Mean Field Langevin Actor-Critic: Faster Convergence and Global Optimality beyond Lazy Learning

Kakei Yamamoto [1] **Kazusato Oko** [2 3] **Zhuoran Yang** [4] **Taiji Suzuki** [2 3]

Abstract

This work explores the feature learning capabilities of deep reinforcement learning algorithms in the pursuit of optimal policy determination. We particularly examine an over-parameterized neural actor-critic framework within the mean-field regime, where both actor and critic components undergo updates via policy gradient and temporal-difference (TD) learning, respectively. We introduce the *mean-field Langevin TD learning* (MFLTD) method, enhancing mean-field Langevin dynamics with proximal TD updates for critic policy evaluation, and assess its performance against conventional approaches through numerical analysis. Additionally, for actor policy updates, we present the *mean-field Langevin policy gradient* (MFLPG), employing policy gradient techniques augmented by Wasserstein gradient flows for parameter space exploration. Our findings demonstrate that MFLTD accurately identifies the true value function, while MFLPG ensures linear convergence of actor sequences towards the globally optimal policy, considering a Kullback-Leibler divergence regularized framework. Through both time particle and discretized analysis, we substantiate the linear convergence guarantees of our neural actor-critic algorithms, representing a notable contribution to neural reinforcement learning focusing on *global optimality* and *feature learning*, extending the existing understanding beyond the conventional scope of lazy training.

[1]MIT, Cambridge, MA [2]The University of Tokyo, Tokyo, Japan [3]Center for Advanced Intelligence Project, RIKEN [4]Yale University, New Haven, CT. Correspondence to: Kakei Yamamoto <kakei@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

In recent years, the field of reinforcement learning (RL) (Sutton & Barto, 2018) including the policy gradient method (Williams, 1992; Baxter et al., 1999; Sutton et al., 1999) and the temporal-difference (TD) learning (Sutton, 1988) has made tremendous progress, with deep reinforcement learning methods. The combination of the actor-critic method (Konda & Tsitsiklis, 1999) and neural networks has demonstrated significant empirical success in challenging applications, such as the game of Go (Silver et al., 2016; 2017) or the human-like feedback alignment (Ouyang et al., 2022). In these empirical successes, the employment of deep neural networks plays an indispensable role — their expressivity enable learning meaningful features that benefit decision-making. However, despite the impressive empirical results, there remain many open questions about the theoretical foundations of these methods. In particular, when viewing deep RL methods as optimization algorithms in the space of neural network policies, it remains elusive how deep RL algorithms learn features during the course of finding the optimal policy.

One source of difficulty in the analysis of neural policy optimization comes from the nonconvexity of the expected total reward over the policy space. Also, TD learning used in the policy evaluation subproblem faces classic challenges (Baird, 1995; Tsitsiklis & Van Roy, 1996) stemming from the bias of semi-gradient optimization (Sutton, 1988). Another source of difficulty is the nonlinearity associated with the neural networks parameterizing both the policy and state-action value functions. The tremendous success of deep RL is attributed to its rich expressive power, which is backed by the nonlinearity of neural networks, which at the same time brings a considerable challenge to the optimization aspect. Unfortunately, the advantages of data-dependent learning of neural networks in the context of RL have only a limited theoretical understanding. Classical theoretical studies of policy optimization and policy evaluation problems, including the actor-critic method, limit their analysis to the case of linear function approximation in both the actor and the critic, where the feature mapping is fixed during learning (Sutton et al., 1999; Kakade, 2001; Bhatnagar et al., 2007; 2009). Recently, some analyses based on the theory

of Neural Tangent Kernel (NTK) (Jacot et al., 2018) are established, which state that an infinite-width neural network is well approximated by a linear function of random features determined by initial parameters under certain conditions (Cai et al., 2019; Wang et al., 2020; Liu et al., 2019). More recent works (Zhang et al., 2020; 2021) establish the study of convergence and optimality of over-parameterized neural networks over lazy training (Chizat et al., 2019), incorporating a mean-field perspective corresponding to NTK. Specifically, by letting the network width be sufficiently large under appropriate conditions in NTK or lazy training regimes, optimality is guaranteed based on the fact that the neural network features are as close as possible to the data-independent initial feature representation. Leahy et al. (2022) In other words, these existing analyses do not fully capture the representation learning aspect of neural RL empowered by the expressivity of neural networks. Thus, in this paper, we aim to address the following question:

Does neural actor-critic provably learn features on the way to the global optima?

We provide an affirmative answer to this question by focusing on the case where both the actor and the critic are represented by an over-parameterized two-layer neural network in the mean-field regime. Under this setting, we propose to update the actor and critic by a variant of policy gradient and TD learning tailored to mean-field neural networks, based on Langevin dynamics. We prove that the critic converges to the correct value function sublinearly and the sequence of actors converges to the globally optimal policy of a Kullback Leibler (KL) divergence regularized objective. More importantly, our theory is beyond the lazy training regime and provably shows that the actor and critic networks perform feature learning in the algorithm.

Our Contributions The main contribution of this paper is to propose the Mean-field Langevin actor-critic algorithm and prove linear convergence and global optimality with *feature learning* (Suzuki, 2019; Ghorbani et al., 2019). We treat the problem of policy improvement and policy evaluation as an optimization over a probability distribution of network parameters with KL-divergence regularization and build convergence analysis based on *mean field Langevin dynamics* (MFLD). Specifically,

1. We introduce the *mean-field Langevin TD learning* (MFLTD) as the policy evaluation component (critic) and show that it converges to the true value function at a sublinear rate. In this algorithm, we employ a double-loop proximity gradient algorithm to resolve the difficulties posed by having semi-gradients. Compared to the existing TD(1) in a basic benchmark, we experimentally test the practicality of this new method.

2. We introduce the *mean-field Langevin policy gradient* (MFLPG) as the policy improvement component (actor) and prove that it converges to the globally optimal policy at a linear convergence rate under KL-divergence regularization, in continuous and discretization case, rep. This algorithm is equivalent to the standard policy gradient in the parameter space with additional injected noises.

At the core of our analysis are (1) the over-parameterization of two-layer neural networks to represent policies and approximate state-action value functions in the mean-field regime, (2) the log-Sobolev-inequality argument to control the local convergence, (3) Techniques for simultaneously controlling the global optimal error and the KL divergence error, an inherent problem arising from the Wasserstein gradient flow in nonconvex objective functions (See the proof of Lemma 9 and Theorem 3), (4) the proximal gradient algorithm for TD learning to prevent convergence breakdown by using the semi-gradient of the mean squared Bellman error. In particular, (1) attributes the problem to the Wasserstein gradient flow and enables the utilization of the convexity of the loss function in the measure space. Furthermore, together with (2), it guarantees linear convergence speed in the presence of globally convergent solutions. Note here that, our whole results are valid with arbitrary regularization parameters. It is worth noting that (3) directly induces global optimality in distribution space, which eliminates constraints on regularization and (4) allows the Bellman error bias not to depend on the scale of the neural network as in lazy training for the first time. To the best of our knowledge, our analysis gives the first global optimality and linear convergence guarantees for the neural policy gradient methods with feature learning, confirming their considerable empirical success. Leahy et al. (2022) analyzed an entropy-regularized policy gradient similar to ours using the Wasserstein gradient flow, but the essence of their analysis is strong convexity due to sufficiently large regularization parameter $\lambda > c > 0$. On the other hand, our convergence analysis allows arbitrary regularization parameters $\lambda > 0$.

Related Works Regarding the convergence and optimality of the actor-critic, there is a need to encompass the two optimization problems of the actor component and the critic component, and in terms of the complexity of each problem, the theoretical research is limited. Regarding TD learning, various approaches mainly utilizing linear function approximation have been made to address the divergence and non-convergence issues arising from semi-gradient (Baird, 1995; Tsitsiklis & Van Roy, 1996). In particular, Capturing neural networks in the NTK regime, Cai et al. (2019) demonstrated sublinear convergence to the true value function, and Zhang et al. (2020) showed such sublinear convergence by attributing this optimization to lazy training. On the other hand, the global convergence of policy gradient methods is limited due to the non-convexity of the objective function, but

Fazel et al. (2018); Yang & Wang (2019) proved the convergence of policy gradient methods to the globally optimal policy in the LQR setting (Fazel et al., 2018; Zhou & Lu, 2023), and Bhandari & Russo (2019); Agarwal et al. (2020) proved convergence to the globally optimal policy in tabular and their own linear settings. Along the line of research, Wang et al. (2020) incorporated Cai et al. (2019) as the critic component, assuming that both the actor and critic are well approximated by linear functions of random features determined by initial parameters. They provided convergence to the globally optimal policy at a sublinear rate. However, these analyses over NTK or lazy training regimes assume that the neural network does not learn features from the input data.

As opposed to the linearization analysis above, we use the following tools of mean-field Langevin theory. In general, gradient method analysis of mean-field neural networks uses the convexity of the objective in the space of probability measures to show its global optimality (Nitanda & Suzuki, 2017; Chizat & Bach, 2018; Mei et al., 2018), MFLD yields to an entropy regularization term in the objective by adding Gaussian noises to the gradient. Within this research stream, our work is closely related to Nitanda et al. (2022); Chen et al. (2023) using convex analysis focusing on the log-Sobolev inequality starting from the Nitanda et al. (2021). There is also a large body of literature analyzing the optimization analysis of supervised learning with over-parameterized neural networks in the mean-field regime (Hu et al., 2021; Chen et al., 2020; Nitanda et al., 2022; Chizat, 2022; Suzuki et al., 2023).

2. Background

The agent interacts with the environment in a discounted Markov decision process (MDP) (Puterman, 2014) given by a tuple $(\mathcal{S}, \mathcal{A}, \gamma, P, r)$. The policy $\pi : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ represents the probability at which the agent takes a specific action $a \in \mathcal{A}$ at a given state $s \in \mathcal{S}$, with the agent receiving a reward $r(s, a)$ when taking an action a at state s, and transitioning to a new state $s' \in \mathcal{S}$ according to the transition probability $P(\cdot|s, a) \in \mathscr{P}(\mathcal{S})$. $\gamma \in (0, 1)$ is the discount factor. Here, we denote the state value function and the state-action value function (Q-function) associated with π by

$$V_\pi(s) = (1-\gamma)\mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau) \mid s_0 = s\right],$$
$$Q_\pi(s, a) = (1-\gamma)\mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau) \mid s_0 = s, a_0 = a\right],$$

where $a_\tau \sim \pi(s_\tau), s_{\tau+1} \sim P(s_\tau, a_\tau)$ for all $\tau \in \mathbb{Z}_{>0}$.

Note that policy π with the transition kernel P induces a Markov chain over state space $\mathcal{S}$, and we make the assumption that every policy π is ergodic, i.e. has a well-defined stationary state distribution ϱ_π and the stationary state-action distribution $\varsigma_\pi = \pi(a|s) \cdot \varrho_\pi(s)$. Moreover, we define the state visitation measure and the state-action visitation measure induced by policy π, respectively, as

$$\nu_\pi(s) = (1-\gamma)\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}(s_\tau = s),$$
$$\sigma_\pi(s, a) = \pi(a|s) \cdot \nu_\pi(s),$$

where $a_\tau \sim \pi(s_\tau), s_{\tau+1} \sim P(s_\tau, a_\tau)$ for all $\tau \in \mathbb{Z}_{>0}$. The visitation measures count the discounted number of steps that the agent visits each s or (s, a) in expectation.

Policy Gradient Here, we define the expected total reward function J_π for all π as

$$J_\pi = (1-\gamma)\mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, a_\tau)\right],$$

where $a_\tau \sim \pi(s_\tau), s_{\tau+1} \sim P(s_\tau, a_\tau)$ for all $\tau \in \mathbb{Z}_{>0}$. The goal of the policy gradient ascent is to maximize J_π by controlling policy π under the reinforcement learning setting defined above, where the optimal policy is denoted by π^*. We parameterize the policy as π_Θ with the vector parameter $\Theta \subset \mathbb{R}^d$ and we further define $J_\Theta = J_{\pi_\Theta}$ for simplicity. The gradient of J_Θ over Θ is introduced by the policy gradient theorem (Sutton et al., 1999) as $\nabla_\Theta J_\Theta = \mathbb{E}_{\nu_{\pi_\Theta}}\left[\int \nabla_\Theta \pi_\Theta(\mathrm{d}a|s) \cdot Q_{\pi_\Theta}(s, a)\right]$. The state-action value function in the above gradient is estimated by the policy evaluation problem.

Temporal-Difference Learning In temporal-difference (TD) learning, we parameterize a Q-function as Q_Ω and aim to estimate Q_π by minimizing the mean-squared Bellman error (MSBE):

$$\min_\Omega \mathrm{MSBE}(\Omega) = \mathbb{E}_{\varsigma_\pi}\left[(Q_\Omega(s, a) - \mathcal{T}^\pi Q_\Omega(s, a))^2\right],$$

where $\mathcal{T}^\pi$ is the Bellman evaluation operator associated with policy π, which is defined by $\mathcal{T}^\pi Q(s, a) = \mathbb{E}\left[(1-\gamma)r(s, a) + \gamma Q(s', a') \mid s' \sim P(s, a), a' \sim \pi(s')\right]$, and Q_Ω is a Q-function parameterized with parameter $\Omega \subset \mathbb{R}^d$. The most common TD-learning algorithm is TD(0), which, in the population version, updates Ω via the semi-gradient $\mathbb{E}_{\varsigma_\pi}\left[(Q_\Omega(s, a) - \mathcal{T}^\pi Q_\Omega(s, a)) \cdot \nabla_\Omega Q_\Omega(s, a)\right]$.

3. Mean-field Langevin Actor Critic

In this section, we introduce a particle-based double-loop neural actor-critic method with the policy and Q-function parameterized by neural networks in discrete time and the convergence analysis in the mean-field limit. We first introduce the parameterization of actor and critic below.

Parameterization of Policy and Q-Function For notational simplicity, we assume that $\mathcal{S} \times \mathcal{A} \in \mathbb{R}^D$ with $D \geq 2$ and that $\|(s,a)\| \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ without loss of generality. We parameterize a function $h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ using a two-layer neural network with width m and d-dimentional parameters $\Theta = (\theta^1, \ldots, \theta^m) \in \mathbb{R}^{d \times m}$ where it holds that $d = D + 2$, which is denoted by $\mathrm{NN}(\Theta; m)$,

$$f_\Theta(s,a) = \frac{1}{m} \sum_{i=1}^{m} h_{\theta^i}(s,a),$$
$$h_\theta(s,a) = R \cdot \beta(b) \cdot \sigma(w^\top (s,a,1)), \quad \theta = (w,b), \quad (1)$$

where $h_\theta(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the nonlinear transformation function, $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, $\beta : \mathbb{R} \to (-1,1)$ is a bounded function that represents the second layer weights with the bound $R > 0$. We now introduce the parameterization of the policy π and the Q-function Q with neural networks in the mean-field regimes respectively. Let $f_\Theta = \mathrm{NN}(\Theta; m)$, $f_\Omega = \mathrm{NN}(\Omega; M)$. Then we denote the policy and Q-function by π_Θ and Q_Ω, which are given by

$$\pi_\Theta(a|s) \propto \exp\left(-f_\Theta(s,a)\right), \quad Q_\Omega(s,a) = f_\Omega(s,a),$$

where the definition yields $\int \pi_\Theta(a|s) \mathrm{d}a = 1$ for all $s \in \mathcal{S}$.

Mean-field Limit By taking mean-field limit $m \to \infty$, we obtain the policy π_ρ and the Q-function Q_q induced by the weight distributions $\rho, q \in \mathcal{P}_2$, respectively:

$$\pi_\rho(a|s) \propto \exp\left(-\mathbb{E}[h_\theta(s,a)]\right), Q_q(s,a) = \mathbb{E}[h_\omega(s,a)],$$

where the expectations are evaluated over $\theta \sim \rho$, $\omega \sim q$, resp. We now impose the following assumption on the two-layer neural network h_θ.

Assumption 1 (Regularity of the neural network.)**.** *For the neural network h_θ defined in Eq. (1), we assume the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is uniformly bounded, L_1-Lipschitz continuous, and L_2-smooth. Besides, we assume the second weight function $\beta : \mathbb{R} \to (-1,1)$ is an odd function which is L_3-Lipschitz continuous and L_4-smooth.*

Without loss of generality, we can assume $\sigma \in (-1,1)$, which implies that the neural network h_θ is bounded by $R > 0$. Assumption 1 is a mild regularity condition except for the boundary of the neural network. Assumption 1 can be satisfied by a wide range of neural networks, e.g., $\beta(\cdot) = \tanh(\cdot/R)$ and $\sigma(\cdot) = \tanh(\cdot)$. We further define $J : \rho \mapsto J[\rho] := J_{\pi_\rho}$ as a functional over ρ.

Remark 1. *Assumptions 1 and 3 ensure our neural network model class is sufficiently rich, covering widely used activation functions like a sigmoid and hyperbolic tangent. This approach aligns with common practices in analytical research Agazzi & Lu (2020); Zhang et al. (2020; 2021); Leahy et al. (2022) Additionally, this assumption is validated in cases where kernels are smooth and light-tailed, like the RBF kernel as mentioned in Suzuki et al. (2023) for applications such as MMD and KSD estimation.*

3.1. Actor Update: Mean-field Langevin Policy Gradient

We aim to minimize the regularized negative expected total rewards $J[\rho]$ over the probability distribution together. The regularized objective can be written as follows:

$$\min_\rho \{\mathcal{F}[\rho] = -J[\rho] + \frac{\lambda}{2}\mathbb{E}_\rho[\|\theta\|_2^2] + \lambda \mathrm{Ent}[\rho] + \lambda Z\},$$

where $\lambda > 0$ is a regularization parameter, and $Z = \frac{1}{2}\ln(2\pi)$ is the normalization constant.

Remark 2. *The L^2-regularization $\mathbb{E}_\rho[\|\theta\|_2^2]$ helps to induce log-Sobolev inequality. This is due to the fact that $\|\theta\|_2^2$ is strongly convex, see Section B.1 especially Proposition 2 for details over log-Sobolev inequality. The entropy regularization term is required by adding Gaussian noise to the gradient, allowing global convergence analysis under less restrictive settings (Mei et al., 2019b). Adding these terms introduces a slight optimization bias of order $\mathcal{O}(\lambda)$. These regularization terms also have statistical benefits to smooth the problem. Note that we can rewrite the objective functional $\mathcal{F}$ as $\min_\rho\{\mathcal{F}[\rho] = -J[\rho] + \lambda \cdot \mathrm{KL}(\rho\|\nu)\}$ where $\nu = \mathcal{N}(0, I_d)$ is a standard Gaussian distribution.*

In the sequel, we introduce the policy gradient over the measure space to construct the MFLD. Let the objective subtracted by the entropy be $F[\rho] := -J[\rho] + \frac{\lambda}{2} \cdot \mathbb{E}_\rho[\|\theta\|_2^2]$.

Proposition 1 (Policy Gradient)**.** *For the distribution ρ over the policy parameter θ, we have*

$$\frac{\delta F}{\delta \rho}[\rho](\theta) = \mathbb{E}_{\sigma_{\pi_\rho}}[A_{\pi_\rho} \cdot h_\theta] + \frac{\lambda}{2}\|\theta\|_2^2, \quad (2)$$

where $\frac{\delta F}{\delta \rho}[\rho](\theta)$ is the first-variation of $F[\rho]$ in Definition 1, and A_{π_ρ} is the advantage function defined by $A_{\pi_\rho}(s,a) = Q_{\pi_\rho}(s,a) - \int \pi_\rho(\mathrm{d}a'|s) \cdot Q_{\pi_\rho}(s,a')$.

See Appendix D.1 for the proof. In practice, we do not get the true advantage function A_{π_ρ}, but instead use the estimator $A_t(s,a) = Q_t(s,a) - \int \pi_t(\mathrm{d}a'|s) \cdot Q_t(s,a')$ with Q_t obtained from critic. Let the initial distribution $\rho_0 = \mathcal{N}(0, I_d)$. Then we update ρ_t according to the following McKean-Vlasov stochastic differential equation, which solves the following Fokker-Planck equation over time $t \in \mathbb{R}_{\geq 0}$:

$$\mathrm{d}\theta_t = -\nabla \widetilde{\frac{\delta F}{\delta \rho}}[\rho_t](\theta_t) \cdot \mathrm{d}t + \sqrt{2\lambda} \cdot \mathrm{d}W_t, \quad (3)$$

$$\partial_t \rho_t = \lambda \cdot \Delta \rho_t + \nabla \cdot \left(\rho_t \cdot \nabla \widetilde{\frac{\delta F}{\delta \rho}}[\rho_t]\right). \quad (4)$$

where $\widetilde{\frac{\delta F}{\delta \rho}}[\rho_t](\theta) = \mathbb{E}_{\sigma_{\pi_\rho}}[A_t \cdot h_\theta] + \frac{\lambda}{2}\|\theta\|_2^2$ is the approximated policy gradient and $\{W_t\}_{t \geq 0}$ is the Brownian motion in $\mathbb{R}^d$ with $W_0 = 0$.

Time and Space Discretization To implement our approach, we represent $\rho_\Theta = \frac{1}{m}\sum_{i=1}^m \delta_{\theta^i}$ as a mixture of m particles $\Theta = \{\theta^i\}_{i\in[m]}$, which corresponds to a neural network with m neurons. Let T be the number of iterations. We perform a discrete-time update at each k-th step of a noisy policy gradient method, where the policy parameter $\Theta_k = \{\theta_k^i\}_{i\in[m]}$ is updated for all $k \in [T]$ as

$$\theta_{k+1}^i = (1-\eta\lambda)\,\theta_k^i - \eta\,\mathbb{E}_{\sigma_{\pi_k}}[A_k \nabla h_{\theta_k^i}] + \sqrt{2\lambda\eta}\xi_k^i, \quad (5)$$

where we define $\rho_k = \rho_{\Theta_k}, \pi_k = \pi_{\Theta_k}$ and denote a learning rate by $\eta > 0$. A_k is an approximation of an advantage function A_{π_k} and ξ_k^i is an i.i.d. random variable $\xi_k^i \sim \mathcal{N}(0, I_d)$. Note that, for each k-step, the agent uniformly sample $L \in [T_{\mathrm{TD}}]$ and adopt $Q^{(L)}$ as Q_k from the estimated Q-functions $\{Q^{(l)}\}_{l\in[T_{\mathrm{TD}}]}$ obtained by MFLTD (Algorithm 2). See Algorithm 1 for more detail. We denote a learning rate by $\eta > 0$. The discrete version of the MFLPG can be attributed to the MFLDs in Eq. (3) by taking the mean-field limit $m, k \to \infty, \eta \to 0$ being $t = \eta k$.

Algorithm 1 Mean-field Langevin Policy Gradient

Input: $\theta_0^i \leftarrow N(0, I_d)$ for all $i \in [m]$ and $\pi_0(\cdot) \leftarrow \pi_{\Theta_0}$.
1: **for** $k = 0$ to $T-1$ **do**
2: Given the current policy π_k, run Algorithm 2 and uniformly sample $L \in [T_{\mathrm{TD}}]$: $Q_k \leftarrow Q^{(L)}$
3: Calculate $A_k = Q_k - \langle \pi_k, Q_k\rangle$ and update with the i.i.d. noise $\xi_k^i \sim \mathcal{N}(0, I_d)$ for all $i \in [m]$ by $\theta_{k+1}^i \leftarrow (1-\eta\lambda)\,\theta_k^i - \eta\,\mathbb{E}_{\sigma_{\pi_k}}[A_k \cdot \nabla h_{\theta_k^i}] + \sqrt{2\lambda\eta}\xi_k^i$
4: $\pi_{k+1} \leftarrow \pi_{\Theta_{k+1}}$
5: **end for**
Output: π_T

3.2. Critic Update: Mean-field Langevin TD Learning

In this section, we introduce the Mean Field Langevin Temporal Difference (MFLTD) method to address the challenges of TD learning in the mean-field regime, especially for optimizing two-layer neural networks. The core issue in TD learning is the semi-gradient of the mean-square Bellman error, which may not always converge due to its non-monotonic descent in the mean-field context. This complexity arises from optimizing over probability measures rather than direct parameter adjustments, similar to navigating a Wasserstein gradient flow instead of a conventional L_2 descent.

MFLTD is a novel double-loop algorithm designed to ensure monotonic objective reduction in each outer loop iteration, akin to proximal gradient methods. The inner loop approximates the true value function by solving a majorization problem that overestimates the mean squared error. This guarantees that the objective function, $\mathcal{L}_l$, is convex over the space of probability distributions. As a result, stationary points of the inner objective yield values at least a constant factor of the expected squared error $\frac{1-\gamma}{2}\mathbb{E}[(Q_q - Q_\pi)^2]$ plus an error term bounded by $O(\lambda)$. This structure systematically reduces the expected squared error with each iteration, enhancing the algorithm's efficiency and clarity.

Inner Loop Update The inner loop is based on the KL-divergence regularized MFLD analysis in (Nitanda et al., 2022; Chizat, 2022). In the mean-field view, we minimize the objective $\min_q\{\mathcal{L}_l[q] = L_l[q] + \lambda_{\mathrm{TD}}\,\mathrm{Ent}[q]\}$ where λ_{TD} is a regularization parameter and $L_l[q]$ is defined, for $l \in [0, T_{\mathrm{TD}}]$, by

$$L_l[q] = \mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}^\pi Q^{(l)}) \cdot (Q_q - Q_\pi)] \quad (6)$$
$$+ \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_q)^2] + \frac{\lambda_{\mathrm{TD}}}{2}\mathbb{E}_q[\|\omega\|_2^2] + \lambda_{\mathrm{TD}} Z,$$

where $Z = \frac{1}{2}\ln(2\pi)$ is the normalization constant and, on the right-hand side, the first term is the linearized surrogate TD error and the second one is the proximal control term. We obtain the MFLD at time s as

$$\mathrm{d}\omega_s = -\nabla\frac{\delta L_l}{\delta q}[q_s](\omega_s)\cdot \mathrm{d}t + \sqrt{2\lambda_{\mathrm{TD}}}\cdot \mathrm{d}W_s,$$

where $\{W_s\}_{s\geq 0}$ is the Brownian motion in $\mathbb{R}^d$ with $W_0 = 0$. Let $x' = (s', a')$ be the next state and action of $x = (s, a)$. To understand the intuition behind the proximal semi-gradient, we have the gradient of first variation of L_l as

$$\nabla\frac{\delta L_l}{\delta q}[q](\omega)$$
$$= \mathbb{E}_{\varsigma_\pi}\left[\left(\overline{Q}_q^{(l)}(x) - (1-\gamma)r(x) - \gamma Q^{(l)}(x')\right)\nabla h_\omega(x)\right] + \lambda_{\mathrm{TD}}\omega,$$

where we define the averaged Q-function by $\overline{Q}_q^{(l)} = (Q_q - \gamma\, Q^{(l)})/(1-\gamma)$ and the expectation is obtained under $(x, x') \sim \varsigma_\pi$.

Outer Loop Update The last iterate Q_{Ω_N} of the previous inner loop is given in the outer loop as $Q^{(l+1)}$. See Algorithm 2 for the discretization algorithm of MFLTD. We remark that considering that the inner-loop algorithm converges to the optimum at the exponential rate, the computational complexity of the inner-loop does not become a bottleneck in implementation. In this regard, the results in Section 5 offer valuable insights.

Algorithm 2 Mean-field Langevin TD Learning

Input: $\omega_0^j \leftarrow N(0, I_d)$ for all $j \in [M]$ and the policy π.
1: **for** $l = 0$ to $T_{\mathrm{TD}} - 1$ **do**
2: **for** $n = 0$ to $N-1$ **do**
3: Average Q-function: $\overline{Q}_n^{(l)} = \frac{1}{1-\gamma}(Q_{\Omega_n} - \gamma\, Q^{(l)})$
4: Update with the i.i.d. noise $\xi_n^j \sim \mathcal{N}(0, I_d)$ for all $j \in [M]$:
$\omega_{n+1}^j \leftarrow \omega_n^j - \eta_{\mathrm{TD}} \cdot \nabla\frac{\delta L_l}{\delta q}[q_n](\omega_n^j) + \sqrt{2\lambda_{\mathrm{TD}}\eta_{\mathrm{TD}}}\xi_n^j$
5: **end for**
6: $Q^{(l)} \leftarrow Q_{\Omega_N}$
7: **end for**
Output: $\{Q^{(l)}\}_{l\in[T_{\mathrm{TD}}]}$

4. Main Results

In this section, we present the results of our investigation into the theoretical support of the mean-field Langevin actor-critic. First of all, we base our analysis on the regularity condition that the reward is bounded.

Assumption 2 (Regularity Condition on Reward)**.** *We assume that there exists an absolute constant $R_r > 0$ such that $R_r = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |r(s,a)|$. As a result, we have $|V_\pi(s)| \leq R_r, |Q_\pi(s,a)| \leq R_r, |J_\pi| \leq R_r$ and $|A_\pi(s,a)| \leq 2R_r$ for all π and $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

Combining Assumption 1 and 2 yields that $2R_r \leq R$ by setting $R > 0$ large enough. Such a regularity condition is commonly used in the literature (Liu et al., 2019; Wang et al., 2020). In what follows, we introduce the following regularity condition on the state-action value function Q_π.

Assumption 3 (Value Function Class)**.** *We define for $R, M > 0$*

$$\mathcal{F}_{R,M} = \Big\{ \int \beta' \cdot \sigma(w^\top(s,a,1)) \cdot \rho'(\mathrm{d}\beta', \mathrm{d}w) \\ : \mathrm{KL}(\rho'\|\nu) \leq M,\ \rho' \in \mathscr{P}((-R,R)\times\mathbb{R}^{d-1}) \Big\} \quad (7)$$

which is equivalent to the function class of $\mathbb{E}_{\theta\sim\rho}[h_\theta]$ for $\rho \in \mathcal{P}_2$. We assume that $Q_\pi(x), A_\pi(x) \in \mathcal{F}_{R,M}$ for any π.

As will be further explained in Appendix B.2, we note that Assumption 3 is a natural regularity condition on Q_π, A_π, as $\mathcal{F}_{R,M}$ captures a rich family of functions, which is a subset of the Barron class (Barron, 1993). Indeed, by making the neural network radius R, M sufficiently large, $\mathcal{F}_{R,M}$ asymptotically approaches the Barron class and captures a rich function class by the universal approximation theorem (Barron, 1993; Pinkus, 1999). Also, as long as smoothness and boundedness of networks are assumed (Assumption 1), every network can be included in the above class at least with a small modification. Similar regularity condition is a commonly used concept in literature (Farahmand et al., 2016; Yang & Wang, 2019; Liu et al., 2019; Wang et al., 2020).

4.1. Mean-field Langevin TD Learning

In the continuous-time limit, the next step is obtained by $q^{(l+1)} = q_S$ where we define S as the inner-loop run-time. Regarding the outer-loop update, we obtain the following one-step descent lemma.

Lemma 1 (One-Step Descent Lemma for MFLTD)**.** *Let $q_*^{(l+1)}$ be the inner-loop optimal distribution for any inner step l. For $\{Q^{(l)}\}_{l\in[T_{\mathrm{TD}}]}$ in Algorithm 2 with the TD update in Line 2, it holds that*

$$\frac{\gamma(2-\gamma)}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}\left[(\Delta Q^{(l+1)})^2 - (\Delta Q^{(l)})^2\right] \\ \leq -\frac{1-\gamma}{2}\|\Delta Q^{(l+1)}\|_{\varsigma_\pi}^2 + \frac{2R}{1-\gamma}\|Q^{(l+1)} - Q_*^{(l+1)}\|_{\varsigma_\pi} \\ + \lambda_{\mathrm{TD}}\,\mathrm{KL}(q^{(l+1)}\|q_*^{(l+1)}) + \lambda_{\mathrm{TD}}\,\mathrm{KL}(q_\pi\|\nu), \quad (8)$$

where we define that $\Delta Q^{(l)} = Q^{(l)} - Q_\pi$, and denote by $Q_*^{(l+1)}$ the Q-function $Q_{q_*^{(l+1)}}$, and $q^{(l+1)}, q_\pi$ are the weight distributions inducing $Q^{(l+1)}, Q_\pi$, resp.

See Appendix C.1 for the proof. The existence of q_π is guaranteed by Assumption 3. Lemma 1 illustrates the one-step descent behavior. The second and third terms of the right-hand side of Eq. (8) represent non-asymptotic errors obtained through the inner loop, and it exponentially decreases with an increase in the run-time S of the inner loop. The key to the proof of Lemma 1 is the use of geometric features related to the norm of the Bellman equation operator in Lemma 11. The shrinking norm suppresses errors in the semi-gradient direction that deviates from the true gradient direction. In what follows, Combining Proposition 5 and Lemma 1 allows us to establish the global convergence theorem for the MFLTD as

Theorem 1 (Global Convergence of the MFLTD)**.** *Under Assumption 1, 2, and 3, the outputs $\{Q^{(l)}\}_{l\in[T_{\mathrm{TD}}]}$ of Algorithm 2 satisfies, for $\lambda_{\mathrm{TD}} > 0$ and the inner runtime $S > 0$, that*

$$\frac{1}{T_{\mathrm{TD}}}\sum_{l=1}^{T_{\mathrm{TD}}}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] \leq \frac{8\gamma R^2}{(1-\gamma)^2 T_{\mathrm{TD}}} \\ + C_1\lambda_{\mathrm{TD}}^{-\frac{1}{2}}e^{(-\alpha\lambda_{\mathrm{TD}}S)} + C_2 e^{(-2\alpha\lambda_{\mathrm{TD}}S)} + C_3\lambda_{\mathrm{TD}},$$

where $C_1, C_2, C_3 > 0$ are the absolute constants such that $C_1 = \frac{8\sqrt{3}R^3}{(1-\gamma)^{5/2}}, C_2 = \frac{24R^4}{(1-\gamma)^2}, C_3 = \frac{2M}{1-\gamma}$, and we define α as a LSI constant in Definition 2.

See Appendix C.2 for the proof. Theorem 1 implies that if the inner-loop error goes to zero, then Q-function converges to the true state-action value function Q_π at the time-averaged sublinear rate $\mathcal{O}(1/T_{\mathrm{TD}})$ with the regularization bias $\mathcal{O}(\lambda_{\mathrm{TD}})$. Therefore, setting the parameters suitably yields the following corollary.

Corollary 1. *Under the same conditions as Theorem 1, let $S = -\frac{3\log\lambda_{\mathrm{TD}}}{2\alpha\lambda_{\mathrm{TD}}}$ and $\lambda_{\mathrm{TD}} = T_{\mathrm{TD}}^{-1}$, uniform sampling $L \in [T_{\mathrm{TD}}]$ yields that*

$$\mathbb{E}_{L,\varsigma_\pi}[(Q^{(L)} - Q_\pi)^2] \leq \frac{C_0 + C_1 + C_3}{T_{\mathrm{TD}}} \wedge \frac{C_2}{T_{\mathrm{TD}}^3},$$

where we denote $C_0 = \frac{8\gamma R^2}{(1-\gamma)^2}$.

See the Appendix C.4 for the proof. Corollary 1 shows that, given a policy π, $Q^{(L)} \to Q_\pi$ as $T_{\mathrm{TD}} \to \infty$ at rate $\mathcal{O}(1/T_{\mathrm{TD}})$. This result is in perfect agreement with the convergence rate $\mathcal{O}(1/T_{\mathrm{TD}})$ that Cai et al. (2019) obtains from TD learning in the NTK regime. Note that the previous work on TD learning in the mean-field regime has the $1/\alpha$-rate bias with α being the network scaling factor (Zhang et al., 2020). Thus, this is the first time that global convergence has been demonstrated in a domain that takes advantage of the data-dependent advantage of neural networks.

4.2. Mean-field Langevin Policy Gradient

We introduce the analysis of global convergence of the MFLPG, under the mean-field limit and the discretized setting, rep. First, we lay out a moment condition as

Assumption 4 (Moment Condition on Radon-Nikodym Derivative)**.** *We assume that there exists absolute constants $\kappa, \iota > 0$ such that for any $t \in \mathbb{R}_{\geq 0}$*

$$\textit{(i)}\ \left\|\mathrm{d}\sigma_t/\mathrm{d}\varsigma_t\right\|_{\varsigma_t,2} \leq \iota, \qquad \textit{(ii)}\ \left\|\mathrm{d}\sigma^*/\mathrm{d}\sigma_t\right\|_{\sigma_t,2} \leq \kappa,$$

where $\frac{\mathrm{d}\sigma_t}{\mathrm{d}\varsigma_t}$ and $\frac{\mathrm{d}\sigma^*}{\mathrm{d}\sigma_t}$ are the Radon-Nikodym derivatives.

Note that when the MDP starts at the stationary distribution ς_t, the state-action visitation measures σ_t are identical to ς_t. Regarding Assumption 4-(i), if the induced Markov state-action chain rapidly reaches equilibrium, this assumption also holds true. The same requirement is imposed by Liu et al. (2019); Wang et al. (2020). Meanwhile, the optimal moment condition in Assumption 4-(ii) asserts that the concentrability coefficients are upper-bounded. This regularity condition is a commonly used concept in literature (Farahmand et al., 2016; Chen & Jiang, 2019; Liu et al., 2019; Wang et al., 2020) and is utilized to guarantee the global optimality in Theorem 2.

What we follow, since MFLD can be basically attributed to the Wasserstein gradient flow, the convergence to the stationary point is guaranteed. We define the proximal Gibbs distribution $\widehat{\rho}_t$ by

$$\widehat{\rho}_t \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}[\rho_t]\right),$$

which the Wasserstein gradient flow in Eq. (4) goes toward unless we have critic errors. Motivated by the log-Sobolev-inequality argument in (Nitanda et al., 2022), the convergence in KL divergence between $\rho_t, \widehat{\rho}_t$ are yielded as

Lemma 2 (Convergence to Stationary Point of MFLPG)**.** *Under Assumption 1,2, 3, and 4-(i) we obtain for any $t \geq 0$ that*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] \leq -\alpha\lambda^2 \cdot \mathrm{KL}(\rho_t||\widehat{\rho}_t) \\ + 2R^2(L_1+L_3)^2\iota^2\,\mathbb{E}_{\varsigma_t}[(Q_t - Q_{\pi_t})^2],\end{aligned} \tag{9}$$

where $\alpha > 0$ is the LSI constant of $\widehat{\rho}_t$ and Q_t is the Q-function estimator given by the critic.

See Appendix D.2 for the proof. The second term on the right-hand side of Eq. (9) is the policy evaluation error given by the result in Corollary 1. Lemma 2 implies that, if the critic error is zero, then ρ_t goes toward $\widehat{\rho}_t$ as $t \to \infty$. The Entropy-sandwich-argument (derived from Lemma 3.4 (Chizat, 2022)) yields that if J is the convex functional then the performance difference $\mathcal{F}[\rho_t] - \min_{\rho\in\mathcal{P}_2}\mathcal{F}[\rho]$ is upper-bounded by $\mathrm{KL}(\rho_t\|\widehat{\rho}_t)$, which concludes the global optimality of the stationary point. By contrast, we obtain the global convergence by utilizing the one-point convexity of J_π at the global optimum π^*, which is established by Proposition 3 in Kakade & Langford (2002).

Lemma 3 (Global Optimality of Stationary Point)**.** *Assume the same conditions as Lemma 2 and Assumption 4-(ii), and under $\mathrm{KL}(\nu\|\widehat{\rho}_t) \leq M$ for all $t \geq 0$, we have for all $t \geq 0$ and $\lambda > 0$ that*

$$\max_\pi J_\pi - J[\rho_t] \leq \lambda\,\mathrm{KL}(\rho_t||\widehat{\rho}_t) + \widetilde{C}_1\sqrt{\lambda}, \tag{10}$$

where we denote $\widetilde{C}_1 = \frac{1}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2 + 2M$.

See Appendix D.3 for the proof. Lemma 3 implies that $J[\rho_t] \to \max_\pi J_\pi$ with a regularization bias $\mathcal{O}(\sqrt{\lambda})$ as ρ_t gets much closer to $\widehat{\rho}_t$. In what follows, we establish the global optimality and the convergence rate of the MFLPG.

Theorem 2 (Global Optimality and Convergence of the MFLPG)**.** *We set $T_{\mathrm{TD}} = \frac{1}{\alpha\lambda^{3/2}}$. Under the same conditions as Lemma 3, the continuous MFLPG yields for all $t \in \mathbb{R}_{\geq 0}$ and $\lambda > 0$ that*

$$\max_\pi J_\pi - J[\rho_t] \leq 2R\exp(-\alpha\lambda t) + \mathcal{O}(\lambda^{1/2}).$$

See Appendix D.4 for the proof. Theorem 2 demonstrates that the MFLPG achieves linear convergence in continuous time with regularization bias $\mathcal{O}(\lambda^{1/2})$, significantly overwhelming the $\mathcal{O}(t^{-1/2})$ convergence rate typical of NTK regime (Wang et al., 2020). Furthermore, we highlight an enhanced convergence without bias through a time-varying $\lambda_t = \mathcal{O}(1/\log t)$ strategy, given in the annealing argument in Chizat (2022).

Theorem 3 (The Convergence Analysis for Fully Discretized Particle)**.** *Suppose that $\left|\frac{\delta^2 J}{\delta\rho^2}[\rho](\theta,\theta')\right| \leq L(1 + c(\|\theta\|^2 + \|\theta'\|^2))$ for any $\theta, \theta' \in \mathbb{R}^d$ and define $\delta_{\mathrm{TD}}^2 = \mathcal{O}(\mathbb{E}_{\varsigma_{\pi_k}}[(Q_k - Q_{\pi_k})^2])$. Under the same conditions as Theorem 2, run Algorithm 1, where the actor update is given in Eq. (5). It holds for all $m \in \mathbb{Z}_{\geq 0}$, $T \in \mathbb{Z}_{\geq 0}$, $\eta \geq 0$, and $\lambda > 0$ that*

$$\begin{aligned}\max_\pi J_\pi - \mathbb{E}[J_{\Theta_T}] \leq 2R\exp(-\alpha\lambda\eta T/2) \\ + \eta\left(\widetilde{C}_{\lambda,1}\frac{1}{m} + \widetilde{C}_{\lambda,2}(\eta^2 + \lambda\eta) + \widetilde{C}\alpha\lambda^{3/2} + \delta_{\mathrm{TD}}^2\right),\end{aligned} \tag{11}$$

where $\widetilde{C}_{\lambda,1} = \mathcal{O}\left(1 \vee \frac{\lambda}{\eta}\right)$, $\widetilde{C}_{\lambda,2} = \mathcal{O}\left(\frac{1}{\lambda} \vee \frac{1}{\eta}\right)$ *and* $\widetilde{C} = \mathcal{O}(1)$. *Especially, setting* $\eta \asymp \lambda = \mathcal{O}(\frac{1}{m})$, $T_{\mathrm{TD}} = \Omega(m)$, *and* $T = \Omega(m^3 e^m)$ *yields for any* $m \in \mathbb{Z}_{\geq 0}$ *particles that*

$$\max_{\pi} J_\pi - \mathbb{E}[J_{\Theta_T}] \leq \mathcal{O}\left(\frac{1}{m^2}\right). \tag{12}$$

See Appendix E for the proof. Eq. (11) implies that the expected J_{Θ_k} goes toward $\max_\pi J_\pi$ biased by the finite width error $\mathcal{O}(\lambda/m \vee \eta/m)$ and the time-discretization error depending on η, as $T \to \infty$. In addition, Eq. (12) demonstrates that we can evaluate the expected performance error $\max_\pi J_\pi - \mathbb{E}[J_{\Theta_T}] = \mathcal{O}(m^{-2})$ with the suitable parameter setting. This result is an indicator of how well the neural network performs as its width is increased, i.e., $m \to \infty$, and it is faster than the rate $\mathcal{O}(m^{-1/16})$ derived from the existing policy gradient research in NTK regime (Wang et al., 2020). Although it is hard to understand explicitly the impact of training in the mean-field regime on feature learning, training relies on a fixed set of features in NTK scaling, limiting expressiveness to the network's width. In contrast, MF scaling allows feature bases in the first layer to adjust dynamically, enhancing the training data's adaptability. In fact, MF networks have a superior representation ability over finite-width NTK networks, which struggle to approximate characteristic functions (Suzuki, 2019; Ghorbani et al., 2019; Damian et al., 2022). See Appendix B.3 for more detail.

5. Numerical Analysis

In this section, we conducted a numerical experiment to compare the Critic component, which is based on the proposed MFLTD, against the existing TD(1) algorithm that utilizes the Bellman error semi-gradient. Additionally, we demonstrated how the learning performance differs when using a neural network that follows the NTK with a representation that is independent of input data and dependent on initial values. Specifically, we performed learning on the CartPole-v1 environment provided by OpenAI's Gym and implemented the estimation of the state-action value function during optimal policy selection. In this experiment, we used a neural network with 256 neurons, ran 4000 episodes with a discounted factor of $\gamma = 0.99$, and employed a learning rate of $\eta = 0.0001$ for MFLTD. Notably, we conducted MFLTD's inner loop with a step size of $K = 10$, repeated it $T_{\mathrm{TD}} = 400$ times in the outer loop, and sampled using one episode for each inner step. Furthermore, we applied Gaussian noise of magnitude induced by the entropy regularization parameter $\lambda = 0.001$, following Algorithm 2, along with L_2 regularization. To assess the difference in performance due to representation learning covered by Mean-field analysis, we also implemented NTK-TD with a double-loop setup where representations are fixed at initial values, similar to MFLTD. Additionally, to address the primary weakness of our proposed algorithm, the double-loop, we examined its impact on computational complexity compared to single-loop TD(1).

Figure 1 presents the average and standard deviation of each learning process conducted ten times. From this figure, we observe that MF training with features independent of initial values outperforms when compared with an equal number of neurons, primarily due to increased expressiveness gained through feature learning. Furthermore, while the single-loop results are faster in regions of lower accuracy under the same computational load and time, they exhibit decreased speed in regions of higher accuracy, ultimately demonstrating that our proposed double-loop method approximates the true value function more effectively.

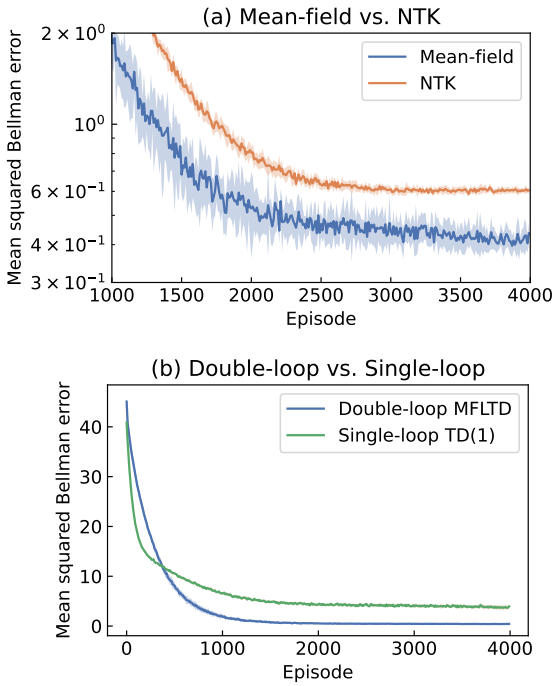

Figure 1. Comparison of Time Evolution of Mean Squared Bellman Error Between Algorithms for TD Learning Near the Optimal Policy in the Game Model "CartPole-v1".

6. Conclusion

We studied neural policy optimization in the mean-field regime and provided the first global optimality guarantee and the linear convergence rate for a neural actor-critic algorithm in the presence of feature learning. For both actor and critic, we attributed their updates to the MFLD and analyzed their evolutions as the optimization of corresponding probability measures under mean-field limit and time-and-particle discretized version, resp. We provide theoretical guarantees for global convergence to global optimality, and empirical experiments that validate the superiority of the proposed algorithm in policy evaluation.

Acknowledgements

KY was partially supported by JST CREST (JPMJCR2115) and Takenaka Scholarship Foundation. KO was partially supported by JST, ACT-X Grant Number JPMJAX23C4, JAPAN. TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbe, E., Adsera, E. B., and Misiakiewicz, T. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623. PMLR, 2023.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66. PMLR, 09–12 Jul 2020.

Agazzi, A. and Lu, J. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. pp. 1–24, 2020.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946. Curran Associates, Inc., 2022.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Morgan Kaufmann, 1995.

Bakry, D. and Émery, M. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*, pp. 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg.

Barron, A. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994. ISSN 1573-0565.

Baxter, J., Bartlett, P., and Weaver, L. Direct gradient-based reinforcement learning: II. Gradient ascent algorithms and experiments. Technical report, Research School of Information Sciences and Engineering, Australian National University, 1999.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *CoRR*, abs/1906.01786, 2019.

Bhatnagar, S., Ghavamzadeh, M., Lee, M., and Sutton, R. S. Incremental natural actor-critic algorithms. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20, pp. 105—-112. Curran Associates, Inc., 2007.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009. ISSN 0005-1098.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32(NeurIPS): 1–22, 2019.

Chen, F., Ren, Z., and Wang, S. Uniform-in-time propagation of chaos for mean field langevin dynamics, 2023. URL https://arxiv.org/abs/2212.03050.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13363–13373. Curran Associates, Inc., 2020.

Chizat, L. Mean-field Langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3036–3046. Curran Associates, Inc., 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2933–2943. Curran Associates, Inc., 2019.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993. doi: 10.1162/neco.1993.5.4.613.

E, W., Ma, C., and Wu, L. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019. ISSN 1539-6746.

Farahmand, A., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, 17(139):1–66, 2016.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1467–1476. PMLR, 2018.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9108–9118. Curran Associates, Inc., 2019.

Holley, R. and Stroock, D. W. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

Hu, K., Ren, Z., Šiška, D., and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043 – 2065, 2021.

Huang, B., Lu, C., Leqi, L., Hernández-Lobato, J. M., Glymour, C., Schölkopf, B., and Zhang, K. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, pp. 9260–9279. PMLR, 2022.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580. Curran Associates, Inc., 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Kakade, S. M. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1531–1538. MIT Press, 2001.

Klusowski, J. M. and Barron, A. R. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1009–1014. MIT Press, 1999.

Le Lan, C., Tu, S., Oberman, A., Agarwal, R., and Bellemare, M. G. On the generalization of representations in reinforcement learning. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4132–4157. PMLR, 28–30 Mar 2022.

Leahy, J.-M., Kerimkulov, B., Siska, D., and Szpruch, L. Convergence of policy gradient for entropy regularized MDPs with neural network approximation in the mean-field regime. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12222–12252. PMLR, 17–23 Jul 2022.

Li, Z., Ma, C., and Wu, L. Complexity measures for neural networks with general activation functions using path-based norms. *arXiv preprint arXiv:2009.06132*, 2020.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10564–10575. Curran Associates, Inc., 2019.

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

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. 99(1):1–77, 2019a.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2388–2464. PMLR, 25–28 Jun 2019b.

Menz, G. and Schlichting, A. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809 – 1884, 2014.

Mousavi-Hosseini, A., Park, S., Girotti, M., Mitliagkas, I., and Erdogdu, M. A. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh*

International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=6taykzqcPD.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Nitanda, A., Wu, D., and Suzuki, T. Particle dual averaging: Optimization of mean field neural network with global convergence rate analysis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19608–19621. Curran Associates, Inc., 2021.

Nitanda, A., Wu, D., and Suzuki, T. Convex analysis of the mean field Langevin dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9741–9757. PMLR, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063. MIT Press, 1999.

Suzuki, T. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Suzuki, T., Wu, D., and Nitanda, A. Convergence of mean-field langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. *arXiv preprint*, 2023.

Telgarsky, M. Feature selection and low test error in shallow low-rotation reLU networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=swEskiem99.

Tsitsiklis, J. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 9, pp. 1075–1081. MIT Press, 1996.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

Weinan, E., Ma, C., and Wu, L. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Sci. China Math*, 2019.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement Learning*, pp. 5–32, 1992.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=edJ_HipawCa.

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6995–7004. PMLR, 2019.

Zhang, Y., Cai, Q., Yang, Z., Chen, Y., and Wang, Z. Can temporal-difference and Q-learning learn representation? a mean-field theory. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19680–19692. Curran Associates, Inc., 2020.

Zhang, Y., Chen, S., Yang, Z., Jordan, M., and Wang, Z. Wasserstein flow meets replicator dynamics: A mean-field analysis of representation learning in actor-critic. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15993–16006. Curran Associates, Inc., 2021.

Zhou, M. and Lu, J. Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees. *Journal of Machine Learning Research*, 24 (222):1–34, 2023.

A. Notations

We denote by $\mathscr{P}(\mathscr{X})$ the set of distribution measures over the measurable space $\mathscr{X}$. Given a distribution measure function $\mu \in \mathscr{P}(\mathscr{X})$, the expectation with respect to μ as $\mathbb{E}_{\theta\sim\mu}[\cdot]$ or simply $\mathbb{E}_\mu[\cdot], \mathbb{E}_\theta[\cdot]$ when the random variable and distribution are obvious from the context. In addition, for $\mu \in \mathscr{P}(\mathscr{X})$ and $p > 0$, we define $\|f(\cdot)\|_{\mu,p} = (\int_\Theta |f|^p \mathrm{d}\mu)^{\frac{1}{p}}$ as the $L^p(\mu)$-norm of f. We define $\|f(\cdot)\|_{\mu,\infty} = \inf\{C \geq 0 : |f(x)| \leq C \text{ for } \mu\text{-almost every } x\}$ as the $L^\infty(\mu)$-norm of f. We write $\|f\|_{\mu,p}$ for notational simplicity when the variable of f is obvious from the context. Especially, the $L_2(\mu)$-norm is denoted by $\|\cdot\|_\mu$. For a vector $v \in \mathbb{R}^d$ and $p > 0$, we denote by $\|v\|_p$ the L^p-norm of v. Given two distribution measures $\mu, \rho \in \mathscr{P}(\mathscr{X})$, we denote the Radon–Nikodým derivative between μ and ρ by $\frac{\mathrm{d}\mu}{\mathrm{d}\rho}$. $\mathrm{KL}(\cdot\|\cdot)$ stands for the Kullbuck-Leibler divergence as $\mathrm{KL}(\mu\|\rho) = \int \mathrm{d}\mu \ln \frac{\mathrm{d}\mu}{\mathrm{d}\rho}$, and also $\mathrm{I}(\cdot\|\cdot)$ stands for the Fisher divergence as $\mathrm{I}(\mu\|\rho) = \int \mathrm{d}\mu \|\nabla_\theta \ln \frac{\mathrm{d}\mu}{\mathrm{d}\rho}\|_2^2$. Also, we define the entropy $\mathrm{Ent}[\cdot]$ by $\mathrm{Ent}[\mu] = \int \mathrm{d}\mu \ln \mu$. Let $\mathcal{P}_2 \subset \mathscr{P}(\mathbb{R}^d)$ be the space of probability density functions such that both the entropy and second moment are finite.

B. Additional Remarks

B.1. Logarithmic Sobolev Inequality

In this paper, we extend the convergence analysis of a nonlinear Fokker-Planck equation, mean-field Langevin dynamics to the context of reinforcement learning. The analysis is based on the KL-divergence regularization (Mei et al., 2019a; Hu et al., 2021; Chen et al., 2020) and the induced log-Sobolev inequality (Nitanda et al., 2022; Chizat, 2022). Below are some mathematical tools necessary for them. Particularly, in the MFLD convergence analysis, it is important to make use of the following proximal Gibbs distribution defined as follows. To define the MFLD of functional F, we first introduce the first variation of functionals as

Definition 1 (First-variation of Functionals)**.** Let $F : \mathcal{P}_2 \to \mathbb{R}$ and we suppose there is a functional $\frac{\delta F}{\delta \rho} : \mathcal{P}_2 \times \mathbb{R}^d \ni (\rho, \theta) \mapsto \frac{\delta F}{\delta \rho}[\rho](\theta) \in \mathbb{R}$ such that for any $\rho, \rho' \in \mathcal{P}_2$,

$$\left.\frac{\mathrm{d}F(\rho + \epsilon \cdot (\rho' - \rho))}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = \int \frac{\delta F}{\delta \rho}[\rho](\theta)(\rho' - \rho)(\mathrm{d}\theta),$$

for all $\rho \in \mathcal{P}_2$. If there exists a functional $\frac{\delta F}{\delta \rho}[\rho](\theta)$, we say that F is differentiable at ρ.

Note that any first variation of a functional is invariant with respect to a constant shift. In what follows, we define the proximal Gibbs distribution (PGD) with a first variation, as

Definition 2 (Proximal Gibbs Distribution (PGD))**.** Let $\rho \in \mathcal{P}_2$ and $\lambda > 0$ the temperature. We define the Gibbs distribution with potential function $-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}$ around ρ for any $\theta \in \mathbb{R}^d$ by

$$\widehat{\rho}(\theta) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}[\rho](\theta)\right).$$

We call $\widehat{\rho}(\theta)$ the proximal Gibbs distribution of the functional F around ρ.

The convergence analysis of ρ_t over the objective F heavily depends on the relationship between the PGD around ρ_t, $\widehat{\rho}_t$ and the optimal distribution ρ^*. Regarding the convergence rate of MFLDs, the key analysis is depending on the following logarithmic Sobolev inequality.

Definition 3 (Logarithmic Sobolev Inequality (LSI))**.** We define that a distribution measure $\rho \in \mathcal{P}_2$ satisfies a logarithmic Sobolev inequality with constant $\alpha > 0$, which is called $\mathrm{LSI}(\alpha)$ in short, if and only if, for any smooth function $\Psi : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\rho[\Psi^2] < \infty$, it holds that

$$\mathbb{E}_\rho[\Psi^2 \ln(\Psi^2)] - \mathbb{E}_\rho[\Psi^2] \cdot \ln(\mathbb{E}_\rho[\Psi^2]) \leq \frac{2}{\alpha}\mathbb{E}_\rho[\|\nabla \Psi\|_2^2],$$

which is equivalent to the condition that for all $\mu \in \mathcal{P}_2$ absolutely continuous w.r.t. ρ, it holds

$$\mathrm{KL}(\mu\|\rho) \leq \frac{1}{2\alpha}\mathrm{I}(\mu\|\rho).$$

In particular, the LSI holds uniformly for the PGD over the mean-field neural network condition, given some appropriate boundedness assumptions. The result is achieved by leveraging two well-known facts. Firstly, it is established that strongly log-concave densities satisfy the LSI with a dimension-free constant, up to the spectral norm of the covariance. For instance, Bakry & Émery (1985) showed the following lemma:

Lemma 4 (Bakry & Émery (1985))**.** *If $\rho \propto \exp(-f(\theta))$ is a smooth probability density with $f : \mathbb{R}^d \to \mathbb{R}$ and there exists $c > 0$ such that the Hessian matrix of f satisfies $\nabla^2 f \succeq c \cdot I_d$, then the distribution $\rho(\theta)\mathrm{d}\theta$ satisfies the LSI with constant c.*

It is worth noting that, for example, the Gaussian distribution $\nu \sim \mathcal{N}(0, I_d)$ satisfies Lemma 4 with the LSI constant $c = 1$. That is, $\nu \sim \mathcal{N}(0, I_d)$ satisfies $\mathrm{LSI}(1)$. In addition to that, preservation of LSI under bounded perturbation has been demonstrated in Holley & Stroock (1987) as

Lemma 5 (Holley & Stroock (1987))**.** *If ρ is a distribution on $\mathbb{R}^d$ that satisfies the LSI with constant $c > 0$, and for a bounded function $f : \mathbb{R}^d \to \mathbb{R}$, the distribution ρ_f is defined as*

$$\rho_f(\theta) \propto \exp(f(\theta)) \cdot \rho(\theta),$$

then ρ_f satisfies the LSI with a constant $c/\exp(4|f|_\infty)$.

Combined with the previous example of Lemma 4, ν_f with some uniformly bounded potential function f satisfies Lemma 5. These lemmas lead to the important fact that follows. Under the definition of the two-layer neural network in mean-field regime and Assumption 1, the PGD of each function appearing in this paper satisfies the LSI with an absolute constant α. Specifically, we have

Proposition 2 (LSI Constant of PGD)**.** *Let the first-variation of a function L, $\frac{\delta L}{\delta \rho}$ be uniformly bound by $C > 0$, and $F = L + \frac{\lambda}{2} \cdot \mathbb{E}_\rho[\|\theta\|_2^2]$ with $\lambda > 0$. Then we have that the PGD around ρ, $\widehat{\rho}$ satisfies the LSI with a constant $\alpha = \frac{1}{\exp(\frac{4C}{\lambda})}$.*

In our case, the boundness of each first-variation is guaranteed by the neural network's boundness in Assumption 1 and the reward's boundness in Assumption 2. It is worth noting that the exponential dependence on the LSI constant may be inevitable in the most general setting (Menz & Schlichting, 2014).

Definition 4 (Inequalities of Poincaré and Talagrand)**.** Consider a probability measure ρ within the space $\mathcal{P}_2$. It fulfills the Poincaré inequality with a constant $\alpha > 0$ if, for any function $f : \mathbb{R}^d \to \mathbb{R}$ that is smoothly defined, the variance under ρ is given by

$$\mathrm{Var}_\rho(f) = \mathbb{E}_\rho[f^2] - (\mathbb{E}_\rho[f])^2 \leq \frac{1}{\alpha}\mathbb{E}_\rho[\|\nabla f\|_2^2].$$

Additionally, ρ adheres to Talagrand's inequality with a constant $\alpha > 0$ when, for every measure $\mu \in \mathcal{P}_2$ that is absolutely continuous relative to ρ, the squared 2-Wasserstein distance between μ and ρ is constrained by

$$\frac{\alpha}{2} W_2^2(\mu, \rho) \leq \mathrm{KL}(\mu \| \rho).$$

If ν is in compliance with the Logarithmic Sobolev Inequality (LSI) characterized by a constant α, then it inherently satisfies the Poincaré inequality with the identical constant.

B.2. On the Function Class

In Assumption 3, we considered the class of measures with the bounded KL divergence and some regularity condition. We first note that, as we let M and R large, then $\mathcal{F}_{R,M}$ contain a wider class of neural networks. What is worth mentioning is the relation to the so-called Barron class. As we increase M and R, we can approximate a neural network in the Barron class with arbitrary accuracy.

Barron (1993; 1994) showed that a neural network with a sigmoid activation function can avoid the curse of dimensionality (Weinan et al., 2019) if the Fourier transform of the function f satisfies certain integrability conditions, and he defined a function class with good properties that can be approximated universally (Barron, 1993; Pinkus, 1999). Particularly, we name the function class as the Barron class and denote it as $\mathcal{B}_\mathcal{F}$, such that

$$\int_{\mathbb{C}^d} \|\omega\|_1^2 \cdot |\widehat{f}(\omega)| \mathrm{d}\omega < \infty,$$

where $\widehat{f}$ is the Fourier transform of a function f.

One pleasant aspect of considering the Barron class is that one of the biggest contributions of feature learning, the avoidance of the curse of dimensionality inherent in neural networks, theoretically arises (Weinan et al., 2019). The Barron class is also closely related to the avoidance of the curse of dimensionality in other function spaces such as in the mixed Besov space (Suzuki, 2019).

A similar analysis of function classes has been developed (Klusowski & Barron, 2016; E et al., 2019) and, in particular, the following derivations of the Barron class are known:

Definition 5 (Barron Class (Li et al., 2020))**.** The Barron class is defined as

$$\mathcal{B}_\infty = \left\{ \int_{\mathbb{R}^d} \beta(w) \cdot \sigma(w^\top (x,1)) \cdot \rho(\mathrm{d}w) \ : \ \rho \in \mathscr{P}(\mathbb{R}^d), \ \inf_\rho \|\beta(w) \cdot (\|w\|_1 + 1)\|_{\rho,\infty} < \infty \right\},$$

The Barron norm for any $f \in \mathcal{B}_\infty$ is defined by $\|f\|_{\mathcal{B}_\infty} = \inf_\rho \|\beta(w) \cdot (\|w\|_1 + 1)\|_{\rho,\infty}$. In addition, we define the R-Barron space by $R' > 0$ by

$$\mathcal{B}_{R'} = \{ f \in \mathcal{B}_\infty \ : \ \|f\|_{\mathcal{B}_\infty} \leq R' \}.$$

Note that the R-Barron space $\mathcal{B}_{R'}$ corresponds to the function class $\mathcal{F}_{R,M}$ targeted by our neural network. That is, our function class can approximate an element of the Barron class with any degree of accuracy as a set. Although the R-Barron space and the Barron class cannot be directly compared, they are closely related and can be adequately covered by a sufficiently large R'.

Finally, we remark that based on Assumption 1, which guarantees the smoothness and boundedness of neural networks, it is very easy for such a network to satisfy Eq. (7) with some R and M at least with a small modification. For any such neural network, if we consider convolution of the corresponding measure with a Gaussian of small variance, this does not change the output of the network very much due to the smoothness and boundedness, this smoothens the distribution and as a result, guarantees that the modified neural network belongs to our class of measures.

B.3. More Related Works on Feature Learning: MF vs. NTK

(1) The performance difference due to representational capabilities In NTK scaling, training relies on a fixed set of features, limiting expressiveness to the network's width. In contrast, MF scaling allows feature bases in the first layer to dynamically adjust, enhancing adaptability to the training data. For instance, representing specific polynomials in a high-dimensional space requires exponentially more neurons in NTK scaling, as initial layer weights may not align with key dimensions. However, in MF scaling, the network's width need not increase with the dimensionality of the data, sidestepping this optimization issue, as discussed by Damian et al. (2022). This adaptability gives MF networks a superior representation ability over finite-width NTK networks, which struggle to approximate common learning theory functions.

(2) Sample complexity with respect to the dimension. The MF network with infinite width also has an advantage over the NTK network with infinite width in terms of sample complexity in high-dimensional problems. If you look at the learned network in MF, the first layer parameters typically align with the important directions. But not in NTK, because NTK weight is not allowed to travel so much. As a result, the MF learned network typically has low dimensionality and, therefore, requires fewer samples. Specific examples are parity function (Telgarsky, 2023), polynomials of few relevant dimensions (Damian et al., 2022), low dimensionality (Mousavi-Hosseini et al., 2023), hierarchical functions (Abbe et al., 2023), and single-index models(Ba et al., 2022). For example, in order to learn the 2-parity in a d-dimensional space, $\mathcal{O}(d^2/\epsilon)$ sample is required for NTK but $\mathcal{O}(d/\epsilon)$ for MF (Telgarsky, 2023) to achieve the accuracy. However, the dimension dependency will be absorbed by just looking at T or T_{TD}. This is not because MF has less (or more) representation ability and the whole hypothesis class is smaller (or larger) but because the parameters can gain the aligned structure as a result of training.

(3) Reinforcement learning unique feature learning trends. Feature learning is very useful in the context of RL. In particular, RL agents often encounter state-action samples possessing low-dimensional structures. For instance, when training a reinforcement learning model to control a robotic arm, the feasible states are not uniformly distributed across the entire space but are constrained to specific subspaces due to their mechanical limitations. Additionally, in cases where images are used as state inputs, the learning data can often be projected onto lower-dimensional spaces. This reduction

in dimensionality plays a crucial role not only in the implicit learning within NN but is also explicitly addressed in some recent studies (Dayan, 1993; Yang et al., 2021; Huang et al., 2022; Le Lan et al., 2022). In response to these research trends, recent studies (Zhang et al., 2020; 2021) confront the implicit representation learning optimization in neural networks through the utilization of Lazy training. On the other hand, our study aims to revisit the significance of MF networks in this context. Through our research, we expect to demonstrate how the efficiency of representation learning, already established in supervised learning paradigms as mentioned earlier, can similarly benefit reinforcement learning.

C. Mean-field Langevin TD Learning

C.1. Proof of Lemma 1

Proof. From the definition of $L_l[\cdot]$ in Eq. (6), for $s \in [0, T_{\mathrm{TD}}]$ we have

$$\begin{aligned}\mathcal{L}_l[q] =& L_l[q] + \lambda_{\mathrm{TD}} \cdot \mathrm{Ent}[q] \\ =& \mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q_q - Q_\pi)] + \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_q)^2] + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q\|\nu),\end{aligned} \tag{13}$$

where we ignore a constant without loss of generality. The inner algorithm performs a gradient descent of $\mathcal{L}_l$ over Wasserstein metric. We evaluate the difference of the objective function L_l between the optimum of the true objective function q_π, and $q^{(l+1)}$ which is ideally the optimum of the majorization problem, from above and below, respectively. For $l \in \mathbb{N}$ we have

$$\begin{aligned}\mathcal{L}_l[q_\pi] - \mathcal{L}_l[q^{(l+1)}] =& -\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] + \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] \\ & - \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q^{(l+1)})^2] + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi\|\nu) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q^{(l+1)}\|\nu).\end{aligned} \tag{14}$$

In what follows, we upper bound the first term on the right-hand side of Eq. (14) by each difference of Q-functions without any transition kernels. For simplicity, we define that $\Delta Q^{(l)} = Q^{(l)} - Q_\pi$, I is an identity operator, and $\mathcal{P} : L^2(\varsigma_\pi)(\mathcal{S} \times \mathcal{A}) \to L^2(\varsigma_\pi)(\mathcal{S} \times \mathcal{A})$ as the linear operator such that $\mathcal{P}Q(s,a) = \int \mathrm{d}s' P(s'|s,a) \int \mathrm{d}a' \pi(a'|s') Q(s,a)$, $Q \in L^2(\varsigma_\pi)(\mathcal{S} \times \mathcal{A})$. Focusing on the fact that we can reformulate the first term on the right-hand side of Eq. (14) as $\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] = \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l)}]$, it holds that

$$\begin{aligned}&\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)} - \Delta Q^{(l)})(I - \gamma\mathcal{P})(\Delta Q^{(l+1)} - \Delta Q^{(l)})] \\ &=\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l+1)}] + \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I - \gamma\mathcal{P})\Delta Q^{(l)}] \\ &\quad - \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l)}] - \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I - \gamma\mathcal{P})\Delta Q^{(l+1)}] \\ &=\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l+1)}] + \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I - \gamma\mathcal{P})\Delta Q^{(l)}] \\ &\quad + \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I - \gamma\mathcal{P}^*)\Delta Q^{(l+1)}] - \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I - \gamma\mathcal{P})\Delta Q^{(l+1)}] \\ &\quad - 2\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l)}] \\ &=\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l+1)}] + \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I - \gamma\mathcal{P})\Delta Q^{(l)}] \\ &\quad + \gamma \cdot \mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(\mathcal{P}^* - \mathcal{P})\Delta Q^{(l)}] - 2\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I - \gamma\mathcal{P})\Delta Q^{(l)}],\end{aligned} \tag{15}$$

where $\mathcal{P}^*$ is the adjoint operator of $\mathcal{P}$. As for each term of Eq. (15), we have the following inequalities:

$$\begin{aligned}&\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)} - \Delta Q^{(l)})(I - \gamma\mathcal{P})(\Delta Q^{(l+1)} - \Delta Q^{(l)})] \\ &=\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})(I - \gamma\mathcal{P})(Q^{(l+1)} - Q^{(l)})] \\ &=\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2] - \gamma \cdot \mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)}) \cdot \mathcal{P}(Q^{(l+1)} - Q^{(l)})] \\ &\le\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2] + \gamma \cdot \mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2]^{1/2} \cdot \|\mathcal{P}(Q^{(l+1)} - Q^{(l)})\|_{\varsigma_\pi,2} \\ &\le\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2] + \gamma \cdot \mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2]^{1/2} \cdot \mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2]^{1/2} \\ &=(1+\gamma) \cdot \mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)})^2],\end{aligned} \tag{16}$$

where the first inequality follows from Hölder's inequality and the second one follows from Lemma 11. In exactly the same way, we have

$$\begin{aligned}
&-\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I-\gamma\mathcal{P})\Delta Q^{(l+1)}]\\
&\quad=-\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]+\gamma\cdot\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}\cdot\mathcal{P}\Delta Q^{(l+1)}]\\
&\quad\leq-\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]+\gamma\cdot\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]^{1/2}\cdot\|\mathcal{P}(\Delta Q^{(l+1)})\|_{\varsigma_\pi,2}\\
&\quad\leq-\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]+\gamma\cdot\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]^{1/2}\cdot\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]^{1/2}\\
&\quad=-(1-\gamma)\cdot\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2],
\end{aligned}\tag{17}$$

where the first inequality follows from Hölder's inequality and the second one follows from Lemma 11. From the same discussions, we also have $-\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I-\gamma\mathcal{P})\Delta Q^{(l)}]\leq-(1-\gamma)\cdot\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l)})^2]$. In addition, from the fact that $\mathbb{E}_{\varsigma_\pi}[Q(\mathcal{P}^*-\mathcal{P})Q]=0$ for all $Q\in L^2(\varsigma_\pi)(\mathcal{S}\times\mathcal{A})$, it holds that

$$\begin{aligned}
&-\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(\mathcal{P}^*-\mathcal{P})\Delta Q^{(l)}]\\
&\quad=-\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(\mathcal{P}^*-\mathcal{P})\Delta Q^{(l)}]+\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(\mathcal{P}^*-\mathcal{P})\Delta Q^{(l+1)}]\\
&\quad=\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(\mathcal{P}^*-\mathcal{P})(Q^{(l+1)}-Q^{(l)})]\\
&\quad\leq\frac{1}{2}\cdot\frac{2(1-\gamma)}{\gamma}\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]+\frac{1}{2}\cdot\frac{\gamma}{2(1-\gamma)}\|(\mathcal{P}^*-\mathcal{P})(Q^{(l+1)}-Q^{(l)})\|_{\varsigma_\pi,2}^2\\
&\quad\leq\frac{1-\gamma}{\gamma}\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]+\frac{\gamma}{1-\gamma}\|\mathcal{P}(Q^{(l+1)}-Q^{(l)})\|_{\varsigma_\pi,2}^2\\
&\quad\leq\frac{1-\gamma}{\gamma}\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)})^2]+\frac{\gamma}{1-\gamma}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)}-Q^{(l)})^2],
\end{aligned}\tag{18}$$

where the first inequality follows from Young's inequality with an arbitrary constant $\frac{2(1-\gamma)}{\gamma}>0$, and the last one follows from Lemma 11. Combining Eq. (16), (17), (18), and (15), we obtain that

$$\begin{aligned}
&-\mathbb{E}_{\varsigma_\pi}[(Q^{(l)}-\mathcal{T}Q^{(l)})\cdot(Q^{(l+1)}-Q_\pi)]\\
&\quad=-\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I-\gamma\mathcal{P})\Delta Q^{(l)}]\\
&\quad=-\frac{1}{2}\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(I-\gamma\mathcal{P})\Delta Q^{(l+1)}]-\frac{1}{2}\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l)}(I-\gamma\mathcal{P})\Delta Q^{(l)}]\\
&\qquad-\frac{\gamma}{2}\mathbb{E}_{\varsigma_\pi}[\Delta Q^{(l+1)}(\mathcal{P}^*-\mathcal{P})\Delta Q^{(l)}]\\
&\qquad+\frac{1}{2}\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l+1)}-\Delta Q^{(l)})(I-\gamma\mathcal{P})(\Delta Q^{(l+1)}-\Delta Q^{(l)})]\\
&\quad\leq-\frac{1-\gamma}{2}\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l)})^2]+\frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)}-Q^{(l)})^2].
\end{aligned}\tag{19}$$

Plugging Eq. (19) into Eq. (14), we obtain the following upper-bound of Eq. (14) as

$$\begin{aligned}
\mathcal{L}_l[q_\pi]-\mathcal{L}_l[q^{(l+1)}]\leq&\frac{\gamma(2-\gamma)}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(l)})^2]\\
&+\lambda_{\mathrm{TD}}\cdot\mathrm{KL}(q_\pi\|\nu)-\lambda_{\mathrm{TD}}\cdot\mathrm{KL}(q^{(l+1)}\|\nu),
\end{aligned}\tag{20}$$

In what follows, we give a lower bound on the difference for majorization objectives using the strong convexity of L_l. From the definition of $\mathcal{L}_l$ in Eq. (13), it holds that

$$\frac{\delta\mathcal{L}_l}{\delta q}[q](\omega)=\mathbb{E}_{\varsigma_\pi}[(Q^{(l)}-\mathcal{T}Q^{(l)})\cdot h_\omega]+\frac{1}{1-\gamma}\mathbb{E}_{\varsigma_\pi}[(Q_q-Q^{(l)})\cdot h_\omega]+\lambda_{\mathrm{TD}}\cdot\ln\frac{q}{\nu}.\tag{21}$$

In what we follow, we control the error induced by the difference between the last iterate of inner-loop dynamics, $q^{(l+1)}$, and the optimal distribution of $\mathcal{L}^{(l)}$, $q_*^{(l+1)}$.It holds from Eq. (21) that

$$\begin{aligned}
&-\int \frac{\delta \mathcal{L}_l}{\delta q}[q^{(l+1)}](\mathrm{d}q_\pi - \mathrm{d}q^{(l+1)}) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q^{(l+1)}) \\
=&\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] + \frac{1}{1-\gamma}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] \\
&- \lambda_{\mathrm{TD}} \cdot \int \ln \frac{q^{(l+1)}}{\nu}(\mathrm{d}q_\pi - \mathrm{d}q^{(l+1)}) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q^{(l+1)}) \\
=&\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] + \frac{1}{1-\gamma}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] \\
&- \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| \nu) + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q^{(l+1)} \| \nu).
\end{aligned} \tag{22}$$

Plugging Eq. (22) into Eq. (14), we have

$$\begin{aligned}
\mathcal{L}_l[q_\pi] - \mathcal{L}_l[q^{(l+1)}] =& -\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] \\
&+ \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] - \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q^{(l+1)})^2] \\
&+ \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| \nu) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q^{(l+1)} \| \nu) \\
=&\frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] - \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q^{(l+1)})^2] \\
&+ \frac{1}{1-\gamma} \cdot \mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q^{(l)}) \cdot (Q^{(l+1)} - Q_\pi)] \\
&+ \int \frac{\delta \mathcal{L}_l}{\delta q}[q^{(l+1)}](\mathrm{d}q_\pi - \mathrm{d}q^{(l+1)}) + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q^{(l+1)}) \\
=&\frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q_\pi)^2] \\
&+ \int \frac{\delta \mathcal{L}_l}{\delta q}[q^{(l+1)}](\mathrm{d}q_\pi - \mathrm{d}q^{(l+1)}) + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q^{(l+1)})
\end{aligned}. \tag{23}$$

We evaluate the inner-loop error bound in the sequel by establishing Lemma 6.

Lemma 6 (Inner-Loop Error Bound)**.** *Under assumptions of Proposition 5, for any $l \in \mathbb{N}$, $s > 0$, we have*

$$\begin{aligned}
&-\int \frac{\delta \mathcal{L}_l}{\delta q}[q_s](\mathrm{d}q_\pi - \mathrm{d}q_s) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_s) \\
&\leq \frac{2R}{1-\gamma}\left(\mathbb{E}_{\varsigma_\pi}[(Q_{q_s} - Q_*^{(l+1)})^2]\right)^{\frac{1}{2}} + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_s \| q_*^{(l+1)}).
\end{aligned}$$

Proof. See Appendix C.3 for a detailed proof. □

Lemma 6 guarantees that the optimality error $-\int \frac{\delta \mathcal{L}_l}{\delta q}[q_s](\mathrm{d}q_\pi - \mathrm{d}q_s)$ is biased by at most the KL divergence $\lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_s)$. Recall that the inner dynamics is stopped at the time $s = S > 0$ when we set the next outer iterate $q^{(l+1)}$ as that $q^{(l+1)} = q_S$. Combining Eq. (20), Eq. (23), and Lemma 6, we finish the proof of Lemma 1. □

C.2. Proof of Theorem 1

Proof. Before jumping to the proof of Theorem 1, we evaluate the mean-squared error between the Q-function Q_{q_s} induced by q_s and the global optimal Q-function $q_*^{(l+1)} = Q_{q^*}$ over the L^2-norm. In the sequel, we provide the following convergence lemma about the mean squared error of Q-functions.

Lemma 7 (Linear Convergence of the Mean Squared Error of Q-functions)**.** *Under the same assumption of Theorem 1, for $l \in \mathbb{N}$ we have*

$$\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q_*^{(l+1)})^2] \leq \frac{4(3-2\gamma)R^4}{(1-\gamma)\lambda_{\mathrm{TD}}} \cdot \exp(-2\alpha\lambda_{\mathrm{TD}} S).$$

where we denote $Q_*^{(l+1)} = Q_{q_*^{(l)}}$ with the global optimal distribution $q_*^{(l+1)}$ of the inner objective $\mathcal{L}_l$ and the definition of each variable follows that of Proposition 5.

Proof. For any parameter distributions $q, q' \in \mathcal{P}_2$, we first upper bound the Q-function difference with the Wasserstein distance.

$$\begin{aligned}(Q_q(x) - Q_{q'}(x))^2 =& \left(\int h_\omega(x)(\mathrm{d}q(\omega) - \mathrm{d}q'(\omega))\right)^2 \\ \leq& R^2 \cdot \|q - q'\|_1^2 \\ \leq& 2R^2 \cdot \mathrm{KL}(q\|q'),\end{aligned} \tag{24}$$

where $R > 0$ is an absolute constant defined in Assumption 1 and the last inequality follows from Pinsker's inequality. Combining eq. (24) and Proposition 5, we obtain that

$$\begin{aligned}\mathbb{E}_{\varsigma_\pi}[(Q_{q_s} - Q_*^{(l+1)})^2] \leq& 2R^2 \cdot \mathrm{KL}(q_s\|q_*^{(l+1)}) \\ =& \frac{2R^2}{\lambda_{\mathrm{TD}}} \exp\left(-2\alpha\lambda_{\mathrm{TD}} s\right) \cdot \left(\mathcal{L}_l[q_0] - \mathcal{L}_l[q_*^{(l+1)}]\right).\end{aligned}$$

To control the right-hand side of the inequality, we evaluate the objective difference $\mathcal{L}_l[q_0] - \mathcal{L}_l[q_*^{(l+1)}]$. It holds from Assumption 1, 2, that

$$\begin{aligned}\mathcal{L}_l[q_0] - \mathcal{L}_l[q_*^{(l+1)}] =& \mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}Q^{(l)}) \cdot (Q_0 - Q_*^{(l+1)})] \\ &+ \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_0)^2] + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_0\|\nu) \\ &- \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_*^{(l+1)})^2] - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_*^{(l+1)}\|\nu) \\ \leq& 4R^2 + \frac{1}{2(1-\gamma)}4R^2 \\ =& \frac{2(3-2\gamma)R^2}{1-\gamma},\end{aligned}$$

where we use $\mathrm{KL}(q_0\|\nu) = 0$. We conclude the proof of Lemma 7. □

By Lemma 1, we have

$$\begin{aligned}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q_\pi)^2] \leq& \frac{\gamma(2-\gamma)}{(1-\gamma)^2}\mathbb{E}_{\varsigma_\pi}\left[(\Delta Q^{(l)})^2 - (\Delta Q^{(l+1)})^2\right] \\ &+ \frac{4R}{(1-\gamma)^2}(\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q_*^{(l+1)})^2])^{\frac{1}{2}} \\ &+ \frac{2\lambda_{\mathrm{TD}}}{1-\gamma} \cdot \mathrm{KL}(q^{(l+1)}\|q_*^{(l+1)}) + \frac{2\lambda_{\mathrm{TD}}}{1-\gamma}\mathrm{KL}(q_\pi\|\nu).\end{aligned}$$

Combining Lemma 7 and Proposition 5, by the same argument of the proof of Lemma 7, it holds that

$$\begin{aligned}\mathbb{E}_{\varsigma_\pi}[(Q^{(l+1)} - Q_\pi)^2] \leq& \frac{\gamma(2-\gamma)}{(1-\gamma)^2}\mathbb{E}_{\varsigma_\pi}\left[(\Delta Q^{(l)})^2 - (\Delta Q^{(l+1)})^2\right] \\ &+ \frac{8\sqrt{3}R^3}{(1-\gamma)^{\frac{5}{2}}\lambda_{\mathrm{TD}}^{\frac{1}{2}}} \cdot \exp\left(-\alpha\lambda_{\mathrm{TD}} S\right) \\ &+ \frac{24R^4}{(1-\gamma)^2} \cdot \exp\left(-2\alpha\lambda_{\mathrm{TD}} S\right) + \frac{2\lambda_{\mathrm{TD}} M}{1-\gamma}.\end{aligned} \tag{25}$$

Telescoping (25) for $s = 0, \ldots, T_{\mathrm{TD}} - 1$, we obtain

$$\begin{aligned}\frac{1}{T_{\mathrm{TD}}}\sum_{s=1}^{T_{\mathrm{TD}}}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] \leq & \frac{\gamma(2-\gamma)}{(1-\gamma)^2 T_{\mathrm{TD}}}\mathbb{E}_{\varsigma_\pi}\left[(\Delta Q^{(0)})^2 - (\Delta Q^{(T_{\mathrm{TD}})})^2\right] \\ & + \frac{8\sqrt{3}R^3}{(1-\gamma)^{\frac{5}{2}}\lambda_{\mathrm{TD}}^{\frac{1}{2}}} \cdot \exp\left(-\alpha\lambda_{\mathrm{TD}} S\right) \\ & + \frac{24R^4}{(1-\gamma)^2} \cdot \exp\left(-2\alpha\lambda_{\mathrm{TD}} S\right) + \frac{2\lambda_{\mathrm{TD}} M}{1-\gamma} \\ \leq & \frac{\gamma(2-\gamma)}{(1-\gamma)^2 T_{\mathrm{TD}}}\mathbb{E}_{\varsigma_\pi}\left[(\Delta Q^{(0)})^2\right] \\ & + \frac{8\sqrt{3}R^3}{(1-\gamma)^{\frac{5}{2}}\lambda_{\mathrm{TD}}^{\frac{1}{2}}} \cdot \exp\left(-\alpha\lambda_{\mathrm{TD}} S\right) \\ & + \frac{24R^4}{(1-\gamma)^2} \cdot \exp\left(-2\alpha\lambda_{\mathrm{TD}} S\right) + \frac{2\lambda_{\mathrm{TD}} M}{1-\gamma}.\end{aligned}$$

Recall that $Q \leq R$ for any Q-function Q from Assumption 1 and 3, we have

$$\mathbb{E}_{\varsigma_\pi}[(\Delta Q^{(0)})^2] = \mathbb{E}_{\varsigma_\pi}[(Q^{(0)} - Q_\pi)^2] \leq 4R^2.$$

Therefore, we have

$$\begin{aligned}\frac{1}{T_{\mathrm{TD}}}\sum_{s=1}^{T_{\mathrm{TD}}}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] \leq & \frac{8\gamma R^2}{(1-\gamma)^2 T_{\mathrm{TD}}} + \frac{8\sqrt{3}R^3}{(1-\gamma)^{\frac{5}{2}}\lambda_{\mathrm{TD}}^{\frac{1}{2}}} \cdot \exp\left(-\alpha\lambda_{\mathrm{TD}} S\right) \\ & + \frac{24R^4}{(1-\gamma)^2} \cdot \exp\left(-2\alpha\lambda_{\mathrm{TD}} S\right) + \frac{2\lambda_{\mathrm{TD}} M}{1-\gamma}\end{aligned}$$

which concludes the proof of Theorem 1. □

C.3. Proof of Lemma 6

Proof. We first present some lemmas on convergence properties. In specific, we prove the convergence of the parameter distribution q_s to the global optimal distribution q^* in the inner-loop MFLD and also Using the two convergence lemmas above, we evaluate the error derived from the inner-loop algorithm.

$$\begin{aligned}& -\int \frac{\delta \mathcal{L}_l}{\delta q}[q_s](\mathrm{d}q_\pi - \mathrm{d}q_s) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_s) \\ & \quad = -\int \left(\frac{\delta \mathcal{L}_l}{\delta q}[q_s] - \frac{\delta \mathcal{L}_l}{\delta q}[q_*^{(l+1)}]\right)(\mathrm{d}q_\pi - \mathrm{d}q_s) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_s),\end{aligned} \tag{26}$$

where the last equality follows from the optimal condition with the stationary point $q_*^{(l+1)}$ as

$$\frac{\delta \mathcal{L}_l}{\delta q}[q_*^{(l+1)}] = \text{const.}$$

For the first term on the right-hand side of Eq. (26), the difference of the first-variations of $\mathcal{L}_l$ satisfies from the definition in Eq. (6) that

$$\frac{\delta \mathcal{L}_l}{\delta q}[q_s] - \frac{\delta \mathcal{L}_l}{\delta q}[q_*^{(l+1)}] = \frac{1}{1-\gamma}\mathbb{E}_{\varsigma_\pi}\left[(Q_{q_s} - Q_*^{(l+1)}) \cdot h_\omega\right] + \lambda_{\mathrm{TD}} \cdot \ln \frac{q_s}{q_*^{(l+1)}}. \tag{27}$$

Plugging Eq. (27) into Eq. (26), we have

$$
\begin{aligned}
-\int \frac{\delta \mathcal{L}_l}{\delta q}[q_s](\mathrm{d}q_\pi - \mathrm{d}q_s) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_s) \\
=& \frac{1}{1-\gamma}\mathbb{E}_{\varsigma_\pi}\left[(Q_{q_s} - Q_*^{(l+1)}) \cdot (Q_{q_s} - Q_\pi)\right] \\
&- \lambda_{\mathrm{TD}} \cdot \int \ln \frac{q_s}{q_*^{(l+1)}}(\mathrm{d}q_\pi - \mathrm{d}q_s) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_s) \\
\leq& \frac{1}{1-\gamma}\mathbb{E}_{\varsigma_\pi}\left[(Q_{q_s} - Q_*^{(l+1)})^2)\right]^{\frac{1}{2}} \cdot \|Q_{q_s} - Q_\pi\|_{\varsigma_\pi,2} \\
&+ \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_s \| q_*^{(l+1)}) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_*^{(l+1)}) \\
\leq& \frac{2R}{1-\gamma}\mathbb{E}_{\varsigma_\pi}\left[(Q_{q_s} - Q_*^{(l+1)})^2)\right]^{\frac{1}{2}} + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_s \| q_*^{(l+1)}) - \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_\pi \| q_*^{(l+1)}) \\
\leq& \frac{2R}{1-\gamma}\mathbb{E}_{\varsigma_\pi}\left[(Q_{q_s} - Q_*^{(l+1)})^2)\right]^{\frac{1}{2}} + \lambda_{\mathrm{TD}} \cdot \mathrm{KL}(q_s \| q_*^{(l+1)}),
\end{aligned}
$$

where the second inequality follows from Assumption 1 and 2.

□

C.4. Proof of Corollary 1

Proof. The results of Corollary 1 naturally follows from the convergence rate of Theorem 1, which concludes that

$$
\frac{1}{T_{\mathrm{TD}}}\sum_{l=1}^{T_{\mathrm{TD}}} \mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_\pi)^2] \leq \frac{C_0}{T_{\mathrm{TD}}} + C_1 \lambda_{\mathrm{TD}}^{-\frac{1}{2}} e^{(-\alpha\lambda_{\mathrm{TD}} S)} + C_2 e^{(-2\alpha\lambda_{\mathrm{TD}} S)} + C_3 \lambda_{\mathrm{TD}}.
$$

letting $S = -\frac{3\log\lambda_{\mathrm{TD}}}{2\alpha\lambda_{\mathrm{TD}}}$ yields that

$$
\begin{aligned}
\lambda_{\mathrm{TD}}^{-\frac{1}{2}} e^{-\alpha\lambda_{\mathrm{TD}} S} &= \lambda_{\mathrm{TD}}, \\
e^{-2\alpha\lambda_{\mathrm{TD}} S} &= \lambda_{\mathrm{TD}}^3.
\end{aligned}
$$

Since we set the regularization parameter for the inner loop as $\lambda_{\mathrm{TD}} = \frac{1}{T_{\mathrm{TD}}}$, sampling $L \in [T_{\mathrm{TD}}]$ uniformly, it holds that $\mathbb{E}_{L,\varsigma_\pi}[(Q^{(L)} - Q_\pi)^2] = \mathcal{O}(T_{\mathrm{TD}}^{-1} \wedge T_{\mathrm{TD}}^{-3})$. □

D. Mean-field Langevin Policy Gradient

D.1. Proof of Proposition 1

Proof. By Proposition 4, we have for all π_ρ that

$$
\begin{aligned}
\mathrm{d}J[\rho] =& \mathbb{E}_{\nu_{\pi_\rho}}\left[\int \mathrm{d}\pi_\rho(\mathrm{d}a) \cdot Q_{\pi_\rho}(a)\right] \\
=& \mathbb{E}_{\nu_{\pi_\rho}}\left[\int \left(-\mathrm{d}f_\rho(a) + \int \pi_\rho(\mathrm{d}a')\mathrm{d}f_\rho(a')\right)\pi_\rho(\mathrm{d}a) \cdot Q_{\pi_\rho}(a)\right] \\
=& \mathbb{E}_{\nu_{\pi_\rho}}\left[-\int \pi_\rho(\mathrm{d}a)\mathrm{d}f_\rho(a) \cdot Q_{\pi_\rho}(a) + \left(\int \pi_\rho(\mathrm{d}a')\mathrm{d}f_\rho(a')\right) \cdot \left(\int \pi_\rho(\mathrm{d}a) Q_{\pi_\rho}(a)\right)\right] \\
=& -\mathbb{E}_{\nu_{\pi_\rho}}\left[\int \pi_\rho(\mathrm{d}a)\mathrm{d}f_\rho(a) \cdot \left(Q_{\pi_\rho}(a) - \int \pi_\rho(\mathrm{d}a') Q_{\pi_\rho}(a')\right)\right] \\
=& -\mathbb{E}_{\sigma_{\pi_\rho}}\left[\mathrm{d}f_\rho \cdot A_{\pi_\rho}\right] \\
=& \int \mathrm{d}\rho(\mathrm{d}\theta)\mathbb{E}_{\sigma_{\pi_\rho}}\left[-h_\theta \cdot A_{\pi_\rho}\right].
\end{aligned}
$$

In addition, we have

$$\mathrm{d}(\mathbb{E}_\rho[\|\theta\|^2]) = \int \mathrm{d}\rho(\mathrm{d}\theta)\|\theta\|^2.$$

Recalling $F[\rho] = -J[\rho] + \frac{\lambda}{2}\mathbb{E}_\rho[\|\theta\|^2]$, we obtain that

$$\mathrm{d}F[\rho] = -\,\mathrm{d}J[\rho] + \frac{\lambda}{2}\mathrm{d}(\mathbb{E}_\rho[\|\theta\|^2]) = \int \mathrm{d}\rho(\mathrm{d}\theta)\left(\mathbb{E}_{\sigma_{\pi_\rho}}\left[h_\theta \cdot A_{\pi_\rho}\right] + \frac{\lambda}{2}\|\theta\|^2\right). \tag{28}$$

From the definition of the first-variation of $F[\rho]$ in Definition 1, it holds that

$$\mathrm{d}F[\rho] = \int \mathrm{d}\rho(\mathrm{d}\theta)\frac{\delta F}{\delta \rho}[\rho]. \tag{29}$$

Comparing Eq. (28) and Eq. (29), we obtain Eq. (2). □

D.2. Proof of Lemma 2

Proof. First of all, we define the proximal Gibbs distribution of $\mathcal{F}$ around ρ_t by $\widehat{\rho}_t \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}[\rho_t]\right)$. We can obtain the time derivative of $\mathcal{F}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] = \int \frac{\delta \mathcal{F}}{\delta \rho}[\rho_t]\partial_t \rho_t(\mathrm{d}\theta). \tag{30}$$

Since we have $\mathcal{F}[\rho] = F[\rho] + \lambda \cdot \mathrm{Ent}[\rho]$, it holds that

$$\begin{aligned}\frac{\delta \mathcal{F}}{\delta \rho}[\rho] =& \frac{\delta F}{\delta \rho}[\rho] + \lambda \cdot \ln \rho \\ =& -\lambda \cdot \ln \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}[\rho]\right) + \lambda \cdot \ln \rho \\ =& \lambda \cdot \ln \frac{\rho}{\widehat{\rho}} - \lambda \cdot \ln Z_\lambda,\end{aligned} \tag{31}$$

where we define by $Z_\lambda > 0$ the normalization constant of $\widehat{\rho}_t \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}[\rho_t]\right)$, i.e., we have

$$Z_\lambda = \int \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta \rho}[\rho_t]\right)\mathrm{d}\theta.$$

We note that the first variance can ignore the shift of constants. On the other hand, from the definitions we have the following Fokker-Planck equation about the time evolution of ρ_t:

$$\begin{aligned}\partial_t \rho_t =& \lambda \cdot \Delta \rho_t + \nabla \cdot \left(\rho_t \cdot \nabla \widetilde{\frac{\delta \mathcal{F}}{\delta \rho}}[\rho_t]\right) \\ =& \lambda \cdot \nabla\left(\rho_t \cdot \nabla \ln \frac{\rho_t}{\widehat{\rho}_t}\right) + \nabla \cdot \left(\rho_t \cdot \left(\nabla \widetilde{\frac{\delta \mathcal{F}}{\delta \rho}}[\rho_t] - \nabla \frac{\delta \mathcal{F}}{\delta \rho}[\rho_t]\right)\right).\end{aligned} \tag{32}$$

Plugging Eq. (31) and (32) into Eq. (30), it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] \leq -\frac{\lambda^2}{2}\int \rho_t(\mathrm{d}\theta)\left\|\nabla \ln \frac{\rho_t}{\widehat{\rho}_t}\right\|_2^2 + \frac{1}{2}\int \rho_t(\mathrm{d}\theta)\left(\left\|\nabla \widetilde{\frac{\delta \mathcal{F}}{\delta \rho}}[\rho_t] - \nabla \frac{\delta \mathcal{F}}{\delta \rho}[\rho_t]\right\|_2^2\right). \tag{33}$$

For the first term on the right-hand side of Eq. (33), it holds from the LSI of $\widehat{\rho}_t$ with the LSI constant $\alpha > 0$ that

$$\begin{aligned}-\frac{\lambda^2}{2}\int \rho_t(\mathrm{d}\theta)\left\|\nabla \ln \frac{\rho_t}{\widehat{\rho}_t}\right\|_2^2 =& -\frac{\lambda^2}{2}\mathrm{I}(\rho_t||\widehat{\rho}_t) \\ \leq& -\alpha\lambda^2 \cdot \mathrm{KL}(\rho_t||\widehat{\rho}_t).\end{aligned}$$

Note that α depends on λ at the order $\mathcal{O}(\exp(-1/\lambda))$. See Proposition 2 for the detail of the construction of the LSI constant.

In the sequel, to bound the second term on the right-hand side of Eq. (33),

Lemma 8. *Under Assumption 1, and 4-(I), it holds for any $\theta \in \mathbb{R}^d$ that*

$$\int \rho_t(\mathrm{d}\theta) \left\| \nabla \widetilde{\frac{\delta \mathcal{F}}{\delta \rho}}[\rho_t](\theta) - \nabla \frac{\delta \mathcal{F}}{\delta \rho}[\rho_t](\theta) \right\|^2 \leq 4R^2(L_1+L_3)^2 \iota^2 \, \mathbb{E}_{\varsigma_t} \left[(Q_t - Q_{\pi_t})^2 \right].$$

Proof. We obtain the difference of advantage functions as

$$\begin{aligned}
\nabla \widetilde{\frac{\delta \mathcal{F}}{\delta \rho}}[\rho_t] - \nabla \frac{\delta \mathcal{F}}{\delta \rho}[\rho_t] &= \mathbb{E}_{\sigma_t}[\nabla h_\theta(s,a) \cdot (A_t(s,a) - A_{\pi_t}(s,a))], \\
&= \mathbb{E}_{\sigma_t}\left[\nabla h_\theta(s,a) \cdot (Q_t(s,a) - Q_{\pi_t}(s,a))\right] \\
&\quad - \mathbb{E}_{\sigma_t}\left[\nabla h_\theta(s,a) \cdot \left(\int \pi_t(\mathrm{d}a'|s) Q_t(s,a') - V_{\pi_t}(s) \right)\right],
\end{aligned}$$

where we denote by $A_t(s,a) = Q_t(s,a) - \int \pi_t(\mathrm{d}a') Q_t(s,a')$ the advantage function estimator with the Q-function given by the critic at time t. The second term in the right-hand side can be transformed as

$$\begin{aligned}
\mathbb{E}_{\sigma_t}\left[\nabla h_\theta(a) \cdot \left(\int \pi_t(\mathrm{d}a') Q_t(a') - V_{\pi_t} \right)\right] &= \mathbb{E}_{\nu_t}\left[\int \pi_t(\mathrm{d}a) \nabla h_\theta(a) \cdot \left(\int \pi_t(\mathrm{d}a') \left(Q_t(a') - Q_{\pi_t}(a')\right) \right)\right] \\
&= \mathbb{E}_{\nu_t}\left[\int \pi_t(\mathrm{d}a) \left(Q_t(a) - Q_{\pi_t}(a)\right) \cdot \left(\int \pi_t(\mathrm{d}a') \nabla h_\theta(a') \right)\right] \\
&= \mathbb{E}_{\sigma_t}\left[\left(\int \pi_t(\mathrm{d}a') \nabla h_\theta(a') \right) \cdot \left(Q_t(a) - Q_{\pi_t}(a)\right)\right],
\end{aligned}$$

where we exchange a with a' in the second equality. Since the neural network h_θ is assumed to be $R(L_1+L_3)$-Lipschitz continuous in Assumption 1, we can evaluate the squared expectation as

$$\begin{aligned}
\int \rho_t(\mathrm{d}\theta) \left\| \nabla \widetilde{\frac{\delta \mathcal{F}}{\delta \rho}}[\rho_t] - \nabla \frac{\delta \mathcal{F}}{\delta \rho}[\rho_t] \right\|^2 &= \int \rho_t(\mathrm{d}\theta) \left\| \mathbb{E}_{\sigma_t}\left[\left(\nabla h_\theta - \int \pi_t(\mathrm{d}a') \nabla h_\theta \right) \cdot (Q_t - Q_{\pi_t}) \right] \right\|^2 \\
&\leq 4R^2(L_1+L_3)^2 \mathbb{E}_{\sigma_t}\left[|Q_t - Q_{\pi_t}| \right]^2,
\end{aligned}$$

where the inequality follows from Hölder's inequality. Assumption 4-(i) bounds the moment of the Radon-Nikodim derivative, it holds that

$$\begin{aligned}
\mathbb{E}_{\sigma_t}\left[|Q_t - Q_{\pi_t}| \right]^2 &= \mathbb{E}_{\varsigma_t}\left[\frac{\mathrm{d}\sigma_t}{\mathrm{d}\varsigma_t} \cdot |Q_t - Q_{\pi_t}| \right]^2 \\
&\leq \left\| \frac{\mathrm{d}\sigma_t}{\mathrm{d}\varsigma_t} \right\|_{\varsigma_t, 2}^2 \mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2 \right] \\
&\leq \iota^2 \mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2 \right],
\end{aligned}$$

where $\frac{\mathrm{d}\sigma_t}{\mathrm{d}\varsigma_t}$ is the Radon-Nikodim derivative between the state-action visitation measure σ_t and the stationary state-action distribution ς_t corresponding to the same policy π_t, and the first inequality follow from Hölder's inequality. Combining the inequalities yields the proof of Lemma 8. □

Combining all of them, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}[\rho_t] \leq -\alpha \lambda^2 \cdot \mathrm{KL}(\rho_t || \widehat{\rho}_t) + 2R^2(L_1+L_3)^2 \iota^2 \, \mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2 \right],$$

which concludes the proof of Lemma 2. □

D.3. Proof of Lemma 3

Proof. It is well known that the expected total reward function $J[\pi]$ has non-convexity, which makes the optimization of the expected total rewards much more difficult. To access the problem, we make use of a proposition to prove the one-point convexity of $J[\pi]$ at the global optimum π^*. This proposition is established by Kakade & Langford (2002).

Proposition 3 (Expected Total Rewards Difference (Kakade & Langford, 2002))**.** *For all π, π', it holds that*

$$(1-\gamma)\cdot(J[\pi'] - J[\pi]) = \mathbb{E}_{\sigma_{\pi'}}\left[A_\pi\right]$$

where $\sigma_{\pi'}$ and $\nu_{\pi'}$ are the state-action visitation measure and the state visitation measure induced by policy π', respectively.

In our analysis, we utilize Proposition 3 as a one-point convexity of the expected total rewards to prove the global optimality of the stationary point of the MFLPG. In specific, we first evaluate the left-hand side of Eq. (10), the performance difference. Let $\pi^* = \arg\max J_\pi$ be the globally optimal policy of the expected total reward function J and further define the globally optimal expected total reward by $J^* = J_{\pi_*}$. By Proposition 3, it holds for any $t \in \mathbb{R}_{\geq 0}$ that

$$J^* - J[\rho_t] = (1-\gamma)^{-1}\mathbb{E}_{\sigma_*}\left[A_{\pi_t}\right] = (1-\gamma)^{-1}\mathbb{E}_{\sigma_t}\left[\frac{\mathrm{d}\sigma_*}{\mathrm{d}\sigma_t}\cdot A_{\pi_t}\right], \tag{34}$$

where $\mathbb{E}_{\sigma_*}[\cdot] = \mathbb{E}_{\sigma_{\pi^*}}[\cdot]$, hereafter.

On the other hand, we evaluate the first term on the right-hand side of Eq. (10) as

$$\begin{aligned}\lambda\,\mathrm{KL}(\rho_t||\widehat{\rho}_t) =& \lambda\int\rho_t(\mathrm{d}\theta)\ln\frac{\rho_t}{\widehat{\rho}_t}\\ =& -\lambda\int\ln\frac{\rho_t}{\widehat{\rho}_t}\left(\frac{1}{\sqrt{\lambda}}\widehat{\rho}_t - \rho_t\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\widehat{\rho}_t\|\rho_t\right).\end{aligned}$$

To bound the right-hand side, consider the following functional for any $\rho \in \mathcal{P}_2$:

$$-\lambda\int\ln\frac{\rho_t}{\widehat{\rho}_t}\left(\frac{1}{\sqrt{\lambda}}\rho - \rho_t\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\|\rho_t\right) = -\sqrt{\lambda}\,\mathrm{KL}(\rho\|\widehat{\rho}_t) + \lambda\,\mathrm{KL}(\rho_t\|\widehat{\rho}_t),$$

whose maximizer is explicitly $\rho = \widehat{\rho}_t$. Then we have

$$\begin{aligned}\lambda\,\mathrm{KL}(\rho_t||\widehat{\rho}_t) \geq& \max_{\rho\in\mathcal{P}_2}\left\{-\lambda\int\ln\frac{\rho_t}{\widehat{\rho}_t}\left(\frac{1}{\sqrt{\lambda}}\rho - \rho_t\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\|\rho_t\right)\right\}\\ \geq& -\lambda\int\ln\frac{\rho_t}{\widehat{\rho}_t}\left(\frac{1}{\sqrt{\lambda}}\rho - \rho_t\right)\mathrm{d}\theta - \sqrt{\lambda}\mathrm{KL}\left(\rho\|\rho_t\right)\\ =& \int\mathbb{E}_{\sigma_t}\left[A_{\pi_t}\cdot h_\theta\right]\left(\rho_t - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta - (\sqrt{\lambda}-\lambda)\ln Z_\lambda\\ &-\int\ln\frac{\rho_t}{\nu}\left(\sqrt{\lambda}\rho - \lambda\rho_t\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\|\rho_t\right),\end{aligned} \tag{35}$$

where you set $\rho \in \mathcal{P}_2$ arbitrarily. Recall $\frac{\delta F}{\delta\rho}[\rho_t](\theta) = \mathbb{E}_{\sigma_t}[A_{\pi_t}h_\theta] - \lambda\ln\nu(\theta)$, then it holds that

$$\begin{aligned}\mathrm{KL}(\nu\|\widehat{\rho}_t) =& \int\mathrm{d}\nu\ln\nu - \int\mathrm{d}\nu\ln\widehat{\rho}_t\\ =& \frac{1}{\lambda}\mathbb{E}\left[A_{\pi_t}\int h_\theta\mathrm{d}\nu\right] + \ln Z_\lambda\\ =& \ln Z_\lambda,\end{aligned}$$

where we use $\int h_\theta\mathrm{d}\nu = 0$, since h_θ is an odd function. On the right-hand side of Eq. (35), The first term holds that

$$\begin{aligned}\int\mathbb{E}_{\sigma_t}\left[A_{\pi_t}h_\theta\right]\left(\rho_t - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta =& \mathbb{E}_{\sigma_t}\left[A_{\pi_t}\left(\int h_\theta\rho_t(\mathrm{d}\theta) - \frac{1}{\sqrt{\lambda}}\int h_\theta\rho(\mathrm{d}\theta)\right)\right]\\ =& \mathbb{E}_{\sigma_t}\left[A_{\pi_t}\left(f_t - \frac{f_\rho}{\sqrt{\lambda}}\right)\right],\end{aligned}$$

where $f_\rho = \int h_\theta \rho(\mathrm{d}\theta)$.

On the other hand, the rest on the right-hand side of Eq. (35) can be evaluated by

$$\begin{aligned}
-\int \ln \frac{\rho_t}{\nu}\left(\sqrt{\lambda}\rho - \lambda\rho_t\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\|\rho_t\right) =& -\int \ln \frac{\rho_t}{\nu}\left(\sqrt{\lambda}\rho - \lambda\rho_t\right)(\mathrm{d}\theta) - \sqrt{\lambda}\int \rho(\mathrm{d}\theta)\ln\frac{\rho}{\rho_t} \\
=& \lambda \int \mathrm{d}\rho_t \ln \frac{\rho_t}{\nu} - \sqrt{\lambda}\int \mathrm{d}\rho \ln \frac{\rho}{\nu} \\
=& \lambda\,\mathrm{KL}(\rho_t\|\nu) - \sqrt{\lambda}\,\mathrm{KL}(\rho\|\nu).
\end{aligned}$$

Combining all of them, we can evaluate the left-hand side of Eq. (35) by

$$\lambda\,\mathrm{KL}(\rho_t||\widehat{\rho}_t) \geq \mathbb{E}_{\sigma_t}\left[A_{\pi_t}\left(f_t - \frac{f_\rho}{\sqrt{\lambda}}\right)\right] + \lambda\,\mathrm{KL}(\rho_t\|\nu) - \sqrt{\lambda}\mathrm{KL}(\rho\|\nu) - \sqrt{\lambda}\mathrm{KL}(\nu\|\widehat{\rho}_t), \tag{36}$$

where we can take any $\rho \in \mathcal{P}_2$ arbitrarily.

Lemma 9 (Error of Global Optimality)**.** *Under the same conditions as Theorem 2, for $t \geq 0$, there exists $f_\rho \in \mathcal{F}_{R,M}$ such as*

$$J^* - J[\rho_t] - \mathbb{E}_{\sigma_t}\left[A_{\pi_t}\left(f_t - \frac{f_\rho}{\sqrt{\lambda}}\right)\right] \leq \frac{1}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2\sqrt{\lambda}. \tag{37}$$

Especially, $f_\rho = -A_{\pi_t} \in \mathcal{F}_{R,M}$ satisfies the inequality above.

Proof. Plugging Eq. (34) to Eq. (37), the left-hand side of Eq. (37)

$$\begin{aligned}
J^* - J[\rho_t] - \mathbb{E}_{\sigma_t}\left[A_{\pi_t}\left(f_t - \frac{f_\rho}{\sqrt{\lambda}}\right)\right] &= \mathbb{E}_{\sigma_t}\left[A_{\pi_t}\cdot\left(\frac{f_\rho}{\sqrt{\lambda}} - f_t + (1-\gamma)^{-1}\frac{\mathrm{d}\sigma_*}{\mathrm{d}\sigma_t}\right)\right] \\
&= \frac{1}{\sqrt{\lambda}}\langle A_{\pi_t}, f_\rho\rangle_{\sigma_t} - \langle A_{\pi_t}, f_t\rangle_{\sigma_t} + (1-\gamma)^{-1}\left\langle A_{\pi_t}, \frac{\mathrm{d}\sigma_*}{\mathrm{d}\sigma_t}\right\rangle_{\sigma_t},
\end{aligned} \tag{38}$$

where $\langle\cdot,\cdot\rangle_{\sigma_t}$ denotes inner product which introduces $L_{\sigma_t,2}$ norm. Recall that we have $A_{\pi_t} \in \mathcal{F}_{R,M}$ from Assumption 3 and that the second weight function $\beta(\cdot)$ is an odd function given by Assumption 1, then we can adapt $f_\rho = -A_{\pi_t}$. it holds for the first term on the right-hand side in Eq. (38) that

$$\frac{1}{\sqrt{\lambda}}\langle A_{\pi_t}, f_\rho\rangle_{\sigma_t} = -\frac{1}{\sqrt{\lambda}}\|A_{\pi_t}\|_{\sigma_t,2}^2$$

The second term on the right-hand side in Eq. (38) can be bounded by

$$\begin{aligned}
-\langle A_{\pi_t}, f_t\rangle_{\sigma_t} &\leq |\langle A_{\pi_t}, f_t\rangle_{\sigma_t}| \\
&\leq \|A_{\pi_t}\|_{\sigma_t,2}\cdot\|f_t\|_{\sigma_t,2} \\
&\leq R\,\|A_{\pi_t}\|_{\sigma_t,2},
\end{aligned} \tag{39}$$

where $R > 0$ is an absolute constant defined in Assumption 1. Meanwhile, the third term on the right-hand side in Eq. (38) can be bounded by

$$\begin{aligned}
(1-\gamma)^{-1}\left\langle A_{\pi_t}, \frac{\mathrm{d}\sigma_*}{\mathrm{d}\sigma_t}\right\rangle_{\sigma_t} &\leq (1-\gamma)^{-1}\|A_{\pi_t}\|_{\sigma_t,2}\left\|\frac{\mathrm{d}\sigma_*}{\mathrm{d}\sigma_t}\right\|_{\sigma_t,2} \\
&\leq \frac{\kappa}{1-\gamma}\|A_{\pi_t}\|_{\sigma_t,2},
\end{aligned} \tag{40}$$

where the first inequality follows from Jensen's inequality, the second inequality follows from Assumption 4, and $\kappa > 0$ is an absolute constant defined in Assumption 4-(ii). By plugging Eq. (39) and (40) into Eq. (38), we have

$$\mathbb{E}_{\sigma_t}\left[A_{\pi_t}\cdot\left(\frac{f_\rho}{\sqrt{\lambda}} - f_t + (1-\gamma)^{-1}\frac{\mathrm{d}\sigma_*}{\mathrm{d}\sigma_t}\right)\right] \leq -\frac{1}{\sqrt{\lambda}}\|A_{\pi_t}\|_{\sigma_t,2}^2 + \left(R + \frac{\kappa}{1-\gamma}\right)\|A_{\pi_t}\|_{\sigma_t,2}. \tag{41}$$

Considering that the right-hand side of Eq. (41) is a quadratic function for $\|A_{\pi_t}\|_{\sigma_t,2}$ as $-ax^2+bx \leq \frac{b^2}{4a}$ for any $x, a, b > 0$, it can be bounded by

$$-\frac{1}{\sqrt{\lambda}}\|A_{\pi_t}\|_{\sigma_t,2}^2 + \left(R + \frac{\kappa}{1-\gamma}\right)\|A_{\pi_t}\|_{\sigma_t,2} \leq \frac{\sqrt{\lambda}}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2.$$

□

Combining Eq. (36) and (37), we finish the proof of Lemma 3 as

$$J^* - J[\rho_t] \leq \lambda \operatorname{KL}(\rho_t||\widehat{\rho}_t) + \frac{\sqrt{\lambda}}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2 + 2\sqrt{\lambda}M - \lambda \operatorname{KL}(\rho_t\|\nu), \tag{42}$$

where we use $\operatorname{KL}(\rho\|\nu) \leq M$ from the definition of $\rho \in \mathcal{F}_{R,M}$. □

D.4. Proof of Theorem 2

Proof. Combining Lemma 2 and Eq. (42) in the proof of Lemma 3, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] &\leq -\alpha\lambda^2 \operatorname{KL}(\rho_t||\widehat{\rho}_t) + 2R^2(L_1+L_3)^2\iota^2\,\mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2\right] \\
&\leq -\alpha\lambda\left(J^* - J[\rho_t] - \frac{\sqrt{\lambda}}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2 - 2\sqrt{\lambda}M + \lambda \operatorname{KL}(\rho_t\|\nu)\right) + 2R^2(L_1+L_3)^2\iota^2\,\mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2\right] \\
&\leq -\alpha\lambda\left(\mathcal{F}[\rho_t] + J^* - \frac{\sqrt{\lambda}}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2 - 2\sqrt{\lambda}M\right) + 2R^2(L_1+L_3)^2\iota^2\,\mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2\right].
\end{aligned}$$

According to Corollary 1 with $T_{\mathrm{TD}} = \Omega(1/\alpha\lambda^{3/2})$, the last term on the right-hand side can be bounded by

$$2R^2(L_1+L_3)^2\iota^2\,\mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2\right] = \mathcal{O}(\alpha\lambda^{3/2}),$$

and define $\varepsilon(\lambda)$ as

$$\varepsilon(\lambda) = \frac{\sqrt{\lambda}}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2 + 2\sqrt{\lambda}M + \frac{2R^2(L_1+L_3)^2\iota^2}{\alpha\lambda}\mathbb{E}_{\varsigma_t}\left[(Q_t - Q_{\pi_t})^2\right] = \mathcal{O}(\sqrt{\lambda}).$$

Hence, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] \leq -\alpha\lambda\left(\mathcal{F}[\rho_t] + J^* - \varepsilon(\lambda)\right).$$

We obtain from a straightforward application of the Grönwall's inequality that

$$J^* - J[\rho_t] \leq \exp(-2\alpha\lambda t)\cdot\left(\mathcal{F}[\rho_0] + J^* - \varepsilon(\lambda)\right) + \varepsilon(\lambda) - \lambda \operatorname{KL}(\rho_t\|\nu).$$

Recalling $\rho_0 = \nu \sim \mathcal{N}(0, I_d)$ yields that

$$\mathcal{F}[\rho_0] = -J[\rho_0] + \lambda \operatorname{KL}(\rho_0\|\nu) = -J[\rho_0].$$

Considering $\varepsilon(\lambda), \lambda \operatorname{KL}(\rho_t\|\nu) \geq 0$, we obtain that

$$J^* - J[\rho_t] \leq \exp(-2\alpha\lambda t)\left(J^* - J[\rho_0]\right) + \varepsilon(\lambda),$$

which concludes the proof of Theorem 2 with $|J| \leq R$ given by Assumption 2. □

E. Time and Space Discretized Mean-field Langevin Policy Gradient: Proof of Theorem 3

We control the discretization error by the argument analogous to Chen et al. (2023). For the finite particle setting, we analyse the dynamics under the distribution space $\mathcal{P}_2^m$ of m particles $\Theta = \{\theta^i\}_{i\in[m]} \in \mathbb{R}^{d\times m}$, instead of the conventional single distribution space on $\theta \in \mathbb{R}^d$.

Hereafter, we denote the distribution of $\Theta = \{\theta^i\}_{i\in[m]} \in \mathbb{R}^{d\times m}$ by $\rho^m \in \mathcal{P}_2^m$. Corresponding to $\mathcal{P}_2^m$, we introduce the following objective:

$$\mathcal{F}^m[\rho^m] = \mathbb{E}_{\Theta\sim\rho^m}\left[-J_\Theta + \frac{\lambda}{2}\|\Theta\|^2\right] + \frac{\lambda}{m}\mathrm{Ent}(\rho^m).$$

We also define two types of proximal distribution for ρ_Θ as

$$\widehat{\rho}_\Theta(\theta) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta\rho}[\rho_\Theta](\theta)\right),$$
$$\widehat{\rho}^m(\Theta) \propto \exp\left(-\frac{m}{\lambda}F[\rho_\Theta]\right).$$

We have the continuous dynamics to consider the discretized update of each step for any $t' \in [0,\eta]$ as

$$\mathrm{d}\theta^i_{k,t'} = -\nabla\widetilde{\frac{\delta F}{\delta\rho}}[\rho_k](\theta^i_k)\cdot\mathrm{d}t' + \sqrt{2\lambda}\cdot\mathrm{d}W^i_{k,t'}, \tag{43}$$

where $\theta^i_{k,0} = \theta^i_k$ and $\theta^i_{k,\eta} = \theta^i_{k+1}$ and $\{W^i_{k,t'}\}_{t'\in[0,\eta]}$ is an i.i.d. Brownian motion with $W^i_{k,0} = 0$. Note that the dynamics correspond to the McKean-Vlasov SDE with the bias term fixed in Eq. (4). We define $\rho_{k,t'} \in \mathcal{P}_2$ and $\rho^m_{k,t'} \in \mathcal{P}_2^m$ as distributions induced by $\Theta_{k,t'} = \{\theta^i_{k,t'}\}_{i\in[m]}$. In addition, we define $\Theta^{-i} = \{\theta^j\}_{j\neq i}$ and $\rho^i(\cdot|\Theta^{-i})$ as the conditioned distribution of θ^i conditioned by Θ^{-i}. Hence, the time derivative of $\mathcal{F}^m[\rho^m_{k,t'}]$ can be evaluated as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}^m[\rho^m_{k,t'}] \leq -\frac{\lambda^2}{2}\frac{1}{m}\sum_{i=1}^m\int\rho^m_{k,t'}(\mathrm{d}\Theta)\left\|\nabla_i\ln\frac{\rho^m_{k,t'}}{\widehat{\rho}^m}(\Theta)\right\|_2^2 \tag{44a}$$
$$+\frac{1}{m}\sum_{i=1}^m\int\rho^m_{k,t'}(\mathrm{d}\Theta)\left(\left\|\nabla\widetilde{\frac{\delta\mathcal{F}}{\delta\rho}}[\rho_k](\theta^i) - \nabla\frac{\delta\mathcal{F}}{\delta\rho}[\rho_k](\theta^i)\right\|_2^2\right) \tag{44b}$$
$$+\frac{1}{m}\sum_{i=1}^m\int\rho^m_{k,t'}(\mathrm{d}\Theta)\rho^m_{k,0}(\mathrm{d}\Theta')\left(\left\|\nabla_i\frac{\delta\mathcal{F}}{\delta\rho}[\rho_{k,t'}](\theta^i) - \nabla_i\frac{\delta\mathcal{F}}{\delta\rho}[\rho_{k,0}](\theta'^i)\right\|^2\right), \tag{44c}$$

which follows from a similar argument as one in the proof in Theorem 2. The first term (44a) is the main term to decide the convergence rate. By the leave-one-out argument in the proof of Theorem 2.1 of Chen et al.(2022), the first term of the right-hand side can be upper bounded by

$$-\frac{1}{m}\sum_{i=1}^m\int\rho^m_{k,t'}(\mathrm{d}\Theta)\left\|\nabla_i\ln\frac{\rho^m_{k,t'}}{\widehat{\rho}^m}(\Theta)\right\|_2^2 \leq -\frac{\alpha}{m}\sum_{i=1}^m\int\rho^m_{k,t'}(\mathrm{d}\Theta)\,\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\middle\|\,\widehat{\rho}_{\Theta^{-i}}\right) + \frac{1}{m}\widetilde{C}_{\lambda,3}, \tag{45}$$

where $\widetilde{C}_{\lambda,3} = \mathcal{O}(\lambda^{-2}\vee\lambda^{-1}\eta^{-1})$. This inequality is given by Proposition 6. To control the first term on the right-hand side of Eq. (45), we use the following lemma.

Lemma 10 (Global Optimal Error for Discretized MFLPG)**.** *Under the same conditions as Theorem 3, it holds for all $k\in[T], t'\in[0,\eta], i\in[m]$, and $\lambda > 0$ that*

$$-\frac{\lambda}{m}\sum_{i=1}^m\int\rho^m_{k,t'}(\mathrm{d}\Theta)\,\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\middle\|\,\widehat{\rho}_{\Theta^{-i}}\right)$$
$$\leq -J^* + \mathbb{E}_{\Theta\sim\rho^m_{k,t'}}[J[\rho_\Theta]] + \sqrt{\lambda}\left[\frac{1}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2 + 2M\right] + \frac{1}{m}\widetilde{C}_{\lambda,4} - \lambda\,\mathrm{KL}\left(\rho^m_{k,t'}\middle\|\,\mu^m\right)$$

where $\widetilde{C}_{\lambda,4} = \mathcal{O}(\lambda^{-5/2}\vee\lambda^{-3/2}\eta^{-1})$.

See Appendix E.1 for the proof. Combining Eq. (45) and Lemma 10 finishes the upper-bound of the term (44a).

The second term (44b) can be evaluated with Lemma 8 for continuous version of MFLPG as

$$\frac{1}{m}\sum_{i=1}^{m}\int \rho_{k,t'}^{m}(\mathrm{d}\Theta)\left(\left\|\nabla\frac{\widetilde{\delta\mathcal{F}}}{\delta\rho}[\rho_k](\theta^i)-\nabla\frac{\delta\mathcal{F}}{\delta\rho}[\rho_k](\theta^i)\right\|_2^2\right)\leq 4R^2L_1^2L_3^2\iota^2\,\mathbb{E}_{\varsigma_{\pi_k}}\left[(Q_k-Q_{\pi_k})^2\right]:=\delta_{\mathrm{TD},k}^2,$$

which corresponds to the policy evaluation error given by MFLTD.

Moreover, the third term (44c) can be bounded by

$$\frac{1}{m}\sum_{i=1}^{m}\int \rho_{k,t'}^{m}(\mathrm{d}\Theta)\rho_{k,0}^{m}(\mathrm{d}\Theta')\left(\left\|\nabla_i\frac{\delta\mathcal{F}}{\delta\rho}[\rho_{k,t'}](\theta^i)-\nabla_i\frac{\delta\mathcal{F}}{\delta\rho}[\rho_{k,0}](\theta'^i)\right\|^2\right)\leq \widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta),$$

with constant $\widetilde{C}_{\lambda,2}=\mathcal{O}(\lambda^{-1}\vee\eta^{-1})$. The bound follows from Proposition 7, which controls the time discretization error.

Combining all of them, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}^m[\rho_{k,t'}^m] &\leq \frac{\alpha\lambda}{2}\left(-J^*+\mathbb{E}_{\Theta\sim\rho_{k,t'}^m}[J[\rho_\Theta]]+\sqrt{\lambda}\left[\frac{1}{4}\left(R+\frac{\kappa}{1-\gamma}\right)^2+2M\right]+\frac{1}{m}\widetilde{C}_{\lambda,4}-\lambda\,\mathrm{KL}\left(\rho_{k,t'}^m\middle\|\mu^m\right)\right)\\
&\quad+\frac{\lambda^2\widetilde{C}_{\lambda,3}}{m}+\delta_{\mathrm{TD},k}^2+\widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta)\\
&\leq -\frac{\alpha\lambda}{2}\left(\mathcal{F}^m[\rho_{k,t'}^m]+J^*\right)+\alpha\lambda^{3/2}\left[\frac{1}{8}\left(R+\frac{\kappa}{1-\gamma}\right)^2+M\right]+\frac{\alpha\lambda}{2m}\widetilde{C}_{\lambda,4}\\
&\quad+\frac{\lambda^2\widetilde{C}_{\lambda,3}}{m}+\delta_{\mathrm{TD},k}^2+\widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta)\\
&\leq -\frac{\alpha\lambda}{2}\left(\mathcal{F}^m[\rho_{k,t'}^m]+J^*\right)+\widetilde{C}\alpha\lambda^{3/2}+\frac{\widetilde{C}_{\lambda,1}}{m}+\widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta)+\delta_{\mathrm{TD},k}^2,
\end{aligned}$$

where we define each constants as follows

$$\begin{aligned}
\widetilde{C}&=\frac{1}{8}\left(R+\frac{\kappa}{1-\gamma}\right)^2+M=\mathcal{O}(1),\\
\widetilde{C}_{\lambda,1}&=\lambda^2\widetilde{C}_{\lambda,3}+\frac{\alpha\lambda}{2}\widetilde{C}_{\lambda,4}=\mathcal{O}\left(1\vee\frac{\lambda}{\eta}\right),\\
\widetilde{C}_{\lambda,2}&=\mathcal{O}\left(\frac{1}{\lambda}\vee\frac{1}{\eta}\right).
\end{aligned}$$

By applying the Grönwall's lemma, we obtain the following one-step update:

$$\begin{aligned}
\mathcal{F}^m[\rho_{k+1}^m]+J^* &\leq \exp(-\alpha\lambda\eta/2)\left(\mathcal{F}^m[\rho_k^m]+J^*\right)\\
&\quad+\frac{1-\exp(-\alpha\lambda\eta/2)}{\alpha\lambda/2}\left(\widetilde{C}\alpha\lambda^{3/2}+\frac{\widetilde{C}_{\lambda,1}}{m}+\widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta)+\delta_{\mathrm{TD},k}^2\right)\\
&\leq \exp(-\alpha\lambda\eta/2)\left(\mathcal{F}^m[\rho_k^m]+J^*\right)+\eta\left(\widetilde{C}\alpha\lambda^{3/2}+\frac{\widetilde{C}_{\lambda,1}}{m}+\widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta)+\delta_{\mathrm{TD},k}^2\right),
\end{aligned}$$

where the second inequality follows from the fact that $1-e^{-x}\leq x$ for any $x\in\mathbb{R}$. This reduction holds at every iteration k. Hence, if $\delta_{\mathrm{TD},k}^2=\delta_{\mathrm{TD}}^2$ then we arrive at the desired result for all $k\in[T]$:

$$J^*-\mathbb{E}_{\Theta\sim\rho_k^m}[J_\Theta]\leq\exp(-\alpha\lambda\eta k/2)\left(J^*-\mathbb{E}_{\Theta\sim\rho_0^m}[J_\Theta]\right)+\eta\left(\widetilde{C}\alpha\lambda^{3/2}+\frac{\widetilde{C}_{\lambda,1}}{m}+\widetilde{C}_{\lambda,2}(\eta^2+\lambda\eta)+\delta_{\mathrm{TD}}^2\right).$$

Recall that $J_\Theta \leq R$ from Assumption 2, we have for all $T \in \mathbb{Z}_{\geq 0}$ that

$$J^* - \mathbb{E}[J_{\Theta_T}] \leq 2R\exp(-\alpha\lambda\eta T/2) + \eta\left(\widetilde{C}\alpha\lambda^{3/2} + \frac{\widetilde{C}_{\lambda,1}}{m} + \widetilde{C}_{\lambda,2}(\eta^2 + \lambda\eta) + \delta_{\mathrm{TD}}^2\right),$$

which concludes the result of Theorem 3.

E.1. Proof of Lemma 10

Proof. We bound the KL divergence $\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\middle\|\widehat{\rho}_{\Theta^{-i}}\right)$ with the same argument as the proof of Theorem 2 as

$$\begin{aligned}\lambda\,\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\middle\|\widehat{\rho}_{\Theta^{-i}}\right) &= \lambda\int\rho^i(\mathrm{d}\theta|\Theta^{-i})\ln\frac{\rho^i(\theta|\Theta^{-i})}{\widehat{\rho}_{\Theta^{-i}}(\theta)}\\ &= -\lambda\int\ln\frac{\rho^i(\theta|\Theta^{-i})}{\widehat{\rho}_{\Theta^{-i}}(\theta)}\left(\frac{1}{\sqrt{\lambda}}\widehat{\rho}_{\Theta^{-i}}(\theta) - \rho^i(\theta|\Theta^{-i})\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\widehat{\rho}_{\Theta^{-i}}(\theta)\,\middle\|\rho^i(\theta|\Theta^{-i})\right).\end{aligned}$$

To bound the right-hand side, consider the following functional for any $\rho \in \mathcal{P}_2$:

$$\begin{aligned}&-\lambda\int\ln\frac{\rho^i(\theta|\Theta^{-i})}{\widehat{\rho}_{\Theta^{-i}}(\theta)}\left(\frac{1}{\sqrt{\lambda}}\rho - \rho^i(\theta|\Theta^{-i})\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\,\middle\|\rho^i(\theta|\Theta^{-i})\right)\\ &\quad= -\sqrt{\lambda}\,\mathrm{KL}\left(\rho\|\widehat{\rho}_{\Theta^{-i}}(\theta)\right) + \lambda\,\mathrm{KL}\left(\rho^i(\theta|\Theta^{-i})\middle\|\widehat{\rho}_{\Theta^{-i}}(\theta)\right),\end{aligned}$$

whose maximizer is explicitly $\rho = \widehat{\rho}_{\Theta^{-i}}(\theta)$. Then we have for $\Theta \in \mathbb{R}^{d\times m}$ that

$$\begin{aligned}&\lambda\,\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\middle\|\widehat{\rho}_{\Theta^{-i}}\right)\\ &\geq \max_{\rho\in\mathcal{P}_2}\left\{-\lambda\int\ln\frac{\rho^i(\theta|\Theta^{-i})}{\widehat{\rho}_{\Theta^{-i}}(\theta)}\left(\frac{1}{\sqrt{\lambda}}\rho - \rho^i(\theta|\Theta^{-i})\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\,\middle\|\rho^i(\theta|\Theta^{-i})\right)\right\}\\ &\geq -\lambda\int\ln\frac{\rho^i(\theta|\Theta^{-i})}{\widehat{\rho}_{\Theta^{-i}}(\theta)}\left(\frac{1}{\sqrt{\lambda}}\rho - \rho^i(\theta|\Theta^{-i})\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\,\middle\|\rho^i(\theta|\Theta^{-i})\right)\\ &= -\int\frac{\delta J}{\delta\rho}[\rho_{\Theta^{-i}}](\theta)\left(\rho^i(\theta|\Theta^{-i}) - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta - (\sqrt{\lambda}-\lambda)\mathrm{KL}(\nu\|\widehat{\rho}_{\Theta^{-i}})\\ &\quad-\int\ln\frac{\rho^i(\theta|\Theta^{-i})}{\nu}\left(\sqrt{\lambda}\rho - \lambda\rho^i(\theta|\Theta^{-i})\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\,\middle\|\rho^i(\cdot|\Theta^{-i})\right),\end{aligned}\tag{46}$$

where we can set $\rho \in \mathcal{P}_2$ arbitrarily and we denote $\pi^{-i} = \pi_{\rho_{\Theta^{-i}}}$ for simple notation. On the right-hand side of Eq. (46), The first term holds that

$$\begin{aligned}&-\int\frac{\delta J}{\delta\rho}[\rho_{\Theta^{-i}}](\theta)\left(\rho^i(\theta|\Theta^{-i}) - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta\\ &\geq -\int\frac{\delta J}{\delta\rho}[\rho_\Theta](\theta)\left(\rho^i(\theta|\Theta^{-i}) - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta - \int\left|\frac{\delta J}{\delta\rho}[\rho_\Theta](\theta) - \frac{\delta J}{\delta\rho}[\rho_{\Theta^{-i}}](\theta)\right|\left(\rho^i(\theta|\Theta^{-i}) + \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta,\end{aligned}\tag{47}$$

The first term on the right-hand side can be evaluated as

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m\int\left\{-\int\frac{\delta J}{\delta\rho}[\rho_\Theta](\theta)\left(\rho^i(\theta|\Theta^{-i}) - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta\right\}\mathrm{d}\rho_{k,t'}^m(\mathrm{d}\Theta)\\ &\quad= -\int\int\frac{\delta J}{\delta\rho}[\rho_\Theta](\theta)\left(\rho_\Theta(\theta) - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta\,\mathrm{d}\rho_{k,t'}^m(\mathrm{d}\Theta)\\ &\quad= \int\int\mathbb{E}_{\sigma_{\pi_\Theta}}[A_{\pi_\Theta}\,h_\theta]\left(\rho_\Theta - \frac{1}{\sqrt{\lambda}}\rho\right)\mathrm{d}\theta\,\mathrm{d}\rho_{k,t'}^m(\mathrm{d}\Theta)\\ &\quad= \int\mathbb{E}_{\sigma_{\pi_\Theta}}\left[A_{\pi_\Theta}\left(\int h_\theta\rho_\Theta\,\mathrm{d}\theta - \frac{1}{\sqrt{\lambda}}\int h_\theta\rho\,\mathrm{d}\theta\right)\right]\mathrm{d}\rho_{k,t'}^m(\mathrm{d}\Theta)\\ &\quad= \int\mathbb{E}_{\sigma_{\pi_\Theta}}\left[A_{\pi_\Theta}\left(f_\Theta - \frac{1}{\sqrt{\lambda}}f_\rho\right)\right]\mathrm{d}\rho_{k,t'}^m(\mathrm{d}\Theta).\end{aligned}$$

To control the error of global optimality, the same argument as Lemma 9 yields that

$$\int \mathbb{E}_{\sigma_{\pi_\Theta}}\left[A_{\pi_\Theta}\left(f_\Theta - \frac{1}{\sqrt{\lambda}} f_\rho\right)\right] \mathrm{d}\rho^m_{k,t'}(\mathrm{d}\Theta) \geq J^* - \mathbb{E}_{\Theta\sim\rho^m_{k,t'}}[J[\rho_\Theta]] - \frac{\sqrt{\lambda}}{4}\left(R + \frac{\kappa}{1-\gamma}\right)^2. \tag{48}$$

The third term on the right-hand side of Eq. (47) can be bounded as

$$\begin{aligned}
&\int \left|\frac{\delta J}{\delta \rho}[\rho_\Theta](\theta) - \frac{\delta J}{\delta \rho}[\rho_{\Theta^{-i}}](\theta)\right| \left(\rho^i(\theta|\Theta^{-i}) + \frac{1}{\sqrt{\lambda}}\rho\right) \mathrm{d}\theta \\
&= \int \int_0^1 \left| -\frac{1}{m}\frac{\delta^2 J}{\delta \rho^2}[p_\epsilon](\theta, \theta^i) + \frac{1}{m(m-1)}\sum_{j\neq i}\frac{\delta^2 J}{\delta \rho^2}[p_\epsilon](\theta, \theta^j)\right| \mathrm{d}\epsilon \left(\rho^i(\theta|\Theta^{-i}) + \frac{1}{\sqrt{\lambda}}\rho\right) \mathrm{d}\theta \\
&\leq \frac{L}{m}\int \left|2 + c\left(2\|\theta\|^2 + \|\theta^i\|^2 + \frac{1}{m-1}\sum_{j\neq i}\|\theta^j\|^2\right)\right| \left(\rho^i(\theta|\Theta^{-i}) + \frac{1}{\sqrt{\lambda}}\rho\right) \mathrm{d}\theta \\
&\leq \frac{(1+1/\sqrt{\lambda})L}{m}\left\{2 + c\left(\|\theta^i\|^2 + \frac{1}{m-1}\sum_{j\neq i}\|\theta^j\|^2\right)\right\} + \frac{2L}{m}\left(\frac{1}{\sqrt{\lambda}}\int \|\theta\|^2 \,\mathrm{d}\rho + \|\theta^i\|^2\right)
\end{aligned}$$

where we define p_ϵ as a connecting distribution $p_\epsilon = \epsilon\rho_\Theta + (1-\epsilon)\rho_{\Theta^{-i}}$ for $\forall \epsilon \in [0,1]$ and use the assumption on the second variation of $J[\rho]$. By taking the averaged expectation over each $i \in [m]$ and the randomness of $\Theta \sim \rho^m_{k,t'}$, we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \int \left\{\frac{(1+1/\sqrt{\lambda})L}{m}\left\{2 + c\left(\|\theta^i\|^2 + \frac{1}{m-1}\sum_{j\neq i}\|\theta^j\|^2\right)\right\} + \frac{2L}{m}\left(\frac{1}{\sqrt{\lambda}}\int \|\theta\|^2 \,\mathrm{d}\rho + \|\theta^i\|^2\right)\right\} \mathrm{d}\rho^m_{k,t'}(\mathrm{d}\Theta) \\
&= \frac{(1+1/\sqrt{\lambda})L}{m}\left(2 + \frac{2c}{m}\sum_{i=1}^m \mathbb{E}[\|\theta^i\|^2]\right) + \frac{2L}{m}\left(\frac{1}{\sqrt{\lambda}}\int \|\theta\|^2 \,\mathrm{d}\rho + \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\|\theta^i\|^2]\right),
\end{aligned}$$

where we take the expectation over $\Theta \sim \rho^m_{k,t'}$. We can bound the moment of each particle $\mathbb{E}[\|\theta^i\|^2]$ as

$$\begin{aligned}
\mathbb{E}[\|\theta^i_{k+1}\|^2] &= \mathbb{E}\left[\left\|\theta^i_k + \eta\nabla\widetilde{\frac{\delta F}{\delta \rho}}[\rho_k](\theta^i_k) + \sqrt{2\eta\lambda}\xi^i_k\right\|^2\right] \\
&= \mathbb{E}\left[\left\|(1-\lambda\eta)\theta^i_k - \eta\,\mathbb{E}_{\sigma_{\pi_k}}[A_k\,\nabla h_{\theta^i_k}] + \sqrt{2\eta\lambda}\xi^i_k\right\|^2\right] \\
&\leq \frac{1}{1-\lambda\eta}(1-\lambda\eta)^2\mathbb{E}\left[\|\theta^i_k\|^2\right] + \frac{1}{\lambda\eta}\mathbb{E}\left[\left\|-\eta\,\mathbb{E}_{\sigma_{\pi_k}}[A_k\,\nabla h_{\theta^i_k}] + \sqrt{2\eta\lambda}\xi^i_k\right\|^2\right] \\
&\leq (1-\lambda\eta)\mathbb{E}\left[\|\theta^i_k\|^2\right] + \frac{2\eta}{\lambda}\mathbb{E}\left[\left\|\mathbb{E}_{\sigma_{\pi_k}}[A_k\,\nabla h_{\theta^i_k}]\right\|^2\right] + 4\mathbb{E}\left[\left\|\xi^i_k\right\|^2\right] \\
&\leq (1-\lambda\eta)\mathbb{E}\left[\|\theta^i_k\|^2\right] + \frac{2\eta}{\lambda}(R^2L_1L_3)^2 + 4d,
\end{aligned}$$

where the first inequality follows from the fact that $(a+b)^2 \leq \frac{a^2}{1-c} + \frac{b^2}{c}$ for any $a, b, c \in \mathbb{R}$ and the last inequality follows from Assumption 1 and 2. Applying the Grönwall's inequality, we obtain that

$$\begin{aligned}
\mathbb{E}[\|\theta^i_k\|^2] &\leq (1-\lambda\eta)^k\mathbb{E}\left[\|\theta^i_0\|^2\right] + \frac{2}{\lambda^2}(R^2L_1L_3)^2 + \frac{4d}{\lambda\eta}, \\
&\leq d + \frac{2R^4L_1^2L_3^2}{\lambda^2} + \frac{4d}{\lambda\eta} =: \widetilde{C}_{\lambda,5},
\end{aligned} \tag{49}$$

where we use the initial distribution being a normal standard distribution $\theta^i_0 \sim \mathcal{N}(0, I_d)$. In the sequel, we bound the moment over ρ. Note that $f_\rho \in \mathcal{F}_{R,M}$ yields that $\mathrm{KL}(\rho\|\nu) \leq M$. From Lemma 4 and Definition 4, it holds that

$$W_2^2(\rho,\nu) := \inf_{\varphi\in\Phi(\rho,\nu)}\int \|\theta - \theta'\|^2\varphi(\mathrm{d}\theta, \mathrm{d}\theta') \leq 2\mathrm{KL}(\rho\|\nu) \leq 2M,$$

where $\Phi(\rho,\nu)$ denotes the set of joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal laws ρ and ν on the first and second factors, respectively. Recall $\|\theta\|^2 \le 2\|\theta'\|^2 + 2\|\theta-\theta'\|^2$, then taking the expectation over such a $\varphi \in \Phi(\rho,\nu)$ yields that

$$\begin{aligned}\mathbb{E}_{\theta\sim\rho}[\|\theta\|^2] &\le 2\mathbb{E}_{\theta'\sim\nu}[\|\theta'\|^2] + 2\int \|\theta-\theta'\|^2 \varphi(\mathrm{d}\theta,\mathrm{d}\theta') \\ &\le 2(1+2M). \end{aligned} \tag{50}$$

Combining Eq. (49) and (50), it holds that

$$\begin{aligned}&\frac{(1+1/\sqrt{\lambda})L}{m}\left(2+\frac{2c}{m}\sum_{i=1}^m \mathbb{E}[\|\theta^i\|^2]\right) + \frac{2L}{m}\left(\frac{1}{\sqrt{\lambda}}\int \|\theta\|^2\,\mathrm{d}\rho + \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\|\theta^i\|^2]\right) \\ &\le \frac{L}{m}\left(1+\frac{1}{\sqrt{\lambda}}\right)\left(2+2c\,\widetilde{C}_{\lambda,5}\right) + \frac{2L}{m}\left(\frac{2(1+2M)}{\sqrt{\lambda}} + \widetilde{C}_{\lambda,5}\right), \\ &= \frac{\widetilde{C}_{\lambda,4}}{m}, \end{aligned} \tag{51}$$

where we define $\widetilde{C}_{\lambda,4} = 2L\left(1+\frac{1}{\sqrt{\lambda}}\right)\left(1+c\,\widetilde{C}_{5,\lambda}\right) + 2L\left(\frac{2(1+2M)}{\sqrt{\lambda}} + \widetilde{C}_{5,\lambda}\right) = \mathcal{O}(\lambda^{-5/2} \vee \lambda^{-3/2}\eta^{-1})$.

On the other hand, the rest on the right-hand side of Eq. (46) can be evaluated under the averaged expectation taken over all $i \in [m]$ and the randomness of $\Theta \sim \rho^m_{k,t'}$ by

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[-\int \ln\frac{\rho^i(\theta|\Theta^{-i})}{\nu}\left(\sqrt{\lambda}\rho - \lambda\rho^i(\theta|\Theta^{-i})\right)\mathrm{d}\theta - \sqrt{\lambda}\,\mathrm{KL}\left(\rho\,\|\rho^i(\cdot|\Theta^{-i})\right)\right] \\ &= \frac{\lambda}{m}\sum_{i=1}^m \mathbb{E}\left[\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\|\,\nu\right)\right] - \sqrt{\lambda}\,\mathrm{KL}(\rho\|\nu). \end{aligned}$$

The first term on the right-hand side can be bounded as

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\|\,\nu\right)\right] &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\mathrm{Ent}\left[\rho^i(\cdot|\Theta^{-i})\right]\right] - \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \ln\nu(\theta^i)\right] \\ &\ge \mathrm{KL}\left(\rho^m_{k,t'}\|\,\mu^m\right), \end{aligned}$$

where the inequality follows from Lemma 3.6 in Chen et al. (2023). In addition, $\rho, \widehat{\rho}_{\Theta^{-i}} \in \mathcal{F}_{R,M}$ yields that

$$\frac{\lambda}{m}\sum_{i=1}^m \mathbb{E}\left[\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\|\,\nu\right)\right] - \sqrt{\lambda}\,\mathrm{KL}(\rho\|\nu) \ge \mathrm{KL}\left(\rho^m_{k,t'}\|\,\mu^m\right) + 2\sqrt{\lambda}\,M. \tag{52}$$

Combining Eq. (48), (51), and (52) yields that

$$\begin{aligned}&-\frac{\lambda}{m}\sum_{i=1}^m \int \rho^m_{k,t'}(\mathrm{d}\Theta)\,\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\|\,\widehat{\rho}_{\Theta^{-i}}\right) \\ &\le -J^* + \mathbb{E}_{\Theta\sim\rho^m_{k,t'}}[J[\rho_\Theta]] + \frac{\sqrt{\lambda}}{4}\left(R+\frac{\kappa}{1-\gamma}\right)^2 + 2\sqrt{\lambda}\,M + \frac{\widetilde{C}_{\lambda,4}}{m} - \lambda\,\mathrm{KL}\left(\rho^m_{k,t'}\|\,\mu^m\right), \end{aligned}$$

which concludes the proof of Lemma 10. □

F. Auxiliary Lemmas

This section is devoted to presenting the related propositions and lemmas used in the proof.

First of all, we present the policy gradient theorem presented by Sutton et al. (1999) as Proposition 4, which provides the gradient of the expected total reward function J with a policy π_Θ parameterized by Θ. Refer to the original paper for the proof.

Proposition 4 (Policy Gradient Theorem (Sutton et al., 1999))**.** *For any MDP, it holds that*

$$\nabla J[\pi_\Theta] = \mathbb{E}_{\sigma_{\pi_\Theta}}\left[\int \nabla \pi_\Theta(\mathrm{d}a|s)\cdot Q_{\pi_\Theta}(s,a)\right],$$

where σ_{π_θ} is the state visitation measure.

Second, we provide the basic proposition to access the inner-loop convergence of MFLTD. For l-th outer step, we define the global optimal distribution of an inner-loop MFLD by $q_*^{(l+1)}$. It holds that

Proposition 5 (Linear Convergence of Inner-Loop MFLD)**.** *Under Assumption 1, 2, and 3, if we run the noisy gradient descent which is the inner-loop algorithm in Algorithm 2 for all l-step, we obtain for all $s \geq 0, l \in [0, T_{\mathrm{TD}}-1]$ and $\lambda_{\mathrm{TD}} > 0$ that*

$$\lambda_{\mathrm{TD}}\,\mathrm{KL}(q_s\|q_*^{(l+1)}) \leq \mathcal{L}_l[q_s] - \mathcal{L}_l[q_*^{(l+1)}] \leq \exp\left(-2\alpha\lambda_{\mathrm{TD}}s\right)\cdot\left(\mathcal{L}_l[q_0] - \mathcal{L}_l[q_*^{(l+1)}]\right),$$

where α is the LSI constant induced by λ_{TD}.

Proof. For the proof of Proposition 5, we apply Nitanda et al. (2022)'s convex optimization analysis. First of all, we prove the convexity of L_l over the parameter distribution. From the definition of L_l, we can reformulate L_l as

$$L_l[q] = \int U\mathrm{d}q + \left(\int V\mathrm{d}q\right)^2, \tag{53}$$

where U, V do not depend on q but only θ. Eq. (53) results that the objective function L_l is convex in terms of a functional with a probability distribution q as a variable, where we define the convexity condition for functional F and $q, q' \in \mathcal{P}_2$ as

$$F[q'] \geq F[q] + \int \frac{\delta F}{\delta q}[q](\theta)(\mathrm{d}q' - \mathrm{d}q).$$

Therefore, Noting that the MFLD used in the inner algorithm of the MFLTD is a simple Wasserstein gradient flow with the convex objective functional L_l, it holds from Theorem 1 in (Nitanda et al., 2022) that

$$\mathcal{L}_l[q_s] - \mathcal{L}_l[q_*^{(l+1)}] \leq \exp\left(-2\alpha\lambda_{\mathrm{TD}}s\right)\cdot\left(\mathcal{L}_l[q_0] - \mathcal{L}_l[q_*^{(l+1)}]\right),$$

where α is the LSI constant induced by λ_{TD}, which is given in Proposition 2. In addition, the proof of the first inequality of the statement Proposition 5 follows from Proposition 1. in (Nitanda et al., 2022), from which we have

$$\lambda_{\mathrm{TD}}\,\mathrm{KL}(q_s\|q_*^{(l+1)}) \leq \mathcal{L}_l[q_s] - \mathcal{L}_l[q_*^{(l+1)}].$$

We finish the proof of Proposition 5. □

In the sequel, we introduce the following basic lemma about the norm of the transition operator, which is useful for a geometric property of the semi-gradient of the Bellman error.

Lemma 11 (Transition Operator Norm)**.** *Let the linear operator $\mathcal{P} : L^2(\varsigma_\pi)(\mathcal{S}\times\mathcal{A}) \to L^2(\varsigma_\pi)(\mathcal{S}\times\mathcal{A})$ be the transition operator satisfying $\mathcal{P}Q(s,a) = \int \mathrm{d}s' P(s'|s,a)\int \mathrm{d}a'\pi(a'|s')Q(s',a'),\ Q \in L^2(\varsigma_\pi)(\mathcal{S}\times\mathcal{A})$. Then the operator norm of $\mathcal{P}$ is no more than* 1.

Proof. For all $Q(x) \in L^2(\varsigma_\pi)(\mathcal{S}\times\mathcal{A}), x = (s,a) \in \mathcal{S}\times\mathcal{A}$, we have that

$$\begin{aligned}
\|\mathcal{P}Q(x)\|_{\varsigma_\pi,2} =& \mathbb{E}_{x\sim\varsigma_\pi}\left[\left(\int \mathrm{d}(\pi\otimes P)(x'|x)Q(x')\right)^2\right]\\
\leq& \mathbb{E}_{x\sim\varsigma_\pi}\left[\int \mathrm{d}(\pi\otimes P)(x'|x)Q(x')^2\right]\\
=& \mathbb{E}_{x\sim\varsigma_\pi}\left[Q(x)^2\right]\\
=& \|Q(x)\|_{\varsigma_\pi,2},
\end{aligned}$$

where the inequality follows Jensen's inequality and the second equality follows the fact that ς_π is the stationary distribution under the transition probability. □

What we follow, we introduce some propositions to control each error given in the proof of Theorem 3.

Proposition 6 (Log-Sobolev Inequality for Discretized Mean-field Langevin Dynamics)**.** *We define* $\Theta^{-i} = \{\theta^j\}_{j \neq i}$ *and* $\rho^i(\cdot|\Theta^{-i})$ *as the conditioned distribution of* θ^i *conditioned by* Θ^{-i}*. Under the same conditions as Theorem 3, it holds that*

$$-\frac{1}{m}\sum_{i=1}^{m}\int \rho^m_{k,t'}(\mathrm{d}\Theta)\left\|\nabla_i \ln \frac{\rho^m_{k,t'}}{\widehat{\rho}^m}(\Theta)\right\|_2^2 \leq -\frac{\alpha}{m}\sum_{i=1}^{m}\int \rho^m_{k,t'}(\mathrm{d}\Theta)\,\mathrm{KL}\left(\rho^i(\cdot|\Theta^{-i})\middle\|\,\widehat{\rho}_{\Theta^{-i}}\right) + \frac{\widetilde{C}_{\lambda,3}}{m},$$

where $\widetilde{C}_{\lambda,3} = \mathcal{O}(\lambda^{-2} \vee \lambda^{-1}\eta^{-1})$.

Proof. The proof of Proposition 6 is analogous to Chen et al. (2023). See the proof of Theorem 2.6 in Chen et al. (2023) for more detail. □

Proposition 7 (Time Discretization Error for MFLD)**.** *The distribution contour* $\rho^m_{k,t'}$ *for* $t' \in [0,\eta]$ *follows the dynamics in Eq.* (43)*. Under the same conditions as Theorem 3, it holds that*

$$\frac{1}{m}\sum_{i=1}^{m}\int \rho^m_{k,t'}(\mathrm{d}\Theta)\rho^m_{k,0}(\mathrm{d}\Theta')\left(\left\|\nabla_i \frac{\delta \mathcal{F}}{\delta \rho}[\rho_{k,t'}](\theta^i) - \nabla_i \frac{\delta \mathcal{F}}{\delta \rho}[\rho_{k,0}](\theta'^i)\right\|^2\right) \leq \widetilde{C}_{\lambda,2}(\eta^2 + \lambda\eta)$$

where $\widetilde{C}_{\lambda,2} = \mathcal{O}(\lambda^{-1} \vee \eta^{-1})$.

Proof. The proof of Proposition 7 is given by Suzuki et al. (2023). See the proof of Lemma 2 in Suzuki et al. (2023) for more detail. □