Lecture 8 : Shallow NNs in the feature learning regime

- In previous lecture: lazy training: a non-linear model $\theta \to \phi(\theta)$ can behave as a linear one when its curvature around initialisation vanishes → analysis through TK / NTK.

→ For $\phi(\theta)$ = NN (shallow) of the form

$$f(\theta) = \alpha(m) \sum_{j=1}^{m} \phi(\theta_j) \qquad \theta_j \sim \mu \text{ , lazy regime}$$

as soon as $m \cdot \alpha(m) \to \infty$.

→ Today Active regime $\alpha(m) = 1/m$.

Q1: What is the functional space?
Q2: Learning dynamics?

Shallow NN and Variation-Norm spaces.

→ with choice $\alpha(m) = 1/m$, the model becomes

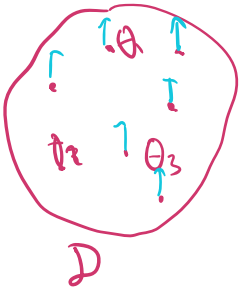
$$f(x; \theta_1 \dots \theta_m) = \frac{1}{m} \sum_{j=1}^{m} \phi(x; \theta_j) \qquad x \in \mathbb{R}^d$$

$$\phi(x;\theta) = c \cdot \sigma(\langle x, a \rangle + b) \qquad \theta = (a; b; c) \in \mathbb{R}^{d+1+1} =: D$$

c: output weight, a: input weight, b: bias, $\bar{\theta} \in \bar{D}$

$$c \cdot \sigma(\langle (x;1); \bar{\theta} \rangle)$$

→ f is totally characterised by the "point cloud"



$\theta_1 \cdots \theta_m$.

$\theta_1 \to \theta_{\alpha(1)}$

$\vdots \to \vdots$

$\theta_m \to \theta_{\alpha(m)}$

$\alpha: \{1,\dots,m\} \to \{1,\dots,m\}$.

→ I can completely describe f via the empirical measure

$$\mu_m = \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_j} \in \mathcal{P}(D)$$

δ : Dirac Mass.

→

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

Integral representation of f.

→ Q: Associated function class? Consider a non-negative regularisation term $V: D \to \mathbb{R}$ $V(\theta) = \|\theta\|^2$ [weight decay]

$V(\theta) = |c| \cdot \|\bar{\theta}\|$ [Path norm]

→ Define

$$\underset{(V)}{F} = \left\{ f(x) = \int_D \phi(x;\theta)\, d\mu(\theta) \text{ for } \mu \in \mathcal{P}(D) \text{ such that } \int_D V(\theta)\, d\mu(\theta) < +\infty \right\}.$$

→ Fact: F is a normed (= Banach) space, with norm

$$\|f\| = \inf \left\{ \int V\, d\mu \;;\; f(x) = \int \phi(x;\theta)\, d\mu \right\}$$

$\|f\|_V = \inf\{ \}$

↳ Ex: For $V(c;\bar\theta) = |c|\,\|\bar\theta\|$ this is called the variation-norm space [Bach' 17]

a.k.a. "Barron" space.

Q: How does this space compare with the RKHS?

A Write $\mathcal{D} = \mathbb{R}\times\bar{\mathcal{D}}$ and assume $\bar{\mathcal{D}} = \mathbb{S}^d$ unit $(d+1)$-dim. sphere.

→ Fix an arbitrary measure Π on $\bar{\mathcal{D}}$, and consider

$$\mathcal{F}_\Pi = \left\{ f(x) = \int g(\bar\theta)\,\sigma(\langle \tilde{x}, \bar\theta\rangle)\,d\Pi(\bar\theta)\,; \quad g\in L_2(\bar{\mathcal{D}};\Pi) \right\} \quad \left(\int |g(\bar\theta)|^2 d\Pi < +\infty\right)$$

($\tilde{x} = (x,1)$)

↳ We saw that $\mathcal{F}_\Pi$ is an RKHS, with kernel given by

$$K(x,x') = \mathbb{E}_{\bar\theta\sim\Pi}\left\{ \sigma(\langle x,\bar\theta\rangle)\,\sigma(\langle x',\bar\theta\rangle)\right\}$$

→ If $f\in\mathcal{F}_\Pi$, let's define $\mu\in P(\mathcal{D})$

$$\mu(c,\bar\theta) = \delta(c - g(\bar\theta))\cdot\Pi(\bar\theta)$$

(1) $$f(x) = \int_{\mathcal{D}=\mathbb{R}\times\bar{\mathcal{D}}} \phi(x;\theta)\,d\mu(\theta) \qquad V(c,\bar\theta) = |c|$$

(ii) $\|f\|_F \leq \int V(c,\bar{\theta})\, d\mu = \int_{\bar{D}} |g(\bar{\theta})|\, d\pi(\bar{\theta}) =$

$= \mathbb{E}_{\bar{\theta}\sim\pi}\left[|g(\bar{\theta})|\right]$

⇒ By Jensen's ineq. $|\mathbb{E}X|^2 \leq \mathbb{E}X^2$

→ This means that $f \in \mathcal{F}$

$\|f\|_F \leq \|f\|_{\mathcal{F}_\pi} \quad \forall \pi$

→ Therefore, $\mathcal{F} = \bigcup_{\pi \in P(\bar{D})} \mathcal{F}_\pi$

→ Ex $f(x) = \sigma(\langle \bar{x}, \bar{\theta}^\bullet \rangle)$ → Realised as a Dirac for $\mu = d_{\bar{\theta}^\bullet}$ so $\|f\|_F = 1$

↳ Not in the RKHS $\mathcal{F}_\pi$ for $\pi \ll$ unif. measure.

→ Remark: Instance of a "separation" between two different hypotheses spaces

→ Remark: Same thing holds for $V(\theta) = \|\theta\|^2$ under homogeneous activations ($\sigma(tx) = t\sigma(x)$)

Q: Is it still possible to learn efficiently in this bigger

space?

↳ In terms of generalisation, all the Rademacher-based analysis still work (by replacing $\|f\|_{\mathcal{F}_{TT}}$ by $\|f\|_{\mathcal{F}}$).

↳ What about optim?

Shallow NNs (in active regime) as particle interaction systems

→ Focus on least-squares regression:

$$f_\theta = \frac{1}{m}\sum_{j=1}^{m}\phi(\theta_j)$$

$$\min_{\theta_1 \cdots \theta_m} L(\theta_1 \cdots \theta_m) = \| f_\theta - f^* \|_\nu^2 + \lambda \frac{1}{m}\sum_{j=1}^{m} V(\theta_j)$$

→ By expanding the square, $\langle f_\theta, f_\theta \rangle_\nu = \|f_\theta\|_\nu^2 + \|f^*\|_\nu^2 - 2\langle f_\theta, f^* \rangle_\nu$

and using the linear $f_\theta = \frac{1}{m}\sum_{j=1}^{m}\phi(\theta_j)$, $\langle f, g \rangle_\nu := \frac{1}{n}\sum_{i=1}^{n} f(x_i) g(x_i)$

we can write L as

$$L(\theta_1 \cdots \theta_m) = C - \frac{2}{m}\sum_{j=1}^{m} F(\theta_j) + \frac{1}{m^2}\sum_{i,j} K(\theta_i, \theta_j), \text{ with}$$

$$F(\theta) := \langle \phi(\theta), f^* \rangle_\nu - \frac{\lambda}{2} V(\theta)$$

$$K(\theta, \theta') = \langle \phi(\theta), \phi(\theta') \rangle_\nu$$

↳ Interpret L as the Hamiltonian/Energy of a

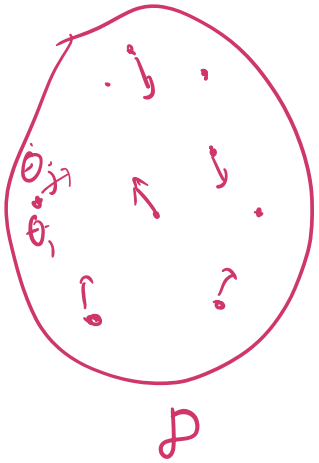
system of m particles $\{\theta_1 \cdots \theta_m\}$ in $\mathcal{D}$.

$F \longleftrightarrow$ external field.

$K \longleftrightarrow$ interaction kernel.

→ Gradient flow with respect to $\theta_1 \cdots \theta_m$ gives us

$$\dot{\theta}_j = -\frac{m}{2} \nabla_{\theta_j} L(\theta_1 \cdots \theta_m) = +\nabla F(\theta_j) - \frac{1}{m} \sum_{j=1}^{m} \nabla K(\theta_j, \theta_j)$$



↳ This is Lagrangian perspective. (tracking each particle).

↳ This system of m d-dim ODEs is very complex. L is non-convex wrt $\theta_1 \cdots \theta_m$.

→ Can we aim to obtain a "simple" collective behavior?

→ let's consider the Eulerian perspective instead.

$$\vec{\theta} = (\theta_1 \cdots \theta_m) \in \mathcal{D}^m \longleftrightarrow \mu = \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_j} \in \mathcal{P}(\mathcal{D})$$

$$f(x) = \int_{\mathcal{D}} \phi(x; \theta) \, d\mu$$

→ The Energy/ loss now becomes.

$$L(\theta_1 \cdots \theta_m) = \text{const} - \frac{2}{m} \sum_{j=1}^{m} F(\theta_j) + \frac{1}{m^2} \sum_{i,j} K(\theta_i, \theta_j), \text{ with}$$

$$\mathcal{L}[\mu] = -2\int_D F(\theta)\,d\mu + \iint_{D\times D} K(\theta,\theta')\,d\mu(\theta)\,d\mu(\theta')$$

$\mu \in \mathcal{P}(D)$.

$$\left\| \int \phi\,d\mu - f^* \right\|_V^2 + \int V(\theta)\,d\mu.$$

→ $\mathcal{L}$ is convex with respect to μ.

↳ If $\mu_0, \mu_1 \in \mathcal{P}(D)$ and

$$\mu_t = t\mu_1 + (1-t)\mu_0$$

then $\mathcal{L}[\mu_t] \le t\mathcal{L}[\mu_1] + (1-t)\mathcal{L}[\mu_0]$

→ Is this convexity "felt" by our dynamics. ?

Relationship between GD/GF and a choice of metric ?

Proximal View point of GD

→ The standard (Euclean) GD step

$$\theta_{t+1} = \theta_t - \eta\nabla f(\theta_t)$$

$$= \operatorname*{argmin}_\theta \left\{ f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta}\|\theta-\theta_t\|^2 \right\}$$

linear approx of f at θ_t ; proximity term.

→ Proximity can be measured by different metrics!

→ The extension of GD to general (non-Euclien) metrics is given by Mirror Descent [Nemirovski & Yudin 83]

$$\theta_{t+1} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{\eta_t} D(\theta; \theta_t) \right\}$$

$$\approx \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ f(\theta) + \frac{1}{\eta_t} D(\theta; \theta_t) \right\}$$

Bregman Divergence.

→ Back to our setting in measure space.

$\vec{\theta}_t = (\theta_1 \cdots \theta_m) \longleftrightarrow \mu_t$

(Euclean GD ↓)

$\vec{\theta}_{t+1} \longleftrightarrow \mu_{t+1}$

what kind of descent does this update correspond to?

W_2^2

→ Recall that $\dot{\theta}_j = +\nabla F(\theta_j) - \frac{1}{m} \sum_{j'=1}^{m} \nabla K(\theta_j, \theta_{j'})$

using $\mu_t = \frac{1}{m} \sum_{j=1}^{m} \delta_{\theta_j(t)}$, we have

→ $\dot{\theta}_j(t) = \nabla F(\theta_j(t)) - \int_D \nabla K(\theta_j^{(t)}, \theta) \, d\mu_t(\theta)$

$:= -\nabla G(\theta_j(t); \mu_t)$, with

$G(\theta; \mu) = -F(\theta) + \int K(\theta, \theta') \, d\mu(\theta')$

Recall $\mathcal{L}[\mu] = -2\int_D F(\theta)\, d\mu + \iint_{D\times D} K(\theta,\theta')\, d\mu\, d\mu(\theta')$

$$\frac{\delta \mathcal{L}}{\delta \mu(\theta)} \Rightarrow -2F(\theta) + 2\int_D K(\theta,\theta')\, d\mu(\theta')$$

G is the <u>first variation</u> of $\mathcal{L}$ w.r.t. μ

$$G = \frac{\delta \mathcal{L}}{\delta \mu}.$$

→ In physics, G is the <u>instantaneous</u> potential of the system.

→ Consider a (smooth) test function

$$\chi : D \to \mathbb{R}$$

$$\int_D \chi(\theta)\, d\mu_t(\theta) = \frac{1}{m}\sum_{j=1}^{m} \chi(\theta_j(t))$$

(time derivative)

$$\int_D \chi(\theta)\, \partial_t d\mu_t(\theta) = \frac{1}{m}\sum_{j=1}^{m} \left\langle \nabla\chi(\theta_j(t)),\ \dot{\theta}_j(t) \right\rangle$$

(time derivative)

$$= -\frac{1}{m}\sum_{j=1}^{m} \left\langle \nabla\chi(\theta_j(t)),\ \nabla G(\theta_j(t), \mu_t) \right\rangle$$

$$= -\int_{\mathcal{D}} \langle \nabla \mathcal{K}(\theta), \nabla G(\theta; \mu_t) \rangle \, d\mu_t(\theta)$$

↳ This is the continuity equation / transport equation / Liouville equation.

$$\partial_t \mu_t = \operatorname{div}\left(\nabla G(\theta; \mu_t) \cdot \mu_t \right)$$

↳ Time-dependent velocity field.

PDE in the space of measures.

Q: Is this PDE a gradient flow? wrt which metric?

↳ The proximal interpretation of this dynamics is

$$\mu_{t+1} = \underset{\mu \in \mathcal{P}(\mathcal{D})}{\operatorname{argmin}} \left\{ \mathcal{L}[\mu] + \frac{1}{2\eta} W_2^2(\mu_t, \mu) \right\}$$

$\mathcal{D}$ metric space $\rightsquigarrow$ $\mathcal{P}(\mathcal{D})$

Remarks: . This formulation of SNN dynamics was developed in [Chizat, Bach '18]
[Mei, Nguyen, Montanari '18]
[Rotskoff, EVE '18]
[Sirignano, Spiliopoulos '18]

- Description is exact

- Adding noise to the dynamics (GF → Langevin dynamics)
  ↓
  McKean-Vlasov equation.

→ Two main questions

(i) under what conditions this PDE converges to the global optimum of $\mathcal{L}$ (convergence in t)

(ii) Effect of over-parametrisation (convergence in m)?