Wednesday 07/28: TA session 1 (THEODOR MISIAKIEWICZ) (STANFORD)

Complement to ANDREA MONTANARI's lectures

① Today: Implicit regularization

② Friday: "Benign overfitting in linear regression" [Bartlett, Long, Lugosi, Tsigler, 2019]

③ Next Tuesday: Going beyond the linear regime (?)

① IMPLICIT REGULARIZATION (or BIAS)

ANDREA's lectures: linear regime

→ implicit regularization of GD: minimum ℓ_2-norm interpolating solution

⇒ More generally: IR plays a crucial role in the generalization properties of overparametrized models

Setting:

- Supervised learning: $\{(y_i, x_i)\}_{i \le n}$ data

$x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ (or $y_i \in \{\pm 1\}$)

- Parametric family of models: $\mathcal{F} = \{ f(\cdot, \theta) : \theta \in \mathbb{R}^p \}$

$f(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}$

$x \longmapsto f(x, \theta)$ (θ: parameters, $\theta \in \mathbb{R}^p$; x: input)

- Learning: * loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

* Empirical Risk Minimization (ERM):

$$\text{minimize } \hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i, \theta))$$

Modern approach: GD / SGD θ^t

$\hat{R}_n(\theta^t) \longrightarrow 0$ ("INTERPOLATION")

$\sum_{i=1}^{n} f(x_i, \theta_*)$

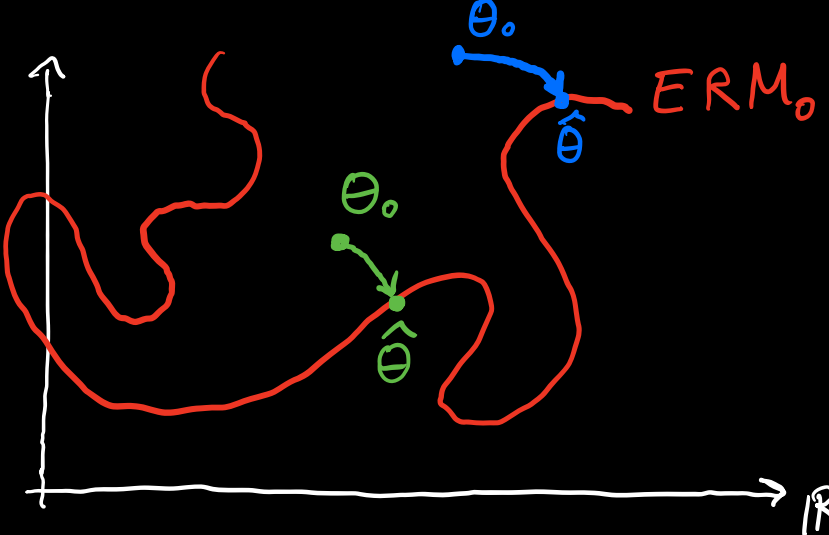
$(\langle x_i, \theta_* \rangle - \langle x_i, \theta \rangle)^2$

$\| X(\theta_* - \theta) \|_2^2 + \lambda \|\theta\|_2^2$ (X: $n \times p$ matrix, θ: p vector)

Overparametrized models: $\hat{f}(x_i, \theta) = y_i$

$n \ll p$ $\quad \theta = \theta_* + v$

$\mathrm{ERM}_0 := \{ \theta \in \mathbb{R}^p : \hat{R}_n(\theta) = 0 \}$ is large ($p - n$ subspace, $Xv = 0$)



Empirical minimizers generalize more or less well

→ Opt algo introduce a bias in this choice

$\Rightarrow$ Opt algorithm selects solution $\hat{\theta} \in \mathrm{ERM}_0$

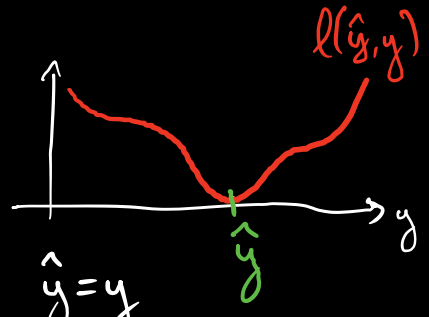
- This bias is really understood in only a few models.
- Here focus on linear models.

"Characterizing implicit bias in terms of optimization geometry"
— Gunasekar, Lee, Soudry, Srebro (2018)

- Linear models: $f(x, \theta) = \langle x, \theta \rangle \qquad x, \theta \in \mathbb{R}^p$
- ERM: $\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle)$

2 family of losses:

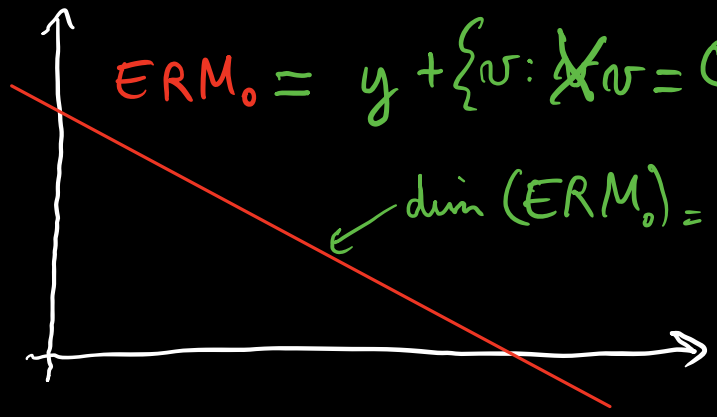
① Regression setting: $\ell(y, \hat{y}) \geq 0$
$= 0$ iff $\hat{y} = y$



$$ERM_0 = \{\theta \in \mathbb{R}^p : \langle x_i, \theta \rangle = y_i \quad \forall i \leq n\}$$
$$= \{\theta \in \mathbb{R}^p : X\theta = y\}$$

$y = (y_1 \dots y_n) \in \mathbb{R}^n$

$X = \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times p}$
$\leftarrow p \rightarrow$

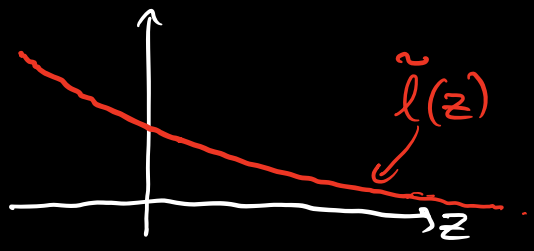


(if X full row rank)

② Classification setting: $y \in \{-1, +1\}$ $\quad \ell(\hat{y}, y) = \tilde{\ell}(\hat{y} y)$

$\tilde{\ell}(z) > 0$, $\ell(z) \to 0$ iff $z \to \infty$

monotonically decreasing



e.g.: $\tilde{\ell}(z) = e^{-z}$ (exp loss)

$\tilde{\ell}(z) = \log(1 + e^{-z})$ (logistic)

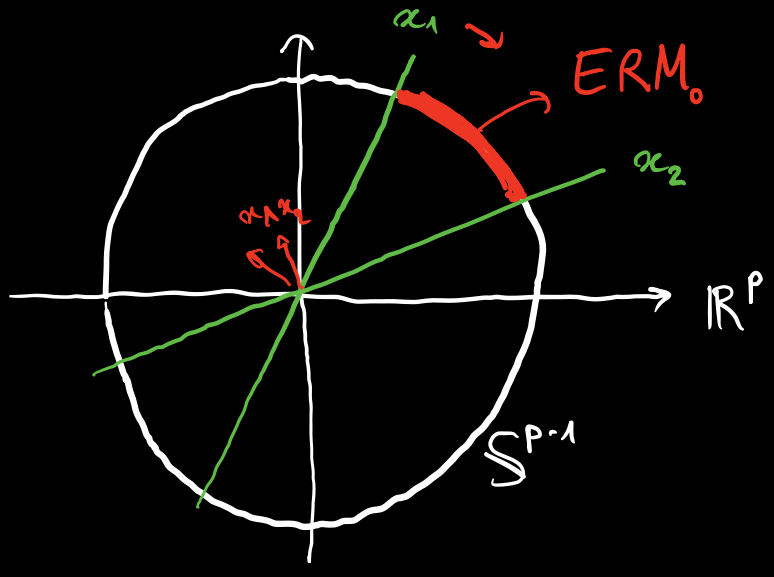
→ In this case: $\hat{R}_n(\theta) > 0 \quad \forall \theta$ finite

→ Prediction: $\hat{f}(x, \theta) = \underline{\text{sign}(\langle \theta, x \rangle)} = \text{sign}\left(\langle \underbrace{\frac{\theta}{\|\theta\|_2}}_{= \xi}, x \rangle\right)$

$\xi \in \mathbb{S}^{p-1}$: unit sphere in p dimension

→ $\hat{f}(x_i, \theta)\, y_i \geq 0$

$ERM_0' := \{\xi \in \mathbb{S}^{p-1} : \langle \xi, x_i \rangle y_i \geq 0 \ \forall i \leq n\}$



① Regression setting: (assume ℓ is differentiable)

Mirror descent (MD):

- potential: $\psi: \mathbb{R}^p \to \mathbb{R}$ differentiable + strictly convex
- Bregman divergence w.r.t. ψ:

$$D_\psi(\theta, \theta_0) := \psi(\theta) - \psi(\theta_0) - \langle \nabla\psi(\theta_0), \theta - \theta_0 \rangle \geq 0$$

$D_\psi(\cdot, \theta_0)$: strictly convex with unique minimizer at $\theta = \theta_0$

- MD algo:
  - initialization θ_0
  - step size η_t
  - update:

$$\theta^{t+1} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \eta_t \langle \theta, \nabla \hat{R}_n(\theta_t) \rangle + D_\psi(\theta, \theta^t) \right\}$$

Example: 1) $\psi(\theta) = \frac{1}{2}\|\theta\|_2^2$ $\quad D_\psi(\theta, \theta_0) = \frac{1}{2}\|\theta - \theta_0\|_2^2$

$$\theta^{t+1} = \theta^t - \eta_t \nabla \hat{R}_n(\theta^t)$$ Gradient Descent

2) $\theta \in \mathbb{R}^p_{>0}$ (positive orthant) $\quad \Delta = \{\alpha \in \mathbb{R}^p_{>0} : 1^T\alpha = 0\}$

$$\psi(\theta) = \sum_{i=1}^p \theta_i \log \theta_i \quad \leftarrow \alpha \log \alpha$$

* Lots of work on conditions such that MD cv to global optimum

* Question here: which point do we converge to?

Proposition: Assume $\{\theta : X\theta = y\} \neq \emptyset$. ($\exists$ interpolator)

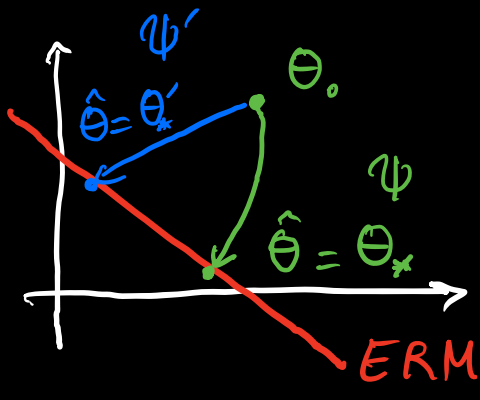
Consider MD with initialization θ^0.

Assume $\hat{R}_n(\theta^t) \xrightarrow[t\to\infty]{} 0$ and $(\theta^t)_{t\geq 1}$ remains bounded

Then:

$$\lim_{t\to\infty} \theta^t = \theta_* := \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \{D_\psi(\theta,\theta_0) : \hat{R}_n(\theta) = 0\}$$

Rmk:



$\psi' \neq \psi$

Proof: KKT conditions of $\theta_* = \operatorname{argmin}\{D_\psi(\theta,\theta_0) : \hat{R}_n(\theta) = 0\}$ ($X\theta = y$)

Lagrangian: $h(\theta, u) = D_\psi(\theta,\theta_0) - \langle u, X\theta - y\rangle$

$D_\psi(\theta,\theta_0) = \psi(\theta) - \psi(\theta_0) - \langle \nabla\psi(\theta_0), \theta - \theta_0\rangle$

$\longrightarrow \nabla_\theta h(\theta, u) = \nabla\psi(\theta) - \nabla\psi(\theta_0) - X^T u$

RHS:

KKT:
$$\begin{cases} \nabla\psi(\theta_*) - \nabla\psi(\theta_0) = X^T u \\ X\theta_* = y \end{cases} \qquad (I)$$

MD: $\theta^{t+1} = \operatorname{argmin}\left\{ \eta_t \langle \theta, \nabla \hat{R}_m(\theta^t) \rangle + D_\psi(\theta, \theta^t) \right\}$

KKT: $\nabla\psi(\theta^{t+1}) - \nabla\psi(\theta^t) + \eta_t \nabla \hat{R}_m(\theta^t) = 0$

$$\nabla \hat{R}_m(\theta) = \frac{1}{m} \sum_{i=1}^m \ell'(y_i, \langle \theta, a_i \rangle)\, a_i = X^T r(\theta)$$

$$r(\theta) = (r_1(\theta), \dots, r_m(\theta)) \qquad r_i(\theta) = \frac{1}{m} \ell'(y_i, \langle a_i, \theta \rangle)$$

So we can write: $\nabla\psi(\theta^{t+1}) - \nabla\psi(\theta^t) = -\eta_t X^T r(\theta^t)$

$$\longrightarrow \nabla\psi(\theta^t) - \nabla\psi(\theta^0) = X^T \left(- \sum_{s=0}^{t-1} \eta_s r(\theta^s) \right) \qquad (*)$$

v^t

By compactness: θ^t cv along a subsequence $\theta^{t_k} \to \theta^s \in \mathbb{R}^p$

By differentiability: $\nabla\psi(\theta^{t_k}) \to \nabla\psi(\theta^s)$

By (*): $v^{t_k} \to v^s$ + $v^{t_k} \in \operatorname{Im}(X^T)$ closed

$\Rightarrow v^s \in \operatorname{Im}(X^T)$

$v^{t_k} \to X^T u^s$

$v^s = X^T u^s$

$$\begin{cases} \nabla\Psi(\theta^s) - \nabla\Psi(\theta^0) = X^T u^s \\ X\theta^s = y \end{cases}$$

$R(\theta_t) \to 0$ assumption

(because $R(\theta^s) = 0$)

$\equiv$ KKT condition (I) (of θ_*)

D_Ψ strictly convex $\Rightarrow$ unique solution $\theta^s = \theta_*$

Every subsequence θ^{t_k} cv to θ_* $\Rightarrow$ $\theta^t \to \theta_*$ $\square$

Summary: if $\hat{R}_n(\theta^t) \to 0$ + $\{\theta^t\}$ bounded) then

$$\theta^t \xrightarrow[t\to\infty]{} \theta_* = \operatorname*{argmin}_{\theta} \{ D_\Psi(\theta, \theta^0) : \hat{R}_n(\theta) = 0 \}$$

"Minimum Bregman divergence interpolating solution"

Hence $\hat{\theta}$ θ^* depends on:

1. Algorithm (choice of Ψ)
2. Initialization (choice of θ^0)
3. Parametrization

Parametrization example: $F : \mathbb{R}^p \to \mathbb{R}^p$, $\theta \mapsto A\theta$ $\quad A \in \mathbb{R}^{p\times p}$ invertible

→ approach 1: GD on θ init $\theta^0 = 0$

→ approach 2: GD on $\tilde{\theta} = A\theta$, $\tilde{\theta}^0 = 0$

$$\hat{\theta}^{(1)} = \operatorname{argmin}\{\|\theta\|_2^2 : X\theta = y\} \quad (1)$$

$$\hat{\tilde{\theta}}^{(2)} = \operatorname{argmin}\{\|\tilde{\theta}\|_2^2 : XA^{-1}(\tilde{\theta}) = y\}$$

$$\hat{\theta}^{(2)} = A^{-1}(\hat{\tilde{\theta}}^{(2)})$$

$$= \operatorname{argmin}\{\|A\theta\|_2 : X\theta = y\} \quad (2)$$

$X^\top$, $AX^\top$

Unless A orthogonal $\quad \hat{\theta}^{(2)} \neq \hat{\theta}^{(1)}$ $\quad (1) \neq (2)$

Other example: $F(\theta) = (\sqrt{|\theta_1|}, \dots, \sqrt{|\theta_p|})$

→ $\theta^0 = \alpha 1$, $\alpha > 0$ $\quad$ was shown: $\theta^t \xrightarrow{(GD)} \operatorname{argmin}\{\|\theta\|_1 : X\theta = y\}$

[Gunasekar et al, 2017]

Kernel method

Rmk: ANDREA's lectures

Linear regime:

$$f(x,\theta) \simeq f_{lin}(x,a,\theta_0) = f(x,\theta_0) + \langle \nabla_\theta f(x,\theta_0), a\rangle$$

$a = \theta - \theta_0$, $a \in \mathbb{R}^p$

$a \in \mathbb{R}^p \to a \in L^2(\mathbb{R}^d)$

$f_{lin}(x,a) = \int \sigma(\langle x,\theta\rangle)\, a(\theta)\, d\theta$

Denote: $y'_i = y_i - f(x_i, \theta_0)$

$$\Phi = \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix} \begin{bmatrix} \nabla f(x_1, \theta_0)^T \\ \vdots \\ \nabla f(x_n, \theta_0)^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

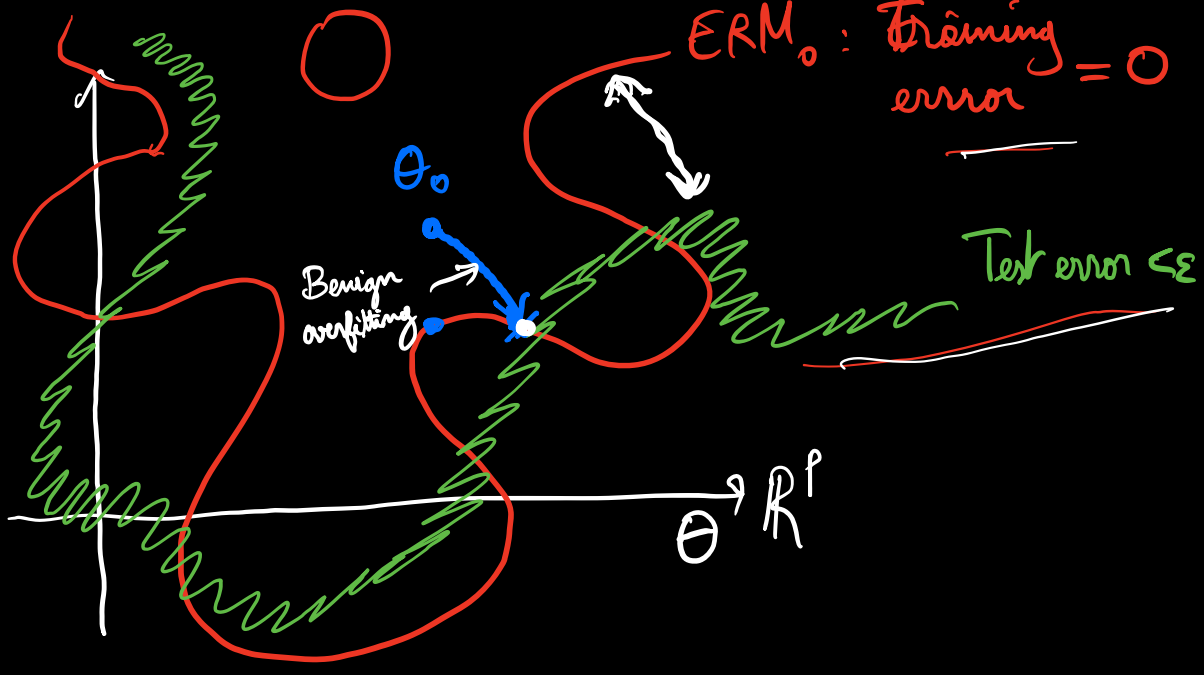
$\longleftarrow p \longrightarrow$

GD w/ 0: $\hat{a} = \operatorname{argmin} \{ \|a\|_2^2 : y' = \Phi a \}$

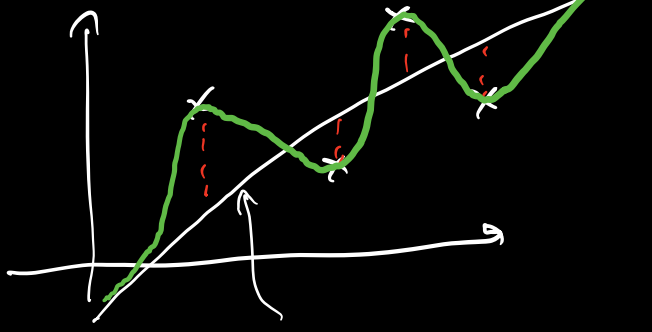
"Ridgeless" regression solution $\hat{a}(0^+)$

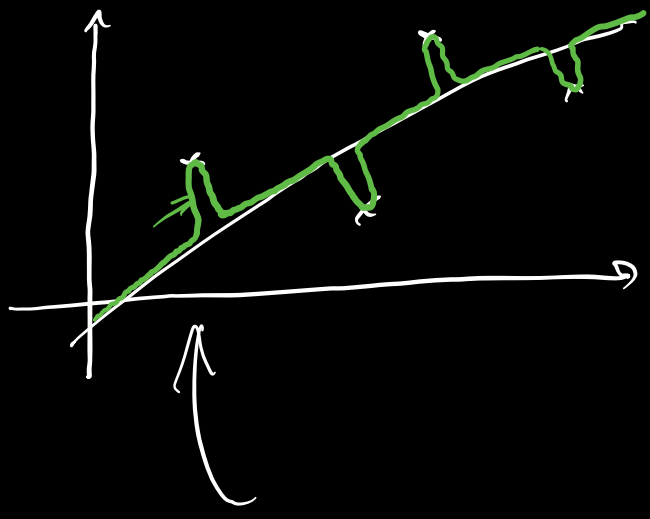
where $\hat{a}(\lambda) = \operatorname{argmin}_{a \in \mathbb{R}^p} \{ \|y' - \Phi a\|_2^2 + \lambda \|a\|_2^2 \}$

$\lambda \searrow 0^+$



$\hat{\theta} \rightarrow R_{test}(\hat{\theta}) = \mathbb{E}[\ell(y_{new}, f(x_{new}, \hat{\theta}))]$





$$\hat{f}(x,\theta) = \langle x,\theta \rangle$$

$$f_*(x,\theta_*) = \langle x,\theta_* \rangle$$

$$y_i = f_*(x_i;\theta_*) + \varepsilon_i$$

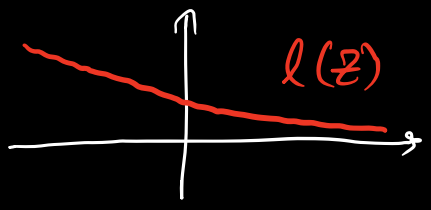
$$\hat{f}(x,\hat{\theta}) = \langle x,\hat{\theta} \rangle + \Delta(x)$$

small spikes

$$\mathbb{E}[\Delta(x)^2] \ll 1$$

② Classification setting:

Setting: $\{(y_i, x_i)\}_{i \le n}$ $\quad y_i \in \{-1, +1\}$



$$\hat{f}(x, \theta) = \operatorname{sign}(\langle \theta, x \rangle)$$

$$\hat{R}_n(\theta) = \sum_{i=1}^{n} \ell(\langle \theta, x_i \rangle y_i)$$

Steepest descent (SD):

- $\|\cdot\|$ norm but not necessarily ℓ_2-norm
- SD algo:
  - initialization θ^0
  - step-sizes η_t
  - update:

$$\theta^{t+1} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \eta_t \langle \nabla \hat{R}_n(\theta^t), \theta \rangle + \frac{1}{2} \|\theta - \theta^t\|^2 \right\}$$

Rewrite: $\theta^{t+1} = \theta^t + \eta_t v^t$

where $v^t = \underset{v \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \langle \nabla \hat{R}_n(\theta^t), v \rangle + \frac{1}{2} \|v\|^2 \right\}$

Dual norm: $\|x\|_* := \sup_{\|v\| \le 1} \langle x, v \rangle$ $\quad u_t = -\nabla \hat{R}_n(\theta^t)$

$$\Rightarrow \min_v \left\{ \frac{1}{2}\|v\|^2 - \langle u_t, v \rangle \right\} = \min_{t \ge 0} \min_{\|v\| = t} \left\{ \frac{1}{2} t^2 - \langle v, u_t \rangle \right\}$$

$$= \min_t \left\{ \frac{1}{2} t^2 - t \|u_t\|_* \right\} = -\frac{1}{2} \|\ \|_*^2 \quad \text{iff } t = \|u_t\|_*$$

$$\Rightarrow \quad v^t = \operatorname*{argmax}_{v} \left\{ \langle v, u_t \rangle : \|v\| \le \|u_t\|_* \right\}$$

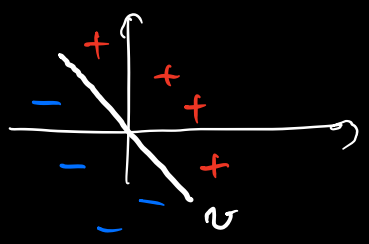
SD update:

$$\begin{cases} \theta^{t+1} = \theta^t - \eta_t v^t \\ v^t = \operatorname*{argmax}_{\|v\| \le \|\nabla \hat{R}_n(\theta^t)\|_*} \left\{ \langle v, \nabla \hat{R}_n(\theta^t) \rangle \right\} \end{cases}$$

Setting:

- Data is separable:

$$\exists\, v \in \mathbb{R}^d, \; y_i \langle x_i, v \rangle > 0 \quad \forall i$$



- $\ell(z) = e^{-z}$

Thm: Assume separable data + $\|x_i\|_* \le B \quad \forall i \le n$

$$\eta_t \le \left[\frac{1}{B^2 R(\theta^t)} \wedge \eta_{\max} \right]$$

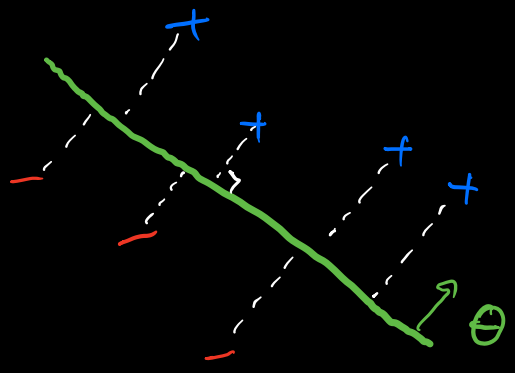
Then

$$\lim_{t \to \infty} \min_{i \le n} \frac{y_i \langle \theta^t, x_i \rangle}{\|\theta^t\|} = \max_{\|\theta\| \le 1} \min_{i \le n} y_i \langle \theta, x_i \rangle$$

Further, if max on RHS is uniquely achieved at θ_*

then $\theta^t \to \theta_*$ as $t \to \infty$

Rmk: * if $\|\cdot\| = \|\cdot\|_2$ then RHS = max margin classification



* Margin: $\gamma = \max_{\|\theta\|=1} \min_{i \le m} \langle \theta, \tilde{x}_i \rangle$ $(\tilde{x}_i = y_i x_i)$

(exercise)

$$= \min_{\pi \in \Delta_{m-1}} \|\tilde{X}^T \pi\|_*$$

where $\Delta_{m-1} = \{\pi \in \mathbb{R}^m : \pi \ge 0, \langle 1, \pi \rangle = 1\}$

e.g.: $\hat{R}_m(\theta) = \sum_{i=1}^m e^{-y_i \langle x_i, \theta \rangle} = \sum_{i=1}^m \pi_i(\theta)$

$\pi_i(\theta) = e^{-y_i \langle x_i, \theta \rangle}$

$$\nabla \hat{R}_m(\theta) = -\tilde{X}^T \pi(\theta)$$

$$\frac{\nabla \hat{R}_m(\theta)}{\hat{R}_m(\theta)} = -\frac{\tilde{X}^T \pi(\theta)}{\langle 1, \pi(\theta) \rangle}$$

$$\Rightarrow \frac{\|\nabla \hat{R}_m(\theta)\|_*}{\hat{R}_m(\theta)} \ge \gamma$$

Proof:

- WLOG: $y_i = +1$
- Denote for simplicity $R(\theta) := \hat{R}_n(\theta) = \sum_{i=1}^{n} e^{-\langle x_i, \theta \rangle}$

We will consider continuous time gradient flow:

$$\dot{\theta}_t = v_t \quad : \quad \langle v_t, -\nabla R(\theta_t) \rangle = \|v_t\|^2 = \|\nabla R(\theta_t)\|_*^2$$

Lemma:

(1) $\int_0^\infty \|\nabla R(\theta_t)\|_*^2 \, dt < \infty$

(2) $R(\theta_t) \to 0$

(3) $\int_0^\infty \|\nabla R(\theta_t)\|_* \, dt = \infty$

Proof: (1) $\frac{d}{dt} R(\theta_t) = \langle \nabla R(\theta_t), v_t \rangle = -\|\nabla R(\theta_t)\|_*^2$

$$\int_0^\infty \|\nabla R(\theta_t)\|_*^2 \, dt \le R(\theta_0) < \infty$$

(2) (1) $\Rightarrow \|\nabla R(\theta_t)\|_* \to 0$

+ separability assumption: $\exists v : \langle v, x_i \rangle > 0 \ \forall i$

$$\langle v, R(\theta_t) \rangle = \sum_{i=1}^{n} e^{-\langle x_i, \theta_t \rangle} \langle x_i, v \rangle \ge R(\theta^t) \cdot \min_{i \le n} \langle v, x_i \rangle$$

$$\nabla R(\theta_t) \to 0 \implies R(\theta_t) \to 0 \quad (\text{i.e. } \min_i \langle x_i, \theta_t \rangle \to \infty)$$

(3) $\|\theta^t\| \le \|\theta^0\| + \int_0^t \|v_s\| ds$

$= \|\theta_0\| + \int_0^t \|\nabla R(\theta_s)\|_* ds$

$\lim_{t\to\infty} \|\theta^t\| = \infty \implies \int_0^\infty \|\nabla R(\theta_s)\|_* ds = \infty$ □

Proof of thm: $\frac{d}{dt} R(\theta^t) = -\|\nabla R(\theta^t)\|_*^2$

Hence $-\frac{d}{dt} \log R(\theta^t) = \frac{\|\nabla R(\theta^t)\|_*^2}{R(\theta^t)}$

$\implies -\log R(\theta^T) = -\log R(\theta^0) + \int_0^T \frac{\|\nabla R(\theta^t)\|_*^2}{R(\theta^t)} dt$

with $-\log\left(\sum_{i=1}^n e^{-\langle x_i, \theta^T\rangle}\right) \le -\log\left(e^{-\langle x_i, \theta^T\rangle}\right) = \langle x_i, \theta_T^T\rangle$

$\implies \begin{cases} \min_{i\le n} \langle x_i, \theta^T\rangle \ge -\log R(\theta^0) + \int_0^T \frac{\|\nabla R(\theta^t)\|_*^2}{R(\theta^t)} dt \\ \|\theta^T\| \le \|\theta^0\| + \int_0^T \|\nabla R(\theta^t)\|_* dt \end{cases}$

$\implies \min_{i\le n} \frac{\langle x_i, \theta^T\rangle}{\|\theta^T\|} \ge \frac{-\log R(\theta^0) + \int_0^T \frac{\|\nabla R(\theta^t)\|_*^2}{R(\theta^t)} dt}{\|\theta^0\| + \underbrace{\int_0^T \|\nabla R(\theta^t)\|_* dt}_{\to\infty \text{ as } T\to\infty}}$

$\ge \frac{\int_0^T \|\nabla R(\theta^t)\|_* \underbrace{\frac{\|\nabla R(\theta^t)\|_*}{R(\theta^t)}}_{\ge\delta} dt}{\int_0^T \|\nabla R(\theta^t)\|_* dt} - o_T(1)$

$$\geq \gamma - o_T(1)$$

and $\min_{i \leq n} \frac{\langle x_i, \theta^T \rangle}{\|\theta^T\|} \leq \max_{\|\theta\| \leq 1} \min_{i \leq n} \langle x_i, \theta \rangle = \gamma$

Hence $\lim_{T \to \infty} \left[\min_{i \leq n} \frac{\langle x_i, \theta^T \rangle}{\|\theta^T\|} \right] = \gamma$ $\square$