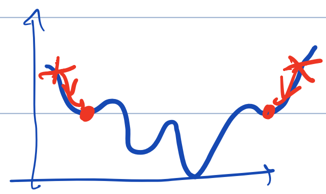
Lecture 5 : $\boxed{\text{Lazy regime}}$

[Jacot, Gabriel, Hongler, 2018]
[Du et al. 2019] [Chizat, Bach, 2019] [Oymak et al. 2020] ...

To train NNs, need to solve a highly non-convex problem :

$$\min_{W_L \dots W_1} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - W_L \circ \sigma \circ \dots \circ \sigma W_1 x_i \right)^2$$

Typically, expect to be hard



get trapped in local minima

→ to the point where most of the ML community abandoned NNs in the 2000s
(until AlexNet 2012)

Empirical observations:

1) as # parameters ↗, easier to optimize: landscape appears to simplify greatly, enabling local search algo (e.g. SGD) to find global optima reliably.

"tractability via overparametrization"

2) Overparametrization does not harm generalization (Lecture 3 & 4)

This lecture:

* Example of such a mechanism of tractability via overparametrization

"Lazy" or "linear" regime

* General mechanism that has nothing to do a priori with neural nets

* In this regime: NN behave as a linear model

$\longrightarrow$ neural tangent model

$\longrightarrow$ Neural tangent kernel in the limit

Lazy optimization regime

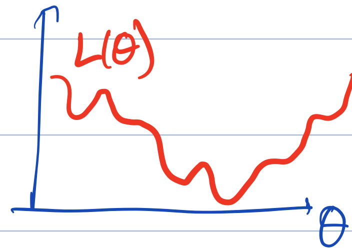
High-level intuition: very simple idea

1 data point (y, x) $\quad y \in \mathbb{R},\ x \in \mathbb{R}$

1-param model $f(x;\theta)$ $\quad \theta \in \mathbb{R}$

$$L(\theta) = (y - f(x;\theta))^2$$

$\rightarrow$ might be very complicated



Initialize dynamics at θ^0

Idea: we can always rescale output of our model $\alpha f(x;\theta)$ $\quad \alpha > 0$ and $\alpha \to \infty$ to linearize the model and make the problem quadratic on a small neighborhood of θ_0

$\longrightarrow$ strongly convex $\rightarrow$ optimization is easy

2^{nd} order Taylor expansion:

assume $\neq 0$

$$f(x;\theta) = f(x;\theta_0) + f'(x;\theta_0)(\theta - \theta_0) + \frac{f''(\tilde{\theta})}{2}(\theta - \theta_0)^2$$

$$L_\alpha(\theta) = (y - \alpha f(x;\theta))^2$$

For simplicity: $v_\alpha := \alpha f(x;\theta_0) - y$

$f_0' := f'(x;\theta_0)$

$f'' := \frac{1}{2} f''(x;\tilde{\theta})$ $\qquad u := \theta - \theta_0$

$$L_\alpha(\theta) = \left(v_\alpha + \alpha f_0' u + \alpha f'' u^2\right)^2$$

$$= v_\alpha^2 + 2 v_\alpha f_0' (\alpha u) + (f_0')^2 (\alpha u)^2 + 2 v_\alpha f'' (\alpha u^2) + 2 f_0' f'' (\alpha^2 u^3) + (f'')^2 (\alpha^2 u^4)$$

Assume $f(x;\theta_0) \ll y$, in fact $f(x;\theta_0) = 0$
so that $v_\alpha = v_0$

Consider a small neighborhood $\theta - \theta_0 \lesssim \frac{1}{\alpha}$

$$\tilde{u} = \frac{\theta - \theta_0}{\alpha}$$

$$L_\alpha(\theta_0 + \alpha \tilde{u}) = v_0^2 + 2 v_0 f_0' \tilde{u} + (f_0')^2 \tilde{u}^2 + \frac{2 v_0 f'' \tilde{u}^2}{\alpha} + \frac{2 f_0' f'' \tilde{u}^3}{\alpha} + \frac{(f'')^2 \tilde{u}^4}{\alpha^2}$$

$$\alpha \to \infty \quad L_\alpha(\theta_0 + \alpha \tilde{u}) \approx (v_0 + f_0' \tilde{u})^2$$

$\rightarrow$ quadratic model : θ_t = gradient flow

1) $$\boxed{L_\alpha(\theta_t) \lesssim L_\alpha(\theta_0)\, e^{-\frac{(b_0')^2}{2} t}}$$

$\hookrightarrow$ we started from a very complicated optimization landscape and proved convergence to global optima $L_\alpha(\hat{\theta}) = 0$ exponentially fast

2) Throughout the dynamics :

$$\boxed{\alpha f(x;\theta^t) \simeq \alpha f(x;\theta^0) + \alpha f'(x;\theta^0)\,(\theta^t-\theta_0)}$$

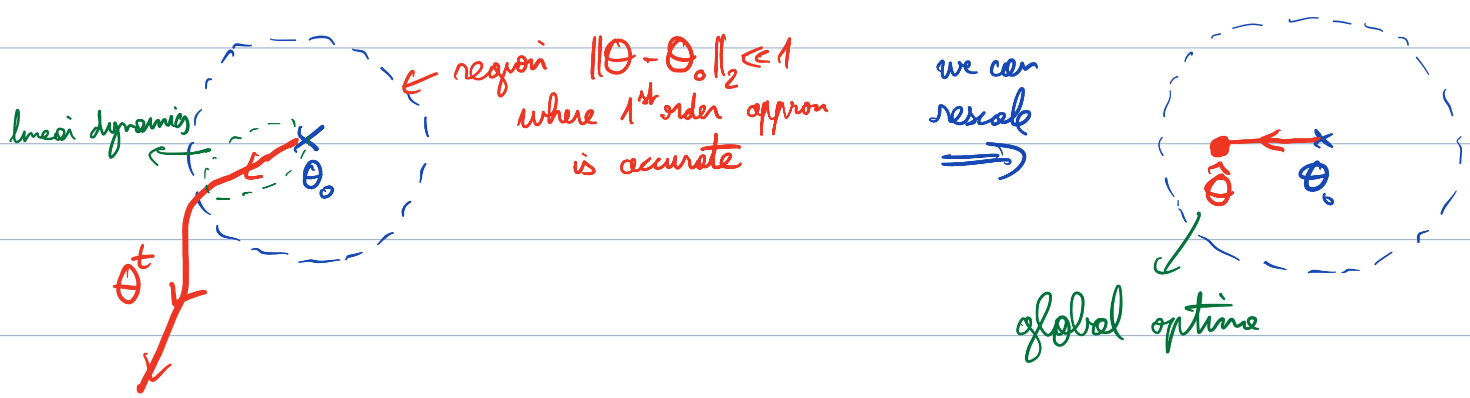
$\hookrightarrow$ behave as linear model (linear in parameter θ)

$\rightarrow$ we can replace our non-linear model by a linear model during training & at test time

$\rightarrow$ behave effectively as a linear model

Rmk: Another way of thinking about it:

$$f(x;\theta^t) = f(x;\theta_0) + f'(x;\theta_0)(\theta^t-\theta_0) + O(|\theta^t-\theta_0|^2)$$



3) A consequence is that $\|\theta_t - \theta_0\|_2 \ll \theta_0$

$\rightarrow$ weights barely move

[Chizat, Bach, 2019] "Lazy regime"

Several papers showed global cv before (2018)
but 2) & 3) (weights barely moving + linearization of the NN)
were kind of hidden in the proof
$\Longrightarrow$ these have important consequences on learning in this regime $\rightarrow$ essentially linear regression
which temper the achievement of showing global cv
and whether this is a good model for what happens
with NNs

$\boxed{\text{Next}}$ precise conditions for global convergence

A meta theorem for global CV in lazy regime

$(y_i, x_i)_{i \leq n}$ data

param model $f(x;\theta)$ $\quad \theta \in \mathbb{R}^p$

$$\hat{R}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f(x_i;\theta)\right)^2 = \frac{1}{2n} \| y - f_n(\theta) \|_2^2$$

$$f_n : \mathbb{R}^p \to \mathbb{R}^n \qquad f_n(\theta) = \begin{pmatrix} f(x_1;\theta) \\ \vdots \\ f(x_n;\theta) \end{pmatrix} \in \mathbb{R}^n$$

Denote $\Phi_n(\theta) := D f_n(\theta) \in \mathbb{R}^{n \times p}$ the Jacobian

$$L_{n,0} := \sup_{\theta \neq \theta_0} \frac{\| \Phi(\theta) - \Phi(\theta_0) \|_{op}}{\| \theta - \theta_0 \|_2}$$

Let's do a warm-up:

$$\boxed{\text{Condition so that there exists an interpolating solution in a small neighborhood of } \theta_0}$$

We want to find $f_m(\theta) = y$

By Taylor's expansion:

$$\theta_t = (1-t)\theta_0 + t\theta$$

$$y - f_m(\theta_0) = f_m(\theta) - f_m(\theta_0)$$

$$(*) \qquad = \Phi(\theta_0)(\theta - \theta_0) + \underbrace{\int_0^t \left(\Phi(\theta_t) - \Phi(\theta_0)\right)(\theta - \theta_0)\, dt}_{e(\theta)}$$

Note that:

$$\| e(\theta) \|_2 \leq L_{m,0} \| \theta - \theta_0 \|_2^2$$

Rearranging the terms in $(*)$

$$\theta = \theta_0 + \Phi(\theta_0)^+ \left(y - f_m(\theta_0) + \delta\right)$$

where $\delta = - \underbrace{e\big(\theta_0 + \Phi(\theta_0)^+ (\overbrace{y - f_m(\theta_0) + \delta}^{=: \tilde{y}})\big)}_{=: F(\delta)}$

(fixed pt equation)

$\delta = F(\delta)$ When does such a solution exists?

<u>Use Brouwer's fixed point theorem</u>

$$\|F(\delta)\|_2 \leqslant L_{m,0} \left(\|\Phi_0^+ \tilde{y}\|_2 + \|\Phi_0^+\|_{op} \|\delta\|_2 \right)^2$$

F maps the ball of radius r into the ball of radius

$$L_{m,0} \left(\|\Phi_0^+ \tilde{y}\|_2 + \|\Phi_0^+\|_{op}\, r \right)^2$$

Taking $r := \dfrac{\|\Phi_0^+ \tilde{y}\|_2}{\|\Phi_0^+\|_{op}}$,

then if $L_{m,0} \leqslant \dfrac{1}{4 \|\Phi_0^+ \tilde{y}\|_2 \, \|\Phi_0^+\|_{op}}$

F maps ball of radius r to ball of radius r

and there exists a fixed pt δ_*

<u>i.e.</u> $\exists$ interpolating solution $\theta_* = \theta_0 + \Phi_0^+ (y - f_m(\theta_0) + \delta_*)$

order $\ll \|y - f_m(\theta_0)\|_2$

$\hookrightarrow$ approximately solut° of $y = f_m(\theta_0) + \Phi_0 (\theta - \theta_0)$

of linear model

Global CV of gradient flow

$$=: \Phi(\theta_t) = Df_m(\theta_t)$$

<u>Gradient flow</u>: $\begin{cases} \dot{\theta}_t = -\nabla_\theta \hat{R}(\theta) = \frac{1}{m} \Phi_t \left(y - f_m(\theta_t)\right) \\ \text{initialization at } \theta_0 \end{cases}$

easy to extend analysis to discrete dynamics

We will compare this dynamic to the dynamics on the linearized model

$$f_{lin}(x;\theta) = \underbrace{f(x;\theta_0) + \langle \theta - \theta_0 , \nabla_\theta f(x;\theta_0) \rangle}_{1^{st} \text{ order Taylor expansion}}$$

$$\hat{R}_{lin}(\theta) = \frac{1}{2m} \left\| y - f_m(\theta_0) - \Phi_0 (\theta - \theta_0) \right\|_2^2$$

$$\frac{d}{dt} \bar{\theta}_t = -\nabla_\theta \hat{R}_{lin}(\bar{\theta}_t) = \frac{1}{m} \Phi_0^T \left(y - f_m(\theta_0) - \Phi_0 (\bar{\theta}_t - \theta_0)\right)$$

Will assume $p \geq m$ $\quad$ $\text{rank}(\Phi_0) = m$

↳ overparametrized

$$\mathrm{ERM}_0^{lin} = \left\{ \theta : \Phi_0(\theta - \theta_0) = y - f_n(\theta_0) \right\}$$

$$\bar{\theta}_t \longrightarrow \bar{\theta}_\infty = \operatorname{argmin} \left\{ \|\bar{\theta} - \theta_0\|_2 \ : \bar{\theta} \in \mathrm{ERM}_0^{lin} \right\}$$

Lecture 3

<u>Notations</u> :

$$L_n := \sup_{\theta_1 \neq \theta_2} \frac{\| Df_n(\theta_1) - Df_n(\theta_2) \|_{op}}{\| \theta_1 - \theta_2 \|_2}$$

$$\sigma_{min} := \sigma_{min}\left(Df_n(\theta_0) \right)$$

$$\sigma_{max} := \sigma_{max}\left(Df_n(\theta_0) \right)$$

$$\| Df(\theta) \|_{op} = \sup_{a \in \mathbb{R}^p} \frac{\| a^\top \nabla f(\cdot, \theta) \|_{L^2(\mathbb{P})}}{\| a \|_2}$$

$$\mathrm{Lip}(Df) = \sup_{\theta_1 \neq \theta_2} \frac{\| \nabla f(\cdot, \theta_1) - \nabla f(\cdot, \theta_2) \|_{op}}{\| \theta_1 - \theta_2 \|_2}$$

Thm: [Bartlett, Montanari, Rakhlin, '21]

Assume

$$L_m \, \|y - f_m(\theta_0)\|_2 < \frac{1}{4} \sigma_{min}^2 \qquad (*)$$

Then $\forall t > 0$

(1) $$\hat{R}(\theta_t) \leq \hat{R}(\theta_0) \, e^{-\lambda_0 t} \qquad \lambda_0 := \frac{\sigma_{min}^2}{2m}$$

(2) $$\|\theta_t - \theta_0\|_2 \leq \frac{2}{\sigma_{min}} \|y - f_m(\theta_0)\|_2$$

$$\|\theta_t - \bar{\theta}_t\|_2 \leq C \, L_m \frac{\sigma_{max}^2}{\sigma_{min}^2} \|y - f_m(\theta_0)\|_2^2$$

(3) $$\|f(\theta_t) - f_{lin}(\bar{\theta}_t)\|_{L^2(P)} = \mathbb{E}_x\left[\left(f(x;\theta_t) - f_{lin}(x;\bar{\theta}_t)\right)^2\right]^{1/2}$$

$$\leq C \left\{ \frac{\mathrm{Lip}(Df)}{\sigma_{min}^2} + \|Df(\theta_0)\|_{op} \frac{L_m \, \sigma_{min}^2}{\sigma_{max}^5} \right\} \|Y - f_m(\theta_0)\|_2^2$$

Rmk 1: Condition $(*)$ $\Rightarrow$ condition $\exists$ interpolating solut° earlier

$$L_{m,0} \leq \frac{1}{4 \, \|\Phi_0^+ (y - f_m(\theta_0))\|_2 \, \|\Phi_0^+\|_{op}}$$

$$\|\Phi_0^+(y-f_m(\theta_0))\|_2 \, \|\Phi_0^+\|_{op} \leq \|\Phi_0^+\|_{op}^2 \, \|y-f_m(\theta_0)\|_2$$

$$\hookrightarrow = \frac{1}{\sigma_{min}(\Phi_0)^2}$$

$$L_{m,0} \leq L_m < \frac{\sigma_{min}^2}{4\|y-f_m(\theta_0)\|_2} \leqslant \frac{1}{4\|\Phi_0^+(y-f_m(\theta_0))\|_2 \, \|\Phi_0^+\|_{op}}$$

Rmk 2: (1) shows global cv to 0!! (success!)

(2) shows θ_t stays close to θ_0

and $\theta_t \simeq \overline{\theta}_t$ dynamics where we replace $f \to f_{lin}$

(3) From a statistical perspective, this is the most important: shows that $f(\cdot;\theta^t)$ and $f_{lin}(\cdot;\overline{\theta}^t)$ behave the same on test data

$\Rightarrow$ show that learning $\equiv$ to learning with linear model $f_{lin}(\cdot;\theta)$

Rmk 3: Rescaling $\quad f_\alpha(\cdot\,;\theta) = \alpha f(\cdot\,;\theta)$

assume $\quad f(x;\theta_0) = 0$

$$\left(\begin{aligned} &\mathrm{Lip}(Df_{\alpha,m}) = \alpha L_m \\ &\|y - f_\alpha(\theta_0)\|_2 = \|y\|_2 \\ &\sigma_{\min}(Df_{\alpha,m}) \simeq \alpha \end{aligned}\right)$$

Condition become

$$\alpha \ll \alpha^2$$

i.e $\boxed{\alpha \gg 1}$

Application: 2-layer neural networks

$$f(x;\Theta) = \frac{\alpha}{\sqrt{M}} \sum_{j=1}^{M} a_j \, \sigma(\langle w_j, x\rangle) \qquad x \underset{iid}{\sim} \mathrm{Unif}(S^{d-1})$$

To simplify, fix $a_j \in \{\pm 1\}$

$$\Theta = (w_1, \dots, w_M) \in \mathbb{R}^{Md} \qquad p = Md$$

$$w_1^0, \dots, w_{\frac{M}{2}}^0 \underset{iid}{\sim} N(0, \mathrm{Id}_d)$$

$$w_{\frac{M}{2}+i}^0 = w_i^0 \qquad \begin{array}{l} a_1 = \dots = a_{M/2} = 1 \\ a_{\frac{M}{2}+1} = \dots = a_M = -1 \end{array}$$

so that $f(x;\Theta^0) = 0$.

$$[D f_n(\Theta_0)]_{i,(jk)} = \frac{\alpha}{\sqrt{M}} a_j \, \sigma'(\langle w_j, x_i\rangle)(x_i)_k$$

Assume y_i are $O(1)$-sub Gaussian.

Lemma: Assume $Md \geq C n \log n$. Then w.h.p

$\|y - f_n(\theta_0)\| \lesssim \sqrt{n}$ $\quad \sigma_{max} \lesssim \alpha(\sqrt{n} + \sqrt{d})$

$\sigma_{min} \gtrsim \alpha\sqrt{d}$ $\quad L_n \lesssim \alpha\sqrt{\frac{d}{M}}(\sqrt{n} + \sqrt{d})$

Condition $\quad L_n \cdot \|y - f_n(\theta_0)\|_2 < \frac{1}{4}\sigma_{min}^2$

$$\alpha\sqrt{\frac{d}{M}}(\sqrt{n} + \sqrt{d})\sqrt{n} \lesssim \alpha^2 d \overset{n \geq d}{\Longrightarrow} \boxed{\alpha \gtrsim \sqrt{\frac{n^2}{Md}}}$$

Corollary: If $n \geq d$ $\quad Md \gtrsim n \log n$ $\quad \alpha \gtrsim \sqrt{\frac{n^2}{Md}}$

Then (1) $\hat{R}_n(\theta_t) \lesssim \sqrt{n}\, e^{-\alpha^2 \frac{d}{n} t}$

(2) $\|\theta_t - \theta_0\|_2 \lesssim \frac{1}{\alpha}\sqrt{\frac{n}{d}} \simeq \frac{1}{\alpha}\sqrt{\frac{n}{Md^2}}\|\theta_0\|_2$

(3) $\|f(\theta_t) - f_{lin}(\bar{\theta}_t)\|_{L^2(\mathbb{P})} \lesssim \frac{1}{\alpha^2}\sqrt{\frac{n^5}{Md^4}}$

RMK: Hence linear theory is accurate when

(1) Fixed α : $p = Md \longrightarrow \infty$

(scaling $\frac{\alpha}{\sqrt{M}}$)

(2) Fixed M : $Md \gtrsim n \log n$ $\qquad \alpha \longrightarrow \infty$

Rmk: For $\alpha = 1$, need $p = Md \gtrsim n^2$

Necessary $p = Md \geqslant n$

bound might be suboptimal.

Rmk: If $\alpha \searrow 0$ with M, we will get different regime

e.g. $\alpha = \frac{\alpha_0}{\sqrt{M}}$ $\qquad f(x;\theta) = \frac{1}{M} \sum_j a_j \sigma(\langle w_j, x \rangle)$

$$\underset{M \to \infty}{=} \int a \, \sigma(\langle w, x \rangle) \, \rho(da \, dw)$$

$\longrightarrow$ non-linear dynamics (Mean-Field regime)

Learning in the Lazy regime

$$f_{lin}(x;\theta) = \underbrace{f(x;\theta_0)}_{\text{offset not trained}} + \langle \nabla f(x;\theta_0), \theta - \theta_0 \rangle$$

<u>NT model</u>: $f_{NT}(x; b) = \langle \nabla f(x;\theta_0), b \rangle$

$$f(x;\theta) = \frac{1}{\sqrt{M}} \sum_{j\in[M]} a_j \, \sigma(\langle w_j, x\rangle)$$

$$f_{NT} = \underbrace{\langle a, \phi_{RF}(x)\rangle}_{\text{linearizat° 2nd layer}} + \underbrace{\langle b, \phi_{NT}(x)\rangle}_{\text{linearizat° 1st layer}}$$

$$\phi_{RF}(x) = \frac{1}{\sqrt{M}} \begin{pmatrix} \sigma(\langle w_1, x\rangle) \\ \vdots \\ \sigma(\langle w_M, x\rangle) \end{pmatrix} \in \mathbb{R}^M \qquad \phi_{NT}(x) = \frac{1}{\sqrt{M}} \begin{bmatrix} \sigma'(\langle w_1, x\rangle)\, x \\ \\ \sigma'(\langle w_M, x\rangle)\, x \end{bmatrix} \in \mathbb{R}^{Md}$$

Associated kernel:

$$K_M(x_1,x_2) = \langle \phi_{RF}(x_1), \phi_{RF}(x_2)\rangle + \langle \phi_{NT}(x_1), \phi_{NT}(x_2)\rangle$$

$$= \frac{1}{M} \sum_{j\in[M]} \sigma(\langle w_j, x_1\rangle)\sigma(\langle w_j, x_2\rangle) + \langle x_1, x_2\rangle \sigma'(\langle w_j, x_1\rangle)\sigma'(\langle w_j, x_2\rangle)$$

$$\xrightarrow[M\to\infty]{WLLN} K(x_1, x_2)$$

$$= \mathbb{E}_\omega\left[\sigma(\langle\omega, x_1\rangle)\sigma(\langle\omega, x_2\rangle)\right] + \langle x_1, x_2\rangle \mathbb{E}_\omega\left[\sigma'(\langle\omega, x_1\rangle)\,\sigma'(\langle\omega, x_2\rangle)\right]$$

"NEURAL TANGENT KERNEL" (NTK)

Jacot, Gabriel, Hongler, 2018.

If $\|x_1\|_2 = \|x_2\|_2 = 1$ $\qquad \omega \sim N(0, Id_d)$

$$K(x_1, x_2) = h(\langle x_1, x_2\rangle) \qquad h(t) = \mathbb{E}\left[\sigma(G_1)\sigma(G_2) + t\sigma'(G_1)\sigma'(G_2)\right]$$

$$\begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix}\right)$$

$M \to \infty$ in this limit: learning with NN in this regime

$\equiv$ kernel method with NTK

$\hookrightarrow$ see next lecture !

Does the lazy regime explain NNs used in practice?

Community got very excited: scaling $\frac{1}{\sqrt{M}}$ seemed to correspond to what is used in practice, but:

<u>Lazy regime:</u>
* $\|\theta^t - \theta^0\|_2 \ll \|\theta^0\|$ weights barely move
* $f(x;\theta^t) \simeq f(x;\theta^0) + \langle \theta^t - \theta^0, \nabla_\theta f(x;\theta^0) \rangle$

effectively behave as a linear model (kernel method)

<u>Extensive litterature checking numerically validity of this regime:</u>

→ weights θ^t don't stay near initializat° (e.g. filters in CNNs)

→ if we replace $f(x;\theta)$ by $f_{lin}(x;\theta)$ ⟹ drop in performance

⚠ Depends on architecture & data set: NTK sometimes match performance

<u>Theory:</u> linear models are much less powerful methods than non-linear NNs (in terms of approximation + generalization capabilities)

⟹ many "separation" results

Proof of Theorem Part (1):

$$y_t := f_n(\theta_t)$$

$$\dot{\theta}_t = -\frac{1}{m}\Phi_t^T(y_t - y_0)$$ [param space]

$$\dot{y}_t = \Phi_t\,\dot{\theta}_t = -\frac{1}{m}K_t(y_t - y)$$ [fct space]

"Kernel" $K_t := \Phi_t\,\Phi_t^T \in \mathbb{R}^{m\times m}$

$$\frac{d}{dt}\underbrace{\|Y_t - Y\|_2^2}_{=2m\hat{R}(\theta_t)} = -\frac{2}{m}\langle Y_t - Y,\ \underbrace{K_t}\,(Y_t - Y)\rangle$$

$$\lambda_{min}(K_0) = \sigma_{min}^2 = \sigma_{min}(\Phi_0)^2$$

$$\sigma_{min}(\Phi_t) \geq \sigma_{min}(\Phi_0) - L_m \cdot \|\theta_t - \theta_0\|_2$$

If $\|\theta_t - \theta_0\|_2 \leq r_* := \frac{\sigma_{min}}{2L_m}$ then $\lambda_{min}(K_t) \geq \left(\frac{\sigma_{min}}{2}\right)^2$

Let $t_* := \inf\{t : \|\theta_t - \theta_0\|_2 > r_*\}$

$$t < t_* \implies \|Y_t - Y\|^2 \leq \|Y_0 - Y\|^2 e^{-\lambda_0 t}$$

$\hookrightarrow \lambda_0 = \frac{\sigma_{min}^2}{2m}$

$\rightarrow$ Need to show $t_* = \infty$

$$\|\dot{\theta}_t\|_2 = \frac{1}{m}\|\Phi_t^T(Y_t - Y_0)\|_2$$

$$\frac{d}{dt}\|Y_t - Y\|_2 = \frac{1}{2\|Y_t - Y\|_2}\frac{d}{dt}\|Y_t - Y\|_2^2$$

$$= -\frac{1}{m}\frac{\langle Y_t - Y, K_t(Y_t - Y)\rangle}{\|Y_t - Y\|_2}$$

$$= -\frac{1}{m}\frac{\|\Phi_t^T(Y_t - Y)\|_2^2}{\|Y_t - Y\|_2}$$

$t \leq t_*$

$\sigma_{min}(\Phi_t) \geq \frac{\sigma_{min}}{2}$

$$\leq -\frac{\sigma_{min}}{2m}\|\Phi_t^T(Y_t - Y)\|_2$$

$\|Y_t - Y_0\|_2 \leq \frac{2}{\sigma_{min}}\|\Phi_t^T(Y_t - Y_0)\|_2$

$$\frac{d}{dt}\|Y_t - Y\|_2 + \frac{\sigma_{min}}{2}\|\dot{\theta}_t\|_2 \leq 0$$

$$\frac{d}{dt}\left[\|Y_t - Y\|_2 + \frac{\sigma_{min}}{2}\|\theta_t - \theta_0\|_2\right] \leq 0$$

$$\Longrightarrow \|Y_t - Y\|_2 + \frac{\sigma_{min}}{2}\|\theta_t - \theta_0\|_2 \leq \|Y_0 - Y\|_2$$

$$\Longrightarrow \|\theta_t - \theta_0\|_2 \leq \frac{2}{\sigma_{min}}\|Y_0 - Y\|_2$$

$L_m\|Y - Y_0\|_2 < \frac{1}{4}\sigma_{min}^2$

If $t_* < \infty$, $\|\theta_{t_*} - \theta_0\|_2 \leq \frac{2}{\sigma_{min}}\|Y - Y_0\|_2 < \frac{\sigma_{min}}{2L_m} = r_*$

$\Longrightarrow t_* = \infty$ This proves (1) $\square$ $\lightning$