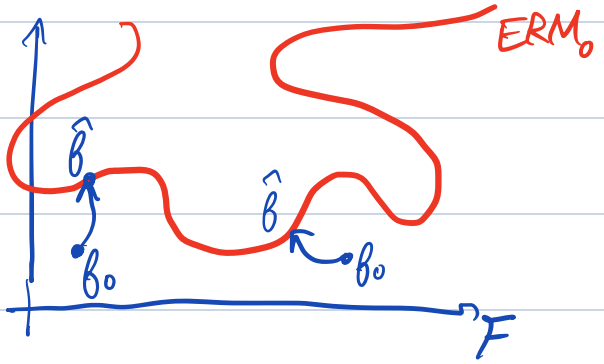
Lecture 3: Implicit / Algorithmic bias

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

$\mathcal{F} = \{ f(\cdot;\theta) : \theta \in \mathbb{R}^p \}$ overparametrized models $p \gg n$

$$ERM_0 := \{ f \in \mathcal{F} : \hat{R}_n(f) = 0 \}$$



→ ERM minimizers generalize more or less well

→ Gpt algo introduce a bias in this choice

⇒ Gpt algorithm selects a solut° $\hat{f} \in ERM_0$

Is that a fair description of what NNs do in practice?

- For LLMs data $= \infty$ (for now): doesn't train until interpolation regime
- Previously: not exactly trained until interpolation but still much closer to interpolation learning than UC
- Regime to "magnify" new properties of NNs: see [Misha Belkin: Fit without fear review paper 2021]

* The 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)

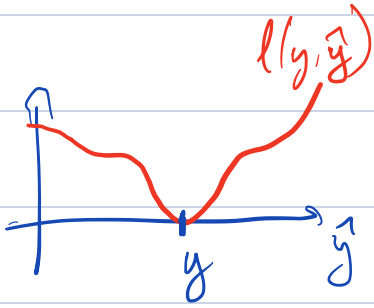
– Linear models: $f(x,\theta) = \langle x, \theta \rangle \qquad x, \theta \in \mathbb{R}^p$

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

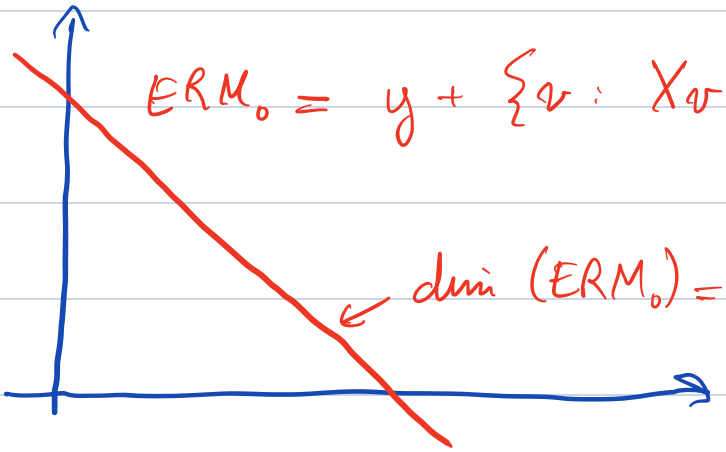
2 Family of losses:

① Regression setting: $y \in \mathbb{R} \qquad \ell(y,\hat{y}) \geq 0$

$= 0$ iff $\hat{y} = y$



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



$X = \begin{bmatrix} - x_1 - \\ \\ - x_n - \end{bmatrix}$ ($n \times p$)

(if X is full rank)

(2) Classification setting $y \in \{+1, -1\}$

$\ell(\hat{y}, y) = \tilde{\ell}(\hat{y}y)$ $\quad \tilde{\ell}(z) > 0$ $\quad \tilde{\ell}(z) \to 0$ iff $z \to \infty$

monotonically decreasing

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

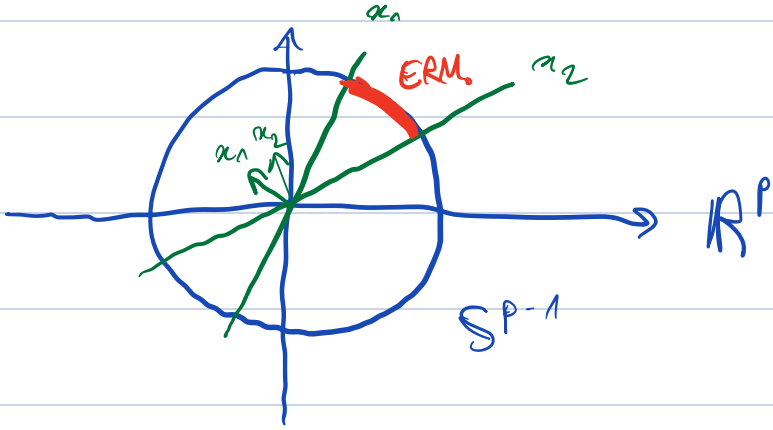
$\tilde{\ell}(z) = \log(1 + e^{-z})$ (logistic)

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

→ Prediction: $\hat{f}(x, \theta) = \operatorname{sign}(\langle \theta, x \rangle) = \operatorname{sign}\left(\left\langle \frac{\theta}{\|\theta\|_2}, x \right\rangle\right)$

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

$$\widetilde{ERM}_0 := \{\xi \in \mathbb{S}^{p-1} : \langle \xi, x_i \rangle y_i \geq 0 \quad \forall i \leq n\}$$



interpolating here means $\operatorname{sign}(\langle \theta, x_i \rangle) = y_i$

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

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

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

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

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

Examples:

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

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

2) $\theta \in \mathbb{R}^p_{>0}$ (positive orthant)

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

Rmk:
* MD can adapt to the geometry of $\mathcal{H}$
* Lots of work on conditions such that MD cv to global optimum
* Question here:
  which point do we converge to?

Theorem [Gunasekar et al., '18]

Assume $\{\theta : X\theta = y\} \neq \emptyset$

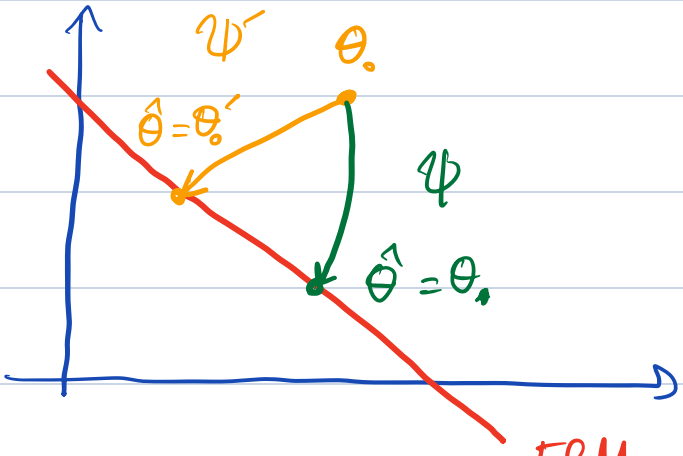
Consider MD with initialization θ^0

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

Then

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

Rmk:



$\psi \neq \psi$

$ERM_* = \{\theta : X\theta = y\}$

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

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

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

$\rightarrow \nabla_\theta h(\theta, u) = \nabla\psi(\theta) - \nabla\psi(\theta_*) - X^T u$

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

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

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

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

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

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

$$\longrightarrow \nabla \Psi(\theta^t) - \nabla \Psi(\theta^0) = X^T \underbrace{\left(- \sum_{s=0}^{t-1} \eta_s r(\theta^s) \right)}_{= v^t} \quad (*)$$

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

By differentiability: $\nabla \Psi(\theta^{t_k}) \longrightarrow \nabla \Psi(\theta^s)$

By (*): $v^{t_k} \longrightarrow v^s$

$v^{t_k} \in \mathrm{Im}(X^T)$ closed

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

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

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

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

$R(\theta^t) \to 0$ by 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_* $\Rightarrow$ $\theta^t \to \theta_*$

□

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: $\hat{\theta} = \theta_*$ depends on ① Algo (choice of ψ)

② Initialization (choice of θ^0)

③ Parametrization

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

$\theta \mapsto A\theta$

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

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

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

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

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

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

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

→ $\theta^* = \alpha 1$ $\quad \alpha > 0$ $\quad$ was shown $\quad \theta^t \xrightarrow[GD]{} \operatorname{argmin} \{ \|\theta\|_1 ; X\theta = y \}$

[Gunasekar et al., 2017]

Next few lectures: $f(x;\theta) = \langle \theta, \Phi(x) \rangle$

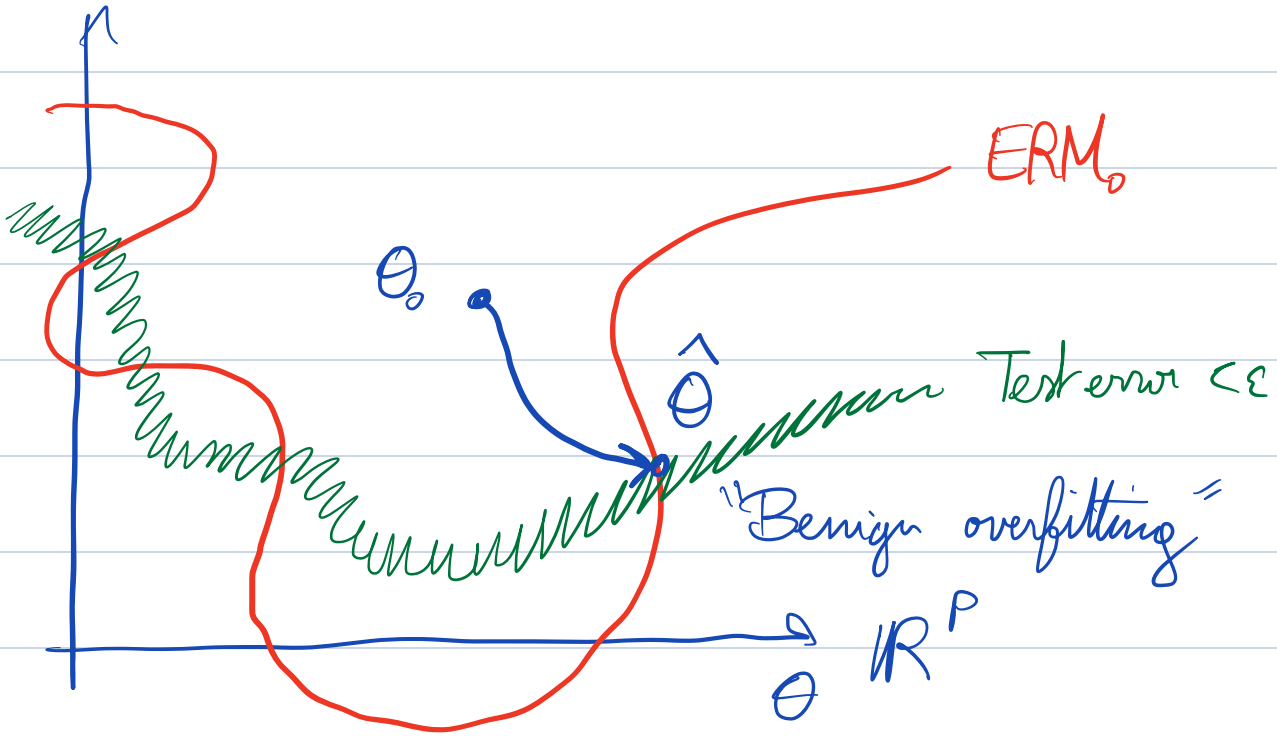
GD cv to $\hat{\theta} = \operatorname{argmin} \{ \|\theta\|_2^2 : y = \Phi(X)\theta \}$

$\theta_0 = 0$

$$\hat{\theta} = (\Phi(X)\Phi(X)^T)^+ \Phi(X)^T y$$

Implicit regularization: which $\hat{\theta} \in ERM_0$

What about generalization? Next lecture!

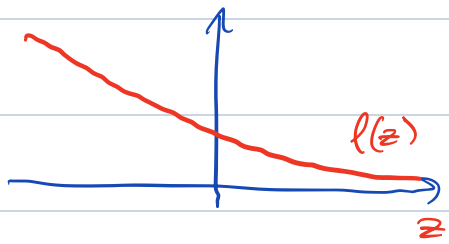


2 CLASSIFICATION SETTING

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

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

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



Steepest descent (SD):

* $\|\cdot\|$ norm (not necessarily $\|\cdot\|_2$- norm)

* **SD algo:**
  - initialization θ^0
  - step sizes η_t
  - update:

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

$v^t = \frac{\theta^{t+1} - \theta^t}{\eta_t}$ this is equivalent to

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

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

Dual norm: $\|x\|_* = \sup_{\|v\| \le 1} \langle x, v \rangle$

$u_t := -\nabla \hat{R}_n(\theta^t)$

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

$$= \min_s \left\{ \frac{1}{2}s^2 - s\|u_t\|_* \right\} = -\frac{1}{2}\|u_t\|_*^2$$

$s = \|u_t\|_*$

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

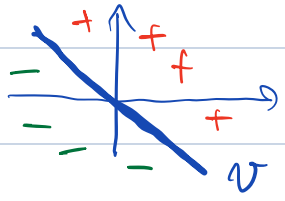
SD update:

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

✗ why minus

✗ Standard GD if $\|\cdot\|_2$

Setting:
- Data is separable: $\exists v \in \mathbb{R}^d$, $y_i \langle x_i, v \rangle > 0$ $\forall i$



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

THM [Gunasekar et al., '18]

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

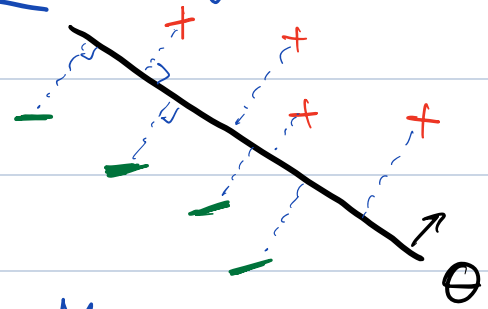
$$\eta_t \le \left[\frac{1}{B^2 R(\theta^t)} \wedge \eta_{\max} \right]$$

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



$\tilde{x}_i = y_i x_i$

* **Margin:** $\gamma = \max_{\|\theta\|=1} \min_{i\le n} \langle \theta, \tilde{x}_i \rangle$

$$= \max_{\|\theta\|=1} \min_{i\le n} \langle e_i, \tilde{X}\theta \rangle$$

$$= \max_{\|\theta\|\le 1} \min_{\pi \in \Delta_{n-1}} \langle \pi, \tilde{X}\theta \rangle$$

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

Slater's condition

$$= \min_{n \in \Delta_{n-1}} \max_{\|\theta\| \le 1} \langle \theta, \tilde{X}^T n \rangle$$

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

Here: $\hat{R}_n(\theta) = \sum_{i=1}^n e^{-y_i \langle x_i, \theta \rangle} = \sum_{i=1}^n n_i(\theta)$

$$\nabla \hat{R}_n(\theta) = - \tilde{X} n(\theta)$$

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

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

Proof:

- WLOG $y_i = +1$
- Denote for simplicity $R(\theta) := \hat{R}_n(\theta) = \sum_{i=1}^n e^{-\langle x_i, \theta \rangle}$
- C^0 time gradient flow:

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

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

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

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

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

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

$\nabla R(\theta^t) \to 0 \Rightarrow R(\theta_t) \to 0$

(i.e. $\min_i \langle x_i, \theta_t \rangle \to \infty$)

(3) $\|\theta^t\| \le \|\theta_0\| + \int_0^t \|v_s\| \, ds$

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

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

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

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

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

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

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

$$\Rightarrow \min_{i\in[n]} \frac{\langle x_i, \theta^T\rangle}{\|\theta^T\|} \geq \frac{-\log R(\theta^0) + \int_0^T \frac{\|\nabla R(\theta^t)\|_*^2}{R(\theta^t)}\, dt}{\|\theta^0\| + \int_0^T \|\nabla R(\theta^t)\|_*\, dt} \xrightarrow[T\to\infty]{} \infty$$

$$\geq \frac{\int_0^T \|\nabla R(\theta^t)\|_* \frac{\|\nabla R(\theta^t)\|_*}{R(\theta^t)}\, dt}{\int_0^T \|\nabla R(\theta^t)\|_*\, dt} - o_T(1) \qquad \frac{\|\nabla R(\theta^t)\|_*}{R(\theta^t)} \geq \gamma$$

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

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