CS 839: Theoretical Foundations of Deep Learning	**Spring 2022**

Lecture 14 Mean Field Analysis of Neural Networks

Instructor: Yingyu Liang *Date: March 10^{th}, 2022* *Scriber: Yiyou Sun*

1 Continuous Setting

Consider the traditional classification task: given input-out data $(x_i, y_i)_{i=1}^n$, $x \in \mathbb{R}^d, y \in \{1, -1\}$. The goal is to find a function $f : \mathbb{R}^d \to \mathbb{R}$, such that:

$$\min_f Q(f) = L(f) + R(f), L(f) = E_{x,y}[l(f(x)), y)],$$

where $l(\cdot)$ is defined to be the loss function and R is a regularization function. Similar to Kernel methods, consider the two-level network given below to represent f:

$$f(\omega, \rho, x) = \int_{\mathbb{R}^d} \sigma(\theta, x)\omega(\theta)\rho(\theta)d\theta \tag{1}$$

where $\sigma(\theta, x) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a known real-valued function, $\omega(\theta) : \mathbb{R}^d \to \mathbb{R}$ is a real value function of θ, and $\rho(\theta)$ is a probability density over θ. For regularizer, we use

$$R(\omega, \rho) = \lambda_1 R_1(\omega, \rho) + \lambda_2 R_2(\rho)$$

, where

$$R_1(\omega, \rho) = \int r_1(\omega(\theta))\rho(\theta)d\theta, r_1(\omega) = |\omega|^2$$
$$R_2(\rho) = \int r_2(\theta)\rho(\theta)d\theta, r_2(\theta) = ||\theta||^2$$

Next we show a discrete NN approximates the continuous one when hidden nodes go to infinity and then drive the evolution rule of $\rho(\theta)$ and $\omega(\theta)$ from the (noisy) GD algorithm when the step size becomes small.

2 Discrete Setting

Consider a finite NN with the following form to approximate $f(\omega, \rho, x)$:

$$\widehat{f}(\mu, \theta, x) = \frac{1}{m}\sum_{j=1}^{m} \mu^j \sigma(\theta_t^j, x) \tag{2}$$

with the regularization term:

$$\widehat{R}_1(\mu,\theta) = \frac{1}{m}\sum_{j=1}^{m} r_1(\mu^j), \widehat{R}_2(\theta) = \frac{1}{m}\sum_{j=1}^{m} r_2(\theta^j), \tag{3}$$

Consider trainwith objective denoted as,

$$\widehat{Q}(u,\theta) = \mathbb{E}_{x,y} l(\widehat{f}(u,\theta,x),y) + \lambda_1 \widehat{R}_1(u,\theta) + \lambda_2 \widehat{R}_2(\theta) \tag{4}$$

We can solve it through the standard (noisy) GD, the algorithm is given by:

Step 0. Initialize $(\mu_0, \theta_0 \sim P_0(\mu,\theta))$
Step 1. Update θ_j by

$$\theta_{t+1}^j = \theta_t^j - \Delta t \nabla_{\theta^j}\left[\widehat{Q}\left(u_t,\theta_t\right)\right] - \sqrt{\lambda_3}\xi_{t+1}^j,$$

where Δt is the step size and $\xi_{t+1}^j \sim N\left(0, \sqrt{2\Delta t} I_d\right)$.
Step 2. Update μ_j by

$$u_{t+1}^j = u_t^j - \Delta t \nabla_{u^j}\left[\widehat{Q}\left(u_t,\theta_t\right)\right] - \sqrt{\lambda_3}\zeta_{t+1}^j,$$

where $\zeta_{t+1}^j \sim N(0, \sqrt{2\Delta t})$.

2.1 Plain GD

We first consider the unnoisy setting where $\lambda_3 = 0$. We have the following Lemma.

Lemma 1. Suppose at time $t \geq 0$, we have $\theta_t^j \sim \rho_t$, and let $u_j^t = \omega_t\left(\theta_t^j\right)$. Assume l' is continuous and σ is twice differentiable. For all x, we have:

$$\lim_{m\to\infty} \widehat{f}\left(u_t,\theta_t,x\right) = f\left(\omega_t,\rho_t,x\right) \tag{5}$$

Furthermore, when $\Delta t \to 0, m \to \infty$, we can derive,

$$\begin{aligned}
\frac{d\rho_t(\theta)}{dt} &= -\nabla_\theta \cdot \left[\rho_t(\theta) g_2\left(t,\theta,\omega_t(\theta)\right)\right] \\
\frac{d\omega_t(\theta)}{dt} &= g_1\left(t,\theta,\omega_t(\theta)\right) - \nabla_\theta\left[\omega_t(\theta)\right] \cdot g_2\left(t,\theta,\omega_t(\theta)\right),
\end{aligned}$$

where ∇_θ means the divergence, g_1 and g_2 satisfy:

$$g_1(t,\theta,u) = -\mathbb{E}_{x,y}\left[l'\left(f\left(\omega_t,\rho_t,x\right),y\right)\sigma(\theta,x)\right] - \lambda_1 \nabla_u\left[r_1(u)\right]$$

$$g_2(t,\theta,u) = -\mathbb{E}_{x,y}\left[l'\left(f\left(\omega_t,\rho_t,x\right),y\right) u \nabla_\theta \sigma(\theta,x)\right] - \lambda_2 \nabla_\theta\left[r_2(\theta)\right]$$

To prove the lemma, we utilize the tool with Fokker-Planck Equation to compute the evolution.

Background with Fokker-Planck Equation Suppose the movement of a particle in m-dimensional space can be characterized by the stochastic differential equation given below:

$$dx_t = g\left(x_t, t\right) d_t + \sqrt{2\beta^{-1}}\Sigma d_{B_t}$$

Let $x_t \sim p(x,t)$, the evolution of $p(x,t)$ is given by:

$$\frac{\partial p(x,t)}{\partial t} = \frac{\Sigma\Sigma^\top}{\beta}\nabla^2 p(x,t) - \nabla\cdot\left[p(x,t) g\left(x_t,t\right)\right]$$

.

Proof of Lemma 1. Let the $p_t(\theta,\mu)$ as the joint distribution for (θ,μ):

$$(\theta_t^j, u_t^j) \sim p_t(\theta,u) = \rho_t\delta\left(u=\omega_t(\theta)\right)$$

We can rewrite $f\left(\omega_t,\rho_t,x\right)$ as:

$$f\left(\omega_t,\rho_t,x\right) = \int_{\mathbb{R}^{d+1}} \sigma(\theta,x) p_t(\theta,u) d\theta du$$

By the Law of the Large number, when $m\to\infty$,

$$\widehat{f}\left(u_t,\theta_t,x\right) \to f\left(\omega_t,\rho_t,x\right)$$

Now we denote

$$\widehat{g}_2(t,\theta,u) = -\mathbb{E}_{x,y}\left[l'\left(\widehat{f}\left(u_t,\theta_t,x\right),y\right) u\nabla_\theta\sigma(\theta,x)\right] - \lambda_2\nabla_\theta\left[r_2(\theta)\right]$$

From the update rule of GD, we have $\theta_{t+1}^j = \theta_t^j + \Delta t\widehat{g}_2\left(t,\theta_t^j,u_t^j\right)$. Let $\Delta t\to 0$, using $u_t^j = \omega_t\left(\theta_t^j\right)$, we have

$$\frac{d\theta_t^j}{dt} = \widehat{g}_2\left(t,\theta_t^j,\omega_t\left(\theta_t^j\right)\right)$$

By applying Fokker-Planck equation,

$$\frac{d\rho_t(\theta)}{dt} = -\nabla_\theta\cdot\left[\rho_t(\theta)\widehat{g}_2\left(t,\theta,\omega_t(\theta)\right)\right]$$

As $m\to\infty$, and because l' is continuous, $\sigma(\theta,x)$ and ρ_t are also second-order smooth, we obtain:

$$\nabla_\theta\cdot\left[\rho_t(\theta)\widehat{g}_2\left(t,\theta,\omega_t(\theta)\right)\right] - \nabla_\theta\cdot\left[\rho_t(\theta) g_2\left(t,\theta,\omega_t(\theta)\right)\right] \overset{\text{a.s.}}{\to} 0$$

To prove the evolution form for $\omega_t(\theta)$, we let:

$$\widehat{g}_1(t,\theta,u) = -\mathbb{E}_{x,y}\left[l'\left(\widehat{f}\left(u_t,\theta_t,x\right),y\right)\sigma(\theta,x)\right] - \lambda_1\nabla_u r_1(u)$$

On one side,

$$
\begin{aligned}
&\omega_{t+\Delta t}\left(\theta_{t+\Delta t}\right) \\
=&\omega_{t+\Delta t}\left(\theta_t+\widehat{g}_2\left(t, \theta_t, \omega_t(\theta)\right) \Delta t+o(\Delta t)\right) \\
=&\omega_t\left(\theta_t+\widehat{g}_2\left(t, \theta_t, \omega_t(\theta)\right) \Delta t+o(\Delta t)\right)+\frac{d \omega_t\left(\theta_t+\widehat{g}_2\left(t, \theta_t, \omega_t(\theta)\right) \Delta t+o(\Delta t)\right)}{d t} \Delta t \\
=&\omega_t\left(\theta_t\right)+\left[\nabla_\theta \omega_t(\theta)\right] \cdot \widehat{g}_2\left(t, \theta_t, \omega_t(\theta)\right) \Delta t+o(\Delta t)+\frac{d \omega_t\left(\theta_t+\widehat{g}_2\left(t, \theta, \omega_t(\theta)\right) \Delta t+o(\Delta t)\right)}{d t} \Delta t .
\end{aligned}
$$

By the update rule $\omega_{t+\Delta t}\left(\theta_{t+\Delta t}\right)=\omega_t\left(\theta_t\right)+\widehat{g}_1\left(t, \theta_t, \omega_t(\theta)\right) \Delta t$, we have:

$$
\frac{d \omega_t\left(\theta_t+\widehat{g}_2\left(t, \theta_t, \omega_t(\theta)\right) \Delta t+o(\Delta t)\right)}{d t}=-\left[\nabla_\theta\left(\omega_t(\theta)\right)\right] \cdot \widehat{g}_2\left(t, \theta_t, \omega_t(\theta)\right)+\widehat{g}_1\left(t, \theta_t, \omega_t(\theta)\right)+o(1)
$$

The proof is finished by Let $\Delta t \rightarrow 0,$ and let $m \rightarrow \infty$.