Deep learning theory Lecture 13

Consider single layer neural networks of the form

$$f_\theta(x) = \frac{1}{\underline{m}} \sum_{j=1}^{m} a_j \sigma(w_j x)$$

Note the different scaling wrt. the NTK regime.

We assume the following statistical model

$$y_j = f_*(x_j) + \varepsilon_j \qquad \varepsilon_j \sim \mathcal{N}(0, \tau^2)$$

We aim to minimize the MSE

$$R(\theta) = \frac{1}{2}\int_{\mathbb{R}^d} |f_\theta(x) - f_*(x)|^2 \, \mathbb{P}(dx) \cancel{+ \tau^2}$$

excess risk

This can be written as

$$\begin{aligned} f_\theta(x) &= \frac{1}{m} \sum_{j=1}^{m} a_j \sigma(w_j x) \\ &= \int_{\mathbb{R}^{d+1+1}} a\sigma(w x) \left(\frac{1}{m} \sum_{j=1}^{m} \delta(a - a_j) \delta(w - w_j) \right) dw \, da \\ &= \int_{\mathbb{R}^{d+2}} a \sigma(w x) \, \mu_m(da, dw) \end{aligned}$$

(mean field)

where $\mu_m(da\, dw) = \frac{1}{m} \sum_{j=1}^{m} \delta(w - w_j) \delta(a - a_j) \, dw \, da$

Note: we can describe the network as a <u>linear</u> function of μ_n (the state of the network)

$$f[\mu](x) = \int a\,\sigma(w\cdot x)\,\mu(da, dw)$$

furthermore, μ is invariant under permutation of neuron indices, and

$$R[\mu] = \int \left(\int a\,\sigma(wx)\,\mu(da\,dw) - f_*(x)\right)^2 \mathbb{P}(dx)$$

is <u>convex</u> in μ (composition of convex functions).

Furthermore, note that at initialization the limit $n \to \infty$ is well-defined: if $a_j, w_j \overset{iid}{\sim} \rho(a, w)$

$$\mu_n \xrightarrow[n\to\infty]{w} \rho$$

The price to pay is that the network lives in an ∞-dimensional space.

<u>The dynamics</u>

The dynamics of each "particle" is

$$\frac{d}{dt} a_j = -\partial_{a_j} \int \left(\frac{1}{m} \sum_l a_l \sigma(w_l x) - f_*(x) \right)^2 \mathbb{P}(dx)$$

$$= -\frac{1}{m} \int \sigma(w_j x) \left(\frac{1}{m} \sum_l a_l \sigma(w_l x) - f_*(x) \right) \mathbb{P}$$

$$= -\frac{1}{m} \int \sigma(w_j x) \left(\int a \sigma(wx) \mu(da\, dw) - f_*(x) \right) \mathbb{P}$$

$$+ \eta \frac{dB_t}{dt}$$

$$\frac{d}{dt} w_j = -\frac{1}{m} x \int a_j \sigma'(w_j x) \left(\int a \sigma(wx) \mu(da\, dw) - f_*(x) \right) \mathbb{P}$$

$$+ \eta \frac{dB_t}{dt}$$

Changing time we have

$$\frac{d}{dt} \begin{pmatrix} a_j \\ w_j \end{pmatrix} = v[\mu] \begin{pmatrix} a_j \\ w_j \end{pmatrix} = D_{\binom{a}{w}} \left(a \sigma(w \cdot x) \left(\dots \right) \mathbb{P}(dx) \right)$$

$$+ \eta \frac{dB_t}{dt}$$

$$v[\mu]\begin{pmatrix} a \\ w \end{pmatrix} = \begin{pmatrix} \int \sigma(wx) \left(\int a' \sigma(w'x) \mu(da'\, dw') - f_*(x) \right) \mathbb{P}(dx) \\ x \int a \sigma'(wx) \left(\int a' \sigma(w'x) \mu(da'\, dw') - f_*(x) \right) \mathbb{P}(dx) \end{pmatrix}$$

This vector field can be evaluated at any point $\binom{a}{\omega}$

The equation that evolves the measure μ is the continuity equation:

$$\partial_t \mu_t = -\operatorname{div}\left(\mu_t \cdot v(\theta)\right) + \frac{1}{2}\eta \Delta \mu_t$$

in this case $v(\theta) = v[\mu_t](\theta)$ (mean-field)

$$\partial_t \mu_t = -\operatorname{div}\left(\mu_t\left(-D_\theta \int a\sigma(w\cdot x)\left(\int a\sigma(w\cdot x)\mu_t - f_*\right)\mathbb{P}(dx)\right)\right) + \frac{1}{2}\eta\Delta\mu_t$$

$$= -\operatorname{div}\left(\mu_t\left(-D_\theta \underbrace{\frac{\delta}{\delta\mu}}_{\text{variational derivative}} R'(\mu_t)\right)\right) \quad \text{(MFPDE)}$$

$\Longrightarrow$ Gradient flow in W_2 space of

$$\frac{1}{2}\int\left(\int a\sigma(w\cdot x)\mu_t(da, dw) - f_*\right)^2 \mathbb{P}(dx) + \eta\int \mu_t \log(\mu_t)$$

For a solution μ_t of MFPDE we write the flow map

$$\Phi_t[\mu_0] = \mu_t$$

Assumptions: σ is bounded, Lipschitz and has Lipschitz derivative σ'

Proposition: Let $\mu_0^{(n)} \xrightarrow[n\to\infty]{} \mu_0 \in \mathcal{P}_2$ then for all $T > 0$

$$\Phi_t[\mu_0^{(n)}] \xrightarrow[n\to\infty]{} \Phi_t[\mu_0] \quad \forall t \in [0,T)$$

Theorem 1: Let μ_0 have full support, then if

$$\mu_t \xrightarrow[t\to\infty]{} \mu_* \ , \ \mu_* \in \operatorname*{argmin}_{\mu \in \mathcal{M}_+^1(\Theta)} R(\mu)$$

Theorem 2: When $\eta > 0$ we have $\mu_t \to \mu_* \in \operatorname{argmin} R_\eta(\mu)$