD/MD/RL/. . . *[T '23, Hu-Ji-T '22]*.

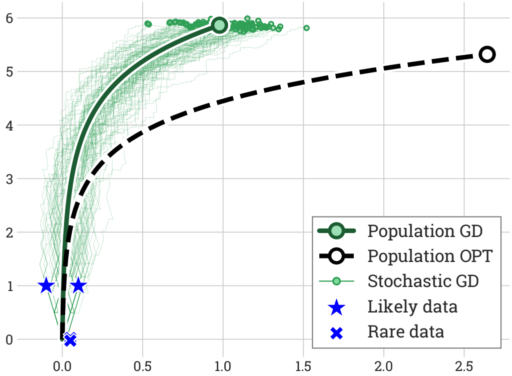
Open #1: intermediate solutions.



GD follows regularization path
For regression *[Efron et al. '04, Rosset-Zhu '07]*;
also GD/MD/RL/... *[T '23, Hu-Ji-T '22]*.

Proof technique: perceptron.

Open #1: intermediate solutions.



GD follows regularization path
For regression *[Efron et al. '04, Rosset-Zhu '07]*;
also GD/MD/RL/. . . *[T '23, Hu-Ji-T '22]*.

Proof technique: perceptron.

Open: tighten gap between GD and optimal path.

Perceptron technique (MD): Smooth/convex $\widehat{\mathcal{R}}$, any z, any $\eta \leq 1/\beta$,

$$\begin{aligned}\|w_{s+1} - z\|^2 - \|w_s - z\|^2 &= 2\eta \left\langle \nabla\widehat{\mathcal{R}}(w_s), z - w_s \right\rangle + \eta^2 \|\nabla\widehat{\mathcal{R}}(w_s)\|^2 \\ &\leq 2\eta \left(\nabla\widehat{\mathcal{R}}(z) - \nabla\widehat{\mathcal{R}}(w_s) + \nabla\widehat{\mathcal{R}}(w_s) - \nabla\widehat{\mathcal{R}}(w_{s+1}) \right),\end{aligned}$$

Perceptron technique (MD): Smooth/convex $\widehat{\mathcal{R}}$, any z, any $\eta \leq 1/\beta$,

$$\begin{aligned}\|w_{s+1} - z\|^2 - \|w_s - z\|^2 &= 2\eta \left\langle \nabla\widehat{\mathcal{R}}(w_s), z - w_s \right\rangle + \eta^2 \|\nabla\widehat{\mathcal{R}}(w_s)\|^2 \\ &\leq 2\eta \left(\nabla\widehat{\mathcal{R}}(z) - \nabla\widehat{\mathcal{R}}(w_s) + \nabla\widehat{\mathcal{R}}(w_s) - \nabla\widehat{\mathcal{R}}(w_{s+1}) \right),\end{aligned}$$

which implies

$$\frac{1}{2\eta t}\|w_t - z\|^2 + \widehat{\mathcal{R}}(w_t) \leq \frac{1}{2\eta t}\|w_0 - z\|^2 + \widehat{\mathcal{R}}(z).$$

Perceptron technique (MD): Smooth/convex $\widehat{\mathcal{R}}$, any z, any $\eta \leq 1/\beta$,

$$\begin{aligned} \|w_{s+1} - z\|^2 - \|w_s - z\|^2 &= 2\eta \left\langle \nabla\widehat{\mathcal{R}}(w_s), z - w_s \right\rangle + \eta^2 \|\nabla\widehat{\mathcal{R}}(w_s)\|^2 \\ &\leq 2\eta \left(\nabla\widehat{\mathcal{R}}(z) - \nabla\widehat{\mathcal{R}}(w_s) + \nabla\widehat{\mathcal{R}}(w_s) - \nabla\widehat{\mathcal{R}}(w_{s+1}) \right), \end{aligned}$$

which implies

$$\frac{1}{2\eta t}\|w_t - z\|^2 + \widehat{\mathcal{R}}(w_t) \leq \frac{1}{2\eta t}\|w_0 - z\|^2 + \widehat{\mathcal{R}}(z).$$

Alternatively: if t is final iterate with $\widehat{\mathcal{R}}(w_t) > \widehat{\mathcal{R}}(z)$, then $\|w_t - w_0\| \leq 2\|z - w_0\|$.

Perceptron technique (MD): Smooth/convex $\widehat{\mathcal{R}}$, any z, any $\eta \le 1/\beta$,

$$\begin{aligned}\|w_{s+1} - z\|^2 - \|w_s - z\|^2 &= 2\eta \left\langle \nabla\widehat{\mathcal{R}}(w_s), z - w_s \right\rangle + \eta^2 \|\nabla\widehat{\mathcal{R}}(w_s)\|^2 \\ &\le 2\eta \left(\nabla\widehat{\mathcal{R}}(z) - \nabla\widehat{\mathcal{R}}(w_s) + \nabla\widehat{\mathcal{R}}(w_s) - \nabla\widehat{\mathcal{R}}(w_{s+1}) \right),\end{aligned}$$

which implies

$$\frac{1}{2\eta t}\|w_t - z\|^2 + \widehat{\mathcal{R}}(w_t) \le \frac{1}{2\eta t}\|w_0 - z\|^2 + \widehat{\mathcal{R}}(z).$$

Alternatively: if t is final iterate with $\widehat{\mathcal{R}}(w_t) > \widehat{\mathcal{R}}(z)$, then $\|w_t - w_0\| \le 2\|z - w_0\|$.

Remark.

- Allows near-initialization analysis of neural networks;
  sample complexity, iterations, width $\frac{1}{\gamma^2_{\text{rkhs}}}$
  (sometimes optimal) *[Ji-T '18]*.
  Fails for squared loss.
- Also grants *consistency* of neural networks:
  fit *any* Borel-measurable $\Pr[y = 1|X = x]$ via early-stopping *[Ji-Li-T '20]*.

Open #2: logistic.

$\|\nabla f(x) - \nabla f(y)\| \le \beta \|x-y\|$

Show benefit over exponential ☹.

f is β-smooth convex,
GD w/ step size $1/\beta$
for any reference solution $\tilde{w}$,
$\exists t$ s.t. $f(w_t) \approx f(\tilde{w})$
and $\|w_t - w_0\| \le 2 \|\tilde{w} - w_0\|$ (2 → 1.2.)

Summary for Linear

* spelling
* $\frac{1}{t}$, $\frac{1}{\ln t}$ rates
  $\ln \Sigma \exp \approx \min$
* OPEN
  * logistic ???
  * early phase / regularization path

Nonlinear cases.

Setup. Feedforward networks

$$x \mapsto F(x; w) := \sigma_L(W_L \sigma_{L-1}(\cdots W_2 \sigma_1(W_1 x) \cdots)),$$

where

- σ_i are coordinate-wise and positive-homogeneous;
- $(W_L, \ldots, W_1)$ are trained.

Nonlinear cases.

Setup. Feedforward networks

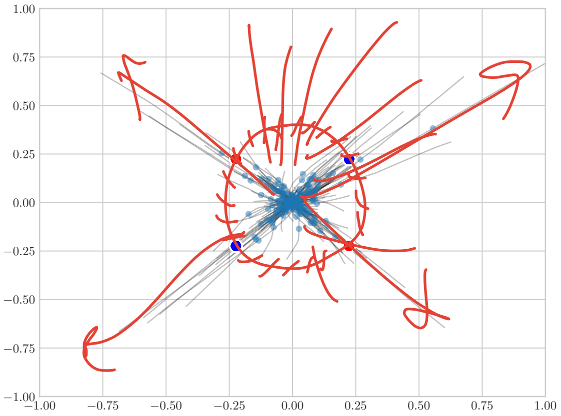
$$x \mapsto F(x; w) := \sigma_L(W_L\sigma_{L-1}(\cdots W_2\sigma_1(W_1x)\cdots)),$$

where

- σ_i are coordinate-wise and positive-homogeneous;
- $(W_L, \ldots, W_1)$ are trained.

Margin concept: still SVM, meaning

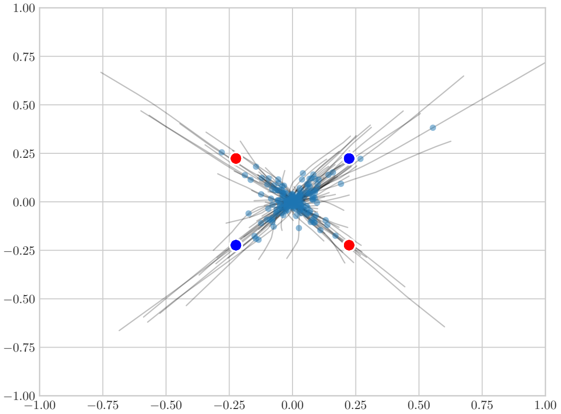
$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|_2^2 \\ \text{subject to} \quad & \forall i \centerdot y_i F(x_i; w) \geq 1. \end{aligned}$$



Motivating example:
2XOR, popularized by Wei-Lee-Liu-Ma '18.

- $x \sim \text{Uniform}\left(\left\{\frac{\pm 1}{\sqrt{d}}\right\}^d\right)$.
- $y = d x_1 x_2$. $\in \{\pm 1\}$
- Dot product kernel lower bound (including ReLU NTK): $\frac{d^2}{\epsilon}$.
- 4 ReLU global max margin solution: $\frac{d}{\epsilon}$.

$\frac{2 \ln d}{\epsilon}$

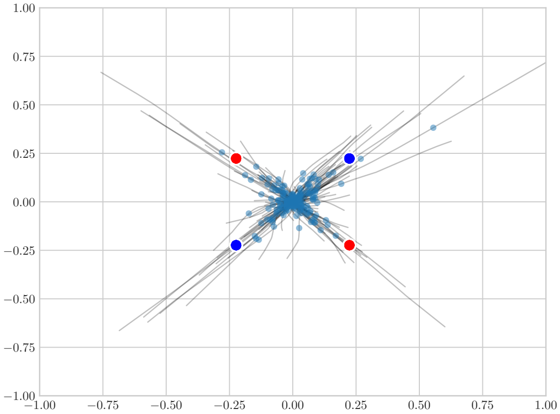


Current status:

- One specialized proof (Margalit Glasgow '23).
- General proofs need modified setting:
  - 2 time scales (Abbe, Bruna, Lee, . . .).
  - Low rotation (Gunasekar, Chizat-Bach, Telgarsky, . . .).
  - Mass concentrates (Chizat-Bach, Telgarsky, . . .).

$$\left(\frac{\exp(y_i f(x_i; w))}{\sum_j \exp(y_j f(x_j; w))}, \dots \right)_{i=1}^{n}$$

assume converges



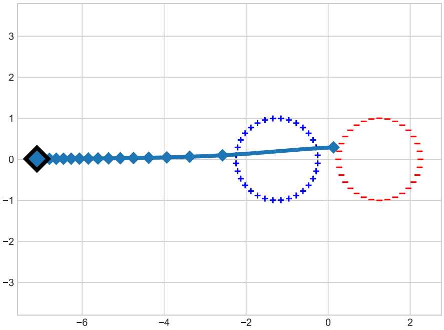
Current status:

- One specialized proof (Margalit Glasgow '23).
- General proofs need modified setting:
  - 2 time scales (Abbe, Bruna, Lee, . . .).
  - Low rotation (Gunasekar, Chizat-Bach, Telgarsky, . . .).
  - Mass concentrates (Chizat-Bach, Telgarsky, . . .).

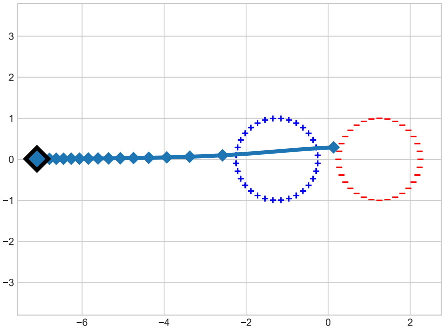
Today:

- Convergence to locally maximal margins.
- Mass concentrates.

Linear case rephrased.



Linear case rephrased.



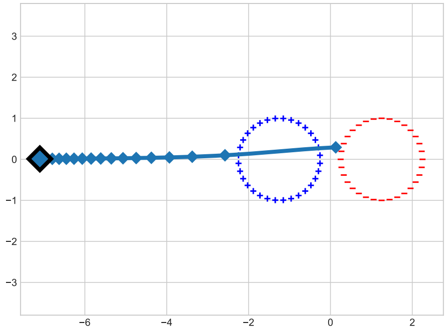
Asymptotics:

$$\frac{w_t}{\|w_t\|} \quad \text{converges,}$$

$$\frac{w_t}{\|w_t\|} \to \text{KKT of } \begin{cases} \min \|w\|^2 \\ \text{s.t. } \; y_i x_i^{\mathsf{T}} w \geq 1 \quad \forall i, \end{cases}$$

$$\left\langle \frac{w_t}{\|w_t\|}, \frac{-\nabla \widehat{\mathcal{R}}(w_t)}{\|\nabla \widehat{\mathcal{R}}(w_t)\|} \right\rangle \to 1.$$

Linear case rephrased.



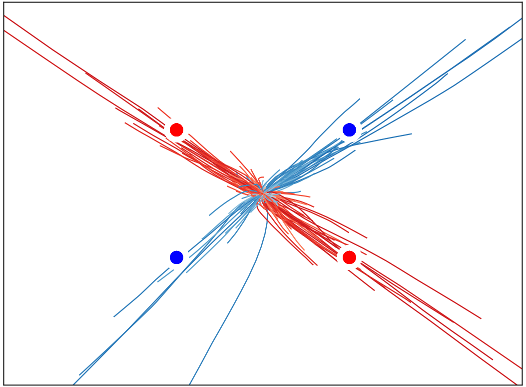
Asymptotics:

$$\frac{w_t}{\|w_t\|} \quad \text{converges,}$$

$$\frac{w_t}{\|w_t\|} \to \text{KKT of } \begin{cases} \min & \|w\|^2 \\ \text{s.t.} & y_i x_i^\top w \geq 1 \quad \forall i, \end{cases}$$

$$\left\langle \frac{w_t}{\|w_t\|}, \frac{-\nabla \widehat{\mathcal{R}}(w_t)}{\|\nabla \widehat{\mathcal{R}}(w_t)\|} \right\rangle \to 1.$$

Homogeneous networks.

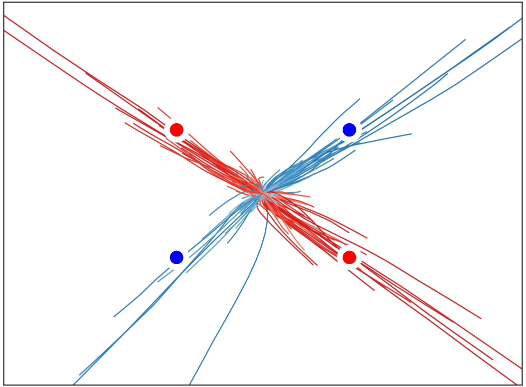


Homogeneous network (e.g., ReLU)

$$x \mapsto \underbrace{\sigma_L(W_L\sigma_{L-1}(\cdots\sigma_1(W_1x)\cdots))}_{F(x;W)},$$

separable ($\inf_t \widehat{\mathcal{R}}(w_t) \leq \frac{1}{n}$).

Homogeneous networks.



Homogeneous network (e.g., ReLU)

$$x \mapsto \underbrace{\sigma_L(W_L \sigma_{L-1}(\cdots \sigma_1(W_1 x)\cdots))}_{F(x;W)},$$

separable ($\inf_t \widehat{\mathcal{R}}(w_t) \le \frac{1}{n}$).

Theorem *[Lyu-Li '19, Ji-T '20]*.

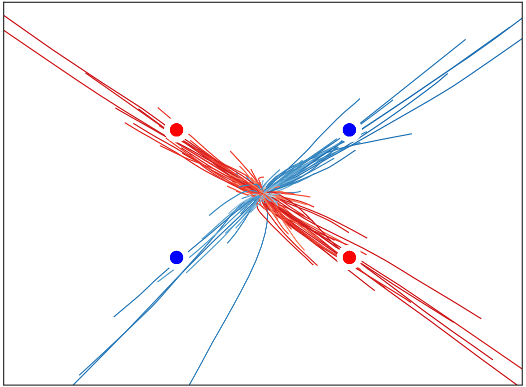
1. Under regularity (e.g., σ_i is ReLU)

$$\frac{w_t}{\|w_t\|} \to \text{KKT of } \begin{cases} \min & \|w\|^2 \\ \text{s.t.} & y_i F(x_i; w) \ge 1 \quad \forall i. \end{cases}$$

2. Under other regularity (e.g., σ_i is ReLU^2)

$$\left\langle \frac{w_t}{\|w_t\|}, \frac{-\nabla\widehat{\mathcal{R}}(w_t)}{\|\nabla\widehat{\mathcal{R}}(w_t)\|} \right\rangle \to 1.$$

Homogeneous networks.



Homogeneous network (e.g., ReLU)

$$x \mapsto \underbrace{\sigma_L(W_L\sigma_{L-1}(\cdots\sigma_1(W_1x)\cdots))}_{F(x;W)},$$

separable ($\inf_t \widehat{\mathcal{R}}(w_t) \leq \frac{1}{n}$).

Theorem *[Lyu-Li '19, Ji-T '20]*.

1. Under regularity (e.g., σ_i is ReLU)

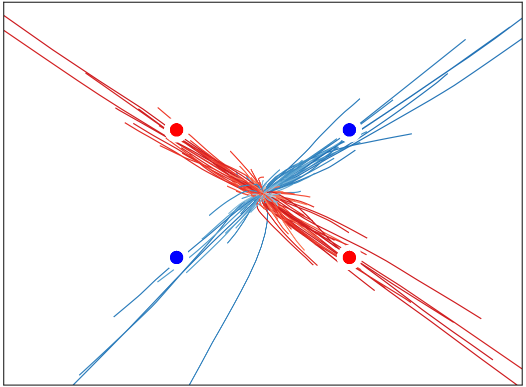
$$\frac{w_t}{\|w_t\|} \to \text{KKT of} \begin{cases} \min & \|w\|^2 \\ \text{s.t.} & y_i F(x_i;w) \geq 1 \quad \forall i. \end{cases}$$

2. Under other regularity (e.g., σ_i is ReLU^2)

$$\left\langle \frac{w_t}{\|w_t\|}, \frac{-\nabla\widehat{\mathcal{R}}(w_t)}{\|\nabla\widehat{\mathcal{R}}(w_t)\|} \right\rangle \to 1.$$

Proof: 1. nonsmooth o-minimality at infinity.
2. Nonlinear version of linear dual.

Homogeneous networks.



Homogeneous network (e.g., ReLU)

$$x \mapsto \underbrace{\sigma_L(W_L\sigma_{L-1}(\cdots\sigma_1(W_1x)\cdots))}_{F(x;W)},$$

separable ($\inf_t \widehat{\mathcal{R}}(w_t) \leq \frac{1}{n}$).

Theorem *[Lyu-Li '19, Ji-T '20]*.

1. Under regularity (e.g., σ_i is ReLU)

$$\frac{w_t}{\|w_t\|} \to \text{KKT of } \begin{cases} \min & \|w\|^2 \\ \text{s.t.} & y_i F(x_i; w) \geq 1 \quad \forall i. \end{cases}$$

2. Under other regularity (e.g., σ_i is ReLU^2)

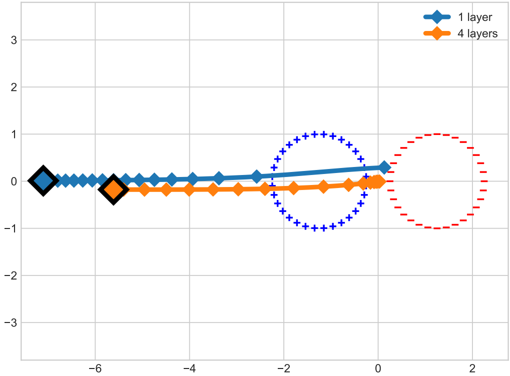
$$\left\langle \frac{w_t}{\|w_t\|}, \frac{-\nabla\widehat{\mathcal{R}}(w_t)}{\|\nabla\widehat{\mathcal{R}}(w_t)\|} \right\rangle \to 1.$$

Proof: 1. nonsmooth o-minimality at infinity.
2. Nonlinear version of linear dual.

Remarks.

- Extends brilliant prior work *[Lyu-Li '18]*.
- KKT of *implicit objective*.
- Decouples linear max margin condition.
- Convenient tool *[Frei, Bartlett, Srebro, Vardi, ...]*.

Deep linear.



Deep linear network:

$$x \mapsto W_L \cdots W_2 W_1 x,$$

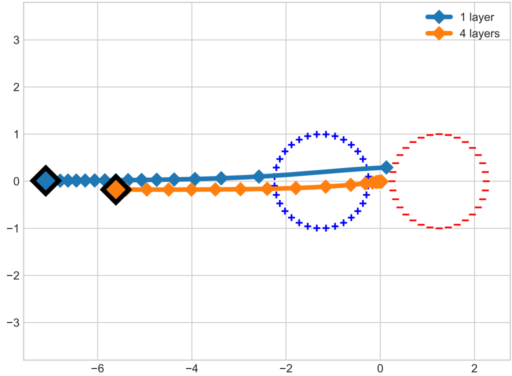
linearly separable.

Corollary *[Ji-T '18, Ji-T '20]*.
Exist unit vectors $(v_L, \ldots, v_0)$ with $v_L = 1$ and $v_0 = \pm$max margin,

$$\frac{W_L(t)\cdots W_1(t)}{\|W_L(t)\cdots W_1(t)\|} \xrightarrow{t\to\infty} \text{max margin}^{\mathsf{T}},$$

$$\forall j\,.\qquad \frac{W_j(t)}{\|W_j(t)\|} \xrightarrow{t\to\infty} v_j v_{j-1}^{\mathsf{T}}.$$

Deep linear.



Deep linear network:

$$x \mapsto W_L \cdots W_2 W_1 x,$$

linearly separable.

Corollary *[Ji-T '18, Ji-T '20]*.
Exist unit vectors $(v_L, \ldots, v_0)$ with $v_L = 1$ and $v_0 = \pm$max margin,

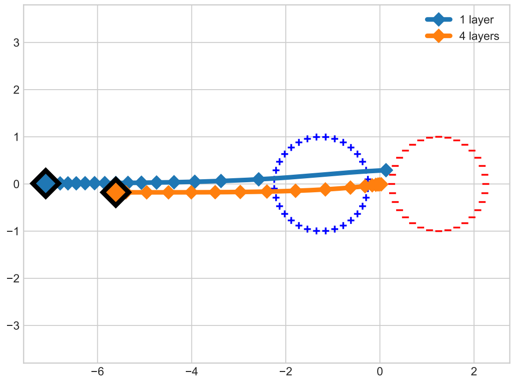
$$\frac{W_L(t)\cdots W_1(t)}{\|W_L(t)\cdots W_1(t)\|} \xrightarrow{t\to\infty} \text{max margin}^{\mathsf{T}},$$

$$\forall j \,.\quad \frac{W_j(t)}{\|W_j(t)\|} \xrightarrow{t\to\infty} v_j v_{j-1}^{\mathsf{T}}.$$

Proof: preceding and linear algebra.

Alternate weaker proof: *[Ji-T '18]*, relation between $W_i W_i^{\mathsf{T}}$ and $W_{i+1}^{\mathsf{T}} W_{i+1}$, plus tons of magic.

Deep linear.



Deep linear network:

$$x \mapsto W_L \cdots W_2 W_1 x,$$

linearly separable.

Corollary *[Ji-T '18, Ji-T '20]*.
Exist unit vectors $(v_L, \ldots, v_0)$ with $v_L = 1$ and $v_0 = \pm$max margin,

$$\frac{W_L(t)\cdots W_1(t)}{\|W_L(t)\cdots W_1(t)\|} \xrightarrow{t\to\infty} \text{max margin}^{\mathsf{T}},$$

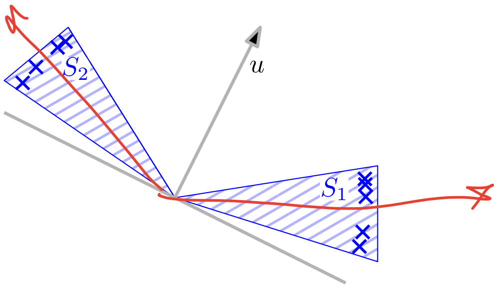
$$\forall j\,. \qquad \frac{W_j(t)}{\|W_j(t)\|} \xrightarrow{t\to\infty} v_j v_{j-1}^{\mathsf{T}}.$$

Proof: preceding and linear algebra.

Alternate weaker proof: *[Ji-T '18]*, relation between $W_i W_i^{\mathsf{T}}$ and $W_{i+1}^{\mathsf{T}} W_{i+1}$, plus tons of magic.

Prior/parallel work:
GD path assumptions.

Mass concentrates.

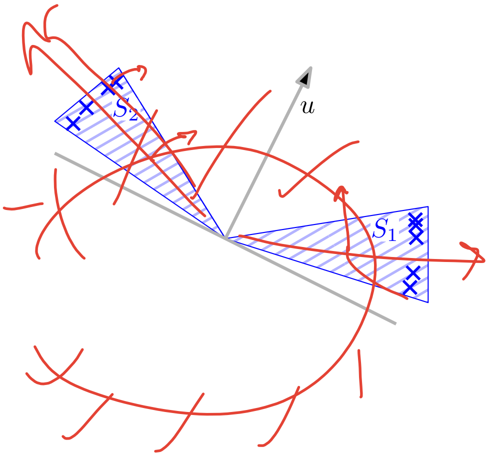


From (Telgarsky '23):
Exists linearly separable data such that:

- Single ReLU is a bad KKT point;
- GF selects good KKT point (two ReLU);
- Small perturbation of data switches good/bad KKT.

$$\mathrm{softmax}(X^\top KQX)$$

Mass concentrates.



From (Telgarsky '23):
Exists linearly separable data such that:

- Single ReLU is a bad KKT point;
- GF selects good KKT point (two ReLU);
- Small perturbation of data switches good/bad KKT.

Related question:

- "Simplicity bias"?

Today's talk:

- **Linear case:**
  - $\frac{1}{t}$ rates, primal/dual proofs.
  - Open: regularization path, logistic.
- **Nonlinear cases:**
  - KKT points and some situations where we escape.
  - Open: reliable general proof technique or intuition!
  - Open: beyond two-layer ReLU...
  - zero- feature learning

Jingfeng Wu Jason Lee

Christos

Samet Oymak

Today's talk:

- **Linear case:**
  - $\frac{1}{t}$ rates, primal/dual proofs.
  - Open: regularization path, logistic.
- **Nonlinear cases:**
  - KKT points and some situations where we escape.
  - Open: reliable general proof technique or intuition!
  - Open: beyond two-layer ReLU...

Thank you!