Lecture 7: Optimization in Shallow NNs I: the "lazy" regime

→ Previous weeks: → Discussing approximation, generalization & optimisation aspects separately.

→ Always tradeoffs between these sources of error.

→ Today: Effect of model parametrization on these tradeoffs.

→ Let $\phi(\theta)$ be a differentiable map $\Theta \xrightarrow{\phi} \mathcal{F}$

θ_0: initial point

Θ: parameter space. $\mathcal{F}$: function space.

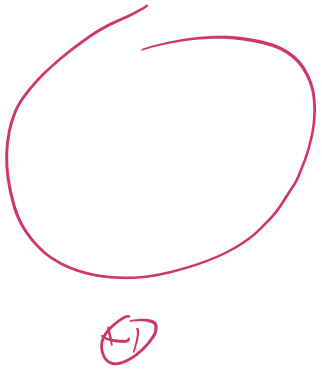
Ex: for a NN: $\phi(\theta): X \subseteq \mathbb{R}^d \to \mathbb{R}$

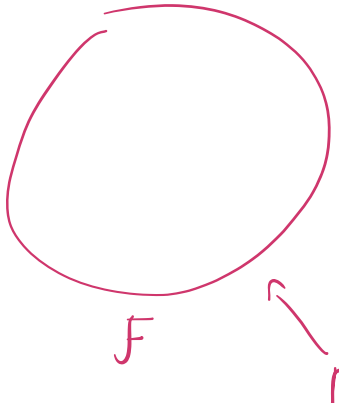
$$x \longmapsto \phi(\theta; x) = W_k \sigma(W_{k-1} \sigma \cdots W_1 x).$$

→ $R: \mathcal{F} \to \mathbb{R}$ risk functional $R(f) = \mathbb{E}_\nu[\ell(f(x), y)]$

→ ℓ is convex w.r.t first argument. $\left(= \|f(x) - f^*\|^2_{L^2(\nu)}\right)$

↳ R is convex wrt f ↳ for least-squares.





↳ If ϕ is linear, then the ERM is also convex.

↳ In that case, first-order methods (GD) can find global min efficiently (dim-free iteration complexity).

↳ RKHS example falls in this category

$$\phi(\theta) = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot) \qquad \text{(linear wrt parameters)}.$$

→ Q: What happens when $\phi(\theta)$ is non-linear?

→ Let us consider gradient-descent with fixed step size η

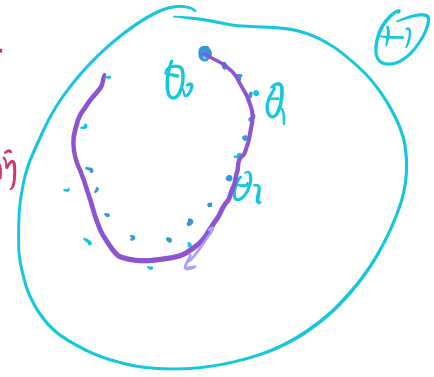
$$\theta_{k+1} = \theta_k - \eta \nabla R(\phi(\theta_k))$$

$R(f) = \mathbb{E}_{x,y}[\ell(f(x), y)]$

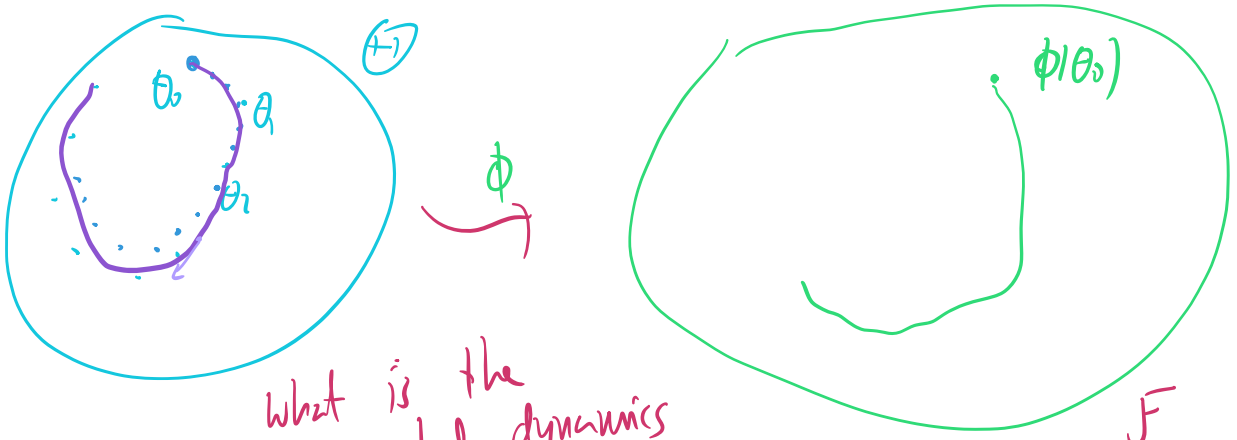
↳ empirical over data!

→ As $\eta \to 0$, gradient descent is approximating an underlying ODE, the gradient-flow:

$$\dot{\theta}(t) = -\nabla_\theta R(\phi(\theta))$$



→



what is the associated dynamics in function space?

→ Let's write down the dynamics in $\mathcal{F}$:

$\phi_t = \phi(\theta_t)$ $\qquad \frac{d}{dt}\theta_t = -\nabla_\theta R(\phi(t))$

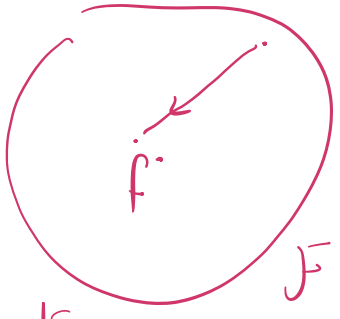
$$\frac{d}{dt}\phi_t = D\phi(\theta_t)^T \cdot \frac{d}{dt}\theta_t$$

$$= -\underbrace{D\phi(\theta_t)^T \cdot D\phi(\theta_t)}_{K_t} \cdot R'(\phi(\theta_t)).$$

R' : functional derivative / first-variation of R in function space. eg $R(f) = \frac{1}{2}\|f - f^*\|_2^2$ $\quad R'(f) = f - f^*$

$K_t = D\phi(\theta(t))^T \cdot D\phi(\theta(t))$. is the tangent kernel.

($K_t : \mathcal{F} \to \mathcal{F}$ maps the "ambient" derivative $R'(\phi_t)$ to a "parameter-aware" update in function space.



- In coordinates,

$$(K_t f)(x) = \int k_t(x,x') f(x') \, d\nu(x')$$

$$k_t(x,x') = \langle \nabla_\theta \phi(\theta_t, x), \nabla_\theta \phi(\theta_t, x') \rangle_{L^2(\theta)}$$

- Let's compute this kernel when $\phi(\theta)$ is a shallow NN of the form $\phi(\theta; x) = \frac{1}{\sqrt{m}} \sum_{i=1}^m g(\theta_i, x)$; $\theta_i \sim \mu_0$ initial distrib.

$$K(x,x') = \frac{1}{m}\sum_{i=1}^{m} \left\langle \nabla_\theta g(\theta,x)\Big|_{\theta=\theta_i}, \nabla_\theta g(\theta,x')\Big|_{\theta=\theta_i} \right\rangle$$

$g(\theta,x) = a\,\Gamma(\langle x,w\rangle)$

at fixed x, x' this is an average of m iid random variables.

$m\to\infty$ ↓

$$\mathbb{E}_{\theta\sim\mu_0}\left[\left\langle \nabla_\theta g(\theta,x), \nabla_\theta g(\theta,x')\right\rangle\right] := \bar{K}(x,x').$$

↳ By choosing the scaling in $1/\sqrt{m}$, the tangent kernel (initial) has a well-defined limit as $m\to\infty$ at time $t=0$.

↳ Q: What happens as we start learning using gradient dynamics in this regime?

Theorem: [Jacot et al, 18, NTK] For all $T>0$, as $m\to\infty$ uniformly in $[0,T]$ and on compacts in $\mathbb{R}^n$,

$$K_t(x,x') \to \bar{K}(x,x'). \qquad \forall x,x'\in\Omega,\ \forall t\in[0,T].$$

$g(\theta,x)$ | **Consequence**: Learning happens in a linear space → The RKHS whose kernel is $\bar{K}$.

$\frac{d}{dt}\phi_t = -K_t\cdot R'(\phi_t)$ becomes (when $\bar{K}>0$)

$\frac{d}{dt}\phi_t = -\bar{K}\cdot(\phi_t - f^*)$ ⇒ global convergence at

exponential rate!

→ How come a non-linear model (shallow NN) behaves like a linear one?

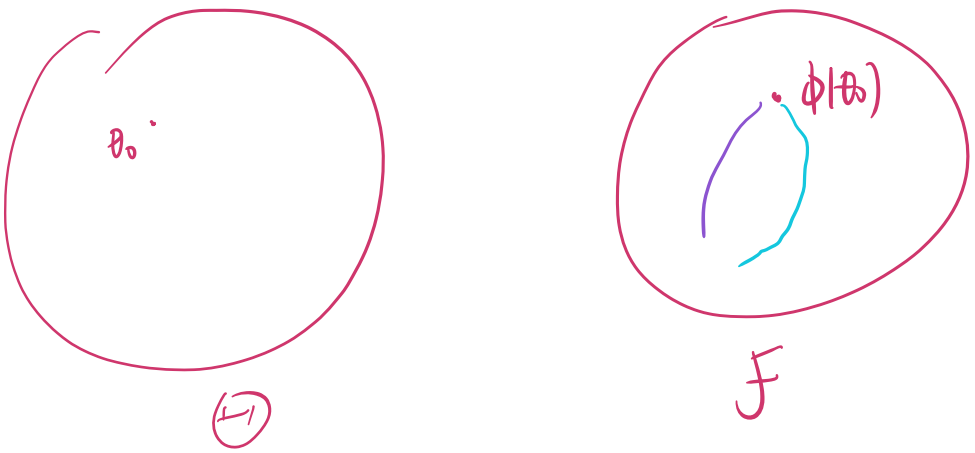
Lazy Dynamics [Chizat, Oyallon, Bach '19]

→ Let $\phi(\theta)$ be a differentiable map $\theta \in \Theta \xrightarrow{\phi} \mathcal{F}$

θ_0: initial point.

→ Consider the linearized tangent model at θ_0:

$$\bar{\phi}(\theta) = \phi(\theta_0) + D\phi(\theta_0) \cdot (\theta - \theta_0)$$



Two dynamics in function space:

(i) $\dot{\theta}_t = -\nabla R(\phi(\theta_t))$; $\phi_t = \phi(\theta_t)$ θ_0

(ii) $\dot{\bar{\theta}}_t = -\nabla R(\bar{\phi}(\bar{\theta}_t))$; $\bar{\phi}_t = \bar{\phi}(\bar{\theta}_t)$ $\bar{\theta}_0 = \theta_0$

Question: When are the learning dynamics under these two models similar?

→ Denote $F(\theta) = R(\phi(\theta))$ and assume θ_0 s.t. $F(\theta_0) > 0$.

→ A gradient step is $\theta_1 = \theta_0 - \eta \nabla F(\theta_0)$

→ Relative change in objective function: $F(\theta_1) \simeq F(\theta_0) + \langle \nabla F(\theta_0), \theta_1 - \theta_0 \rangle$

$$\Delta(F) = \frac{F(\theta_0) - F(\theta_1)}{F(\theta_0)} \simeq \eta \frac{\|\nabla F(\theta_0)\|^2}{F(\theta_0)}$$

→ Relative change in tangent space: $\Delta(D\phi) = \dfrac{\|D\phi(\theta_1) - D\phi(\theta_0)\|}{\|D\phi(\theta_0)\|}$

$$\overset{\text{C-S.}}{\leq} \eta \frac{\|D^2\phi(\theta_0)\| \cdot \|\nabla F(\theta_0)\|}{\|D\phi(\theta_0)\|}$$

→ Lazy Regime: $\Delta(D\phi) \ll \Delta(F)$

tangent space moves much slower than the loss itself.

→ If $R(f) = \|f - f^*\|^2$, $F(\theta) = \|\phi(\theta) - f^*\|^2$

$$\|\nabla F(\theta_0)\| = \|D\phi(\theta_0)^T (\phi(\theta_0) - f^*)\|$$

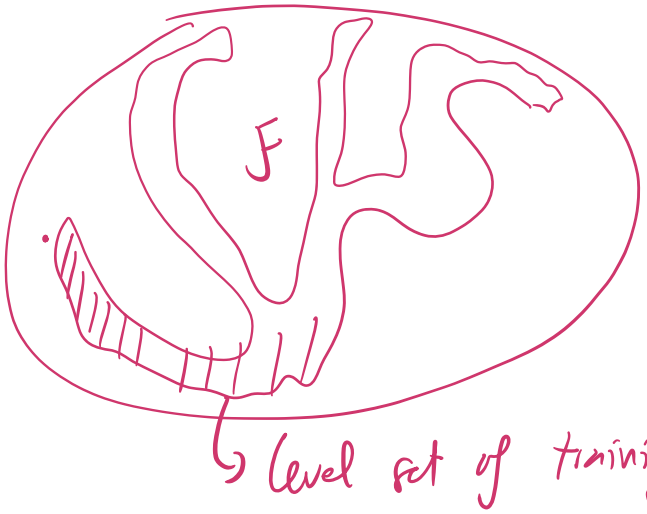
$$\lesssim \|D\phi(\theta_0)\| \cdot \|\phi(\theta_0) - f^*\|$$

→ If we define the relative scale of ϕ at point θ_0 as

$$\kappa_\phi(\theta) := \|\phi(\theta) - f^*\| \cdot \frac{\|D^2\phi(\theta)\|}{\|D\phi(\theta)\|^2}$$, the lazy regime

is equivalent to $\boxed{K_\phi(\theta) \ll 1}$

→ Why the name "lazy"? parameters almost do not move, because loss already converges at linear rate.



↳ Level set of training loss.

Theorem [COB, Thm 2.3 Simplified] Assume ϕ, $D\phi$ are Lipschitz in a neighborhood of θ_0. Let $\theta(t)$, $\bar{\theta}(t)$ be the gradient flows associated with ϕ, $\bar{\phi}$ respectively. Then for $t \leq C_\phi$, it holds

$$\frac{\| \phi(\theta(t)) - \bar{\phi}(\bar{\theta}(t)) \|}{\| \phi(\theta_0) - f^* \|} \lesssim t^2 \cdot \underbrace{K_\phi(\theta_0)}_{\text{↳ relative scaling.}}$$

Remarks:
- Constant C_ϕ is explicit
- In the lazy regime, ($K_\phi(\theta_0) \ll 1$), the non-linear trajectory tracks the linearised traj.
- In the over-parametrised regime, then the result can be extended uniformly in time.

(theorem 2.4).

Proof: See the paper → elementary.

Q. When does the lazy regime happen?

(i) Scaled Models: $\phi_\alpha(\theta) = \alpha \cdot \phi(\theta)$

$$K_{\phi_\alpha}(\theta_0) = \frac{1}{\alpha} \|\alpha\phi(\theta_0) - f^*\| \cdot \frac{\|D^2\phi(\theta_0)\|}{\|D\phi(\theta_0)\|^2}$$

in particular, when $\phi(\theta_0) = 0$, then

$$K_{\phi_\alpha}(\theta_0) = \frac{1}{\alpha} K_\phi(\theta_0).$$

Recall: $K_\phi(\theta) = \|\phi(\theta) - f^*\| \frac{\|D^2\phi(\theta_0)\|}{\|D\phi(\theta_0)\|^2}$

(ii) Homogeneous Models: $\phi(\lambda\theta) = \lambda^r \phi(\theta)$ $\forall \theta$, $\forall \lambda > 0$

same as before: $K_\phi(\lambda \cdot \theta_0) = \lambda^{-r} K_\phi(\theta_0)$. (for 0 init).

(iii) Single hidden layer NN

$$\phi_m(\theta) = \alpha(m) \sum_{i=1}^{m} g(\theta_i), \quad \theta_i \sim \mu \text{ iid}$$

with $\mathbb{E}_\mu[g(\theta)] = 0$, Dg Lipschitz.

$K_m = \mathbb{E} K_{\phi_m}(\theta)$?

- $\mathbb{E} \|\phi_m(\theta)\|^2 = m \cdot \alpha(m)^2 \mathbb{E} \|g(\theta)\|^2$

↳ ϕ_m is a sum of iid terms 0-mean.

- $D\phi_m(\theta) = \alpha(m)\left[Dg(\theta_1), \cdots, Dg(\theta_m)\right]$

$$\frac{D\phi_m(\theta)\, D\phi_m(\theta)^T}{m \cdot \alpha(m)^2} = \frac{1}{m}\sum_{i=1}^{m} Dg(\theta_i)\, Dg(\theta_i)^T$$

$$\downarrow \; m \to \infty$$

$$\mathbb{E}\left[Dg(\theta)\, Dg(\theta)^T\right]$$

$$\Rightarrow \mathbb{E}\,\|D\phi_m(\theta)\|^2 = \mathbb{E}\,\|D\phi_m(\theta)\, D\phi_m(\theta)^T\|$$

$$\sim m\,\alpha(m)^2\,\mathbb{E}\left[Dg(\theta)\, Dg(\theta)^T\right]$$

$$= m\,\alpha(m)^2\,\mathbb{E}\left[\|Dg(\theta)\|^2\right]$$

- $\|D^2\phi_m(\theta)\| = \sup_{u \in \mathbb{R}^{d\times m},\ \|u\|\le 1} \alpha(m) \sum_{i=1}^{m} u_i^T D^2 g(\theta)\, u_i$

$\left(\|A\| = \sup_{\|x\|\le 1} x^T A x\right)$

$$\le \alpha(m) \sup_{\theta_i} \|D^2 g(\theta_i)\|$$

$$\le \alpha(m) \cdot \mathrm{Lip}(Dg)$$

↳ $K_m = \mathbb{E}\,\|\phi(\theta) - f^*\| \cdot \dfrac{\|D^2\phi_m(\theta)\|}{\|D\phi_m(\theta)\|^2}$

$$\lesssim \left(\mathbb{E}\,\|\phi(\theta)\| + \|f^*\|\right) \frac{\mathbb{E}\,\|D^2\phi_m(\theta)\|}{\mathbb{E}\,\|D\phi_m(\theta)\|^2}$$

$$\left(\sqrt{m}\cdot\alpha(m)\cdot A + B\right)\frac{\alpha(m)\cdot C}{m\cdot\alpha(m)^2\cdot D}$$

$$\lesssim \frac{C_1}{\sqrt{m}} + \frac{C_2}{m\cdot\alpha(m)}$$

conclusion: if $m\cdot\alpha(m)\to\infty$ as $m\to\infty$, then the NN enters the lazy regime!

(in particular, NTK scaling $\alpha(m)=\frac{1}{\sqrt{m}}$).

Is this good or bad?

⊕ It guarantees global convergence; optimisation is not the bottleneck.

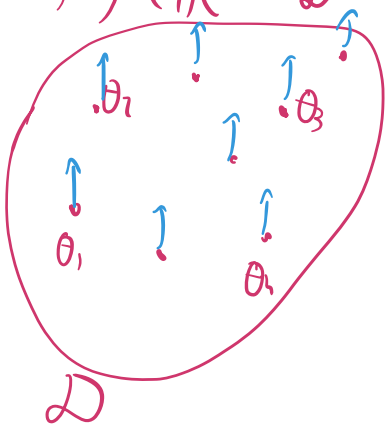
⊖ We are doing linear learning: No features are learnt; RKHS learning → approximation becomes the bottleneck!

Q: What happens when $m\cdot\alpha(m)=\Theta(1)$, ie $\alpha(m)=1/m$?

Shallow NN as particle interaction systems

→ Now our model becomes $\phi(x;\overbrace{\theta_1\cdots\theta_m}^{\theta}) = \frac{1}{m}\sum_{i=1}^{m} g(x;\theta_i)$

$g(x;\theta) = c \cdot V^{-}(\langle x, w \rangle + b)$, $\theta = (a; b; c) \in \mathbb{R}^{\cdot} = D$.



→ Integral Representation:

$$\phi(x;\theta) = \int_D g(x;\theta)\, d\mu_m(\theta)$$

$$\mu_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{\theta_i}$$ the empirical measure associated with $\theta = (\theta_1 \dots \theta_m)$

$$\left(\int f(x)\, d\delta_y(x) = f(y) \right)$$

↳ In next lecture, we will study:

(i) How the associated function is how bigger then the RKHS associated with lazy training

(ii) training dynamics in measure space.

(iii) Natural space where we have representation learning.