CS2910 - Optimization Summer 2023

Lecture C — Mean-Field Networks

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

1 Overview

Nowadays, our course comes to an advanced topic in deep learning theory. In this lecture, we focus on the two-layer neural networks and the efficiency of them.

2 Introduction

Now we introduce the two-layer networks and mean-field networks.

2.1 Two-layer neural networks

Consider the following function called the *two-layer neural network*:

$$f(x; a, W) = a^\top \sigma(Wx) = \sum_{k=1}^{m} a_k \sigma(W_k \cdot x)$$

where $x \in \mathbb{R}^d$ is the input, $W \in \mathbb{R}^{m\times d}$ is the first weight, $a \in \mathbb{R}^m$ is the second weight and $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function. Then we call m the number of neurons and $a_k\sigma(W_k \cdot x)$ the k-th neuron.

To analyze the efficiency of such a network, for some technical reasons (under over-parameterization case), we might let $m \to \infty$ and use some infinite dimensional network to understand it.

2.2 Mean-field networks

Let $x \mapsto \phi(x; W_k)$ denote the k-th neuron. For the two-layer neural network, we have (with some scaling that does not really matter)

$$f(x; \{W_k\}_{k=1}^m) = \frac{1}{m}\sum_{k=1}^{m} \phi(x; W_k) = \int_{\mathbb{R}^d} \phi(x; w)\, d\widehat{\mu}(w)$$

where the probability measure $\widehat{\mu} = \frac{1}{m}\sum_{k=1}^m \delta_{W_k}$ is the empirical distribution of the first-layer neurons.

Now, if we allow $\widehat{\mu}$ can be any (reasonably regular) distribution (not necessary discrete) over $\mathbb{R}^d$, we generalize

$$f(x; \mu) = \int_{\mathbb{R}^d} \phi(x; w)\, d\mu(w)$$

which is a network with potentially infinite neurons.

Example: If we initialize $W_k \sim \mathcal{N}(0, \sigma^2 I_d)$, as $m \to \infty$, the two-layer network $f(x; W)$ will converges to $f(x; \mathcal{N}(0, \sigma^2 I_d))$.

3 Mean Squared Error

Now, we consider the population mean squared error (MSE)

$$L(\mu) = \frac{1}{2}\mathbb{E}_x\left[(f_*(x) - f(x;\mu))^2\right].$$

We apply the gradient flow to (approximately) solve it. The following theorem relates the finite dimensional gradient flow to Wasserstein gradient flow.

Theorem 1 (informal statement of Theorem 2.6 in [CB18])**.** *Under some regularity conditions, as the number of neurons m goes to ∞, the classical gradient flow converges to the Wasserstein gradient flow with respect to L.*

Remark 1. Consider the finite-width network

$$f(x; W) = \frac{1}{m}\sum_{k=0}^{m} \phi(x; W_k).$$

We run the classical gradient descent as:

$$\begin{aligned} \frac{d}{dt}W_k &= -m\nabla_{W_k} L \\ &= \mathbb{E}_x\left[(f_*(x) - f(x))\nabla_{W_k}\phi(x; W_k)\right] \end{aligned}$$

Then let $\mu_{m,0} = \frac{1}{m}\sum_{k=1}^{m}\delta_{W_{k,0}}$ be the empirical distribution of the initialization of the neural network. At time t, let $\mu_{m,t}$ denote the distribution of neurons updated by the classical gradient flow.

On the other hand, let μ_0 be the infinite-width initialization (with $\mu_{m,0} \to \mu_0$ as $m \to \infty$). At time t, let μ_t be the distribution updated by the Wasserstein gradient flow. Then Theorem 1 shows $\mu_{m,t} \to \mu_t$ as $m \to \infty$.

Note that, to make sure the distance between $\mu_{m,t}$ and μ_t is small, we need $\exp(d)$ neurons.

3.1 First variation of MSE

To apply the Wasserstein gradient flow, it is necessary to compute the first variation of MSE. We compute it by definition.

For $\varepsilon > 0$ and a perturbation χ, by elementary calculation

$$\begin{aligned} L(\mu + \varepsilon\chi) &= \frac{1}{2}\mathbb{E}_x\left[(f_*(x) - f(x; \mu + \varepsilon\chi))^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[f_*(x)^2\right] + \frac{1}{2}\mathbb{E}\left[f(x; \mu + \varepsilon\chi)^2\right] - \mathbb{E}\left[f_*(x) f(x; \mu + \varepsilon\chi)\right] \end{aligned}$$

Then taking derivative, we obtain

$$\frac{d}{d\varepsilon}|_{\varepsilon=0}L(\mu+\varepsilon\chi)=\int \mathbb{E}_x\left[(f_*(x)-f(x;\mu))\phi(x;w)\right]\,d\chi(w).$$

Then we know

$$\frac{\delta L}{\delta\mu}[\mu](v)=\mathbb{E}_x\left[(f_*(x)-f(x;\mu))\phi(x;v)\right]=\left\langle f_*(\cdot)-f(\cdot;\mu),\phi(\cdot,v)\right\rangle_{L^2}.$$

3.2 Global convergence

Now we show the convergence property of WGF. Firstly we introduce the universal approximation.

Definition 2. *We say* $\{\sigma(\cdot,v)\}_{v\in\mathbb{R}^d}$ *satisfies the **universal approximation property** if its span is dense in* L^2*.*

Remark 2. The property means the two-layer neural networks can approximate everything.

Theorem 3 (Theorem 3.3 in [CB18]; Theorem 8 in [PN21])**.** *Let* μ_t *be the Wasserstein gradient flow with respect to* L *from* μ*. Suppose that* $\operatorname{supp}(\mu_0)=\mathbb{R}^d$*. Let* $\mu_\infty\triangleq\lim_{t\to\infty}\mu_t$*. Suppose that* σ *is a universal approximation, and* $\phi(\cdot;w)=w_0\sigma(\cdot;w_{1:d})$*. Then under some regularity conditions,* μ_∞ *is a global minimizer of* L*.*

Proof Idea. The whole proof is technically difficult, and we will only show the proof idea. For convenience, assume that $\operatorname{supp}(\mu_\infty)=\mathbb{R}^d$ (this assumption is too strong and to remove it, we need some algebraic topological arguments). Then by descent lemma of WGF, $\phi(\cdot;0)=0$, we have for almost all $v\in\mathbb{R}^d$

$$\frac{\delta L}{\delta\mu}[\mu_\infty]=\left\langle f_*(\cdot)-f(\cdot;\mu_\infty),\phi(\cdot,v)\right\rangle_{L^2}=0.$$

By the universal approximation property, there exists $\{g_m\}$ such that

$$g_m=\sum_{k=1}^m\phi(\cdot,v_k),\quad \lim_{m\to\infty}g_m=f_*-f(\cdot;\mu_\infty).$$

Thus

$$\begin{aligned}0&=\sum_{k=1}^m\left\langle f_*(\cdot)-f(\cdot;\mu_\infty),\phi(\cdot,v_k)\right\rangle_{L^2}\\&=\left\langle f_*(\cdot)-f(\cdot;\mu_\infty),g_m\right\rangle_{L^2}\to\|f(\cdot,\mu_\infty)-f_*(\cdot)\|.\end{aligned}$$

This means μ_∞ is the global minimizer. □

References

[CB18] Lénaïc Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-Parameterized Models Using Optimal Transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3040–3050, Red Hook, NY, USA, 2018. Curran Associates Inc.

[PN21] Huy Tuan Pham and Phan-Minh Nguyen. Global Convergence of Three-layer Neural Networks in the Mean Field Regime, 2021.