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

Lecture 15 Mean Field Analysis II

Instructor: Yingyu Liang *Date: Mar 22th, 2022* *Scriber: Keshu Wu*

1 Problem Settings

Recall the problem setup for a two-layer neural network from the last leture. For the continuous setting:

$$\begin{aligned}
f(\omega,\rho,x) &= \int \sigma(\theta,x)\omega(\theta)\rho(\theta)d\theta \\
\min_f Q(f) &= L(f)+R(f), R(f)=\lambda_1 R_1(f)+\lambda_2 R_2(f) \\
R_1 &= \int r_1(\omega(\theta))\rho(\theta)d\theta, r_1(\omega)=\|\omega\|^2 \\
R_2 &= \int r_2(\omega(\theta))\rho(\theta)d\theta, r_1(\theta)=\|\theta\|^2
\end{aligned}$$

where $\rho(\theta)$ is the probability density over θ.
For the discrete setting:

$$\begin{aligned}
\widehat{f}(u,\theta,x) &= \frac{1}{m}\sum_{j=1}^{m} u^j\sigma(\theta^j,x) \\
\widehat{Q}(u,\theta) &= \widehat{L}(u,\theta)+\lambda_1\widehat{R}_1+\lambda_2\widehat{R}_2 \\
\widehat{R}_1 &= \frac{1}{m}\sum_{j=1}^{m} r_1(u^j) \\
\widehat{R}_2 &= \frac{1}{m}\sum_{j=1}^{m} r_2(\theta^j)
\end{aligned}$$

where u^j is the weight of jth neuron, and θ^j is the weight factor of jth neuron.

2 Discrete NN

Noisy Gradient Descent (NGD) Consider training the constructed network with the objective denoted as $\widehat{Q}(u,\theta)$, and solving it by NGD, the steps are decribed as follows:

- Initialize: Sample θ_0^j, u_0^j (from distribution).
- Update θ_t^j:

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

- Update u_t^j:
$$u_{t+1}^j = u_t^j - \Delta_t \nabla_{u^j} [\widehat{Q}(u_t, \theta_t)] - \sqrt{\lambda_3} \zeta_t^j$$
- $\xi_t^j \sim \mathbb{N}(0, \sqrt{2\Delta t} I_d), \zeta_t^j \sim \mathbb{N}(0, \sqrt{2\Delta t})$

Fokker-Plank Recall the Fokker–Planck equation:

- The stochastic differential equation (SDE): $dX_t = \delta(x_t, t)dt + \sqrt{2\beta^{-1}} \Sigma dB_t$
- Let $p(x,t)$: density of X_t
$$\frac{\partial}{\partial t} = -\nabla \cdot [p(x,t)\delta(X_t, t)] + \frac{\Sigma\Sigma^T}{\beta} \nabla^2 p(x,t)$$

Fokker-Plank gives the evolution of density $p(x,t)$.

Lemma 1 (NGD)**.** $\forall t \geq 0, (\theta_t^j, u_t^j) \sim p_t(\theta, u), j \in [m]$. Then $\forall x$, we have
$$\lim_{n \to \infty} \widehat{f}(u_t, \theta_t, x) = f(\omega_t, \rho_t, x)$$

Furthermore, let $\Delta t \to 0, m \to \infty$, we can derive
$$\frac{dp_t(\theta, u)}{dt} = -\nabla_\theta \cdot [p_t(\theta, u) g_2(t, \theta, u)] - \nabla_u [p_t(\theta, u) g_1(t, \theta, u)] + \lambda_3 \nabla^2 [p_t(\theta, u)] \qquad (1)$$

where g_1, g_2 satisfy
$$g_1 = -E_{x,y}[l'(f(\omega_t, \rho_t, x), y)\sigma(\theta, x)] - \lambda_1 \nabla_u [r_1(u)]$$
$$g_2 = -E_{x,y}[l'(f(\omega_t, \rho_t, x), y) u \nabla_\theta \sigma(\theta, x)] - \lambda_2 \nabla_\theta [r_2(\theta)]$$

and
$$\rho_t(\theta) = \int p_t(\theta, u) du, \omega_t(\theta) = E[u|\theta]$$

Lemma 1 implies that $p_t(\theta, u)$ becomes a diffusion process due to the injection of the random noise.

Proof: The update of (θ_t, u_t) can be written as:
$$\begin{bmatrix} \theta_{t+1} \\ u_{t+1} \end{bmatrix} = \begin{bmatrix} \theta_t \\ u_t \end{bmatrix} + \Delta_t \begin{bmatrix} \widehat{g}_2(t, \theta, u) \\ \widehat{g}_1(t, \theta, u) \end{bmatrix} + \sqrt{2\lambda_3 \Delta t} \begin{bmatrix} N(0, I_d) \\ N(0, 1) \end{bmatrix}$$

Let $\Delta t \to 0$, we have
$$d \begin{bmatrix} \theta_t \\ u_t \end{bmatrix} = \begin{bmatrix} \widehat{g}_2 \\ \widehat{g}_1 \end{bmatrix} dt + \sqrt{2\lambda_3} dBt$$

Let $m \to \infty$, we obtain

$$\widehat{g}_1 \to g_1, \widehat{g}_2 \to g_2$$

We can also write down the evolution of ρ_t and ω_t as follows.

Lemma 2 (Evolution of ρ_t, ω_t)**.** Assume $p_t(\theta, u)$ is as in previous lemma. Then

$$\begin{aligned}
\frac{d\rho_t(\theta)}{dt} &= -\nabla_\theta \cdot [p_t(\theta) g_2(t, \theta, \omega_t(\theta))] + \lambda_3 \nabla^2 \rho_t(\theta) \qquad (2)\\
\frac{d\omega_t(\theta)}{dt} &= g_1(t, \theta, \omega_t(\theta)) - \nabla_\theta \omega_t(\theta) \cdot g_2(t, \theta, \omega_t(\theta))\\
&\quad + \lambda_3 \nabla_\theta^2 \omega_t(\theta) + \frac{2\lambda_3}{\rho_t(\theta)} [\nabla_\theta \rho_t(\theta)][\nabla_\theta \omega_t(\theta))]\\
&\quad - \frac{1}{\rho_t(\theta)} \nabla_\theta \cdot \left[\int p(\theta, u) u [g_2(t, \theta, u) - g_2(t, \theta, \omega_t(\theta))] du \right] \qquad (3)
\end{aligned}$$

Proof: Let $g_{21} : g_2(t, \theta, u) = g_{21}(t, \theta) u - \lambda_2 \nabla_\theta r_2(\theta)$.
For Equation (2):

$$\frac{d\rho_t(\theta)}{dt} = \frac{d}{dt} \int p_t(\theta, u) du = \int \frac{d}{dt} p_t(\theta, u) du$$

According to Equation (1)

$$\begin{aligned}
\frac{d\rho_t(\theta)}{dt} &= -\int \nabla_\theta \cdot [p_t(\theta, u) g_2(t, \theta, u)] du - \int \nabla_u [p_t(\theta, u) g_1(t, \theta, u)] du + \int \lambda_3 \nabla^2 [p_t(\theta, u)] du\\
&= I_1 + I_2 + I_3 \qquad (4)
\end{aligned}$$

Then, for I_1,

$$\begin{aligned}
-I_1 &= \int \nabla_\theta \cdot [p_t (g_{21} \cdot u - \lambda_2 \nabla_\theta r_2(\theta)] \, du\\
&= \nabla_\theta \cdot \left[\int p_t g_{21} u du - \int \lambda_2 p_t \nabla_\theta r_2(\theta) du \right]\\
&= \nabla_\theta \cdot [g_{21} \rho_t \omega_t - \lambda_2 \nabla_\theta r_2(\theta) \rho_t(\theta)]\\
&= \nabla_\theta \cdot [\rho_t(\theta) g_2(t, \theta, \omega_t(\theta))]
\end{aligned}$$

For I_2,

$$\begin{aligned}
I_2 &= \int \nabla_u \left[p_t, g_1 \right] du\\
&= p_t(\theta, u) g_1(t, \theta, u) du|_{u=\infty}^{+\infty}\\
&= 0
\end{aligned}$$

since $|g_1| \leq L_1(C_1\|\theta\| + C_2) + \lambda_2|u|$

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

For I_3,

$$\begin{aligned}
I_3 &= \int \lambda_3 \nabla^2 p_t(\theta, u) du \\
&= \lambda_3 \left[\int \nabla_\theta^2 p_t du + \int \nabla_u^2 p_t du\right] \\
&= \lambda_3 \left[\nabla_\theta^2 \int p_t du + \nabla_u^2 \int p_t du\right] \\
&= \lambda_3 \nabla_\theta^2 \rho_t(\theta)
\end{aligned}$$

Then, plug I_1, I_2, I_3 into Equation (4),

$$\frac{d\rho_t(\theta)}{dt} = -\nabla_\theta \cdot [p_t(\theta) g_2(t, \theta, \omega_t(\theta))] + \lambda_3 \nabla^2 \rho_t(\theta)$$

Thus, Equation (2) is proved.

For Equation (3), define $u_t(\theta) = \omega_t(\theta)\rho_t(\theta) = \int u p_t(\theta, u) du$, then

$$\begin{aligned}
\frac{u_t(\theta)}{dt} &= \int \frac{d}{dt}(u \cdot p_t) du = \int u(\frac{d}{dt} p_t) du \\
\frac{u_t(\theta)}{dt} &= \frac{d\omega_t}{dt} \cdot \rho_t + \omega_t \cdot \frac{d\rho_t}{dt} \\
&= \int -u \nabla_\theta [p_t g_2] du - \int u \nabla_u [p_t g_1] du + \int \lambda_3 u \nabla^2 p_t du \\
&= \int -u \nabla_\theta [p_t g_2] du + I_4 + I_5 \qquad (5)
\end{aligned}$$

For I_4,

$$\begin{aligned}
-I_4 &= \int u \nabla_u [p_t g_1] du \\
&= \int [p_t(\theta, u) g_1(\theta, t, u)] du \\
&= \int p_t(\theta, u) g_{11}(\theta, t) - p_t(\theta, u) \lambda_1 u du \\
&= \rho_t(\theta) g_1(t, \theta, \omega_t(\theta))
\end{aligned}$$

For I_5,

$$\begin{aligned}
I_5 &= \int \lambda_3 u \nabla^2 p_t du \\
&= \lambda_3 \nabla_\theta^2 \left[\int u p_t(\theta, u) du\right] + \lambda_3 \int u \nabla_u [p_t(\theta, u)] du \\
&= \lambda_3 u \nabla^2 p_t du
\end{aligned}$$

Since

$$\begin{aligned}\frac{du_t(\theta)}{dt} &= \rho_t \frac{d\omega_t(\theta)}{dt} + \omega_t \frac{d\omega_t(\rho)}{dt} \\ &= \rho_t \frac{d\omega_t(\theta)}{dt} + \omega_t \{-\nabla_\theta \cdot [\rho(\theta) g_2(t, \theta, \omega_t(\theta))] + \lambda_3 \nabla_\theta^2 \rho_t(\theta)\}\end{aligned}$$

Then, plug I_4, I_5 into Equation (5),

$$\begin{aligned}\frac{d\omega_t(\theta)}{dt} &= g_1(t, \theta, \omega_t(\theta)) - \nabla_\theta \omega_t(\theta) \cdot g_2(t, \theta, \omega_t(\theta)) - \frac{\lambda_3 \omega_t(\theta)}{\rho_t(\theta)} \nabla_\theta^2 \rho_t(\theta) + \frac{\lambda_3}{\rho_t(\theta)} \nabla_\theta^2 \omega_t(\theta) \rho_t(\theta) \\ &\quad - \frac{1}{\rho_t(\theta)} \nabla_\theta \cdot \left[\int p(\theta, u) u g_2(t, \theta, u) - \rho_t(\theta) \omega_t(\theta) g_2(t, \theta, \omega_t(\theta))\right] \\ &= g_1(t, \theta, \omega_t(\theta)) - \nabla_\theta \omega_t(\theta) \cdot g_2(t, \theta, \omega_t(\theta)) + \lambda_3 \nabla_\theta^2 \omega_t(\theta) + \frac{2\lambda_3}{\rho_t(\theta)} [\nabla_\theta \rho_t(\theta)][\nabla_\theta \omega_t(\theta))] \\ &\quad - \frac{1}{\rho_t(\theta)} \nabla_\theta \cdot \left[\int p(\theta, u) u [g_2(t, \theta, u) - g_2(t, \theta, \omega_t(\theta))] du\right]\end{aligned}$$

Thus, Equation (3) is proved.

Please refer to [1] to see more proof details.

References

[1] Cong Fang, Hanze Dong, and Tong Zhang. Over parameterized two-level neural networks can learn near optimal feature representations. 2019.