Mean Field Methods for Two Layer NN – part I.

In this set of two lectures we review the work of Montanari & Mei on two layer NN treated in a "mean field asymptotic limit".

program this week:

1) Two layer NN and classical thms of Cybenko and Barron.
2) Formulation of SGD & test error.
3) Particle system interpretation and idea of Mean Field Analysis.

Next time:

4) Mean field analysis: statics & dynamics.

1) Two layer NN's.

We consider an "hypothesis class" of functions of the type

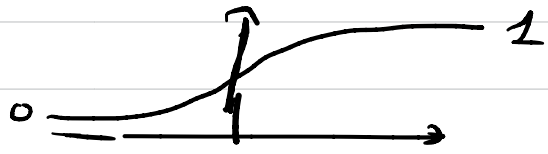
$$\hat{f}(\underline{x};\vartheta) = \frac{1}{N}\sum_{i=1}^{N} a_i\,\sigma\left(\underline{w}_i^T\cdot\underline{x} + b_i\right)$$

- where $\underline{x} \in \mathbb{R}^D$, so $\hat{f}: \mathbb{R}^D \to \mathbb{R}$
- $\underline{\vartheta} = \{\underline{w}_i, a_i, b_i\} = \{\vartheta_i\}_{i=1}^{N}$ a set of parameters (to be learned)

  $\underline{w}_i \in \mathbb{R}^D$, $a_i \in \mathbb{R}$, $b_i \in \mathbb{R}$.

- σ = activation function of sigmoidal type

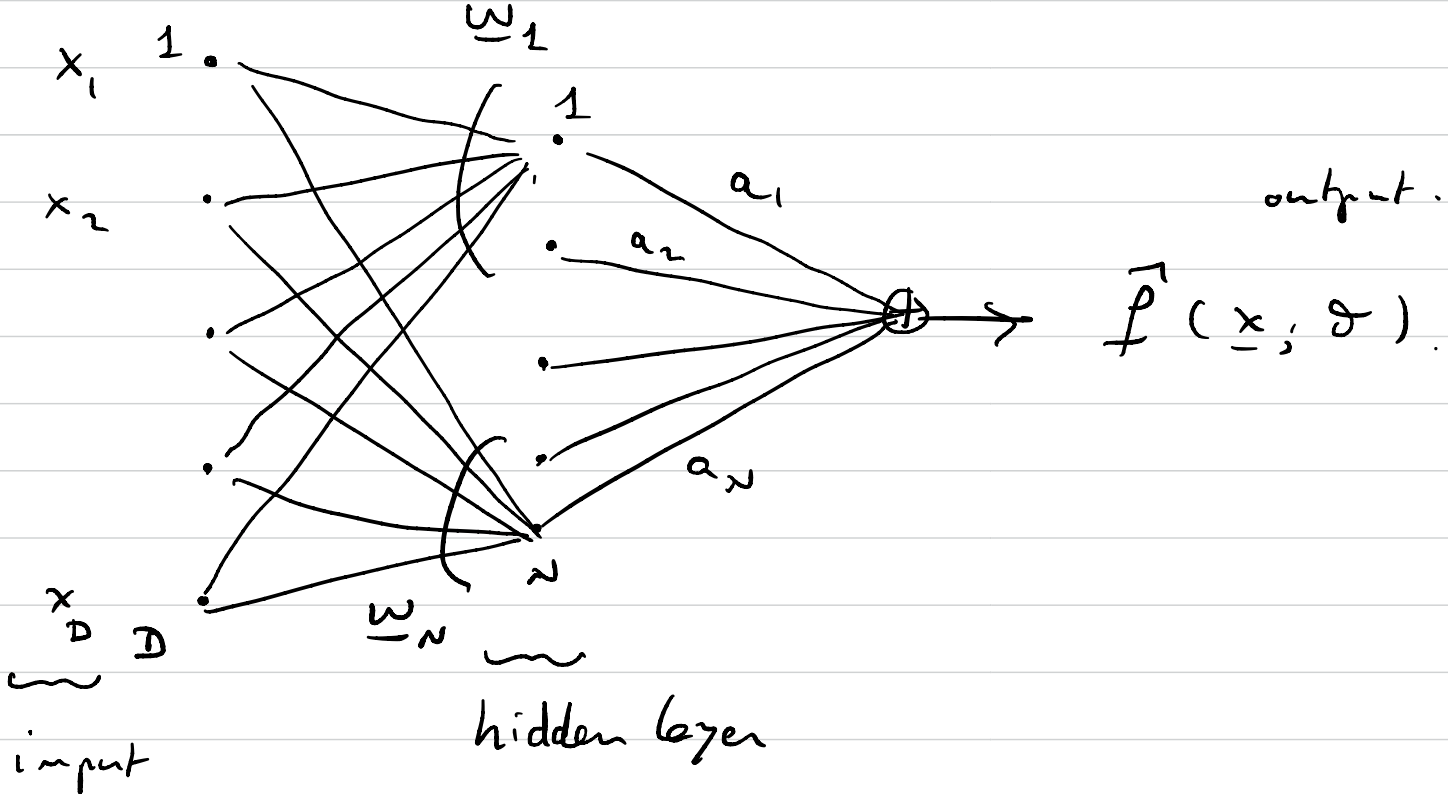
  example $\sigma(x) = \dfrac{1}{1+e^{-2x}}$



= activation of Relu type

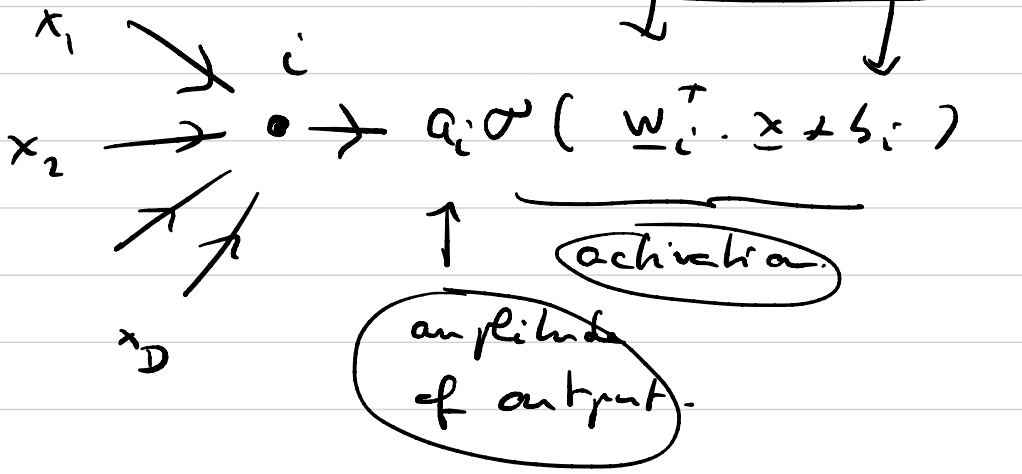
$\sigma(x) = \max(0, x)$.

The NN picture is



For each Neuron i of hidden layer:

weight associated to neuron i and bias



Typically we have training data $(\underline{x}_k, y_k)$ and we want to adjust or learn weights $\underline{\vartheta}$.

Representation Power of NN's.

We state two classic thms about the representability of functions by these two layer one single-hidden layer NN's.

Cybenko 1989 thm:

Let $\mathbb{E}_{x\sim\mathcal{D}}((f(x))^2) < \infty$ i.e $\int dx\, \mathcal{D}(x)|f(x)|^2 < \infty$

Assume $\sigma : \mathbb{R} \to \mathbb{R}$ continuous with

$$\lim_{x\to+\infty} \sigma(x) = +1, \quad \lim_{x\to-\infty} \sigma(x) = 0.$$

Then $\forall \varepsilon > 0$, $\exists N(\varepsilon)$ large enough s.t

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left(f(x) - \frac{1}{N}\sum_{i=1}^{N} a_i \sigma(w_i^T x + b_i)\right)^2\right] \leq \varepsilon$$

The next question is how big should $N(\varepsilon)$ be ?

Barron 1993 thm.

Let $\mathcal{D}(\underline{x})$ have support in $\text{Ball}(\underline{0}, r)$.

Let $f(\underline{x}) = \int_{\mathbb{R}^d} e^{i\omega \cdot \underline{x}} \underbrace{F(\omega)}_{\text{Fourier transf.}} d^D\omega$

and σ continuous with $\lim_{x \to +\infty} \sigma(x) = 1$, $\lim_{x \to -\infty} \sigma(x) = 0$.

Then we can take

$$N(\varepsilon) = \frac{1}{\varepsilon} \left(2r \int \|\omega\|_2 \, F(\underline{\omega}) \, d^D\underline{\omega} \right)^2 .$$

2) Formulation of SGD for two layer NN's.

- Recall the true risk or population risk:

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

here we take

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

and $\mathcal{D}(\underline{x}, y)$ supported on $\mathbb{R}^D \times \mathbb{R}$.

- As we do not know $\mathcal{D}$ we now try as a "proxy" for minimization to run SGD as explained in previous lects. There is no guarantee that we truly minimize of course here since ℓ is not convex w.r.t ϑ. But this is an algorithm.

Stochastic Gradient Descent algo:

$K = 1, 2, \ldots, T$ time steps.

at time k take a fresh sample from training set $(\underline{x}_k, y_k)$ and update

$$\underline{\theta}^{k+1} = \underline{\theta}^{k} + \delta_k \nabla_{\underline{\theta}} \ell(y_k, x_k; \underline{\theta}^k)$$

where $\nabla_{\underline{\theta}} \ell(y_k, x_k; \underline{\theta}^k)$ is an unbiased "stochastic gradient" in the sense that

$$\mathbb{E}_{(x_k, y_k)} \left(\nabla_{\theta} \ell(y_k, x_k; \theta^k) \right)$$

$$= \nabla_{\underline{\theta}} \mathbb{E}_{\theta} \ell(y, x; \underline{\theta}^k)$$

$$= \nabla_{\underline{\theta}} R_{\nu}(\theta) \quad \text{(gradient of true risk)}$$

This can be interpreted as run SGD on

empirical risk $L_S(\theta) = \frac{1}{T} \sum_{k=1}^{T} \ell(y_k, x_k; \theta)$

with a mini-batch of size 1 at each time step k.

#.

Computation of gradient:

$\underline{\theta} = (\underline{\theta}_i)_{i=1}^{N}$, For each $\underline{\theta}_i = (\underline{w}_i, a_i, \zeta_i)$

we have:

$$\nabla_{\underline{\theta}_i} \ell(y_k, x_k; \underline{\theta}) = \nabla_{\underline{\theta}_i} (y_k - \hat{f}(\underline{x}_k, \underline{\theta}))^2$$

$$= \nabla_{\underline{\theta}_i} \left(y_k - \frac{1}{N} \sum_{j=1}^{N} a_j \sigma(\underline{w}_j^T \cdot \underline{x} + \zeta_j) \right)^2$$

$$= 2 \left(y_k - \frac{1}{N} \sum_{j=1}^{N} a_j \sigma(\underline{w}_j^T \cdot \underline{x} + \zeta_j) \right) \cdot \nabla_{\underline{\theta}_i} \left(\frac{1}{N} a_i \sigma(\underline{w}_i^T \cdot \underline{x} + \zeta_i) \right)$$

The SGD update are thus for each

$k = 1 \dots T$ and $i = 1 \dots N$:

$$\underline{\theta}_i^{k+1} = \underline{\theta}_i^k - \frac{2\delta_k}{N} \left(y_k - \frac{1}{N} \sum_{j=1}^{N} a_j^k \sigma(\underline{w}_j^{Tk} \underline{x}_k + b_j^k) \right) \nabla_{\underline{\theta}_i} \left(a_i^k \sigma(\underline{w}_i^{Tk} \underline{x}_k + b_i^k) \right)$$

More explicitly we update for each k the $\underline{w}_i, a_i, b_i$:

$$\underline{w}_i^{k+1} = \underline{w}_i^k - \frac{2\delta_k}{N} \left(y_k - \frac{1}{N} \sum_{j=1}^{N} a_j^k \sigma(\underline{w}_j^{Tk} \underline{x}_k + b_j^k) \right) a_i^k \underline{x} \sigma'(\underline{w}_i^{Tk} \underline{x}_k + b_i^k)$$

$$a_i^{k+1} = a_i^k - \frac{2\delta_k}{N} \left(y_k - \frac{1}{N} \sum_{j=1}^{N} a_j^k \sigma(\underline{w}_j^{Tk} \underline{x}_k + b_j^k) \right) \sigma(\underline{w}_i^{Tk} \underline{x}_k + b_i^k)$$

$$b_i^{k+1} = b_i^k - \frac{2\delta_k}{N} \left(y_k - \frac{1}{N} \sum_{j=1}^{N} a_j^k \sigma(\underline{w}_j^{Tk} \underline{x}_k + b_j^k) \right) a_i^k \sigma'(\underline{w}_i^{Tk} \underline{x}_k + b_i^k)$$

The <u>generalization error</u> is the average error done when a new sample is presented. Here this is the true error/risk:

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

where $(\underline{x}, y) \sim \mathcal{D}$ is the new sample presented (not in training set).

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

$$+ \frac{1}{N^2}\mathbb{E}\left[\sum_{i,j=1}^{N} a_i a_j \sigma\left(\underline{w}_i^T \underline{x} + b_i\right)\sigma\left(\underline{w}_j^T \cdot \underline{x} + b_j\right)\right]$$

$$= \mathbb{E}(y^2) - \frac{2}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}\left[y\, a_i \sigma(\underline{w}_i^T x + b_i)\right]}_{\equiv V(\theta_i) \equiv V(\underline{w}_i, a_i, b_i)}$$

$$+ \frac{1}{N^2}\sum_{i,j=1}^{N} \underbrace{\mathbb{E}\left[a_i a_j \sigma(\underline{w}_i^T \underline{x} + b_i)\sigma(\underline{w}_j^T \cdot \underline{x} + b_j)\right]}_{\equiv U(\theta_i, \theta_j) \equiv U(\underline{w}_i, a_i, b_i, \underline{w}_j, a_j, b_j)}$$

So

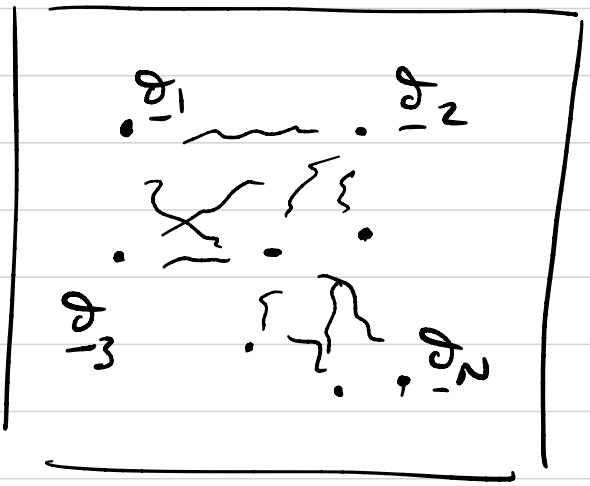
$$\begin{cases} R_N(\theta) = \mathbb{E}(y^2) - \frac{2}{N}\sum_{i=1}^{N} V(\vartheta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\vartheta_i, \vartheta_j) \\ V(\vartheta) = -\mathbb{E}_{\mathcal{D}}\left[y\, a\sigma(\underline{w}\cdot \underline{x} + b) \right] \qquad \vartheta = (\underline{w}, a, b) \\ U(\vartheta, \vartheta') = \mathbb{E}_{\mathcal{D}}\left[a\sigma(\underline{w}\cdot\underline{x} + b)\, a'\sigma(\underline{w}'\underline{x} + b') \right] \end{cases}$$

$$\vartheta = (\underline{w}, a, b) \qquad \vartheta' = (\underline{w}', a', b')$$

We will now discuss a "particle system" interpretation of $R_N(\theta)$ and then of SGD. This interpretation allows to take inspiration from the Mean Field Methods in physics to analyse the $N \to +\infty$ limit of the two layer NN's ($\mathcal{D}$ fixed).

Particle system interpretation.

- $R_N(\underline{\theta})$ is the energy fct of a system of N particles at positions $\underline{\theta}_1, \underline{\theta}_2 \dots \underline{\theta}_N \in \mathbb{R}^{D+2}$



- $V(\underline{\theta})$ external potential energy felt by a particle at position $\underline{\theta}$ (e.g gravity potential, electric potential, ...)
- $U(\underline{\theta}, \underline{\theta}')$ two body interaction potential felt by two particles at positions $\underline{\theta}$ & $\underline{\theta}'$.

Here all pairs particles interact with a strength of order $\frac{1}{N}$.

This allows to use Mean field Methods (analogous to Curie-Weiss Radkin Spin systems)

- SGD: is a kind of dynamics in discrete time steps of the system of particles.

- SGD can be rewritten as:

$$\frac{\underline{\theta}_i^{k+1} - \underline{\theta}_i^{k}}{\varepsilon} = + \nabla_{\theta_i} \Big\{ y_k a_i \sigma\left(\underline{w}_i^k \cdot \underline{x} + b_i^k\right) - \frac{1}{N} \sum_{j=1}^{N} a_i \sigma(\underline{w}_i^k \cdot \underline{x} + b_i^k) a_j \sigma(\underline{w}_j^k \cdot \underline{x} + b_j^k)$$

where $\varepsilon = \frac{2\delta_k}{N}$.

Now the left hand side is a "velocity at time k".

and

$$\underset{x_k, y_k}{\mathbb{E}}\left(\nu_i^k\right) = -\nabla_{\underline{\theta}_i} \Big\{ \underbrace{V(\theta_i^k) + \frac{1}{N} \sum_{j=1}^{N} U(\theta_i^k, \theta_j^k)}_{\text{potential felt by particle } i \text{ due to external pot \& all interactions at time } k} \Big\}$$

$$= -\nabla \text{Potential} = \underline{\underline{\text{Force}}}.$$

$\Rightarrow$ **SGD = Overdamped dynamics.**

- Idea of Mean Field Theory developed next time :

All equations above can be expressed by introducing an empirical particle density at time k.

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

In the limit $N \to +\infty$ we get a continuous particle density at continuous time t :

$$\rho_N^k(\theta) \to \rho(\theta, t)$$

and SGD equations become a PDE for $\rho_N(\theta, t)$. This PDE is often called the Vlasov-Mckean equation. It involve a single density $\rho(\theta, t)$ and potential $V(\theta) + \int d\theta' U(\theta, \theta') \rho(\theta, t)$ can be "Mean Field Potential".

#.

- In a nutshell: in the continuum limit we have a "fluid" with density $\rho(\theta, t)$ and velocity field $\underline{v}(\underline{\theta}, t) = -\underline{\nabla}_{\underline{\theta}} \psi(\theta; \rho(\theta, t))$

$$\psi(\theta; \rho(\theta, t)) = V(\theta) + \int d\theta' U(\theta, \theta') \rho(\theta', t)$$

- Conservation of mass imposes the continuity equation:

$$\frac{\partial}{\partial t}\rho(\theta, t) + \underline{\nabla}_{\underline{\theta}} \cdot \left(\rho(\underline{\theta}, t) \underline{\nabla}_{\underline{\theta}} \psi(\theta; \rho(\theta, t)) \right) = 0$$

(this is the continuous time limit of SGD).

- True risk becomes a functional of $\rho(\theta, t)$:

$$R(\rho) = \int d\theta\, V(\theta) \rho(\theta, t) + \int d\theta \int d\theta' U(\theta, \theta') \cdot \rho(\theta, t) \rho(\theta', t)$$

- The PDE above can be reinterpreted as the solution of a suitable variational problem $\rightarrow$ allows rigorous analysis.