A Mean-Field Analysis of Neural Stochastic Gradient Descent-Ascent for Functional Minimax Optimization

Yuchen Zhu[1] Yufeng Zhang[2] Zhaoran Wang[2] Zhuoran Yang[3] Xiaohong Chen[3]

[1]Georgia Institute of Technology, [2]Northwestern University, [3]Yale University

Abstract

This paper studies minimax optimization problems defined over infinite-dimensional function classes of overparameterized two-layer neural networks. In particular, we consider the minimax optimization problem stemming from estimating linear functional equations defined by conditional expectations, where the objective functions are quadratic in the functional spaces. We address (i) the convergence of the stochastic gradient descent-ascent algorithm and (ii) the representation learning of the neural networks. We establish convergence under the mean-field regime by considering the continuous-time and infinite-width limit of the optimization dynamics. Under this regime, the stochastic gradient descent-ascent corresponds to a Wasserstein gradient flow over the space of probability measures defined over the space of neural network parameters. We prove that the Wasserstein gradient flow converges globally to a stationary point of the minimax objective at a $\mathcal{O}(T^{-1}+\alpha^{-1})$ sublinear rate, and additionally finds the solution to the functional equation when the regularizer of the minimax objective is strongly convex. Here T denotes the time and α is a scaling parameter of the neural networks. In terms of representation learning, our results show that the feature representation induced by the neural networks can deviate from the initial one by a magnitude of $mathcalO(\alpha^{-1})$, measured in terms of the Wasserstein distance. Finally, we apply our general results to concrete examples including policy evaluation, nonparametric instrumental variable regression, and asset pricing.

Functional Conditional Moment Equations and Minimax optimization

Some notations: $X \in \mathcal{X}$ be a vector that includes all the endogenous variables, $Z \in \mathcal{Z}$ denote all the exogenous variables, $W \in \mathcal{W} \subseteq \mathcal{X} \times \mathcal{Z}$ be a subset of variables that may contain both the endogenous and exogenous variables, $\mathcal{F} := \{f : \mathcal{W} \to \mathbb{R}\} \subset L^2(\mathcal{W})$ denote a class of functions defined on $\mathcal{W}$.

In a ***functional conditional moment equation*** problem (Zhu et al., 2024), we aim to find a function $f_0 \in \mathcal{F}$ that solves the following functional equation involving the conditional distribution of X given Z over $\mathcal{F}$:

$$\mathbb{E}_{X|Z}\Big[\Phi(X, Z; f_0)\,\Big|\, Z = z\Big] = 0, \qquad \forall z \in \mathcal{Z}, \tag{1}$$

where $\Phi\colon \mathcal{X} \times \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ is a known functional. For any function $f \in \mathcal{F}$ and any $z \in \mathcal{Z}$, we define a error functional $\bar{\delta}\colon \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ as

$$\bar{\delta}(z; f) := \mathbb{E}_{X|Z}\Big[\Phi(X, Z; f)\,\Big|\, Z = z\Big], \qquad \forall f \in \mathcal{F}, z \in \mathcal{Z}. \tag{2}$$

To solve the functional conditional moment equation, we aim to solve the following regularized functional minimization problem,

$$\min_{f\in\mathcal{F}} J(f) = \mathbb{E}_{\mathcal{D}}\Big[1/2 \cdot \bar{\delta}(z; f)^2 + \lambda\Psi(X, Z; f)\Big]$$

where for any given $(x, z) \in \mathcal{X} \times \mathcal{Z}$, $\Psi(x, z; f) : \mathcal{F} \to \mathbb{R}_+$ is a convex functional of f that maps each function f to a scalar. Moreover, Ψ satisfies

$$\Psi(x, z; 0) = 0, \qquad \Psi(x, z; f) \geq 0, \qquad \forall f \in \mathcal{F}, \tag{3}$$

$$\frac{\delta\Psi(x, z; af_1 + bf_2)}{\delta f} = a \cdot \frac{\delta\Psi(x, z; f_1)}{\delta f} + b \cdot \frac{\delta\Psi(x, z; f_2)}{\delta f}, \qquad \forall f_1, f_2 \in \mathcal{F},\ a, b \in \mathbb{R}. \tag{4}$$

Note that we can turn the functional optimization problem into solving the following **minimax optimization** problem by finding the saddle point of functional $\mathcal{L}(f, g)$ defined below,

$$\min_f \max_g \mathcal{L}(f, g) = \mathbb{E}_{\mathcal{D}}\Big[g(Z) \cdot \Phi(X, Z; f) - 1/2 \cdot g(Z)^2 + \lambda\Psi(X, Z; f)\Big]. \tag{5}$$

Applications of Functional Conditional Moment Equations

Policy Evaluation We consider a Markov decision process given by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S} \subseteq \mathbb{R}^d$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\gamma \in (0, 1)$ is the discount factor. Given a policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$, we aim to estimate value function $V^\pi : \mathcal{S} \to \mathbb{R}$, which satisfies the Bellman equation,

$$\mathbb{E}_{s'|s}\Big[r(s, a) - V^\pi(s) + \gamma \cdot V^\pi(s')\,\Big|\, s\Big] = 0. \tag{6}$$

(6) is a special case of the functional conditional moment equation in (1) by setting the exogenous variable Z to be the current state s, the endogenous variable X to be the next state s' and the function to be estimated $f : \mathcal{S} \to \mathbb{R}$ to be defined on the state space $\mathcal{S}$. $\Phi(X, Z; f) = r + \gamma \cdot f(X) - f(Z)$.

Asset Pricing Asset pricing refers to the process of determining the fair value of financial assets. Consider the Consumption Capital Asset Pricing Model (CCAPM), let C_t denote the consumption level at time t, $c_t \equiv C_t/C_{t-1}$ the consumption growth. The marginal utility of consumption at time t is given by $\mathrm{MU}_t = C_t^{-\gamma_0} f_0(c_t)$, where $\gamma_0 > 0$ is the discount factor, $f_0 : \mathcal{C} \to \mathbb{R}$ is the nonparametric structural demand function, which is an unknown positive function of our interest and is defined on $\mathcal{C}$, the space of consumption growth. The CCAPM model captures the behavior of f_0 through the following equation:

$$\mathbb{E}_{c_{t+1}|c_t}\big[\tilde{r}_{t+1} \cdot f_0(c_{t+1}) - f_0(c_t)\,\big|\, c_t\big] = 0, \tag{7}$$

where the modified return can be further expressed as $\tilde{r}_{t+1} = \delta_0 \cdot r_{t+1} \cdot c_{t+1}^{-\gamma_0}$, $\delta_0 \in (0, 1]$ is the rate of time preference. (7) is a special case of the functional conditional moment equation. We can identify the exogenous variable Z with c_t, the consumption growth at the current time t, and the endogenous variable X with c_{t+1}, the consumption growth at the next time $t + 1$. $\Phi(X, Z; f) = \tilde{r}_{t+1} \cdot f(X) - f(Z)$.

Neural SGDA and Mean-field Limit

We parameterize both f and g with neural networks with width N and parameters $\boldsymbol{\theta} = (\theta^1, \theta^2, \ldots, \theta^N) \in \mathbb{R}^{D\times N}$ and $\boldsymbol{\omega} = (\omega^1, \omega^2, \ldots, \omega^N) \in \mathbb{R}^{D\times N}$

$$f(\cdot; \boldsymbol{\theta}) = \frac{\alpha}{N}\sum_{i=1}^N \phi(\cdot; \theta^i), \quad g(\cdot; \boldsymbol{\omega}) = \frac{\alpha}{N}\sum_{i=1}^N \psi(\cdot; \omega^i). \tag{8}$$

The discrete-time finite width SGDA is,

$$\text{GD}\quad \theta_{k+1}^i = \theta_k^i - \eta\alpha\epsilon \cdot g(z_k; \boldsymbol{\omega}_k) \cdot \nabla_\theta\Phi(x_k, z_k; \phi(\cdot, \theta_k^i)) - \eta\lambda\epsilon \cdot \frac{\delta\Psi(x_k, z_k; f(\cdot, \boldsymbol{\theta}_k))}{\delta f} \cdot \nabla_\theta\phi(x_k; \theta_k^i),$$

$$\text{GA}\quad \omega_{k+1}^i = \omega_k^i + \eta\alpha\epsilon \cdot \Phi(x_k, z_k; f(\cdot, \boldsymbol{\theta}_k)) \cdot \nabla_\omega\psi(z_k; \omega_k^i) - \eta\alpha\epsilon \cdot g(z_k; \boldsymbol{\omega}_k) \cdot \nabla_\omega\psi(z_k; \omega_k^i), \tag{9}$$

Mean-field Limit Passing the network to infinite width limit $N \to +\infty$ and timestep scale $\epsilon \to 0$, function f and g becomes infinite width neural network,

$$f(\cdot; \mu) = \alpha\int \phi(\cdot; \theta)\mu(\mathrm{d}\theta), \quad g(\cdot; \nu) = \alpha\int \psi(\cdot; \omega)\nu(\mathrm{d}\omega). \tag{10}$$

where $\mu(\theta)$ and $\nu(\omega)$ follows the gradient flow,

$$\partial_t\mu_t(\theta) = -\eta \cdot \mathrm{div}_\theta\big(\mu_t(\theta)v^f(\theta; \mu_t, \nu_t)\big),\ \partial_t\nu_t(\omega) = -\eta \cdot \mathrm{div}_\omega\big(\nu_t(\omega)v^g(\omega; \mu_t, \nu_t)\big), \tag{11}$$

and the vector field is given by

$$v^f(\theta; \mu, \nu) = \alpha\mathbb{E}_{\mathcal{D}}\Big[-g(Z; \nu) \cdot \Big\langle\frac{\delta\Phi(X, Z; f(\cdot; \mu))}{\delta f}, \nabla_\theta\phi(\cdot; \theta)\Big\rangle_{L^2} - \lambda \cdot \Big\langle\frac{\delta\Psi(X, Z; f(\cdot; \mu))}{\delta f}, \nabla_\theta\phi(\cdot; \theta)\Big\rangle_{L^2}\Big],$$

$$v^g(\omega; \mu, \nu) = \alpha\mathbb{E}_{\mathcal{D}}\Big[\Phi(X, Z; f(\cdot, \mu)) \cdot \nabla_\omega\psi(Z; \omega) - g(Z; \nu) \cdot \nabla_\omega\psi(Z; \omega)\Big].$$

Convergence of SGDA to Mean-field Limit

We show that the empirical distribution of the parameters $\hat{\mu}_k$ and $\hat{\nu}_k$ weakly converges to the mean-field limit in (11) as the width N goes to infinity and the stepsize scale ϵ goes to zero. Let $\rho_t(\theta, \omega) = \mu_t(\theta) \otimes \nu_t(\omega)$, where (μ_t, ν_t) is the PDE solution to the continuous deterministic dynamics in (11) and $\hat{\rho}_k = N^{-1} \cdot \sum_{i=1}^N \delta_{\theta_k^i} \cdot \delta_{\omega_k^i}$ corresponds to the empirical distribution of $(\boldsymbol{\theta}_k, \boldsymbol{\omega}_k)$.

Let $\{\rho_t\}_{t\geq 0}$ be solution to (11) with $\rho_0 = \mathcal{N}(0, I_D) \otimes \mathcal{N}(0, I_D)$, $\{\hat{\rho}_k\}_{k\geq 0}$ be solution to (9) with $\hat{\rho}_0 = \mathcal{N}(0, I_D) \otimes \mathcal{N}(0, I_D)$. Under mild assumptions, $\widehat{\rho}_{\lfloor t/\epsilon\rfloor}$ converges weakly to ρ_t as $\epsilon \to 0^+$ and $N \to \infty$. It holds for any Lipschitz continuous, bounded function $F : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ that

$$\lim_{\epsilon\to 0^+, N\to\infty} \int F(\theta, \omega)\mathrm{d}\widehat{\rho}_{\lfloor t/\epsilon\rfloor}(\theta, \omega) = \int F(\theta, \omega)\mathrm{d}\rho_t(\theta, \omega).$$

Proof based on **propagation of chaos** type of arguments (Mei et al., 2018; Zhang et al., 2020)

Global Optimality and Convergence of the Mean-Field Limit

We first characterize (μ^*, ν^*) the stationary point of the Wasserstein gradient flow. Turns out that under mild assumptions,

(i) The corresponding function $(f(\cdot; \mu^*), g(\cdot; \nu^*))$ is the saddle point of the objective function $\mathcal{L}(f, g)$ defined in (5).

(ii) There exists a stationary distribution pair (μ^*, ν^*) and constant $\bar{D} > 0$ such that

$$W_2(\mu_0, \mu^*) \leq \alpha^{-1}\bar{D}, \quad W_2(\nu_0, \nu^*) \leq \alpha^{-1}\bar{D}.$$

This result demonstrates that the stationary point of the Wasserstein gradient flow in (11) achieves global optimality as a solution to the minimax objective (5). It allows us to bypass the hardness of solving the nonconvex-nonconcave optimization problem (5) of finding saddle points in the space of neural network parameters $(\boldsymbol{\theta}, \boldsymbol{\omega})$ by searching for a stationary point of the Wasserstein gradient flow instead.

The following result characterizes the **global convergence** of the Wasserstein gradient flow. Let (μ_t, ν_t) be the solution to the Wasserstein gradient flow (11) at time t with $\eta = \alpha^{-2}$ and initial condition $\mu_0 = \nu_0 = \mathcal{N}(0, I_D)$, (f^*, g^*) the saddle point of the minimax objective $\mathcal{L}(f, g)$ in (5). Under mild assumptions, it holds that

$$\inf_{t\in[0,T]} \mathbb{E}_{\mathcal{D}}\Big[\lambda\Psi\big(X, Z; f(\cdot; \mu_t) - f^*(\cdot)\big) + \big(g(Z; \nu_t) - g^*(Z)\big)^2\Big] \leq \mathcal{O}(T^{-1} + \alpha^{-1}). \tag{12}$$

The main takeaway from the results:

- Optimality gap between time t solution and optimal solution decays to zero at a sublinear rate up to an error of order $\mathcal{O}(\alpha^{-1})$
- SGDA induces a data-dependent representation that is significantly different from the initialization, with a richer representation than the NTK regime if $\alpha = \mathcal{O}(1)$ instead of $\mathcal{O}(\sqrt{N})$.

References

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Zhang, Y., Cai, Q., Yang, Z., Chen, Y., and Wang, Z. (2020). Can temporal-difference and q-learning learn representation? A mean-field theory. *arXiv preprint arXiv:2006.04761*.

Zhu, Y., Zhang, Y., Wang, Z., Yang, Z., and Chen, X. (2024). A mean-field analysis of neural stochastic gradient descent-ascent for functional minimax optimization. *arXiv preprint arXiv:2404.12312*.