Tuesday 09/07 THEODOR MISIAKIEWICZ

# MEAN-FIELD THEORY OF NEURAL NETWORKS

Today: quick overview + planning rest of semester

## SETTING:

- Given iid data $\{(x_i, y_i)\}_{1 \le i \le n}$

  e.g. $y_i = f_*(x_i) + \varepsilon_i$

  - covariates $x_i \in \mathbb{R}^d$ $\quad x_i \overset{iid}{\sim} \mathbb{P}$
  - noise $\varepsilon_i$ independent and $\mathbb{E}[\varepsilon_i] = 0$

- Parametric class of functions $f : (x, \theta) \mapsto f(x; \theta)$

- Test error with squared loss:

$$R(\theta) = \mathbb{E}_{y,x}\{(y - f(x,\theta))^2\}$$

E.g., fit $\hat{f}(x, \hat{\theta})$ by minimizing

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2$$

# 2-LAYERS NEURAL NETWORKS:

$$f(x;\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i)$$

$N$ = # of hidden units (neurons)

$\sigma_* : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ activation function

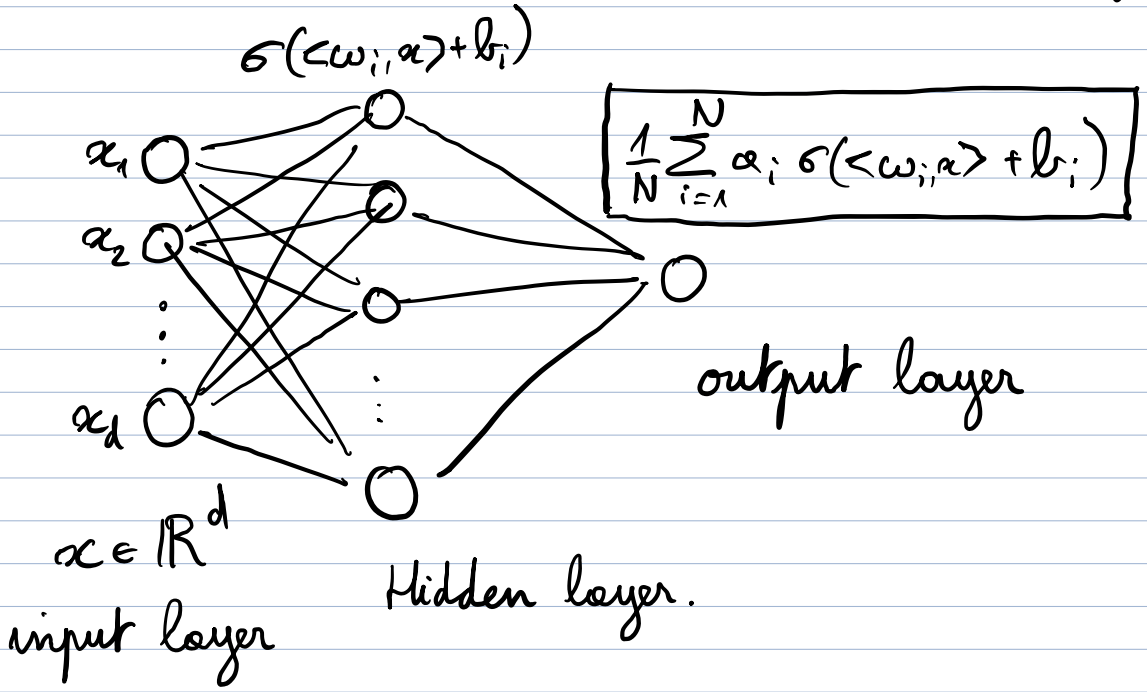
$\theta_i \in \mathbb{R}^D$ parameters

$\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{ND}$

Standard choice : $\sigma_*(x;\theta_i) = a_i\,\sigma(\langle \omega_i, x\rangle + b_i)$

$\theta_i = (a_i, b_i, \omega_i) \in \mathbb{R}^{d+2}$

e.g. $\sigma(x) = \max(x, 0)$, (ReLu) $\qquad \sigma(x) = \dfrac{1}{1+e^{-2x}}$ (sigmoid)



Why 2-layers NNs? ⇒ rich enough class of functions

$\rightarrow \mathbb{E}[f_*(x)^2] < \infty$

**Theorem** [Cybenko, 1989] Take $x \sim \mathbb{P}$, $f_* \in L^2(\mathbb{P})$ and

$\sigma : \mathbb{R} \to \mathbb{R}$ continuous with $\begin{cases} \lim_{x \to \infty} \sigma(x) = 1 \\ \lim_{x \to -\infty} \sigma(x) = 0 \end{cases}$

Then for any $\varepsilon > 0$, there exists $N = N(\varepsilon)$ such that

$$\inf_{\{(a_i, b_i, w_i)\}} \mathbb{E}\left[ \left( f_*(x) - \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle + b_i) \right)^2 \right] \leq \varepsilon$$

⇒ 2-layers NN can approximate any reasonable functions

⇒ From classical theory of universal approximation it is often more insightful to think in terms of empirical distribution:

$$f_N(x; \Theta) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(x; \theta_i) = \int \sigma_*(x; \theta) \, \hat{\rho}_N(d\theta)$$

$$=: f(x; \hat{\rho}_N) \quad \text{with} \quad \hat{\rho}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i}$$

$f(x; \rho) = \int \sigma_*(x; \theta)\, \rho(d\theta)$ for general distribution $\rho \in \mathcal{P}(\mathbb{R}^D)$

# MEAN-FIELD LIMIT : STATICS

Preliminary: look at the population risk

$$R_N(\theta) = \mathbb{E}_{y,x}\left[\left(y - \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\theta_i)\right)^2\right]$$

$$\equiv R_\# + \frac{2}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\theta_i,\theta_j)$$

where
- $R_\# = \mathbb{E}[y^2]$
- $V(\theta) = -\mathbb{E}[y\,\sigma_*(x;\theta)]$
- $U(\theta_1,\theta_2) = \mathbb{E}[\sigma_*(x;\theta_1)\,\sigma_*(x;\theta_2)]$

Rmk: Can think about $R_N(\theta)$ as the energy of $N$ particles in $D$ dimensions moving in an external potential $V(\theta)$ and with pairwise potentials $U(\theta_i,\theta_j)$

→ kernel $U$ is PSD (U is a repulsive interaction)

Again can replace $\hat{\rho}_N \to \rho \in \mathcal{P}(\mathbb{R}^D)$

$$R(\rho) = R_\# + 2\int V(\theta)\,\rho(d\theta) + \int U(\theta_1,\theta_2)\,\rho(d\theta_1)\,\rho(d\theta_2)$$

$U(\theta_1, \theta_2) = \int \sigma_*(x, \theta_1)\, \sigma_*(x; \theta_2)\, \mathbb{P}(dx)$

## Properties of this risk:

① $\rho \mapsto R(\rho)$ convex function on $\mathcal{P}(\mathbb{R}^D)$

$\frac{1}{2}(\rho_1 + \rho_2)$

② Minimizing $R_N(\theta)$ and $R(\rho)$ is not much different

**Prop:** Assume $\exists \varepsilon > 0$ such that for any $\rho \in \mathcal{P}(\mathbb{R}^D)$

if $R(\rho) \le \inf_{\rho} R(\rho) + \varepsilon \implies \int U(\theta, \theta)\, \rho(d\theta) \le K$

Then $\left| \inf_{\theta} R_N(\theta) - \inf_{\rho} R(\rho) \right| \le \frac{K}{N}$

$\inf R_N(\theta) \ge \inf_{\rho} R(\rho)$ $\quad \theta_i \overset{iid}{\sim} \rho$ $\quad |R_N(\theta) - \mathbb{E} R_N(\theta)| \le \frac{1}{N} \int U(\theta, \theta)\, \rho(d\theta)$

$\mathbb{E}[R_N(\theta)] = R(\rho) +$

③ Define functional derivative

$$\Psi(\theta; \rho) \equiv \frac{1}{2} \frac{\delta R}{\delta \rho(\theta)} = V(\theta) + \int U(\theta, \theta')\, \rho(d\theta')$$

("additional energy of adding a single particle at $\theta \in \mathbb{R}^D$")

**Global minima:** distributions $\rho_*$ such that

$$\text{support}(\rho_*) \subseteq \operatorname*{argmin}_{\theta \in \mathbb{R}^D} \Psi(\theta; \rho_*)$$

(→ "Energy cannot decrease by moving mass from $\text{supp}(\rho_*)$" — infinitesimal)

# STOCHASTIC GRADIENT DESCENT

We want to minimize $R_N(\theta)$ : do SGD

$$\theta^{k+1} = \theta^k - \frac{\varepsilon}{2} \nabla_\theta \ell(y_k, x_k; \theta^k)$$

→ consider one-pass over the data : $(y_k, x_k)$ are iid

Get $$\theta_i^{k+1} = \theta_i^k + \varepsilon \nabla_{\theta_i} \sigma_*(x_k; \theta_i^k) \left( y_k - \frac{1}{N} \sum_{i=1}^N \sigma_*(x_k; \theta_i^k) \right)$$

# MEAN-FIELD LIMIT OF SGD : DISTRIBUTIONAL DYNAMICS

We will take both $\varepsilon \to 0$ (continuous time limit)
$N \to \infty$

SGD dynamics describes a set of N particles, with $\theta_i^k$ velocity of particle i :

$$v_i^k = \frac{\theta_i^{k+1} - \theta_i^k}{\varepsilon} = \nabla_{\theta_i} \left( y_k \sigma_*(x_k; \theta_i^k) \right) - \frac{1}{N} \sum_{j=1}^N \left( \nabla_{\theta_i} \sigma_*(x_k; \theta_i^k) \right) \sigma_*(x_k; \theta_j^k)$$

→ taking expectation wrt $y_k, x_k$ (denote $F_k$)

$$\mathbb{E}[v_i^k \mid \mathcal{F}_k] = -\nabla_{\theta_i} V(\theta_i^k) - \frac{1}{N}\sum_{j=1}^{N} \nabla_{\theta_i} U(\theta_i^k, \theta_j^k)$$

→ if at time $k$, density of particles $\approx \varrho_{k\varepsilon}$

$$\mathbb{E}[v_i^k \mid \mathcal{F}_k] \approx v(\theta_i^k; \varrho_t) = -\nabla_{\theta_i} \Psi(\theta_i^k; \varrho_{k\varepsilon})$$

→ Denote $t = k\varepsilon$

→ density $\varrho_t$ should satisfy the continuity equation

$$\partial_t \varrho_t(\theta) + \nabla_\theta \cdot (\varrho_t(\theta)\, v(\theta; \varrho_t)) = 0$$

i.e.
$$\partial_t \varrho_t = \nabla_\theta \cdot (\varrho_t \nabla_\theta \Psi(\theta; \varrho_t))$$

DISTRIBUTIONAL DYNAMICS

Several groups in 2018 showed
[Mei, Montanari, Nguyen], [Chizat, Bach]
[Rotskoff, Vanden-Eijnden], [Sirignano, Spiliopoulos]

$$\lim_{N\to\infty} \lim_{\varepsilon\to 0} \left\| f(\cdot; \varrho^{\lfloor t/\varepsilon \rfloor}_{(N)}) - f(\cdot; \varrho_t) \right\|_{L^2} = 0$$

$$\lim_{N\to\infty} \lim_{\varepsilon\to 0} \left| R_N(\theta^{\lfloor t/\varepsilon \rfloor}) - R(\varrho_t) \right| = 0$$

## NOISY SGD

$$\theta_i^{k+1} = \theta_i^k - \frac{\varepsilon}{2} \nabla_{\theta_i} \ell(y_k, x_k; \theta^k) + \sqrt{\varepsilon/\beta}\, g_i^k$$

$$g_i^k \sim N(0, Id_D)$$

Ger: $$\partial_t \rho_t = \nabla_\theta \cdot (\rho_t \nabla_\theta \Psi(\theta; \rho_t)) + \frac{1}{\beta} \Delta \rho_t$$

+ diffusion term

## GRADIENT FLOW

$$\partial_t \rho_t = \nabla_\theta \cdot (\rho_t \nabla_\theta \Psi(\theta, \rho_t))$$

→ is the gradient flow for $R(\rho)$ in Wasserstein metric

$$\rho_{t+\varepsilon} \approx \underset{\rho \in \mathcal{P}(\mathbb{R}^D)}{\operatorname{argmin}} \left\{ R(\rho) + \frac{1}{2\varepsilon} W_2(\rho, \rho_t)^2 \right\}$$

→ if diffusion term, GF of free energy

$$F(\rho) = \frac{1}{2} R(\rho) - \frac{1}{\beta} S(\rho)$$

with $S(\rho) = -\int \rho(\theta) \log \rho(\theta)\, d\theta$

$$\frac{1}{N} \sum_{i=1}^{N} a_i^{(3)} \sigma\left( \langle w_i^{(2)}, x^{(1)} \rangle \right)$$

$$x^{(1)} = \sigma\left( \langle w_{ij}^{(1)}, x \rangle \right)$$

# TOPICS:

① If anyone is interested in presenting some mathematical topics related to MF limit
(Gradient Flow in Wasserstein spaces etc.)

## Ⓐ Convergence of SGD process to PDE

E.g., if we are interested in the risk

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\Theta^k) - R(\varrho_{k\varepsilon})| \leq Err(T,N,\varepsilon,\dots)$$

E.g., $\sigma_*(x;\Theta) = a\,\sigma(x;\omega)$ $\quad v(\omega) = -\mathbb{E}[c_y\,\sigma(x;\omega)]$

$u(\omega_1,\omega_2) = \mathbb{E}[\sigma(x;\omega_1)\sigma(x,\omega_2)]$

and $|y|, \|\sigma\|_\infty, \|\nabla v(\omega)\|_2, \|\nabla u(\omega_1,\omega_2)\|_2$
$\|\nabla v(\omega)\|_{Lip}, \|\nabla u(\omega_1,\omega_2)\|_{Lip} \leq K$
$+\ \nabla_\omega \sigma(\cdot,\omega)$ $K$-subgaussian

$T = O(1)$, $T = m\varepsilon$, $m = O(\frac{1}{\varepsilon})$

then with proba $\geq 1 - e^{-z^2}$, $\quad \tilde{K} = \text{poly}(K)$

$$Err(T,N,\varepsilon,\dots) \leq \tilde{K} e^{\tilde{K}T^3}\left\{\frac{\sqrt{\log N} + z}{\sqrt{N}} + \left(\sqrt{D + \log N} + z\right)\sqrt{\varepsilon}\right\}$$

[Misiakiewicz, Mei, Montanari, 2019]

Proof: couplings + propagation of chaos + concentration of measure

→ For $T = O(1)$, $N \gg$ cste that only depend on $K$

→ still $e^{KT^3}$ (worst case analysis)

→ Can we do better?

## (B) Global convergence of the PDE

When does $\lim_{t \to \infty} R(\rho_t) = \inf_{\rho} R(\rho)$ ?

**Noisy SGD:**

unique minimizer $\rho_*(\theta) = \frac{1}{Z(T)} \exp\left(-\beta \Psi(\theta; \rho_*)\right)$

Free energy strictly decreasing along PDE path

$F(\rho_t) \longrightarrow \inf_{\rho} F(\rho)$

and for $\beta > \beta_*(\delta)$, $\lim_{t \to \infty} R(\rho_t) \leq \inf_{\rho} R(\rho) + \delta$

## Noiseless SGD

$\rho$ fixed point of PDE iff $\text{support}(\rho) \subseteq \{\theta : \nabla_\theta \psi(\theta; \rho) = 0\}$

[Chizat, Bach, 2018]

Morse-Sard type condition
+ good initialization
+ $\rho_t \to \rho_\infty$ in $W_2$

$\Rightarrow$ $\rho_\infty$ is a global optimizer

[Wojtowytsch, 2020]

extends these results

[Nguyen, Pham, 2020]

replace Morse-Sard type condition by a topology argument

## (C) Rates of convergence of the PDE to minimizer

E.g. in noisy SGD.

Important to transfer results to practical settings

[Javanmard, Mondelli, Montanari, 2019]

$$\sigma_*(x;w) = K^\delta(x - w_i)$$

$$K^\delta(x) = \delta^{-d} K(x/\delta) \qquad K: \mathbb{R}^d \to \mathbb{R}_{\geq 0}$$ 1st order kernel with compact support

Target function: $f_*$ $\alpha$-strongly concave + smooth

Then $R(\rho)$ is $\alpha$-displacement convex

$$R(\rho_t) \leq R(\rho_0) e^{-2\alpha t}$$

$\Rightarrow$ In general, $R(\rho)$ not displacement convex.

## ① Representation vs Optimization

① Representation: $\frac{1}{N}\sum_{i=1}^{N} a_i \sigma(x;w_i) \longrightarrow f(x)$

② Optimization: $f(x;\rho_t) \longrightarrow f(x)$

Is it easier to optimize for some class of functions?

E.g. What are easier to represent?

**Barron-norm:** $f(x) = \int_\Omega a\sigma(\omega^T x)\rho(da, d\omega)$

$= \mathbb{E}_\rho[a\sigma(\omega^T x)]$

$\|f\|_B = \inf_{\rho : f = \mathbb{E}_\rho} \mathbb{E}_\rho[|a|\|\omega\|_1]$

**Prop:** We have

$$\inf_{f_N : N \text{ neurons}} \|f_* - f_N\|_{L^2} \leq \frac{\|f\|_B}{\sqrt{N}}$$

$\exists$ 1-Lipschitz function such that:

(*) $$\inf_{f_N : N \text{ neurons}} \|f_* - f_N\|_{L^2} \geq \frac{c}{N^{1/d}}$$

$$\|f(\cdot; \rho_t)\|_B \leq \int |\theta|^2 \rho_t(d\theta) = o(t) \qquad (**)$$

(second moment ↗ slower than linearly)

(*)+(**) $$R(\rho^t) \geq c\, t^{-\frac{c}{d+2}}$$

[Wojtowytsch, Weinan, 2020]

## (E) Multilayer NNs

Several attempts

[Araujo, Oliveira, Yukimura, 2019]

[Nguyen, Pham, 2020] neuronal embedding

→ degeneracy of the middle layers that collapse into one neuron

→ basically can factorize distribution weights

$$\rho^t = \rho^t_{(0)} \otimes \rho^t_{(1)} \otimes \dots \otimes \rho^t_{(L)}$$

[Fang, Lee, Yang, Zhang, 2020]

attempt to overcome degeneracy using new representation "neural feature flow"

## (F) NTK vs mean-field limit

→ Might be of general interest

Can we compare mean-field and NTK solutions?
Can we study the transition between the two?

[Misiakiewicz, Mei, Montanari, 2019]

NTK appears at beginning of dynamics.

[Chizat, Bach, 2020]

MF find $F_1$ - max margin solution

$\neq$ NTK — $F_2$ - ——————

- How many weeks?
- Can invite speakers