Mean Field Analysis of two layer Neural Networks II.

1) Recap setting from last time

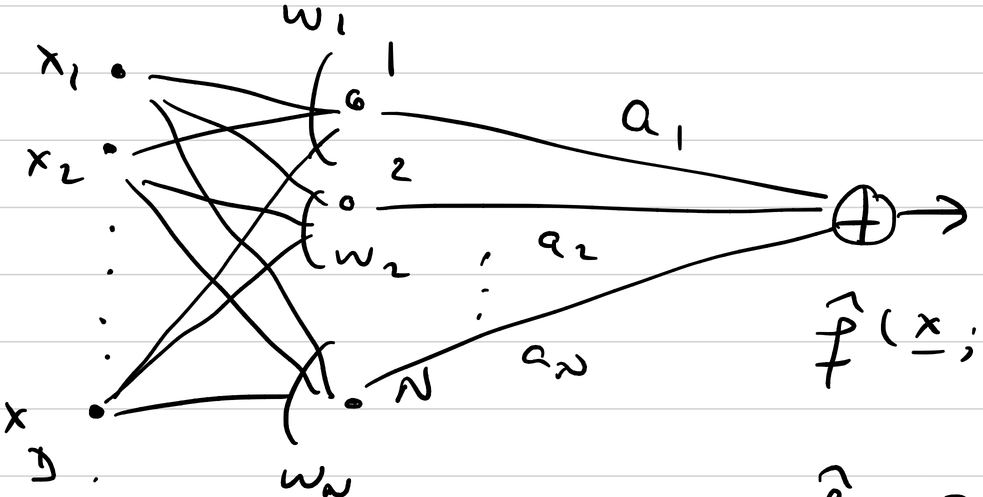
- Training data (x_i, y_i), $i = 1 \dots N$.

e.g $y_i = f(\underline{x}_i) + \xi_i$, $\xi_i \sim \mathcal{N}(0,1)$ iid

$x_i \sim \mathcal{D}_X$. iid

f some "reasonable" fct.

- We want to train a two layer NN:



$$\hat{f}(\underline{x}; \vartheta) = \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\underline{w}_i^T \cdot x + b_i)$$

$$\hat{f} : \mathbb{R}^D \to \mathbb{R}.$$

weights $\vartheta_i = (a_i, b_i, \underline{w}_i) \in \mathbb{R}^{D+2}$

$\vartheta = (\vartheta_1 \dots \vartheta_N)$.

- We want to analyze SGD for true risk:

$$R_N(\theta) = \mathbb{E}_{\mathcal{D}}\left(\ell(\underline{x}, y; \theta) \right)$$

$$\ell(\underline{x}, y; \theta) = (y - \hat{f}(\underline{x}; \theta))^2$$

for $N \to +\infty$, D fixed.

- We argued that in this limit the problem is amenable to analysis of a PDE. The intuition comes interpretation as a "<u>particle system</u>".

Recap SGD iterations: $k = 1, 2, \dots, T$

$$\vartheta^{k+1} = \vartheta^{k} - \delta_k \underline{\nabla}_{\vartheta} \ell(x_k, y_k; \vartheta^k)$$

$\underline{\nabla}_{\vartheta} \ell(x_k, y_k; \vartheta^k) =$ Stochastic gradient since

$$\mathbb{E}_{\vartheta}\left(\nabla_{\vartheta} \ell(x, y; \vartheta^k)\right) = \nabla_{\vartheta} R_N(\vartheta^k).$$ unbiased.

We computed last time:

$$\nabla_{\vartheta_i} \ell(x_k, y_k; \vartheta^k) = \left(y_k - \frac{1}{N} \sum_{i=1}^{N} a_i^k \sigma(\underline{w}_i^k \cdot \underline{x} + s_i^k)\right)$$

$$\cdot \nabla_{\vartheta_i}\left(a_i^k \sigma(\underline{w}_i^k \cdot \underline{x}_k + s_i^k)\right)$$

Recap true risk (or generalization error):

$$R_N(\theta) = \mathbb{E}_{\mathcal{D}}\left[\left(y - \frac{1}{N}\sum_{i=1}^{N} a_i \sigma(\underline{w}_i \cdot \underline{x} + b_i)\right)^2\right]$$

$$= \mathbb{E}(y^2) - \frac{2}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\theta_i, \theta_j)$$

$$\begin{cases} V(\theta) = -\mathbb{E}_{\mathcal{D}}\left[y\, a \sigma(\underline{w}\cdot\underline{x} + \underline{b})\right] \\ U(\theta, \theta') = \mathbb{E}_{\mathcal{D}}\left[a\sigma(\underline{w}\cdot\underline{x}+b)\, a'\sigma(\underline{w}'\cdot\underline{x}+b')\right] \end{cases}$$

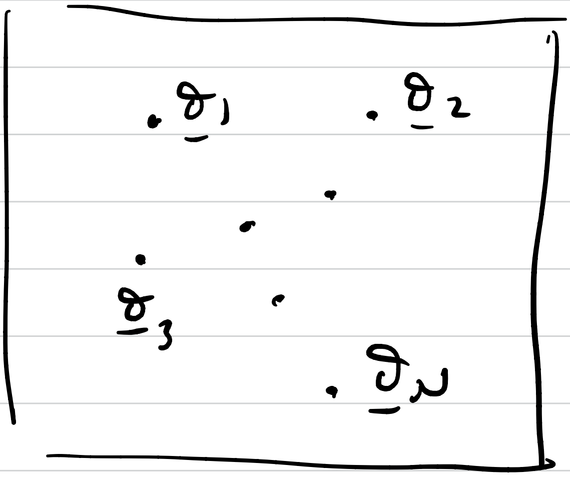
SGD can be re-written as:

$$\frac{\theta_i^{k+1} - \theta_i^{k}}{(2\delta_k/N)} = v_i^k$$

with $\mathbb{E}(v_i^k \mid \underbrace{\mathcal{F}^k}_{\text{the past}}) = \mathbb{E}_{x_k, y_k}(v_i^k)$

$$= -\nabla_\theta\left(V(\theta_i^k) + \sum_{j=1}^{N} U(\theta_i^k, \theta_j^k)\right)$$

Particle system interpretation:



N particles in $\mathbb{R}^{D+2}$.

positions $\underline{\theta}_1^k \dots \theta_N^k$ at time k.

velocities $\underline{v}_1^k \dots v_N^k$.

They feel potentials $V(\theta)$ external and interaction potential $U(\theta, \theta')$ between all pairs.

Potential felt by a particle at θ_i due to external potential and all other interactions is:

$$V(\theta_i) + \sum_{j=1}^{N} U(\theta_i, \theta_j).$$

We introduce the empirical density

$$\rho_N^k(\theta) = \frac{1}{N}\sum_{j=1}^{N} \delta(\theta - \theta_j^k).$$

The potential felt by particle i becomes

$$\psi(\theta_i^k ; \rho_N^k) = V(\theta_i^k) + \int d\theta \, U(\theta_i^k, \theta)\, \rho_N^k(\theta')$$

and the true risk becomes:

$$R_N(\theta^k) = \mathbb{E}(y^2) + 2\int d\theta \, V(\theta)\, \rho_N^k(\theta)$$

$$+ \int d\theta \, d\theta' \, U(\theta, \theta')\, \rho_N^k(\theta)\, \rho_N^k(\theta').$$

2) Mean field theory : statics.

We assume that for $N \to +\infty$ we have

$$\rho_N^k(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta(\theta - \theta_i^k) \longrightarrow \rho(\theta, t).$$

Note that time steps $\frac{2\delta k}{N}$ in SGD tend to zero so it is natural to replace k by a continuous time index t.

Now consider the risk and assume that $\rho(\theta, t)$ has reached (for $t \to +\infty$ say) some stationary or equilibrium density $\rho(\theta)$:

$$R_N(\theta) \xrightarrow[N \to +\infty]{} R(\rho) = \mathbb{E}(y^2) - 2\int d\theta \, V(\theta)\rho(\theta)$$

$$+ \int d\theta \, d\theta' \, U(\theta, \theta') \rho(\theta) \rho(\theta').$$

What is the minimizer ?

We have to minimize under the constraint that $\int d\vartheta\, \rho(\vartheta) = 1$. Introducing a Lagrange parameter:

$$\frac{\delta}{\delta \rho(\vartheta)} \left\{ R(\rho) - \lambda \left(\int d\rho(\vartheta) - 1 \right) \right\} = 0$$

$$\Rightarrow \qquad \frac{\delta R}{\delta \rho(\vartheta)} = \lambda$$

i.e. $$\nabla_\vartheta \left(\frac{\delta R}{\delta \rho(\vartheta)} \right) = 0$$

Now $$\frac{\delta R}{\delta \rho(\vartheta)} = 2\, V(\vartheta) + 2 \int d\vartheta'\, \rho(\vartheta')\, U(\vartheta, \vartheta')$$

so the equilibrium condition is

$$\nabla_{\underline{\vartheta}} \left\{ V(\vartheta) + \int d\vartheta'\, \rho(\vartheta')\, U(\vartheta, \vartheta' \right\} = 0$$

This can be interpreted as saying that the mean field potential defined as

$$\psi(\vartheta;\rho) \equiv \underbrace{V(\vartheta)}_{\text{external pot}} + \underbrace{\int d\vartheta' \rho(\vartheta')\, U(\vartheta,\vartheta')}_{\text{interaction energy due to all other particles.}}$$

should be constant; i.e the force acting at ϑ on the "fluid" vanishes: $-\nabla_{\underline{\vartheta}}\, \psi(\vartheta;\rho) = 0$.

3) Mean field theory: dynamics.

As said above SGD is (with $\varepsilon = \frac{2\delta_k}{N}$)

$$\begin{cases} \dfrac{\theta_i^{k+1} - \theta_i^{k}}{\varepsilon} = v_i^k = \text{stochastic velocity} \\ \qquad\qquad \underline{\text{"overdamped motion"}} \\ \mathbb{E}(v_i^k \mid \text{past}) = - \nabla_{\theta_i^k} \Psi(\theta_i^k ; \rho_N^k) \end{cases}$$

We deduce from here somewhat heuristically the continuity equation for $N \to +\infty$:

$$\frac{\partial}{\partial t} \rho(\theta;t) - \underbrace{\nabla_\theta}_{\text{Divergence}} \cdot \overbrace{\Big(\rho(\theta,t) \overbrace{\underbrace{\nabla_\theta}_{\text{gradient}} \Psi(\theta,\rho)}^{\text{velocity}} \Big)}^{\text{current}} = 0$$

Non linear PDE because $\Psi(\theta,\rho) = V(\theta) + \int d\theta' \rho(\theta',t) U(\theta,\theta')$

It is often attributed to Vlasov and McKean.

$\uparrow$ Plasma physics (Vlasov) $\qquad \uparrow$ Stochastic Processes. (McKean)

Heuristic derivation:

$$\rho_N^{k+1}(\theta) - \rho_N^{k}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\delta(\theta-\theta_i^{k+1}) - \delta(\theta-\theta_i^{k})\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \delta(\theta-\theta_i^{k}+(\theta_i^{k}-\theta_i^{k+1})) - \delta(\theta-\theta_i^{k})$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} \underline{\nabla}_{\theta}\,\delta(\theta-\theta_i^{k}) \cdot (\theta_i^{k}-\theta_i^{k+1})$$

$$\Rightarrow \quad \frac{\rho_N^{k+1}(\theta) - \rho_N^{k}(\theta)}{\varepsilon} \simeq -\frac{1}{N}\sum_{i=1}^{N} \underline{\nabla}_{\theta}\,\delta(\theta-\theta_i^{k}) \cdot \underbrace{\underbrace{\left(\frac{\theta_i^{k+1}-\theta_i^{k}}{\varepsilon}\right)}_{v_i^{k}}}_{-\nabla_{\theta_i^{k}}\psi(\theta_i^{k},\rho_N^{k})}$$

For $N \to +\infty$, time becomes continuous as $\varepsilon \to 0$:

$$\boxed{\frac{\partial}{\partial t}\rho(\theta,t) = \underline{\nabla}_{\theta}\cdot\left(\rho(\theta,t)\,\nabla_{\theta}\psi(\theta,\rho)\right).}$$

Theorem (Montaneri - Mei).

Here stated informally; under suitable hypothesis it is possible to prove that:

$$\sup_{k\in[0,\frac{T}{\varepsilon}]\cap\mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq C e^{CT}\sqrt{\max(\tfrac{1}{N},\varepsilon)}\left(\sqrt{D+\log\frac{N}{\varepsilon}} + z\right)$$

holds with probability $1-e^{-z^2}$, where $\rho_{k\varepsilon}$ is the solution $\rho(\theta,t)$ of PDE for $t=k\varepsilon$.

Remark: This theorem says how far the solution of the PDE (or mean field solution) is from the solution of underlying SGD.

4) Variational formulation of the Nonlinear PDE.

A powerful tool to analyze the solution of the PDE is a variational formulation.

Consider a simpler toy system first: Take gradient flow in $\mathbb{R}^d$:

$$\dot{\underline{x}}(t) = -\underline{\nabla} F(\underline{x}). \qquad \underline{x} \in \mathbb{R}^d.$$

The Euler discretisation would be the "forward scheme"

$$\underline{x}_{k+1} = \underline{x}_k - \varepsilon \underline{\nabla} F(\underline{x}_k)$$

Although iterations look easy to obtain this is pretty unstable and is not the best way to approximate the continuous time evolution.

The backward Euler scheme turns out to be more stable (although harder to implement):

$$x_{k+1} = x_k - \varepsilon \underline{\nabla} f(\underbrace{x_{k+1}})$$

The point now is that this scheme can be formulated in a variational way. Remark that (at least for f convex).

$$\underline{x}_{k+1} = \underset{\underline{x}}{\arg\min} \left\{ f(\underline{x}) + \frac{\|\underline{x} - \underline{x}_k\|^2}{2\varepsilon} \right\}.$$

So here we minimize $f(\underline{x})$ but under the constraint of not moving to far away from current position x_k. (Euclidean norm penalty).

<u>Proof</u> $\underline{\nabla}_x \left\{ f(\underline{x}) + \frac{\|x - x_k\|^2}{\varepsilon} \right\} = \nabla f(\underline{x}) + \frac{x - x_k}{\varepsilon}$

so the extremum is at:

$$\underline{x} = x_k - \varepsilon \nabla f(\underline{x}) \Rightarrow \boxed{x_{k+1} = x}.$$

It is a Minimum if f is convex.

This scheme also garantees that we reach a minimum of f:

$$f(x_{k+1}) + \frac{dist(x_{k+1}, x_k)}{2\epsilon} \underset{\uparrow}{\leq} f(x_k) + \frac{dist(x_k, x_k)}{2\epsilon}$$

Since x_{k+1} is the argmin!

$$\Rightarrow \quad f(x_{k+1}) \leq f(x_k) - \frac{dist(x_{k+1}, x_k)}{2\epsilon}$$

If $x_{k+1} \neq x_k$ then $f(x_{k+1}) < f(x_k)$ we strictly "improve". The sequence $(x_k)_{k \in \mathbb{N}}$ converges as long as f is bounded below.

Going back to the PDE, it is possible to show that:

$$\frac{\partial}{\partial t}\rho(\underline{\theta},t) = \underline{\nabla}\Big(\rho(\theta,t)\,\psi(\theta,\rho_t)\Big)$$

with $\psi(\theta,\rho_t) = \dfrac{\delta R(\rho)}{\delta \rho_t(\theta)}$

is equivalent to $\rho(\theta,t) = \lim_{\varepsilon\to 0} \rho_{(k+1)\varepsilon}$

$$\boxed{\rho_{(k+1)\varepsilon} = \operatorname{argmin}\left[R(\rho) + \frac{W_2^2(\rho,\rho_{k\varepsilon})}{2\varepsilon}\right]}$$

where $W_2^2(\rho,\rho')$ is the squared Wasserstein distance between two distributions:

$$W_2^2(\rho,\rho') \equiv \inf_{\substack{\text{couplings } \gamma \\ \text{of } \rho,\rho'}} \mathbb{E}\left(\|\underline{x}-\underline{y}\|^2\right)$$

where the set of couplings $d\gamma(x,y)$ satisfy

$$\int_Y d\gamma(x,y) = \rho(x)$$

$$\int_X d\gamma(x,y) = \rho'(y)$$

(i.e Marginals equal ρ & ρ').

Idea of proof:

Go back to discretized problem and use that the Wasserstein distance between

$$\rho_N^{k+1}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\delta(\theta - \theta_i^{k+1})$$

$$\rho_N^{k}(\theta') = \frac{1}{N}\sum_{i=1}^{N}\delta(\theta' - \theta_i^{k})$$

is $\inf_\pi \sum_{i=1}^{N} \| \theta_i^{k+1} - \theta_i^{k} \|^2 = \sum_{i=1}^{N} \| \theta_i^{k+1} - \theta_i^{k} \|^2$

↑
for small moves at each step.

Then taking

$$\rho(\vartheta) = \frac{1}{N}\sum_{i=1}^{N} \delta(\vartheta - \vartheta_i)$$

we have

$$R(\rho) + \frac{W_2^2(\rho, \rho_{k\varepsilon})}{2\varepsilon}$$

$$\approx R_N(\vartheta) + \sum_{i=1}^{N} \frac{\| \vartheta_i - \vartheta_i^k \|^2}{2\varepsilon}$$

Minimization over ϑ gives back SGD in Backward Euler scheme;

$$\frac{\underline{\vartheta}_i - \underline{\vartheta}_i^k}{\varepsilon} = -\underline{\nabla}_{\vartheta_i}\left(V(\vartheta_i) + \sum_{j=1}^{N} U(\vartheta_i, \vartheta_j^k) \right)$$

$$\Rightarrow \vartheta_i^{k+1} = \vartheta_i^k - \varepsilon \nabla_{\vartheta^{k+1}} V(\vartheta^{k+1}) + \sum_{j=1}^{N} U(\vartheta_i^{k+1}, \vartheta_j^k)$$

For $\varepsilon \to 0$ this gives back continuity equation. ■