Mean-field Analysis on Two-layer Neural Networks from a Kernel Perspective

Shokichi Takakura [1, 2, *], **Taiji Suzuki** [1, 2] The University of Tokyo [1] AIP RIKEN [2] Now at LY Corp. [*]

Overview

- Formulated the training of neural networks as kernel learning.
- Provided qualitative convergence guarantee of two-layer neural networks in the mean-field regime under mild conditions.
- Proved the superiority of mean-field neural networks to fixed kernel methods.
- Proposed label noise mean-field Langevin dynamics and proved it leads to a "robust" kernel.

Problem Settings

Mean-field Neural Networks

Consider the following two-layer neural networks:

$$f(x; a, \{w_i\}_{i=1}^M) := \frac{1}{M}\sum_{i=1}^M a_i h(x; w_i)$$

In the over-parameterized regime $M \to \infty$, f converges to the following mean-field limit:

$$f(x; P) := \int a h(x; w) \mathrm{d}P(a, w)$$

Another Parametrization

To separate the dynamics of the first layer and the second layer, we consider the following (equivalent) parametrization:

$$f(x; a, \mu) := \int a(w) h(x; w) \mathrm{d}\mu(w),$$

- $a(w)$: the conditional expectation of a
- $\mu(w)$: the marginal distribution of w.

Connection to Kernel Methods

By fixing the distribution μ, the above model is equivalent to the kernel method with the kernel

$$k_\mu(x, x') = \int h(x; w) h(x'; w) \mathrm{d}\mu(w)$$

➜ Training of the first distribution is equivalent to kernel learning = feature learning.

Convergence Analysis

Q. Is it possible to learn the optimal kernel using gradient-based algorithms?

Objective Function

We consider the empirical risk minimization problem with the squared loss and ridge regularization:

$$F(P) = F(a, \mu) = \frac{1}{n}\sum_{i=1}^n (f(x_i; a, \mu) - y_i)^2 + E_\mu\left[\frac{\bar{\lambda}_a}{2} a(w)^2 + \frac{\bar{\lambda}_w}{2}\|w\|_2^2\right]$$

Mean-field Langevin Dynamics (MFLD)

The mean-field Langevin dynamics is

- a continuous limit of noisy gradient descent
- used to solve $\min_\mu L(\mu) + \lambda \mathrm{Ent}(\mu)$ for a certain functional L

$$\mathrm{d}\theta_t = -\nabla \frac{\delta L(\mu)}{\delta \mu}(\theta)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}B_t,$$

⚠ It is difficult to prove the qualitative convergence of MFLD for $\theta = (a, w)$ since each neuron $ah(x; w)$ is not bounded nor Lipschitz continuous w.r.t. (a, w).

Two-timescale Limit

If the learning rate of the second layer is much faster than the first layer, the second layer converges instantly to the optimum a_μ since F is convex w.r.t. a.

- Problem is reduced to minimization on μ
  $$G(\mu) := F(a_\mu, \mu) = \min_a F(a, \mu)$$
- Run the MFLD for $\theta = w$ to solve the minimization of $\mathcal{G}(\mu) = G(\mu) + \lambda \mathrm{Ent}(\mu)$

Main Results

Let μ^* be the optimal distribution and μ_t be the distribution of w at time t. For any $t \geq 0$, we have

$$\mathcal{G}(\mu_t) - \mathcal{G}(\mu^*) \leq \exp(-2\alpha\lambda t)(\mathcal{G}(\mu_0) - \mathcal{G}(\mu^*))$$

➜ Linear convergence to the global optimum

Key Observation

- $G(\mu)$ is convex although $f(x; \mu)$ is not linear in μ

Estimation Error Analysis

Q. Does adapted kernel help generalization?

Barron Space

Barron space is a union of multiple RKHSs:

$$\mathcal{B}_M = \{f(x; a, \mu) \mid \mathrm{KL}(\mu, N(0, 1)) \leq M, \|a\|_{L^2(\mu)} \leq \infty\}$$

$$\|f\|_{\mathcal{B}_M} = \inf_a \{\|a\|_{L^2(\mu)} \mid f(x; a, \mu) = f\}$$

Main Results

Assume that $f^\circ \in \mathcal{B}_M, \|f^\circ\|_{\mathcal{B}_M}^2 \leq R$ for a certain M, R.

- Mean-field neural networks can learn the target function with $O(d \log d)$ samples,
- Any linear estimator (e.g., kernel ridge regression) requires at least $O(d^k)$ samples for a fixed k,

with high probability.

➜ Adapted kernel achieves better sample complexity

Label Noise MFLD

Q. Is there a simple way to obtain a robust kernel?

Label Noise MFLD

Training the second layer with noisy label $\tilde{y}_i = f^\circ(x_i) + \xi_i (\xi_i \sim \mathrm{Unif}([-\tilde{\sigma}, \tilde{\sigma}]))$ implicitly solves the following minimization problem:

$$\tilde{\mathcal{G}}(\mu) = \mathcal{G}(\mu) + \frac{\bar{\lambda}_a \tilde{\sigma}^2}{6n} d(\mu)$$

$d(\mu)$ is *the degrees of freedom* or effective dimension of the kernel k_μ and corresponds to the variance.

➜ Label noise leads to small $d(\mu)$ & avoids overfitting

Numerical Experiments

- Label noise reduces the degrees of freedom
- Label noise improves the generalization error