

---

# Last Iterate Convergence in Monotone Mean Field Games

---

**Noboru Isobe**  
RIKEN AIP  
Tokyo, Japan  
noboru.isobe@riken.jp

**Kenshi Abe**  
CyberAgent  
Tokyo, Japan  
abe\_kenshi@cyberagent.co.jp

**Kaito Ariu**  
CyberAgent  
Tokyo, Japan  
kaito\_ariu@cyberagent.co.jp

## Abstract

In the Lasry–Lions framework, Mean-Field Games (MFGs) model interactions among an infinite number of agents. However, existing algorithms either require strict monotonicity or only guarantee the convergence of averaged iterates, as in Fictitious Play in continuous time. We address this gap with the following theoretical result. First, we prove that the last-iterated policy of a proximal-point (PP) update with KL regularization converges to an equilibrium of MFG under non-strict monotonicity. Second, we see that each PP update is equivalent to finding the equilibria of a KL-regularized MFG. We then prove that this equilibrium can be found using Mirror Descent (MD) with an exponential last-iterate convergence rate. Building on these insights, we propose the Approximate Proximal-Point (APP) algorithm, which approximately implements the PP update via a small number of MD steps. Numerical experiments on standard benchmarks confirm that the APP algorithm reliably converges to the unregularized mean-field equilibrium without time-averaging.

## 1 Introduction

Mean Field Games (MFGs) provide a simple and powerful framework for approximating the behavior of large populations of interacting agents. Formulated initially by Lasry and Lions (2007) and M. Huang et al. (2006), MFGs model the collective behavior of homogeneous agents in continuous time and state settings using partial differential equations (Cardaliaguet and Hadikhanloo 2017; Lavigne and Pfeiffer 2023; Inoue et al. 2023). The formulation of MFGs using Markov decision processes (MDPs) in Bertsekas and Shreve (1978) and Puterman (1994) has enabled the study of discrete-time and discrete-state models (Gomes et al. 2010).

In this context, a player’s policy  $\pi$ , i.e., a probability distribution over actions, induces the so-called mean field  $\mu$ . This mean field  $\mu$ —namely, the distribution of all players’ states—then affects both the state-transition dynamics and the rewards received by every agent. This simple formulation has broadened the applicability of MFGs to Multi-Agent Reinforcement Learning (MARL) (Yang et al. 2018; Guo et al. 2019; Angiuli et al. 2022; Zeman et al. 2023; Angiuli et al. 2024). Moreover, it has become possible to capture interactions among heterogeneous agents (Gao and Caines 2017; Caines and M. Huang 2019).

The applicability of MFGs to MARL drives research into the theoretical aspects of numerical algorithms for MFGs. Under fairly general assumptions, the problem of finding an equilibrium

in MFGs is known to be PPAD-complete (Yardim et al. 2024). Consequently, it is essential to impose assumptions that allow for the existence of algorithms capable of efficiently computing an equilibrium. One such assumption is contractivity (Q. Xie et al. 2021; Anahtarci et al. 2023; Yardim et al. 2023). However, many MFG instances are known to be non-contractive in practice (Cui and Koepl 2021). A more realistic assumption is the Lasry–Lions-type monotonicity employed in Pérolat et al. (2022), F. Zhang et al. (2023), and Yardim and He (2024), which intuitively implies that a player’s reward monotonically decreases as more agents converge to a single state. Under the monotonicity assumption, Online Mirror Descent (OMD) has been proposed and widely adopted (Pérolat et al. 2022; Cui and Koepl 2022; Laurière et al. 2022; Fabian et al. 2023). OMD, especially when combined with function approximation via deep learning, has enabled the application of MFGs to MARL (Yang and Wang 2020; K. Zhang et al. 2021; Cui et al. 2022).

Theoretically, *last-iterate convergence* (LIC) without time-averaging is particularly important in deep learning settings due to the constraints imposed by neural networks (NNs), as it ensures that the policy obtained in the last iteration converges. In NNs, computing the time-averaged policy as in the celebrated Fictitious Play method (Brown 1951; Perrin et al. 2020) may be less meaningful due to nonlinearity in the parameter space. This motivation has spurred significant research into developing algorithms that achieve LIC in finite  $N$ -player games, as seen in, e.g., Mertikopoulos et al. (2018), Piliouras et al. (2022), Abe et al. (2023), and Abe et al. (2024). However, in the case of MFGs, the results on LIC under realistic assumptions are limited. We refer the reader to read § 7 and Appendix A to review the existing results in detail.

We aim to develop a simple method to achieve LIC for MFGs with a realistic assumption. The first result of this paper is the development of a Proximal Point (PP) method using Kullback–Leibler (KL) divergence. We establish a novel convergence result in Theorem 3.1, showing that the PP method achieves LIC under the monotonicity assumption. When attempting to obtain convergence results in MFG, one faces the difficulty of controlling the mean field  $\mu$ , which changes along with the iterative updates of the policy  $\pi$ . We overcome this difficulty using the Łojasiewicz inequality, a classical tool from real analytic geometry.

We further propose the Approximate Proximal Point (APP) method to make the PP method feasible, which can be interpreted as an approximation of it. Here, we show that one iteration of the PP method corresponds to finding an equilibrium of the MFG regularized by KL divergence. This insight leads to the idea of approximating the iteration of PP by regularized Mirror Descent (RMD) introduced by F. Zhang et al. (2023). Our second theoretical result, presented in Theorem 4.3, is the LIC of RMD with an exponential rate. This result is a significant improvement over previous studies that only showed the convergence of the time-averaged policy or convergence at a polynomial rate. In the proof, the dependence of the mean field  $\mu$  on the policy  $\pi$  makes it difficult to readily exploit the Lipschitz continuity of the  $Q$ -function. We address this issue by utilizing the regularizing effect of the KL divergence.

Our experimental results also demonstrate LIC. The APP method can be implemented by making only a small modification to the RMD and experimentally converges to the (unregularized) equilibrium.

In summary, the contributions of this paper are as follows:

#### Contributions

- (i) We present an algorithm based on the celebrated PP method and, for the first time, establish LIC for *non-strictly* monotone MFGs (Theorem 3.1).
- (ii) We show that one iteration of the PP method is equivalent to solving the regularized MFG, which can be solved exponentially fast by RMD (Theorem 4.3).
- (iii) Based on these two theoretical findings, we develop the APP method as an efficient approximation of the PP method (Algorithm 1).

The organization of this paper is as follows: In § 2, we review the fundamental concepts of MFGs. In § 3, we introduce the PP method and its convergence results. In § 4, we present the RMD algorithm and its convergence properties. Finally, in § 5, we propose a combined approximation method, demonstrating its convergence through experimental validation. § 7 provides a review of related works.

## 2 Problem setting and preliminary facts

**Notation:** For a positive integer  $N \in \mathbb{N}$ ,  $[N] := \{1, \dots, N\}$ . For a finite set  $X$ ,  $\Delta(X) := \{p \in \mathbb{R}_{\geq 0}^{|X|} \mid \sum_{x \in X} p(x) = 1\}$ . For a function  $f: X \rightarrow \mathbb{R}$  and a probability  $\pi \in \Delta(X)$ ,  $\langle f, \pi \rangle := \langle f(\bullet), \pi(\bullet) \rangle := \sum_{x \in X} f(x) \pi(x)$ . For  $p^0, p^1 \in \Delta(X)$ , define the KL divergence  $D_{\text{KL}}(p^0, p^1) := \sum_{x \in X} p^0(x) \log(p^0(x)/p^1(x))$ , and the  $\ell^1$  distance as  $\|p^0 - p^1\| := \sum_{x \in X} |p^0(x) - p^1(x)|$ .

### 2.1 Mean-field games

Consider a model based *Mean-Field Game (MFG)* that is defined through a tuple  $(\mathcal{S}, \mathcal{A}, H, P, r, \mu_1)$ . Here,  $\mathcal{S}$  is a finite discrete space of states,  $\mathcal{A}$  is a finite discrete space of actions,  $H \in \mathbb{N}_{\geq 2}$  is a time horizon, and  $P = (P_h)_{h=1}^H$  is a sequence of transition kernels  $P_h: \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ , that is, if a player with state  $s_h \in \mathcal{S}$  takes action  $a_h \in \mathcal{A}$  at time  $h \in [H]$ , the next state  $s_{h+1} \in \mathcal{S}$  will transition according to  $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ . In addition,  $r = (r_h)_{h=1}^H$  is a sequence of reward functions  $r_h: \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \rightarrow [0, 1]$ , and  $\mu_1 \in \Delta(\mathcal{S})$  is an initial probability of state. Note that, in the context of theoretical analysis of the online learning method for MFG (Pérolat et al. 2022; F. Zhang et al. 2023),  $P$  is assumed to be independent of the state distribution. It is reasonable to assume that at any time  $h$ , every state  $s' \in \mathcal{S}$  is reachable:

**Assumption 2.1.** For each  $(h, s') \in [H] \times \mathcal{S}$ , there exists  $(s, a) \in \mathcal{S} \times \mathcal{A}$  such that  $P_h(s' \mid s, a) > 0$ .

Note that it does *not* require that, for any state  $s' \in \mathcal{S}$ , it is reachable by *any* state-action pair  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .

*Remark 2.2.* Our analysis excludes cases in which  $P$  depends on  $\mu$ , as studied in Zeman et al. (2023) and Zeng et al. (2024). These studies also rely on other conditions such as contraction or herding, which differ in nature from our monotonicity assumption. Extending the analysis to a  $\mu$ -dependent  $P$  requires a different approach than that in the existing literature, e.g., Pérolat et al. (2022) and F. Zhang et al. (2023). A full treatment of the case is left for future work.

Given a policy  $\pi$ , the probabilities  $m[\pi] = (m[\pi]_h)_{h=1}^H \in \Delta(\mathcal{S})^H$  of the state is recursively defined as follows:  $m[\pi]_1 = \mu_1$  and

$$m[\pi]_h(s_h) = \sum_{s_{h-1} \in \mathcal{S}, a_{h-1} \in \mathcal{A}} \pi_{h-1}(a_{h-1} \mid s_{h-1}) P_{h-1}(s_h \mid s_{h-1}, a_{h-1}) m[\pi]_{h-1}(s_{h-1}), \quad (2.1)$$

if  $h = 2, \dots, H$ . We aim to maximize the following cumulative reward

$$J(\mu, \pi) := \sum_{(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}} \pi_h(a \mid s) m[\pi]_h(s) r_h(s, a, \mu_h), \quad (2.2)$$

with respect to the policy  $\pi$ , given a sequence of state distributions  $\mu \in \Delta(\mathcal{S})^H$ . The *mean-field equilibrium* defined below means the pair of probabilities  $\mu$  and policies  $\pi$  that achieves the maximum under the constraints (2.1).

**Definition 2.3.** A pair  $(\mu^*, \pi^*) \in \Delta(\mathcal{S})^H \times (\Delta(\mathcal{A})^{\mathcal{S}})^H$  is a *mean-field equilibrium* if it satisfies (i)  $J(\mu^*, \pi^*) = \max_{\pi \in \Delta(\mathcal{S})^H} J(\mu^*, \pi)$ , and (ii)  $\mu^* = m[\pi^*]$ . In addition, set  $\Pi^* \subset (\Delta(\mathcal{A})^{\mathcal{S}})^H$  as the set of all policies that are in mean-field equilibrium.

Under Assumptions 2.4 and 2.5 below, there exists a mean-field equilibrium, see the proof of Saldi et al. (2018, Theorem 3.3.) and Pérolat et al. (2022, Proposition 1.). Note that the equilibrium may not be unique if the inequality given below in Assumption 2.4 is non-strict. In other words, the set  $\Pi^* \subset (\Delta(\mathcal{A})^{\mathcal{S}})^H$  is not a singleton in general. As an illustrative example, one might consider the trivial case where  $r \equiv 0$ . Our goal is to construct an algorithm that approximates a policy in  $\Pi^*$ .

In this paper, we focus on rewards  $r$  that satisfy the following two typical conditions, which are also assumed in Perrin et al. (2020), Perrin et al. (2022), Pérolat et al. (2022), Fabian et al. (2023), and F. Zhang et al. (2023). The first one is *monotonicity* of the type introduced by Lasry and Lions (2007), which means, under a state distribution  $\mu = (\mu_h)_{h=1}^H \in \Delta(\mathcal{S})^H$ , if players choose a strategy—called a policy  $\pi = (\pi_h)_{h=1}^H \in (\Delta(\mathcal{A})^{\mathcal{S}})^H$  to be planned—that concentrates on a state or action, they will receive a small reward.

**Assumption 2.4** (Weak monotonicity of  $r$ ). For all  $\pi, \tilde{\pi} \in (\Delta(\mathcal{A})^S)^H$ , it holds that

$$\sum_{h=1}^H \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left( r_h(s, a, \mu_h^\pi) - r_h(s, a, \mu_h^{\tilde{\pi}}) \right) (\rho_h(s, a) - \tilde{\rho}_h(s, a)) \leq 0, \quad (2.3)$$

where we set  $\mu^\pi = m[\pi]$ ,  $\rho_h(s, a) := \pi_h(a | s) \mu_h^\pi(s)$  and  $\tilde{\rho}_h(s, a) := \tilde{\pi}_h(a | s) \mu_h^{\tilde{\pi}}(s)$ .

A reward  $r$  satisfying [Assumption 2.4](#) is said to be *monotone*. Furthermore,  $r$  is said to be *strictly monotone* if the equality in (2.3) holds only if  $\pi = \tilde{\pi}$ . Although most of the previous papers provide theoretical analysis under strict monotonicity, this excludes the case where the transition is symmetric. We demonstrate that such structures inherently allow the existence of distinct policies generating identical state distributions, leading to the failure of strict monotonicity.

**Example (Failure of strict monotonicity in symmetric transitions).** In general, *symmetry of  $P$*  with respect to states in MFGs violates strict monotonicity, while preserving monotonicity. Consider an MFG with *symmetric transition dynamics*, e.g., consider an MDP defined on  $\mathcal{S} = \{s_1, s_2\}$ ,  $\mathcal{A} = \{a_1, a_2\}$ ,  $H \geq 2$ ,  $\mu_1 = (\frac{1}{2}, \frac{1}{2})$ . For each  $h \in [H]$ , the transition kernels are  $P_h(s' | s, a = a_1) = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}$ ,  $P_h(s' | s, a = a_2) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$ . If we take the policy  $\pi$  such that  $\pi_h(a_1 | s) = 1$ ,  $\pi_h(a_2 | s) = 0$ ,  $\tilde{\pi}_h(a_1 | s) = 0$ ,  $\tilde{\pi}_h(a_2 | s) = 1$ , for all  $s \in \mathcal{S}$  and  $h \in [H]$ , we can see that  $m[\pi]_h = m[\tilde{\pi}]_h = (0.5, 0.5)$  for all  $h$ . Let the reward be of the form  $r_h(s, a, \mu) = R_h(s, a) - f(\mu(s))$  with a non-decreasing function  $f: [0, 1] \rightarrow \mathbb{R}$  such as  $f(x) = x$ , which models a crowd that avoids overcrowding. Then the monotonicity condition holds for the case  $\pi \neq \tilde{\pi}$ . However, strict monotonicity would demand that equality occur *only* if  $\pi = \tilde{\pi}$ . In this example, whenever  $m[\pi] = m[\tilde{\pi}]$  (here the uniform distribution), the above sum is zero even if  $\pi \neq \tilde{\pi}$ . Hence, the game is monotone but *not* strictly monotone. Such phenomena are limitations in games with balanced transitions.  $\square$

The second is the Lipschitz continuity of  $r$  with respect to  $\mu \in (\Delta(\mathcal{S}))^H$ , which is standard in the field of MFGs (Cui and Koepl 2021; Fabian et al. 2023; F. Zhang et al. 2023).

**Assumption 2.5** (Lipschitz continuity of  $r$ ). There exists a constant  $L$  such that for every  $h \in [H]$ ,  $s \in \mathcal{S}$ ,  $a \in \mathcal{A}$ , and  $\mu, \mu' \in \Delta(\mathcal{S})$ :  $|r_h(s, a, \mu) - r_h(s, a, \mu')| \leq L \|\mu - \mu'\|$ .

### 3 Proximal point-type method for MFG

This section presents an algorithm motivated by the Proximal Point (PP) method. Let  $\lambda > 0$  be a sufficiently small positive number, roughly “the inverse of learning rate.” In the algorithm proposed in this paper, we generate a sequence  $((\sigma^k, \mu^k))_{k=0}^\infty \subset (\Delta(\mathcal{A})^S)^H \times \Delta(\mathcal{S})^H$  as

$$\sigma^{k+1} = \arg \max_{\pi \in (\Delta(\mathcal{A})^S)^H} \{J(\mu^{k+1}, \pi) - \lambda D_{m[\pi]}(\pi, \sigma^k)\}, \quad \mu^{k+1} = m[\sigma^{k+1}], \quad (3.1)$$

where  $m$  is defined in (2.1) and  $D_\mu(\pi, \sigma^k) := \sum_h \mathbb{E}_{s \sim \mu_h} [D_{\text{KL}}(\pi_h(s), \sigma_h^k(s))]$  with a probability  $\mu \in \Delta(\mathcal{S})^H$ . If the initial policy  $\pi^0$  has full support, i.e.,  $\min_{(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}} \pi_h^0(a | s) > 0$ , the rule (3.1) is well-defined, see [Proposition C.1](#).

Interestingly, the rule (3.1) is similar to the traditional Proximal Point (PP) method with KL divergence in mathematical optimization and Optimal Transport, see Censor and Zenios (1992) and Y. Xie et al. (2019). Therefore, we also refer to this update rule as the PP method. The well-known (O)MD in P  rolat et al. (2022) can be viewed as a linearization of the objective  $J$  inside (3.1). Consequently, PP—which uses the full, un-linearised  $J$ —is expected to be less sensitive to approximation error, resulting in more robust convergence under non-strict monotonicity than MD. On the other hand, unlike the traditional PP method, our method changes the objective function  $J(\mu^k, \bullet): (\Delta(\mathcal{A})^S)^H \rightarrow \mathbb{R}$  with each iteration  $k \in \mathbb{N}$ . Therefore, it is difficult to derive a theoretical convergence result of our traditional method from traditional theory. See also [Remark 3.3](#).

#### 3.1 Last-iterate convergence result

The following theorem implies the last-iterate convergence of the policies generated by (3.1). Specifically, it shows that under the assumptions above, the sequence of policies converges to the equilibrium set. This result is crucial for the effectiveness of the algorithm in reaching an optimal policy.

**Theorem 3.1.** Let  $(\sigma^k)_{k=0}^\infty$  be the sequence defined by (3.1). In addition to Assumptions 2.1, 2.4, and 2.5, assume that the initial policy  $\pi^0$  has full support, i.e.,  $\min_{(h,s,a) \in [H] \times S \times \mathcal{A}} \pi_h^0(a | s) > 0$ . Then, the sequence  $(\sigma^k)_{k=0}^\infty$  converges to the set  $\Pi^*$  of equilibrium, i.e.,  $\lim_{k \rightarrow \infty} \text{dist}(\sigma^k, \Pi^*) = 0$ , where we set  $\text{dist}(\sigma, \Pi^*) := \inf_{\pi^* \in \Pi^*} \sum_{(h,s) \in [H] \times S} \|\sigma_h(s) - \pi_h^*(s)\|_1$  for  $\sigma \in (\Delta(\mathcal{A})^S)^H$ .

Note that Theorem 3.1 no longer relies on the *strict-monotonicity* imposed in earlier works (Hadikhmaloo and Silva 2019; Elie et al. 2020; Pérolat et al. 2022). Moreover, unlike the continuous-time results of Perrin et al. (2020) and Pérolat et al. (2022), it applies directly to the discrete-time scheme (3.1).

**Proof sketch of Theorem 3.1.** If we accept the next lemma, we can easily prove Theorem 3.1:

**Lemma 3.2.** Suppose Assumption 2.4. Then, for any equilibrium  $(\mu^*, \pi^*)$  it holds that

$$\begin{aligned} D_{\mu^*}(\pi^*, \sigma^{k+1}) - D_{\mu^*}(\pi^*, \sigma^k) &\leq J(\mu^*, \sigma^{k+1}) - J(\mu^*, \pi^*) - D_{\mu^{k+1}}(\sigma^{k+1}, \sigma^k) \\ &\leq J(\mu^*, \sigma^{k+1}) - J(\mu^*, \pi^*). \end{aligned} \quad (3.2)$$

Lemma 3.2 implies that the KL divergence from an equilibrium point to the generated policy becomes smaller as the cumulative reward  $J$  increases. We note that the function  $J(\mu^*, \bullet): (\Delta(\mathcal{A})^S)^H \ni \pi \mapsto J(\mu^*, \pi) \in \mathbb{R}$  is a polynomial, thus real-analytic. Then we apply (Łojasiewicz 1971, §18, Théorème 2) and find that there exist positive constants  $\alpha$  and  $C$  satisfying  $J(\mu^*, \pi) - J(\mu^*, \pi^*) \leq -C(\text{dist}(\pi, \Pi^*))^\alpha$ , for any  $\pi \in (\Delta(\mathcal{A})^S)^H$ . Combining the above two inequalities yields that  $D_{\mu^*}(\pi^*, \sigma^{k+1}) - D_{\mu^*}(\pi^*, \sigma^k) \leq -C(\text{dist}(\sigma^{k+1}, \Pi^*))^\alpha$ . Thus, the telescoping sum of this inequality yields  $\sum_{k=1}^\infty (\text{dist}(\sigma^k, \Pi^*))^\alpha \leq \frac{D_{\mu^*}(\pi^*, \sigma^0)}{C} < +\infty$ , which implies  $\lim_{k \rightarrow \infty} \text{dist}(\sigma^k, \Pi^*) = 0$ .  $\square$

**Remark 3.3** (Challenges in the proof of Theorem 3.1). The technical difficulty in the proof lies in the term  $D_{\mu^{k+1}}(\sigma^{k+1}, \sigma^k)$  in (3.2). If it were not dependent on  $\mu$ , that is,  $D_{\mu^{k+1}} = D_{\mu^*}$ , then LIC would follow straightforwardly from  $D_{\mu^*}(\pi^*, \sigma^{k+1}) - D_{\mu^*}(\pi^*, \sigma^k) \leq -D_{\mu^*}(\sigma^{k+1}, \sigma^k)$ , where we use Definition 2.3 and the second line of (3.2). However,  $D_{\mu^{k+1}}$  changes depending on  $k$ . Therefore, in the above proof, we have made a special effort to avoid using  $D_{\mu^{k+1}}(\sigma^{k+1}, \sigma^k)$ . One may have seen proofs employing the simple argument described above in games other than MFG, such as monotone games (Rosen 1965). The reason why such an argument is possible in monotone games is that the mean field  $\mu$  does not appear. This difference makes it difficult to use the straightforward argument described above in MFGs.

## 4 Approximating proximal point with mirror descent in regularized MFG

As in the PP method, it is necessary to find  $(\mu^{k+1}, \sigma^{k+1})$  at each iteration. However, it is difficult to exactly compute  $(\mu^{k+1}, \sigma^{k+1})$  due to the implicit nature of (3.1). Therefore, this section introduces Regularized Mirror Descent (RMD), which approximates the solution  $(\mu^{k+1}, \sigma^{k+1})$  for each policy  $\sigma^k$ . The novel result in this section is that the divergence between the sequence generated by RMD and the equilibrium decays exponentially as shown in Figure 1.

### 4.1 Approximation of the update rule of PP with regularized MFG

Interestingly, solving (3.1) corresponds to finding an equilibrium for *KL-regularized MFG* introduced in Cui and Koepl (2021) and F. Zhang et al. (2023). We review the settings for the regularized MFG. For each parameter  $\lambda > 0$  and policy  $\sigma \in (\Delta(\mathcal{A})^S)^H$ , which plays the role of  $\sigma^k$  in (3.1), we define the *regularized cumulative reward*  $J^{\lambda, \sigma}(\mu, \pi)$  for  $(\mu, \pi) \in \Delta(S)^H \times (\Delta(\mathcal{A})^S)^H$  to be

$$J^{\lambda, \sigma}(\mu, \pi) := J(\mu, \pi) - \lambda D_{m[\pi]}(\mu, \sigma). \quad (4.1)$$

The assumption of full support is also imposed on  $\sigma$ :

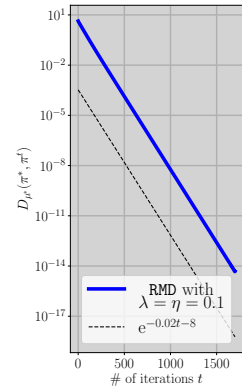


Figure 1: Behavior of RMD.

**Assumption 4.1.** The base  $\sigma$  has full support, i.e., the minimum value given by

$$\sigma_{\min} := \min_{(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}} \sigma_h(a | s)$$

is strictly positive.

For the reward  $J^{\lambda,\sigma}$ , we introduce a *regularized equilibrium*:

**Definition 4.2.** A pair  $(\mu^*, \varpi^*) \in \Delta(\mathcal{S})^H \times (\Delta(\mathcal{A})^S)^H$  is *regularized equilibrium* of  $J^{\lambda,\sigma}$  if it satisfies (i)  $J^{\lambda,\sigma}(\mu^*, \varpi^*) = \max_{\pi \in \Delta(\mathcal{S})^H} J^{\lambda,\sigma}(\mu^*, \pi)$ , and (ii)  $\mu^* = m[\varpi^*]$ .

Specifically,  $(\mu^{k+1}, \sigma^{k+1})$  can be characterized as the regularized equilibrium of  $J^{\lambda,\sigma^k}$  for  $k \in \mathbb{N}$ . Note that the equilibrium is unique under [Assumption 4.1](#), see [Appendix C](#).

In the next subsection, we will introduce RMD using *value functions*, which are defined as follows: for each  $h \in [H]$ ,  $s \in \mathcal{S}$ ,  $a \in \mathcal{A}$ ,  $\mu \in \Delta(\mathcal{S})^H$  and  $\pi \in \Delta(\mathcal{A})^S$ , define the *state value function*  $V_h^{\lambda,\sigma}: \mathcal{S} \times \Delta(\mathcal{S})^H \times (\Delta(\mathcal{A})^S)^H \rightarrow \mathbb{R}$  and the *state-action value function*  $Q_h^{\lambda,\sigma}: \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})^H \times (\Delta(\mathcal{A})^S)^H \rightarrow \mathbb{R}$  as

$$V_h^{\lambda,\sigma}(s, \mu, \pi) := \mathbb{E}_{((s_l, a_l))_{l=h}^H} \left[ \sum_{l=h}^H (r_l(s_l, a_l, \mu_l) - \lambda D_{\text{KL}}(\pi_l(s_l), \sigma_l(s_l))) \right], \quad V_{H+1}^{\lambda,\sigma} \equiv 0, \quad (4.2)$$

$$Q_h^{\lambda,\sigma}(s, a, \mu, \pi) := r_h(s, a, \mu_h) + \mathbb{E}_{s_{h+1} \sim P(s, a, \mu_h)} [V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi)]. \quad (4.3)$$

Here, the discrete-time stochastic process  $((s_l, a_l))_{l=h}^H$  is induced recursively by  $s_h = s$  and  $s_{l+1} \sim P_l(s_l, a_l)$ ,  $a_l \sim \pi_l(s_l)$  for each  $l \in \{h, \dots, H-1\}$  and  $a_H \sim \pi_H(s_H)$ . Note that the objective function  $J^{\lambda,\sigma}$  in [Definition 4.2](#) can be expressed as  $J^{\lambda,\sigma}(\mu, \pi) = \mathbb{E}_{s \sim \mu_1} [V_1^{\lambda,\sigma}(s, \mu, \pi)]$ .

## 4.2 An exponential convergence result

In this subsection, we introduce the iterative method for finding the regularized equilibrium proposed by F. Zhang et al. (2023) as RMD. The method constructs a sequence  $((\pi^t, \mu^t))_{t=0}^\infty \subset (\Delta(\mathcal{A})^S)^H \times \Delta(\mathcal{S})^H$  approximating the regularized equilibrium of  $J^{\lambda,\sigma}$  using the following rule:

$$\begin{aligned} \pi_h^{t+1}(s) &= \arg \max_{p \in \Delta(\mathcal{A})} \left\{ \frac{\eta}{1 - \lambda\eta} \left( \left\langle Q_h^{\lambda,\sigma}(s, \bullet, \pi^t, \mu^t), p \right\rangle - \lambda D_{\text{KL}}(p, \sigma_h(s)) \right) - D_{\text{KL}}(p, \pi_h^t(s)) \right\}, \\ \mu^{t+1} &= m[\pi^{t+1}], \end{aligned} \quad (4.4)$$

where  $\eta > 0$  is another learning rate, and  $Q_h^{\lambda,\sigma}$  is the state-action value function defined in (4.3). We give the pseudo-code of RMD in [Algorithm 2](#). For the sequence of policies in RMD, we can establish the convergence result as follows:

**Theorem 4.3.** Let  $((\mu^t, \pi^t))_{t=0}^\infty \subset \Delta(\mathcal{S})^H \times (\Delta(\mathcal{A})^S)^H$  be the sequence generated by (4.4), and  $(\mu^*, \varpi^*) \in \Delta(\mathcal{S})^H \times (\Delta(\mathcal{A})^S)^H$  be the regularized equilibrium given in [Definition 4.2](#). In addition to [Assumptions 2.4, 2.5, and 4.1](#), suppose that  $\eta \leq \eta^*$ , where  $\eta^* > 0$  is the upper bound of the learning rate defined in (D.5), which only depends on  $\lambda, \sigma, H$  and  $|\mathcal{A}|$ . Then, the sequence  $(\pi^t)_{t=0}^\infty$  satisfies that for  $t \in \mathbb{N}$

$$D_{\mu^*}(\varpi^*, \pi^{t+1}) \leq \left(1 - \frac{\lambda\eta}{2}\right) D_{\mu^*}(\varpi^*, \pi^t),$$

which leads  $D_{\mu^*}(\varpi^*, \pi^t) \leq D_{\mu^*}(\varpi^*, \pi^0) e^{-\lambda\eta t/2}$ . Clearly, the inequality states that an approximate policy  $\pi^t$  satisfying  $D_{\mu^*}(\varpi^*, \pi^t) < \varepsilon$  can be obtained in  $\mathcal{O}(\log(1/\varepsilon))$  iterations.

**Remark 4.4.** While [Theorem 4.3](#) provides an exponentially decreasing bound, the theoretical upper bound  $\eta^*$  on the step size  $\eta$  can be small, see (D.4) and (D.5) in detail.

This metric  $D_{\mu^*}(\varpi^*, \pi^t)$  is widely used in F. Zhang et al. (2023) and Dong et al. (2025) because it provides an upper bound for the so-called exploitability  $\text{Exploit}(\pi) := \max_{\pi'} J(m[\pi], \pi') - J(m[\pi], \pi)$  as  $\text{Exploit}(\pi^t) = \mathcal{O}(\sqrt{D_{\mu^*}(\varpi^*, \pi^t)})$  by the Lipschitz continuity of  $V_1^{\lambda,\sigma}$ . We also note that [Theorem 4.3](#) improves upon the previous results by F. Zhang et al. (2023) and Dong et al. (2025) in the regime with a large number of iterations  $t$ . Indeed, the authors obtained  $D_{\mu^*}(\varpi^*, \frac{1}{T} \sum_{t=1}^T \pi^t) \leq$



$\mathcal{O}(\lambda \log^2 T / \sqrt{T})$  and  $D_{\mu^*}(\varpi^*, \pi^{t+1}) \leq H^3 / \lambda t$ . On the other hand, these bounds for finite  $t$  may be smaller since the constant  $\eta$  inside our exponent could be small.

### 4.3 Intuition for exponential convergence: continuous-time version of RMD

The convergence of  $(\pi^t)_{t=0}^\infty$  can be intuitively explained by considering a continuous limit  $(\pi^t)_{t \geq 0}$  with respect to the time  $t$  of RMD. In this paragraph, we will use the idea of mirror flow (Krichene et al. 2015; Tzen et al. 2023; Deb et al. 2023) and continuous dynamics in games (Taylor and Jonker 1978; Mertikopoulos et al. 2018; Pérolat et al. 2021; Pérolat et al. 2022) to observe the exponential convergence of the flow to equilibrium. According to Deb et al. (2023, (2.1)), the continuous curve of  $\pi$  should satisfy that

$$\frac{d}{dt} \pi_h^t(a | s) = \pi_h^t(a | s) \cdot \left( Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right). \quad (4.5)$$

The flow induced by the dynamical system (4.5) converges to equilibrium *exponentially* as time  $t$  goes to infinity.

**Theorem 4.5.** *Let  $\pi^t$  be a solution of (4.5) and  $\varpi^*$  be a regularized equilibrium defined in Definition 4.2. Suppose that Assumption 2.4. Then*

$$\frac{d}{dt} D_{\mu^*}(\varpi^*, \pi^t) \leq -\lambda D_{\mu^*}(\varpi^*, \pi^t),$$

*for all  $t \geq 0$ . Moreover, the inequality implies  $D_{\mu^*}(\varpi^*, \pi^t) \leq D_{\mu^*}(\varpi^*, \pi^0) \exp(-\lambda t)$ .*

Technically, the non-Lipschitz continuity of the value function  $Q_h^{\lambda, \sigma}(s, a, \bullet, \mu^t)$  in the right-hand side of (4.5) is non-trivial for the existence of the solution  $\pi: [0, +\infty) \rightarrow (\Delta(\mathcal{A})^S)^H$  of the differential equation (4.5), see, e.g., Coddington and Levinson (1984). The proof of this existence and Theorem 4.5 are given in Appendix C.

### 4.4 Proof sketch of the convergence result for RMD

We return from continuous-time dynamics (4.5) to the discrete-time algorithm (4.4). The technical difficulty in the proof of Theorem 4.3 is the non-Lipschitz continuity of the value function  $Q_h^{\lambda, \sigma}$  in (4.4), that is, the derivative of  $Q_h^{\lambda, \sigma}(s, a, \pi, \mu)$  with respect to the policy  $\pi$  can blow up as  $\pi$  approaches the boundary of the space  $(\Delta(\mathcal{A})^S)^H$  of probability simplices. We can overcome this difficulty as shown in the following sketch of proof:

**Proof sketch of Theorem 4.3.** In a similar way to Theorem 4.5, we can obtain the following inequality with a discretization error:

$$D_{\mu^*}(\varpi^*, \pi^{t+1}) - D_{\mu^*}(\varpi^*, \pi^t) \leq -\lambda \eta D_{\mu^*}(\varpi^*, \pi^t) + D_{\mu^*}(\pi^t, \pi^{t+1}), \quad (4.6)$$

where we use a property of KL divergence, see the proof in Appendix D. The remainder of the proof is almost entirely dedicated to showing that the above error term is sufficiently small and bounded compared to the other terms in (4.6). As a result, we obtain the following claim:

**Claim 4.6.** *Suppose that the learning rate  $\eta$  is less than the upper bound  $\eta^*$  in (D.5). Then*

$$D_{\mu^*}(\pi^t, \pi^{t+1}) \leq C \eta^2 D_{\mu^*}(\varpi^*, \pi^t),$$

*where  $C > 0$  is the constant defined in (D.4), which satisfies  $C \eta^* \leq \lambda/2$ .*

The key to proving Claim 4.6 is leveraging another claim that, over the sequence  $(\pi^t)_t$ , the value function  $Q_h^{\lambda, \sigma}$  behaves well, almost as if it were a Lipschitz continuous function, see Lemma D.3 for details. Therefore, applying Claim 4.6 to (4.6) completes the proof.  $\square$

**Remark 4.7** (Challenges in the proof of Theorem 4.3). The technical difficulty in the proof lies in the fact that the  $Q$ -function  $Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t)$  in the algorithm (4.4) depends on the mean field  $\mu^t = m[\pi^t]$ , which is determined forward by (2.1) from *past* times 1 to  $h-1$ . On the other hand,

---

**Algorithm 1:** APP for MFG

---

**Input:** MFG( $\mathcal{S}, \mathcal{A}, H, P, r, \mu_1$ ), initial policy  $\pi^0$ , #iterations  $N$ ,  $\lambda > 0$

- 1 **Initialization:** Set  $k \leftarrow 0$ ,  $\sigma^k \leftarrow \pi^0$ ;
- 2 **while**  $k < N$  **do**
- 3     Compute  $(\mu^{k+1}, \sigma^{k+1})$  by solving

$$\begin{cases} \sigma^{k+1} = \text{RMD}(\text{MFG}, \sigma^k, \lambda, \eta, \sigma^k, \tau), \\ \mu^{k+1} = m[\sigma^{k+1}] \end{cases}$$

Update  $k \leftarrow k + 1$ ;

**Output:**  $\sigma^k (\approx \pi^*)$

---

---

**Algorithm 2:** RMD(MFG,  $\pi^0, \lambda, \eta, \sigma^0, \tau$ )

---

- 1 **Initialization:** Set  $t \leftarrow 0$ ,  $\pi^t \leftarrow \pi^0$ ,  
 $\sigma \leftarrow \sigma^0$ ;
  - 2 **while**  $t < \tau$  **do**
  - 3     Compute  $\mu^t = m[\pi^t]$ ;
  - 4     Compute  $Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t)$  by (4.3);
  - 5     Compute  $\pi^{t+1}$  as  

$$\pi_h^{t+1}(a | s) \propto (\sigma_h(a | s))^{\lambda \eta} (\pi_h^t(a | s))^{1 - \lambda \eta} \cdot \exp(\eta Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t))$$
  - 6     Update  $t \leftarrow t + 1$ ;
  - 7 **return**  $\pi^t (\approx \varpi^*)$
- 

the  $Q$ -function is also determined by the policy from *future* times  $h + 1$  to  $H$  through the dynamic programming principle given by (4.3). As a result, it becomes difficult to apply the backward induction argument, which is known in the context of MDPs and Markov games, to  $Q$ -functions. This difficulty is specific to MFGs and is not seen in other regularized games such as entropy-regularized zero-sum Markov games, where the  $Q$ -function depends only on future policies. Therefore, it is less feasible to directly apply the techniques of existing research, such as Cen et al. (2023), to RMD for MFGs. Our proof instead utilizes the properties of the KL divergence to deal with this difficulty.

#### 4.5 APP: approximating PP updates with RMD

We recall that we need to develop an algorithm that efficiently approximates the update rule of the PP method since the rule (3.1) is intractable. To this end, we employ the *regularized* Mirror Descent (RMD) to solve the (*unregularized*) MFG as a substitute for the rule. Specifically, after repeating the RMD iteration (4.4) a sufficient number of times, we update the base distribution  $\sigma$  using the most recently obtained policy  $\sigma^{k+1}$ . We call this method APP, which is summarized in Algorithm 1. In APP, updating the base seems like a small modification of RMD, but it is crucial for convergence. Without this update, we can only obtain regularized equilibria, which are generally different from our ultimate goal of unregularized equilibria. In fact, Definition 2.3, 4.2 and Assumption 2.4 yield that  $J(\mu^*, \pi^*) - J(\mu^*, \varpi^*) \leq \lambda(D_{\mu^*}(\pi^*, \sigma) - D_{\mu^*}(\varpi^*, \sigma))$ , which roughly implies that the gap between regularized and unregularized equilibria is  $\mathcal{O}(\lambda)$ . Experimental results in Cui and Koepl (2021) also suggest that to find the (unregularized) equilibrium with a regularized algorithm, it is necessary to tune the hyperparameter  $\lambda$  appropriately. Theoretically, the results we have established in Theorems 3.1 and 4.3 provide some convergence guarantees for APP. Empirically, the experimental results in the next section suggest that APP also achieves LIC. We conjecture that the rate of convergence for APP, as predicted by these experiments, may also be derived.

## 5 Numerical experiment

We numerically demonstrate that APP, which is the approximated version of (3.1), can achieve convergence to the mean-field equilibrium. We evaluate the convergence of APP using the Beach Bar Process introduced by Perrin et al. (2020), a standard benchmark for MFGs. In particular, the transition kernel  $P$  in this benchmark gives a random walk on a one-dimensional discretized torus  $\mathcal{S} = \{0, \dots, |\mathcal{S}| - 1\}$ , and the reward is set to be  $r_h(s, a, \mu) = -|a|/|\mathcal{S}| - |s - |\mathcal{S}|/2|/|\mathcal{S}| - \log \mu_h(s)$  with  $a \in \mathcal{A} := \{-1, \pm 0, +1\}$ . Note that this benchmark satisfies the monotonicity assumption in Assumption 2.4. See Appendix F for further details. Figure 2 is a summary of the results of the experiment. The most notable aspect is the convergence of exploitability, as shown in Figure 2b. APP decreases the exploitability with each iteration when we update  $\sigma$ . Figure 2a and 2c illustrate the qualitative validity of the approximation achieved by APP. In this benchmark, the equilibrium is expected to lie at the vertices of the probability simplex. Therefore, RMD, which can shift the equilibrium to the interior of the probability simplex, seems unable to find the mean-field equilibrium



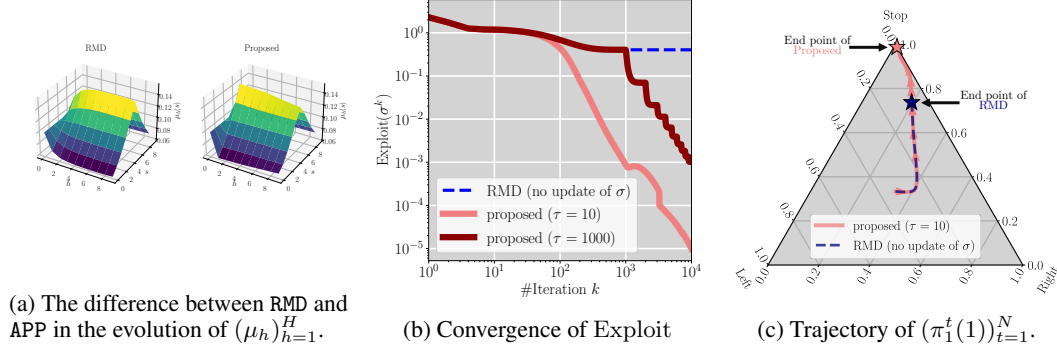


Figure 2: Experimental results for Algorithm 1 for Beach Bar Process

accurately. On the other hand, the sequence  $(\pi^t)_t$  of policies generated by APP shows a behavior that converges to the vertices.

## 6 Limitations

Our results provide the first asymptotic (Theorem 3.1) and exponential (Theorem 4.5) convergence guarantees for PP and RMD in non-strictly monotone (unregularized) MFGs under the model-based setting, assuming that the transition kernels and reward functions are available. This leaves open several important questions. First, we do not consider the more realistic scenario in which the transition of mean-field and reward must be learned from data, nor do we provide any sample-complexity or statistical guarantees, such as those in J. Huang et al. (2024), which would be required for rigorous model-free or data-driven applications. Second, our theoretical advantages are strictly in terms of iteration complexity under monotonicity: we establish faster convergence rates per iteration, but we do not claim any improvements in overall computational cost (for example, the cost of solving each PP subproblem or evaluating  $Q$  in RMD), nor do we analyze how these methods scale with large state or action spaces in practice. Finally, although the proximal-point and mirror-descent structure of PP and RMD makes them, in principle, compatible with nonlinear function approximators, such as NNs, we have not studied approximation errors as in F. Zhang et al. (2023), stability issues, or empirical performance in high-dimensional or highly nonlinear settings.

Establishing LIC for APP remains open. We conjecture that a resolution will require proving a *uniform positive lower bound* on  $\eta^*$  that guarantees LIC for RMD, which will improve the current estimate; see Remark 4.4.

By *synchronous feedback* we mean that, at (outer/inner) iterate  $k$ , all updates use the  $Q$ -function evaluated on the *current* pair  $(\pi^k, \mu^k)$ . In practice, feedback may be delayed or asynchronous. A natural adaptation is to evaluate  $Q$  at a stale iterate, e.g.,  $Q(\pi^{\kappa^{(t)}}, \mu^{\kappa^{(t)}})$  at a delayed index  $\kappa^{(t)}$ . By analogy with asynchronous gradient play in zero-sum games (Ao et al. 2023), we expect last-iterate stability to persist under bounded staleness with a suitably reduced stepsize. A complete analysis in the MFG setting is left for future work.

## 7 Related works

As a result of the focus on the modeling potential of various population dynamics, there has been a significant increase in the literature on computations of equilibria in large-scale MFG, or so-called Learning in MFGs. We refer readers to read (Laurière et al. 2024) as a comprehensive survey of Learning in MFGs. Guo et al. (2019) and Anaharci et al. (2020) developed a fixed-point iteration that alternately updates the mean-field  $\mu$  and policy  $\pi$ , based on the algorithm of MDPs. They showed that this fixed-point iteration achieves LIC under a condition of contraction. However, it is known that the condition of contraction does not hold for many games in Cui and Koepl (2021). In MFGs where the contraction assumption does not hold, it is observed that the fixed-point iteration oscillates in the case of linear-quadratic MFGs (Laurière 2021). Fictitious play, which averages mean fields or policies over time, was developed to prevent this oscillation. Hadikhanloo and Silva (2019), Elie et

al. (2020), and Perrin (2022) showed that the average in fictitious play converges to an equilibrium under the monotonicity assumption in [Assumption 2.4](#). On the other hand, such time averaging has the disadvantage of slowing the experimental rate of convergence observed in Laurière et al. (2024) and making it difficult to scale up using deep learning.

Pérolat et al. (2022) applied Mirror Descent to MFG and developed a scalable method. This method has the practical benefits of being compatible with deep learning and is applicable to variants of variants (Laurière et al. 2022; Fabian et al. 2023). However, the theoretical guarantees are somewhat restrictive, as they often require strong assumptions like contraction for last-iterate convergence. In fact, they showed last-iterate convergence (LIC) of continuous-time algorithms under *strict* monotonicity assumptions. However, results for discrete-time settings or non-strict monotonicity are lacking. In addition to fictitious play and MD, methods using the actor-critic method (Zeng et al. 2024), value iteration (Anahtarci et al. 2020), multi-time scale (Angiuli et al. 2022; Angiuli et al. 2023; Angiuli et al. 2024) and semi-gradient method (C. Zhang et al. 2025) have been developed, but to the best of our knowledge, the theoretical convergence results of these methods require a condition of contraction. See the upper part of [Table 1](#) for details.

Rather than focusing on the algorithm explained above, Cui and Koepl (2021) focused on the problem setting of MFG and aimed to achieve a fast convergence of the algorithms by considering regularization of MFG. This type of regularization is typical in the case of MDPs and two-player zero-sum Markov games, where Mirror Descent achieves exponential convergence (Zhan et al. 2021; Cen et al. 2023). One expects similar convergence results for regularized MFGs, but the fast convergence results without strong assumptions have been limited so far. F. Zhang et al. (2023) and Dong et al. (2025) demonstrated polynomial convergence rates for MD under monotonicity. In addition, the authors in Q. Xie et al. (2021), Mao et al. (2022), Cui and Koepl (2021), and Anahtarci et al. (2023) develop an algorithm that converges polynomially for regularized MFG, and they impose restrictive assumptions such as contraction and strict monotonicity. [Appendix A](#) provides an extensive review comparing existing results in Learning in MFGs.

## 8 Conclusion

This paper proposes the novel method to achieve LIC under the monotonicity ([Assumption 2.4](#)). The main idea behind the derivation of the method is to approximate the PP type method (3.1) using RMD. [Theorem 3.1](#) implies that the PP method achieves LIC, and [Theorem 4.3](#) establish the exponential convergence of RMD. A future task of this study is to prove the convergence rates of the combined method, APP.

## Acknowledgments and Disclosure of Funding

The first author was supported by JSPS KAKENHI Grant Numbers JP22KJ1002 and JP25H01453, as well as by JST ACT-X Grant Number JPMJAX25C2. Kaito Ariu was supported by JSPS KAKENHI Grant Numbers 23K19986 and 25K21291.

## References

- K. Abe, K. Ariu, M. Sakamoto, and A. Iwasaki (2024). “Adaptively Perturbed Mirror Descent for Learning in Games”. In: *ICML*.
- K. Abe, K. Ariu, M. Sakamoto, K. Toyoshima, and A. Iwasaki (2023). “Last-Iterate Convergence with Full and Noisy Feedback in Two-Player Zero-Sum Games”. In: *AISTATS*. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 7999–8028.
- A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun (2022). *Reinforcement Learning: Theory and Algorithms*. online.
- B. Anahtarci, C. D. Kariksiz, and N. Saldi (2020). “Value iteration algorithm for mean-field games”. In: *Syst. Control. Lett.* 143, p. 104744.
- (2023). “Q-Learning in Regularized Mean-field Games”. In: *Dynamic Games and Applications* 13.1, pp. 89–117.
- A. Angiuli, J. Fouque, and M. Laurière (2022). “Unified reinforcement Q-learning for mean field game and control problems”. In: *Math. Control. Signals Syst.* 34.2, pp. 217–271.

- A. Angiuli, J. Fouque, M. Laurière, and M. Zhang (2023). “Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games”. In: *CoRR* abs/2312.06659.
- (2024). “Analysis of Multiscale Reinforcement Q-Learning Algorithms for Mean Field Control Games”. In: *CoRR* abs/2405.17017.
- R. Ao, S. Cen, and Y. Chi (2023). “Asynchronous Gradient Play in Zero-Sum Multi-agent Games”. In: *ICLR*.
- D. P. Bertsekas and S. E. Shreve (1978). *Stochastic optimal control*. Vol. 139. Mathematics in Science and Engineering. The discrete time case. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, pp. xiii+323.
- G. W. Brown (1951). “Iterative Solution of Games by Fictitious Play”. In: *Activity Analysis of Production and Allocation*. Ed. by T. C. Koopmans. New York: Wiley.
- P. E. Caines and M. Huang (2019). “Graphon Mean Field Games and the GMFG Equations:  $\epsilon$ -Nash Equilibria”. In: *CDC. IEEE*, pp. 286–292.
- P. Cardaliaguet and S. Hadikhanloo (2017). “Learning in mean field games: The fictitious play”. In: *ESAIM: COCV* 23.2, pp. 569–591.
- S. Cen, Y. Chi, S. S. Du, and L. Xiao (2023). “Faster Last-iterate Convergence of Policy Optimization in Zero-Sum Markov Games”. In: *ICLR*.
- Y. Censor and S. A. Zenios (1992). “Proximal minimization algorithm with D-functions”. In: *Journal of Optimization Theory and Applications* 73.3, pp. 451–464.
- R. Chill, E. Fasangova, and U. Fakulta (2010). “Gradient Systems”. In: *13th International Internet Seminar*.
- A. Coddington and N. Levinson (1984). *Theory of Ordinary Differential Equations*. International series in pure and applied mathematics. R.E. Krieger.
- K. Cui and H. Koepl (2021). “Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning”. In: *AISTATS*. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 1909–1917.
- (2022). “Learning Graphon Mean Field Games and Approximate Nash Equilibria”. In: *ICLR*.
- K. Cui, A. Tahir, G. Ekinci, A. Elshamhory, Y. Eich, M. Li, and H. Koepl (2022). “A Survey on Large-Population Systems and Scalable Multi-Agent Reinforcement Learning”. In: *CoRR* abs/2209.03859.
- N. Deb, Y.-H. Kim, S. Pal, and G. Schiebinger (2023). *Wasserstein Mirror Gradient Flow as the limit of the Sinkhorn Algorithm*.
- J. Dong, B. Wang, and Y. Yu (2025). “Last-iterate Convergence in Regularized Graphon Mean Field Game”. In: *AAAI*. AAAI Press, pp. 13779–13787.
- R. Elie, J. Pérolat, M. Laurière, M. Geist, and O. Pietquin (2020). “On the Convergence of Model Free Learning in Mean Field Games”. In: *AAAI*, pp. 7143–7150.
- C. Fabian, K. Cui, and H. Koepl (2023). “Learning Sparse Graphon Mean Field Games”. In: *AISTATS*. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 4486–4514.
- (2024). “Learning Mean Field Games on Sparse Graphs: A Hybrid Graphex Approach”. In: *ICLR*.
- S. Gao and P. E. Caines (2017). “The control of arbitrary size networks of linear systems via graphon limits: An initial investigation”. In: *CDC. IEEE*, pp. 1052–1057.
- M. Geist, J. Pérolat, M. Laurière, R. Elie, S. Perrin, O. Bachem, R. Munos, and O. Pietquin (2022). “Concave Utility Reinforcement Learning: The Mean-field Game Viewpoint”. In: *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), pp. 489–497.
- M. Geist, B. Scherrer, and O. Pietquin (2019). “A Theory of Regularized Markov Decision Processes”. In: *ICML*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2160–2169.
- D. A. Gomes, J. Mohr, and R. R. Souza (2010). “Discrete time, finite state space mean field games”. In: *Journal de Mathématiques Pures et Appliquées* 93.3, pp. 308–328.
- X. Guo, A. Hu, R. Xu, and J. Zhang (2019). “Learning Mean-Field Games”. In: *NeurIPS*, pp. 4967–4977.
- X. Guo, A. Hu, and J. Zhang (2024). “MF-OMO: An Optimization Formulation of Mean-Field Games”. In: *SIAM Journal on Control and Optimization* 62.1, pp. 243–270.
- S. Hadikhanloo and F. J. Silva (2019). “Finite Mean Field Games: Fictitious play and convergence to a first order continuous mean field game”. In: *Journal de Mathématiques Pures et Appliquées* 132, pp. 369–397.
- A. Hu and J. Zhang (2024). “MF-OML: Online Mean-Field Reinforcement Learning with Occupation Measures for Large Population Games”. In: *CoRR* abs/2405.00282.

- J. Huang, B. Yardim, and N. He (2024). “On the Statistical Efficiency of Mean-Field Reinforcement Learning with General Function Approximation”. In: *AISTATS*. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 289–297.
- M. Huang, R. P. Malhamé, and P. E. Caines (2006). “Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle”. In: *Communications in Information & Systems* 6.3, pp. 221–252.
- D. Inoue, Y. Ito, T. Kashiwabara, N. Saito, and H. Yoshida (2023). “A fictitious-play finite-difference method for linearly solvable mean field games”. In: *ESAIM: M2AN* 57.4, pp. 1863–1892.
- W. Krichene, A. M. Bayen, and P. L. Bartlett (2015). “Accelerated Mirror Descent in Continuous and Discrete Time”. In: *NIPS*, pp. 2845–2853.
- J.-M. Lasry and P.-L. Lions (2007). “Mean field games”. In: *Jpn. J. Math.* 2.1, pp. 229–260.
- M. Laurière (2021). “Numerical methods for mean field games and mean field type control”. In: *Mean field games*. Vol. 78. Proc. Sympos. Appl. Math. Amer. Math. Soc., Providence, RI, pp. 221–282.
- M. Laurière, S. Perrin, M. Geist, and O. Pietquin (2024). “Learning in Mean Field Games: A Survey”. In: *CoRR* abs/2205.12944v4.
- M. Laurière, S. Perrin, S. Girgin, P. Muller, A. Jain, T. Cabannes, G. Piliouras, J. Perolat, R. Elie, O. Pietquin, and M. Geist (2022). “Scalable Deep Reinforcement Learning Algorithms for Mean Field Games”. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 12078–12095.
- P. Lavigne and L. Pfeiffer (2023). “Generalized Conditional Gradient and Learning in Potential Mean Field Games”. In: *Applied Mathematics & Optimization* 88.3, p. 89.
- S. Leonardos, G. Piliouras, and K. Spendlove (2021). “Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality”. In: *NeurIPS*, pp. 26318–26331.
- S. Łojasiewicz (1971). “Sur les ensembles semi-analytiques.” In: *Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 2*, pp. 237–241.
- W. Mao, H. Qiu, C. Wang, H. Franke, Z. Kalbarczyk, R. K. Iyer, and T. Basar (2022). “A Mean-Field Game Approach to Cloud Resource Management with Function Approximation”. In: *NeurIPS*.
- P. Mertikopoulos, C. Papadimitriou, and G. Piliouras (2018). “Cycles in Adversarial Regularized Learning”. In: *SODA*, pp. 2703–2717.
- J. Pérolat, R. Munos, J. Lespiau, S. Omidshafiei, M. Rowland, P. A. Ortega, N. Burch, T. W. Anthony, D. Balduzzi, B. D. Vyllder, G. Piliouras, M. Lanctot, and K. Tuyls (2021). “From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization”. In: *ICML*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8525–8535.
- J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin (2022). “Scaling Mean Field Games by Online Mirror Descent”. In: *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), pp. 1028–1037.
- S. Perrin (2022). “Scaling up Multi-agent Reinforcement Learning with Mean Field Games and Vice-versa”. PhD thesis. Université de Lille.
- S. Perrin, M. Laurière, J. Pérolat, R. Elie, M. Geist, and O. Pietquin (2022). “Generalization in Mean Field Games by Learning Master Policies”. In: *AAAI*. AAAI Press, pp. 9413–9421.
- S. Perrin, J. Pérolat, M. Laurière, M. Geist, R. Elie, and O. Pietquin (2020). “Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications”. In: *NeurIPS*.
- G. Piliouras, R. Sim, and S. Skoulakis (2022). “Beyond Time-Average Convergence: Near-Optimal Uncoupled Online Learning via Clairvoyant Multiplicative Weights Update”. In: *NeurIPS*.
- M. L. Puterman (1994). *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, pp. xx+649.
- J. B. Rosen (1965). “Existence and uniqueness of equilibrium points for concave  $N$ -person games”. In: *Econometrica* 33, pp. 520–534.
- N. Saldi, T. Başar, and M. Raginsky (2018). “Markov–Nash Equilibria in Mean-Field Games with Discounted Cost”. In: *SIAM Journal on Control and Optimization* 56.6, pp. 4256–4287.
- P. D. Taylor and L. B. Jonker (1978). “Evolutionary stable strategies and game dynamics”. In: *Mathematical Biosciences* 40.1, pp. 145–156.
- B. Tzen, A. Raj, M. Raginsky, and F. R. Bach (2023). “Variational Principles for Mirror Descent and Mirror Langevin Dynamics”. In: *IEEE Control. Syst. Lett.* 7, pp. 1542–1547.
- Q. Xie, Z. Yang, Z. Wang, and A. Minca (2021). “Learning While Playing in Mean-Field Games: Convergence and Optimality”. In: *ICML*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 11436–11447.

- Y. Xie, X. Wang, R. Wang, and H. Zha (2019). “A Fast Proximal Point Method for Computing Exact Wasserstein Distance”. In: *UAI*. Vol. 115. Proceedings of Machine Learning Research. AUAI Press, pp. 433–453.
- Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang (2018). “Mean Field Multi-Agent Reinforcement Learning”. In: *ICML*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 5567–5576.
- Y. Yang and J. Wang (2020). “An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective”. In: *CoRR* abs/2011.00583.
- B. Yardim, S. Cayci, M. Geist, and N. He (2023). “Policy Mirror Ascent for Efficient and Independent Learning in Mean Field Games”. In: *ICML*. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 39722–39754.
- B. Yardim, A. Goldman, and N. He (2024). “When is Mean-Field Reinforcement Learning Tractable and Relevant?”. In: *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems / ACM, pp. 2038–2046.
- B. Yardim and N. He (2024). “Exploiting Approximate Symmetry for Efficient Multi-Agent Reinforcement Learning”. In: *CoRR* abs/2408.15173.
- M. A. u. Zeman, A. Koppel, S. Bhatt, and T. Basar (2023). “Oracle-free Reinforcement Learning in Mean-Field Games along a Single Sample Path”. In: *AISTATS*. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 10178–10206.
- S. Zeng, S. Bhatt, A. Koppel, and S. Ganesh (2024). “A Policy Optimization Approach to the Solution of Unregularized Mean Field Games”. In: *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control – Connections and Perspectives*.
- W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi (2021). “Policy Mirror Descent for Regularized Reinforcement Learning: A Generalized Framework with Linear Convergence”. In: *CoRR* abs/2105.11066.
- C. Zhang, X. Chen, and X. Di (2025). “Stochastic Semi-Gradient Descent for Learning Mean Field Games with Population-Aware Function Approximation”. In: *ICLR*.
- F. Zhang, V. Y. F. Tan, Z. Wang, and Z. Yang (2023). “Learning Regularized Monotone Graphon Mean-Field Games”. In: *NeurIPS*.
- K. Zhang, Z. Yang, and T. Başar (2021). “Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms”. In: *Handbook of Reinforcement Learning and Control*. Cham: Springer International Publishing, pp. 321–384.

## A Detailed Explanation of Related Works

Table 1: Summary of related work on convergence of iterative methods for MFGs

		Assumption	Discrete time	LIC
MFG	Guo et al. (2019)	Contract.	✓	-
	Elie et al. (2020), Hadikhanloo and Silva (2019)	Strict Mono.	✓	-
	Perrin et al. (2020)	Mono.	-	-
	Anahtarci et al. (2020)	Contract.	✓	✓
	Pérolat et al. (2022)	Strict Mono.	-	✓
	Geist et al. (2022)	Concavity	✓	✓
	Angiuli et al. (2022), Angiuli et al. (2023), Angiuli et al. (2024)	Contract.	✓	-
	Yardim et al. (2023)	Contract.	✓	✓
	Zeng et al. (2024)	Herding	✓	-
	C. Zhang et al. (2025)	Contract.	✓	✓
	Ours (Theorem 3.1)	Mono.	✓	✓
Regularized MFG	Q. Xie et al. (2021)	Contract.	✓	-
	Cui and Koepl (2021)	Contract.	✓	✓
	Mao et al. (2022)	Contract.	✓	-
	Anahtarci et al. (2023)	Contract.	✓	✓
	F. Zhang et al. (2023)	Strict Mono.	✓	-
	Dong et al. (2025)	Mono.	✓	✓ w/ poly. rate
	Ours (Theorem 4.5)	Mono.	✓	✓ w/ exp. rate

### A.1 Comparison with literature on MFGs

Based on Table 1, we will discuss the technical contributions made by this paper in Learning in MFGs below.

**Last-iterate convergence (LIC) results for MFGs:** Pérolat et al. (2022) showed that Mirror Descent achieves LIC only under *strictly* monotone conditions, i.e., if the equality in the Lemma E.2 is satisfied only if  $\pi = \tilde{\pi}$ . In contrast, our work establishes LIC even in *non-strictly* monotone scenarios. While the distinction regarding strictness might seem subtle, it is profoundly significant. Indeed, non-strictly monotone MFGs encompass the fundamental examples of finite-horizon Markov Decision Processes. Moreover, in strictly monotone cases, mean-field equilibria become unique. Consequently, as Zeng et al. (2024) also noted, strictly monotone rewards fail to represent MFGs with diverse equilibria.

**Regularized MFGs:** Theorem 4.3, which supports the efficient execution of RMD, is novel in two respects: RMD achieves LIC, and the divergence to the equilibrium decays exponentially. Indeed, one of the few works that analyze the convergence rate of RMD states that the time-averaged policy  $\frac{1}{T} \sum_{t=0}^T \pi^t$  up to time  $T$  converges to the equilibrium in  $\mathcal{O}(1/\varepsilon^2)$  iterations (F. Zhang et al. 2023). Additionally, although it is a different approach from MD, it is known that applying fixed-point iteration to regularized MFG achieves an exponential convergence rate under the assumption that the regularization parameter  $\lambda$  is sufficiently large (Cui and Koepl 2021). In contrast, our work includes the cases if  $\lambda$  is small with  $\eta < \eta^*$ , where we note that  $\eta^*$  depends on  $\lambda$  though (D.5).



**Optimization-based methods for MFGs:** In addition to Mirror Descent and Fictitious Play, a new type of learning method using the characterization of MFGs as optimization problems has been proposed (Guo et al. 2024; Hu and J. Zhang 2024). In this work, the authors establish local convergence of the algorithms without the assumption of monotonicity. Specifically, it is proven that an optimization method can achieve LIC if the initial guess  $\pi^0$  of the algorithm is sufficiently close to the Nash equilibrium, which cannot be verified a priori. In contrast, our convergence results state “global” convergence under the assumption of monotonicity of the reward. We note that the monotonicity can be checked before running the algorithms to ensure convergence of PP and RMD.

**Mean-field-aware methods for MFGs:** The authors in Zeng et al. (2024) and C. Zhang et al. (2025) have recently developed algorithms that sequentially update not only the policy  $\pi$  but also the mean field  $\mu$  and value function. These algorithms have advantages over conventional methods in terms of computational complexity. On the other hand, in theoretical analysis, restrictive assumptions such as contraction are still being used, and there is room for improvement under the monotonicity assumption.

## A.2 Comparison of MFG and related games

In research on the method of learning in games, regularization of games is often studied in order to improve extrapolation. For example, Geist et al. (2019) gave a unified convergence analysis method for regularized MDPs. (Leonardos et al. 2021) also discussed unique regularized equilibria of weighted zero-sum polymatrix games. On the other hand, it is a difficult task to apply the same theoretical analysis methods to MFG as to these games. In Remarks 3.3 and 4.7, we confirmed that the mean field  $\mu$  in MFG can hinder convergence analysis. In the following two paragraphs, we will describe more specifically the difficulty of applying the methods used in other games to MFG.

**Sequential imperfect information game in Pérolat et al. (2021) vs. MFG:** Pérolat et al. (2021) focused on the reaching probability  $\rho^\pi$  over histories in sequential imperfect information games, or extensive-form games. In contrast, we focused on the distribution of states  $\mu = m[\pi]$  in MFGs. The dependency on  $\pi$  is fundamentally different:  $\rho$  depends on  $\pi$  in a linear-like manner, while our  $\mu$  has a highly nonlinear dependency on  $\pi$  thorough the function  $m$  defined in (2.1). Addressing this nonlinearity required novel techniques exploiting the inductive structure of (2.1) with respect to time  $h$ .

**MDP vs. MFG:** The known argument in Zhan et al. (2021, Lemma 6) cannot be directly applied to MFGs. The main reason is that the inner product  $\langle Q^k(s), \pi^{k+1}(s) - p \rangle$  in the right-hand side of the three-point lemma concerns the policy at iteration index  $k + 1$ , not  $k$ . In our analysis (as shown on page 18), this term is transformed into  $\langle Q^k(s), \pi^k(s) - p \rangle$ , which allows us to apply a crucial lemma (Lemma E.4) that holds for MFGs. This transformation is non-trivial and essential for our analysis. In the three-point lemma, the term  $D_{h_s}(\pi^{(k+1)}, \pi^{(k)})$  appears as a discretization error. In contrast, our analysis derives a reverse version  $D_{\mu^*}(\pi^k, \pi^{k+1})$ . This distinction is significant, especially for non-symmetric divergences such as the KL divergence. The reverse order in our analysis is crucial for the theoretical guarantees we provide.

## B Proof of Theorem 3.1

**Proof of Lemma 3.2.** Let  $(\mu^*, \pi^*)$  be a mean-field equilibrium defined in Definition 2.3. By the update rule (3.1) and Lemma E.1, we have

$$\left\langle Q_h^{\lambda, \sigma^k}(s, \bullet, \sigma^{k+1}, \mu^{k+1}) - \lambda \log \frac{\sigma_h^{k+1}(s)}{\sigma_h^k(s)}, (\pi_h^* - \sigma_h^{k+1})(s) \right\rangle \leq 0,$$

for each  $h \in [H]$ ,  $s \in \mathcal{S}$  and  $k \in \mathbb{N}$ , i.e.,

$$\begin{aligned} & D_{\text{KL}}(\pi_h^*(s), \sigma_h^{k+1}(s)) - D_{\text{KL}}(\pi_h^*(s), \sigma_h^k(s)) - D_{\text{KL}}(\sigma_h^{k+1}(s), \sigma_h^k(s)) \\ & \leq \frac{1}{\lambda} \left\langle Q_h^{\lambda, \sigma^k}(s, \bullet, \sigma^{k+1}, \mu^{k+1}), (\sigma_h^{k+1} - \pi_h^*)(s) \right\rangle. \end{aligned} \quad (\text{B.1})$$

Taking the expectation with respect to  $s \sim \mu_h^*$  and summing (B.1) over  $h \in [H]$  yields

$$D_{\mu^*}(\pi^*, \sigma^{k+1}) - D_{\mu^*}(\pi^*, \sigma^k) + D_{\mu^*}(\sigma^{k+1}, \sigma^k)$$

$$\leq \frac{1}{\lambda} \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle Q_h^{\lambda, \sigma^k}(s, \bullet, \sigma^{k+1}, \mu^{k+1}), (\sigma_h^{k+1} - \pi_h^*)(s) \right\rangle \right].$$

By virtue of [Lemmas E.2](#) and [E.4](#), we further have

$$\begin{aligned} & \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle Q_h^{\lambda, \sigma^k}(s, \bullet, \sigma^{k+1}, \mu^{k+1}), (\sigma_h^{k+1} - \pi_h^*)(s) \right\rangle \right] \\ & \leq J^{\lambda, \sigma^k}(\mu^{k+1}, \sigma^{k+1}) - J^{\lambda, \sigma^k}(\mu^{k+1}, \pi^*) - \lambda D_{\mu^*}(\pi^*, \sigma^k) + \lambda D_{\mu^*}(\sigma^{k+1}, \sigma^k) \\ & \leq J^{\lambda, \sigma^k}(\mu^*, \sigma^{k+1}) - J^{\lambda, \sigma^k}(\mu^*, \pi^*) - \lambda D_{\mu^*}(\pi^*, \sigma^k) + \lambda D_{\mu^*}(\sigma^{k+1}, \sigma^k) \\ & \leq J(\mu^*, \sigma^{k+1}) - J(\mu^*, \pi^*) - \lambda D_{\mu^{k+1}}(\sigma^{k+1}, \sigma^k) + \lambda D_{\mu^*}(\sigma^{k+1}, \sigma^k), \end{aligned}$$

where we use the identity  $J^{\lambda, \sigma^k}(\mu^*, \pi) = J(\mu^*, \pi) - \lambda D_{m[\pi]}(\pi, \sigma^k)$  for  $\pi \in (\Delta(\mathcal{A})^S)^H$ , and [Definition 2.3](#).  $\blacksquare$

## C Proof of [Theorem 4.5](#)

**Proof of [Theorem 4.5](#).** Let  $h^*: \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}$  be the convex conjugate of  $h$ , i.e.,  $h^*(y) = \sum_{a \in \mathcal{A}} \exp(y(a))$  for  $y \in \mathbb{R}^{|\mathcal{A}|}$ . From direct computations, we have

$$\begin{aligned} & \frac{d}{dt} D_{\mu^*}(\varpi^*, \pi^t) \\ &= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \frac{d}{dt} D_{\text{KL}}(\varpi_h^*(s), \pi^t(s)) \right] \\ &= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle 1 - \frac{\varpi_h^*(s)}{\pi_h^t(s)}, \frac{d}{dt} \pi_h^t(s) \right\rangle \right] \\ &= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle 1 - \frac{\varpi_h^*(s)}{\pi_h^t(s)}, \pi_h^t(a | s) \left( Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right) \right\rangle \right] \\ &= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle (\pi_h^t - \varpi_h^*)(s), Q_h^{\lambda, \sigma}(s, \bullet, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right\rangle \right] \\ &= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle (\pi_h^t - \varpi_h^*)(s), Q_h^{\lambda, \sigma}(s, \bullet, \pi^t, \mu^t) \right\rangle \right] - \lambda \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle (\pi_h^t - \varpi_h^*)(s), \log \frac{\pi_h^t(s)}{\sigma_h(s)} \right\rangle \right]. \end{aligned}$$

We apply [Lemma E.4](#) for the first term and get

$$\begin{aligned} & \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle (\pi_h^t - \varpi_h^*)(s), Q_h^{\lambda, \sigma}(s, \bullet, \pi^t, \mu^t) \right\rangle \right] \\ &= J^{\lambda, \sigma}(\mu^t, \pi^t) - J^{\lambda, \sigma}(\mu^t, \varpi^*) - \lambda D_{\mu^*}(\varpi^*, \sigma) + \lambda D_{\mu^*}(\pi^t, \sigma). \end{aligned} \tag{C.1}$$

Similarly, we apply [Lemma E.5](#) for the second term and get

$$\sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle (\pi_h^t - \varpi_h^*)(s), \log \frac{\pi_h^t(s)}{\sigma_h(s)} \right\rangle \right] = D_{\mu^*}(\pi^t, \sigma) - D_{\mu^*}(\varpi^*, \sigma) + D_{\mu^*}(\varpi^*, \pi^t). \tag{C.2}$$

Combining [\(C.1\)](#) and [\(C.2\)](#) yields

$$\frac{d}{dt} D_{\mu^*}(\varpi^*, \pi^t) = J^{\lambda, \sigma}(\mu^t, \pi^t) - J^{\lambda, \sigma}(\mu^t, \varpi^*) - \lambda D_{\mu^*}(\varpi^*, \pi^t).$$

By virtue of the definition of mean-field equilibrium and [Lemma E.2](#), we find

$$J^{\lambda, \sigma}(\mu^t, \pi^t) - J^{\lambda, \sigma}(\mu^t, \varpi^*) \leq J^{\lambda, \sigma}(\mu^*, \pi^t) - J^{\lambda, \sigma}(\mu^*, \varpi^*) \leq 0.$$

Therefore, we obtain

$$\frac{d}{dt} D_{\mu^*}(\varpi^*, \pi^t) \leq -\lambda D_{\mu^*}(\varpi^*, \pi^t).$$

**Proposition C.1.** Under Assumptions 2.1, 2.4, and 2.5, there exists a unique maximizer of  $J^{\lambda, \sigma^k}(\mu^k, \bullet): (\Delta(\mathcal{A})^S)^H \rightarrow \mathbb{R}$  for each  $k \in \mathbb{N}$ . ■

Proposition C.1 also leads the uniqueness of the regularized equilibrium introduced in Definition 4.2. To elaborate further: Suppose there are two different regularized equilibria  $(\mu_1^*, \varpi_1^*)$  and  $(\mu_2^*, \varpi_2^*)$ . If we assume  $\varpi_1^* \neq \varpi_2^*$ , the following contradiction arises: From Lemma E.2, we have

$$J^{\lambda, \sigma}(\mu_1^*, \varpi_1^*) + J^{\lambda, \sigma}(\mu_2^*, \varpi_2^*) \leq J^{\lambda, \sigma}(\mu_1^*, \varpi_2^*) + J^{\lambda, \sigma}(\mu_2^*, \varpi_1^*).$$

Additionally, from Proposition C.1, we know that  $J^{\lambda, \sigma}(\mu_1^*, \varpi_1^*) \geq J^{\lambda, \sigma}(\mu_1^*, \varpi_2^*)$  and  $J^{\lambda, \sigma}(\mu_2^*, \varpi_2^*) \geq J^{\lambda, \sigma}(\mu_2^*, \varpi_1^*)$ . Adding these two inequalities gives us

$$J^{\lambda, \sigma}(\mu_1^*, \varpi_1^*) + J^{\lambda, \sigma}(\mu_2^*, \varpi_2^*) \geq J^{\lambda, \sigma}(\mu_1^*, \varpi_2^*) + J^{\lambda, \sigma}(\mu_2^*, \varpi_1^*).$$

Therefore,  $\varpi_1^* = \varpi_2^*$ . Moreover, by the definition of regularized equilibria,  $\mu_1^* = m[\varpi_1^*] = m[\varpi_2^*] = \mu_2^*$ . This contradicts the assumption that the two equilibria are different. Thus, the equilibrium is unique.

The uniqueness of Proposition C.1 itself is a new result. The proof uses a continuous-time dynamics shown in Theorem 4.5, see Appendix C. In the following proof, we employ the same proof strategy as in Chill et al. (2010, Theorem 2.10). Before the proof, set  $v_{s,h}^{\lambda, \sigma}(\pi) := \pi_h(a | s) \left( Q_h^{\lambda, \sigma}(s, a, \pi, m[\pi]) - \lambda \log \frac{\pi_h(a | s)}{\sigma_h(a | s)} \right)$  for  $\pi \in (\Delta(\mathcal{A})^S)^H$ .

**Proof of Proposition C.1.** The existence is shown by a slightly modified version of (F. Zhang et al. 2023, Theorem 2). It remains to prove the uniqueness. Fix the regularized equilibrium  $\varpi^* \in (\Delta(\mathcal{A})^S)^H$ .

First of all, we prove the global existence of (4.5). By the local Lipschitz continuity of the right-hand side of the dynamics (4.5) and Picard–Lindelöf theorem, there exists a unique maximal solution  $\pi$  of (4.5) with the initial condition  $\pi|_{t=0} = \pi^0$ . Namely, there exist  $T \in (0, +\infty]$  and  $\pi: [0, T) \rightarrow \mathbb{R}^{|\mathcal{A}|}$  such that  $\pi$  is differentiable on  $(0, T)$  and it holds that (4.5) for all  $t \in (0, T)$ . Thus, Theorem 4.5 ensures that

$$D_{\mu^*}(\varpi^*, \pi^t) + \lambda \int_0^t D_{\mu^*}(\varpi^*, \pi^\tau) d\tau \leq D_{\mu^*}(\varpi^*, \pi^0) =: c < +\infty,$$

for every  $t \in [0, T)$ . As a result, the trajectory  $\{\pi^t \in (\Delta(\mathcal{A})^S)^H \mid t \in [0, T)\}$  is included in  $K_c := \{\pi \in (\Delta(\mathcal{A})^S)^H \mid D_{\mu^*}(\varpi^*, \pi) \leq c\}$ . Note that  $K_c$  is compact from Pinsker inequality.

Since the right-hand side of (4.5) is continuous on  $K_c$ , we obtain  $\sup_{t \in [0, +\infty)} \|v_{s,h}^{\lambda, \sigma}(\pi^t)\| < +\infty$ .

Thus, the equation (4.5) implies  $\left\| \frac{d\pi^t}{dt} \right\|$  is uniformly bounded on  $[0, T)$ . Hence,  $\pi$  extends to a continuous function on  $[0, T]$ .

To obtain a contradiction, we assume  $T < +\infty$ . Then, there exists the solution  $\pi'$  of (4.5) on a larger interval than  $\pi$  with a new initial condition  $\pi'|_{t'=T} = \pi^T$ , which contradicts the maximality of the solution  $\pi$ .

Therefore, the limit  $\lim_{t \rightarrow \infty} \pi^t$  exists and is equal to  $\varpi^*$ . Here,  $\varpi^*$  is arbitrary, so the regularized equilibrium is unique. ■

## D Proof of Theorem 4.3

We can easily show the following lemma by the optimality of  $\pi^{t+1}$  in (4.4).

**Lemma D.1.** It holds that

$$\left\langle \eta \left( Q_h^{\lambda, \sigma}(s, \bullet, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^{t+1}(s)}{\sigma_h(s)} \right) - (1 - \lambda\eta) \log \frac{\pi_h^{t+1}(s)}{\pi_h^t(s)}, \delta \right\rangle = 0,$$

for all  $\delta \in \mathbb{R}^{|\mathcal{A}|}$  such that  $\sum_a \delta(a) = 0$ .

We next show that  $(\pi^t)_t$  is apart from the boundary of  $\mathcal{A}$  as follows.

**Lemma D.2.** *Let  $(\pi^t)_t$  be the sequence defined by (4.4) and  $\varpi^*$  be the policy satisfies Definition 4.2. Assume that there exist vectors  $w_h^\sigma$  and  $w_h^0(s) \in \mathbb{R}^{|\mathcal{A}|}$  satisfying*

$$\begin{aligned} \lambda H \log \sigma_{\min} \leq w_h^\sigma(a | s) \leq -\lambda H \log \sigma_{\min}, \quad \sigma_h(a | s) &\propto \exp\left(\frac{w_h^\sigma(a | s)}{\lambda}\right), \\ 2\lambda H \log \sigma_{\min} \leq w_h^0(a | s) \leq H, \quad \pi_h^0(a | s) &\propto \exp\left(\frac{w_h^0(a | s)}{\lambda}\right). \end{aligned}$$

for all  $a \in \mathcal{A}$ ,  $\pi^0 \in (\Delta(\mathcal{A})^S)^H$ ,  $h \in [H]$  and  $s \in \mathcal{S}$ . Then, for any  $h \in [H]$ ,  $s \in \mathcal{S}$ , and  $t \geq 0$ , it holds that

$$\max\{\|\log \pi_h^t(s)\|_\infty, \|\log \pi_h^*(s)\|_\infty\} \leq \frac{H(1 - \lambda \log \sigma_{\min})}{\lambda} + \log |\mathcal{A}|.$$

**Proof.** We first show that  $\pi_h^t$  can be written as

$$\pi_h^t(a | s) \propto \exp\left(\frac{w_h^t(a | s)}{\lambda}\right), \quad (\text{D.1})$$

for a vector  $w_h^t(s) \in \mathbb{R}^{|\mathcal{A}|}$  satisfying  $2\lambda H \log \sigma_{\min} \leq w_h^t(a | s) \leq H$ . We prove it by induction on  $t$ . Suppose that there exist  $t \in \mathbb{N}$  and  $w_h^t$  satisfying (D.1). By the update rule (4.4), we have

$$\begin{aligned} \pi_h^{t+1}(a | s) &\propto (\sigma_h(a | s))^{\lambda\eta} (\pi_h^t(a | s))^{1-\lambda\eta} \exp\left(\eta Q_h^{\lambda,\sigma}(s, a, \pi^t, \mu^t)\right) \\ &\propto \exp\left(\frac{\lambda\eta w_h^\sigma(a | s) + (1-\eta\lambda)w_h^t(a | s) + \lambda\eta Q_h^{\lambda,\sigma}(s, a, \pi^t, \mu^t)}{\lambda}\right). \end{aligned}$$

Set  $w_h^{t+1}(a | s) := \lambda\eta w_h^\sigma(a | s) + (1-\eta\lambda)w_h^t(a | s) + \lambda\eta Q_h^{\lambda,\sigma}(s, a, \pi^t, \mu^t)$ , we get  $\pi_h^{t+1}(a | s) \propto e^{\frac{w_h^{t+1}(a | s)}{\lambda}}$ . From Lemma E.3 and the hypothesis of the induction, we get  $2\lambda H \log \sigma_{\min} \leq w_h^{t+1}(a | s) \leq H$ .

Then we have for any  $a_1, a_2 \in \mathcal{A}$ :

$$\frac{\pi_h^t(a_1 | s)}{\pi_h^t(a_2 | s)} = \exp\left(\frac{w_h^t(a_1 | s) - w_h^t(a_2 | s)}{\lambda}\right) \leq \exp\left(\frac{H(1 - \lambda \log \sigma_{\min})}{\lambda}\right).$$

It follows that:

$$\min_{a \in \mathcal{A}} \pi^t(a | s) \geq \exp\left(\frac{-H(1 - \lambda \log \sigma_{\min})}{\lambda}\right) \max_{a' \in \mathcal{A}} \pi_h^t(a' | s) \geq |\mathcal{A}|^{-1} \exp\left(\frac{-H(1 - \lambda \log \sigma_{\min})}{\lambda}\right).$$

Therefore, we have:

$$\|\log \pi_h^t(s)\|_\infty \leq \frac{H(1 - \lambda \log \sigma_{\min})}{\lambda} + \log |\mathcal{A}|.$$

From Lemmas E.1 and E.3, we have for  $\pi_h^*$  and  $a_1, a_2 \in \mathcal{A}$ :

$$\begin{aligned} \frac{\pi_h^*(a_1 | s)}{\pi_h^*(a_2 | s)} &= \exp\left(\frac{Q_h^{\lambda,\sigma}(s, a_1, \pi^t, \mu^t) + w_h^\sigma(a_1 | s) - Q_h^{\lambda,\sigma}(s, a_2, \pi^t, \mu^t) - w_h^\sigma(a_2 | s)}{\lambda}\right) \\ &\leq \exp\left(\frac{H(1 - \lambda \log \sigma_{\min})}{\lambda}\right), \end{aligned}$$

and, we get  $\|\log \pi_h^*(s)\|_\infty \leq \frac{H(1 - \lambda \log \sigma_{\min})}{\lambda} + \log |\mathcal{A}|$ . ■

**Lemma D.3.** *Let  $G_h^{\lambda,\sigma}(s, a, \pi^t, \mu^t) := Q_h^{\lambda,\sigma}(s, a, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)}$ .*

$$\begin{aligned} &\left|G_h^{\lambda,\sigma}(s, a, \pi^t, \mu^t) - G_h^{\lambda,\sigma}(s, a', \pi^t, \mu^t)\right| \\ &\leq 2L \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1 + C^{\lambda,\sigma,H,|\mathcal{A}|} (E_h(a, \pi^t, \varpi^*) + E_h(a', \pi^t, \varpi^*)), \end{aligned}$$

for  $a, a' \in \mathcal{A}$ . Here,

$$C^{\lambda,\sigma,H,|\mathcal{A}|} := 2\lambda |\mathcal{A}| e^{\frac{H(1 - \lambda \log \sigma_{\min})}{\lambda}} + 2(1 + H) - \lambda(1 + 2H) \log \sigma_{\min} + 2\lambda \log |\mathcal{A}|,$$

and

$$E_h(a, \pi^t, \varpi^*) := \mathbb{E} \left[ \sum_{l=h}^H \left\| \pi_l^*(s_l) - \pi_l^t(s_l) \right\|_1 \left| \begin{array}{l} s_h = s, a_h = a, \\ s_{l+1} \sim P_l(s_l, a_l), \\ a_l \sim \varpi_l^*(s_l) \\ \text{for each } l \in \{h, \dots, H\} \end{array} \right. \right].$$

**Proof of Lemma D.3.** We first compute the absolute value as follows:

$$\begin{aligned} & \left| G_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - G_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t) \right| \\ &= \left| \left( Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right) - \left( Q_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a' | s)}{\sigma_h(a' | s)} \right) \right| \\ &\leq \left| \left( Q_h^{\lambda, \sigma}(s, a, \varpi^*, \mu^*) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right) - \left( Q_h^{\lambda, \sigma}(s, a', \varpi^*, \mu^*) - \lambda \log \frac{\pi_h^t(a' | s)}{\sigma_h(a' | s)} \right) \right| \\ &\quad + \left| \left( Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - Q_h^{\lambda, \sigma}(s, a, \varpi^*, \mu^*) \right) - \left( Q_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t) - Q_h^{\lambda, \sigma}(s, a', \varpi^*, \mu^*) \right) \right|. \end{aligned} \tag{D.2}$$

By Lemmas D.2 and E.1, the first term of right-hand side in (D.3) can be computed as

$$\begin{aligned} & \left| \left( Q_h^{\lambda, \sigma}(s, a, \varpi^*, \mu^*) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right) - \left( Q_h^{\lambda, \sigma}(s, a', \varpi^*, \mu^*) - \lambda \log \frac{\pi_h^t(a' | s)}{\sigma_h(a' | s)} \right) \right| \\ &= \left| \left( \lambda \log \frac{\varpi_h^*(a | s)}{\sigma_h(a | s)} - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)} \right) - \left( \lambda \log \frac{\varpi_h^*(a' | s)}{\sigma_h(a' | s)} - \lambda \log \frac{\pi_h^t(a' | s)}{\sigma_h(a' | s)} \right) \right| \\ &\leq \lambda \left( \left| \log \frac{\varpi_h^*(a | s)}{\pi_h^t(a | s)} \right| + \left| \log \frac{\varpi_h^*(a' | s)}{\pi_h^t(a' | s)} \right| \right) \\ &\leq \lambda \left( \frac{1}{\varpi_{\min}^*} + \frac{1}{\min_{a \in \mathcal{A}} \pi_h^t(a | s)} \right) (|\varpi_h^*(a | s) - \pi_h^t(a | s)| + |\varpi_h^*(a' | s) - \pi_h^t(a' | s)|) \\ &\leq 2\lambda |\mathcal{A}| \exp \left( \frac{H(1 - \lambda \log \sigma_{\min})}{\lambda} \right) (|\varpi_h^*(a | s) - \pi_h^t(a | s)| + |\varpi_h^*(a' | s) - \pi_h^t(a' | s)|). \end{aligned} \tag{D.3}$$

By Proposition E.8 and Lemma E.6, the second term is bounded as

$$\begin{aligned} & \left| \left( Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - Q_h^{\lambda, \sigma}(s, a, \varpi^*, \mu^*) \right) - \left( Q_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t) - Q_h^{\lambda, \sigma}(s, a', \varpi^*, \mu^*) \right) \right| \\ &\leq 2L \sum_{l=h}^H \left\| \mu_l^t - \mu_l^* \right\|_1 \\ &\quad + C^{\lambda, \sigma}(\pi^t, \varpi^*) \mathbb{E} \left[ \sum_{l=h+1}^H \left\| \pi_l^*(s_l) - \pi_l^t(s_l) \right\|_1 \left| \begin{array}{l} s_{h+1} \sim P_h(\bullet | s, a), \\ s_{l+1} \sim P_l(s_l, a_l), \\ a_l \sim \varpi_l^*(s_l) \\ \text{for each } l \in \{h+1, \dots, H\} \end{array} \right. \right] \\ &\quad + C^{\lambda, \sigma}(\pi^t, \varpi^*) \mathbb{E} \left[ \sum_{l=h+1}^H \left\| \pi_l^*(s_l) - \pi_l^t(s_l) \right\|_1 \left| \begin{array}{l} s_{h+1} \sim P_h(\bullet | s, a'), \\ s_{l+1} \sim P_l(s_l, a_l), \\ a_l \sim \varpi_l^*(s_l) \\ \text{for each } l \in \{h+1, \dots, H\} \end{array} \right. \right]. \end{aligned}$$

Furthermore,  $C^{\lambda, \sigma}(\pi^t, \varpi^*)$  can be bounded as

$$\begin{aligned} C^{\lambda, \sigma}(\pi^t, \varpi^*) &\leq 2 - \lambda \log \sigma_{\min} + 2\lambda \left( \frac{H(1 - \lambda \log \sigma_{\min})}{\lambda} + \log |\mathcal{A}| \right) \\ &= 2(1 + H) - \lambda(1 + 2H) \log \sigma_{\min} + 2\lambda \log |\mathcal{A}|. \end{aligned}$$

■

**Proof of Theorem 4.3.** Set

$$C := 4H^2 \left( L^2 H^2 + \frac{(C^{\lambda, \sigma, H, |\mathcal{A}|})^2}{|\mathcal{A}| \exp \left( \frac{H(1 - \lambda \log \sigma_{\min})}{\lambda} \right)} \right) \tag{D.4}$$

$$\begin{aligned}
&= 4H^2 \left( L^2 H^2 + \frac{\left( 2\lambda |\mathcal{A}| e^{\frac{H(1-\lambda \log \sigma_{\min})}{\lambda}} + 2(1+H) - \lambda(1+2H) \log \sigma_{\min} + 2\lambda \log |\mathcal{A}| \right)^2}{|\mathcal{A}| e^{\frac{H(1-\lambda \log \sigma_{\min})}{\lambda}}} \right) \\
\eta^* &= \min \left\{ \frac{1}{2H(L + C^{\lambda, \sigma, H, |\mathcal{A}|})}, \frac{\lambda}{2C} \right\}, \tag{D.5}
\end{aligned}$$

where  $C^{\lambda, \sigma, H, |\mathcal{A}|}$  is the constant defined in [Lemma D.3](#). We prove the inequality by induction on  $t$ .

**(I) Base step  $t = 0$ :** It is obvious.

**(II) Inductive step:** Suppose that there exists  $t \in \mathbb{N}$  such that  $\pi^t \in \Omega$ . [Lemma D.1](#) yields that

$$\begin{aligned}
&D_{\mu^*}(\varpi^*, \pi^{t+1}) - D_{\mu^*}(\varpi^*, \pi^t) - D_{\mu^*}(\pi^t, \pi^{t+1}) \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle \log \frac{\pi_h^t(s)}{\pi_h^{t+1}(s)}, (\varpi_h^* - \pi_h^t)(s) \right\rangle \right] \\
&= - \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle \frac{\eta}{1 - \lambda\eta} \left( Q_h^{\lambda, \sigma}(s, \bullet, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^{t+1}(s)}{\sigma_h(s)} \right), (\varpi_h^* - \pi_h^t)(s) \right\rangle \right] \\
&= - \frac{\eta}{1 - \lambda\eta} \underbrace{\sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle Q_h^{\lambda, \sigma}(s, \bullet, \pi^t, \mu^t), (\varpi_h^* - \pi_h^t)(s) \right\rangle \right]}_{=: \text{I}} \\
&\quad + \frac{\lambda\eta}{1 - \lambda\eta} \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \left\langle \log \frac{\pi_h^{t+1}(s)}{\sigma_h(s)}, (\varpi_h^* - \pi_h^{t+1})(s) \right\rangle \right] \\
&\leq - \frac{\eta}{1 - \lambda\eta} (\lambda D_{\mu^*}(\varpi^*, \sigma) - \lambda D_{\mu^*}(\pi^{t+1}, \sigma)) \\
&\quad + \frac{\lambda\eta}{1 - \lambda\eta} (D_{\mu^*}(\varpi^*, \sigma) - D_{\mu^*}(\varpi^*, \pi^{t+1}) - D_{\mu^*}(\pi^{t+1}, \sigma)) \\
&\leq - \frac{\lambda\eta}{1 - \lambda\eta} D_{\mu^*}(\varpi^*, \pi^{t+1}), \tag{D.6}
\end{aligned}$$

where I is bounded from below as follows: By [Lemma E.4](#), we get

$$I = J^{\lambda, \sigma}(\mu^{t+1}, \varpi^*) - J^{\lambda, \sigma}(\mu^{t+1}, \pi^{t+1}) + \lambda D_{\mu^*}(\varpi^*, \sigma) - \lambda D_{\mu^*}(\pi^{t+1}, \sigma).$$

By virtue of the definition of mean-field equilibrium and [Lemma E.2](#), we find

$$J^{\lambda, \sigma}(\mu^{t+1}, \varpi^*) - J^{\lambda, \sigma}(\mu^{t+1}, \pi^{t+1}) \geq J^{\lambda, \sigma}(\mu^*, \varpi^*) - J^{\lambda, \sigma}(\mu^*, \pi^{t+1}) \geq 0.$$

Then, we obtain

$$I \geq \lambda D_{\mu^*}(\varpi^*, \sigma) - \lambda D_{\mu^*}(\pi^{t+1}, \sigma).$$

For the last term  $D_{\mu^*}(\pi^t, \pi^{t+1})$  of the leftmost hand of (D.6), we can employ a similar argument to (Abe et al. 2023, Lemma 5.4), that is, we can estimate  $D_{\mu^*}(\pi^t, \pi^{t+1})$  as follows: Set  $G(a) := G_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) = Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - \lambda \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)}$ . Note that  $\max_{a, a' \in \mathcal{A}} |G(a') - G(a)| \leq$



$\eta^{*-1}$  by Lemma D.3. By the update rule (4.4) and concavity of the logarithmic function  $\log$ , we have

$$\begin{aligned}
& D_{\mu^*}(\pi^t, \pi^{t+1}) \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \sum_{a \in \mathcal{A}} \pi_h^t(a | s) \log \frac{\pi_h^t(a | s)}{\pi_h^{t+1}(a | s)} \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \sum_{a \in \mathcal{A}} \pi_h^t(a | s) \log \frac{\sum_{a' \in \mathcal{A}} (\sigma_h(a' | s))^{\lambda \eta} (\pi_h^t(a' | s))^{1-\lambda \eta} \exp(\eta Q_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t))}{(\sigma_h(a | s))^{\lambda \eta} (\pi_h^t(a | s))^{-\lambda \eta} \exp(\eta Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t))} \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \sum_{a \in \mathcal{A}} \pi_h^t(a | s) \log \frac{\sum_{a' \in \mathcal{A}} \pi_h^t(a' | s) \exp\left(\eta Q_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t) - \lambda \eta \log \frac{\pi_h^t(a' | s)}{\sigma_h(a' | s)}\right)}{\exp\left(\eta Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - \lambda \eta \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)}\right)} \right] \\
&\leq \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \log \sum_{a \in \mathcal{A}} \pi_h^t(a | s) \frac{\sum_{a' \in \mathcal{A}} \pi_h^t(a' | s) \exp\left(\eta Q_h^{\lambda, \sigma}(s, a', \pi^t, \mu^t) - \lambda \eta \log \frac{\pi_h^t(a' | s)}{\sigma_h(a' | s)}\right)}{\exp\left(\eta Q_h^{\lambda, \sigma}(s, a, \pi^t, \mu^t) - \lambda \eta \log \frac{\pi_h^t(a | s)}{\sigma_h(a | s)}\right)} \right]. \tag{D.7}
\end{aligned}$$

If we take  $\eta$  to be  $\eta \leq \eta^*$ , it follows that

$$\eta(G(a') - G(a)) \leq 1,$$

for  $a, a' \in \mathcal{A}$ . Thus, we can use the inequality  $e^x \leq 1 + x + x^2$  for  $x \leq 1$  and obtain

$$\begin{aligned}
& D_{\mu^*}(\pi^t, \pi^{t+1}) \\
&\leq \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \log \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) e^{\eta(G(a') - G(a))} \right] \\
&\leq \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \log \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) \left(1 + \eta(G(a') - G(a)) + \eta^2(G(a') - G(a))^2\right) \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \log \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) \left(1 + (G(a') - G(a))^2\right) \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \log \left(1 + \eta^2 \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) (G(a') - G(a))^2\right) \right] \\
&\leq \eta^2 \sum_{h=1}^H \mathbb{E}_{s \sim \mu_h^*} \left[ \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) (G(a') - G(a))^2 \right].
\end{aligned}$$

By Lemma D.3, we can see that

$$\begin{aligned}
& \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) (G(a') - G(a))^2 \\
&\leq \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) \left( 2L \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1 + C^{\lambda, \sigma, H, |\mathcal{A}|} (E_h(a, \pi^t, \varpi^*) + E_h(a', \pi^t, \varpi^*)) \right)^2 \\
&\leq \sum_{a, a' \in \mathcal{A}} \pi_h^t(a | s) \pi_h^t(a' | s) \left( 8L^2 \left( \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1 \right)^2 + 4 \left( C^{\lambda, \sigma, H, |\mathcal{A}|} \right)^2 (E_h^2(a, \pi^t, \varpi^*) + E_h^2(a', \pi^t, \varpi^*)) \right) \\
&\leq 8L^2 H \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2 + 8 \left( C^{\lambda, \sigma, H, |\mathcal{A}|} \right)^2 \sum_{a \in \mathcal{A}} \pi_h^t(a | s) E_h^2(a, \pi^t, \varpi^*)
\end{aligned}$$

$$\begin{aligned}
&= 8L^2H \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2 + 8 \left( C^{\lambda, \sigma, H, |\mathcal{A}|} \right)^2 \sum_{a \in \mathcal{A}} \frac{\pi_h^t(a | s)}{\varpi_h^*(a | s)} \varpi_h^*(a | s) E_h^2(a, \pi^t, \varpi^*) \\
&\leq 8L^2H \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2 + \frac{8 \left( C^{\lambda, \sigma, H, |\mathcal{A}|} \right)^2}{|\mathcal{A}| \exp \left( \frac{H(1-\lambda \log \sigma_{\min})}{\lambda} \right)} \sum_{a \in \mathcal{A}} \varpi_h^*(a | s) E_h^2(a, \pi^t, \varpi^*) \\
&\leq 8L^2H \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2 + \frac{8H \left( C^{\lambda, \sigma, H, |\mathcal{A}|} \right)^2}{|\mathcal{A}| \exp \left( \frac{H(1-\lambda \log \sigma_{\min})}{\lambda} \right)} \sum_{l=h}^H \mathbb{E}_{s_l \sim \mu_l^*} \left[ \|\pi_l^*(s_l) - \pi_l^t(s_l)\|_1^2 \right] \\
&\leq 8L^2H \sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2 + \frac{4H \left( C^{\lambda, \sigma, H, |\mathcal{A}|} \right)^2}{|\mathcal{A}| \exp \left( \frac{H(1-\lambda \log \sigma_{\min})}{\lambda} \right)} D_{\mu^*}(\varpi^*, \pi^t).
\end{aligned}$$

Moreover, [Lemma E.6](#) bounds  $\sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2$  as

$$\sum_{l=h}^H \|\mu_l^t - \mu_l^*\|_1^2 \leq H \sum_{l=h}^H \sum_{k=0}^{l-1} \mathbb{E}_{s_k \sim \mu_k^*} \left[ \|\pi_k^*(s_k) - \pi_k^t(s_k)\|_1^2 \right] \leq \frac{1}{2} H^2 D_{\mu^*}(\varpi^*, \pi^t).$$

Therefore, we finally obtain

$$D_{\mu^*}(\varpi^*, \pi^{t+1}) \leq (1 - \lambda\eta + C\eta^2) D_{\mu^*}(\varpi^*, \pi^t) \leq \left(1 - \frac{1}{2}\lambda\eta\right) D_{\mu^*}(\varpi^*, \pi^t), \quad (\text{D.8})$$

where we use  $C\eta \leq C\eta^* \leq 1/2$ . ■

## E Useful Lemmas

For Mean-field games, one can write down the *Bellman optimality equation* as follows: for a function  $Q': \mathcal{S} \rightarrow \Delta(\mathcal{A})$ , a policy  $\pi': \mathcal{S} \rightarrow \Delta(\mathcal{A})$ ,  $\sigma': \mathcal{S} \rightarrow \Delta(\mathcal{A})$  and  $s \in \mathcal{S}$  set

$$f_s^{\sigma'}(Q', \pi') = \langle Q'(s), \pi'(s) \rangle - \lambda D_{\text{KL}}(\pi'(s), \sigma'(s)). \quad (\text{E.1})$$

**Lemma E.1.** *Let  $(\mu^*, \varpi^*)$  be equilibrium in the sense of [Definition 4.2](#). Then, it holds that*

$$\varpi_h^*(s) = \arg \max_{p \in \Delta(\mathcal{A})} f_s^{\sigma_h} \left( Q_h^{\lambda, \sigma}(s, \bullet, \varpi^*, \mu^*), p \right) \propto \sigma_h(\bullet | s) \exp \left( \frac{Q_h^{\lambda, \sigma}(s, \bullet, \varpi^*, \mu^*)}{\lambda} \right),$$

for each  $s \in \mathcal{S}$  and  $h \in [H]$ . Moreover,

$$\left\langle Q_h^{\lambda, \sigma}(s, \bullet, \varpi^*, \mu^*) - \lambda \log \frac{\pi_h^*(s)}{\sigma_h(s)}, \delta \right\rangle = 0,$$

for all  $\delta \in \mathbb{R}^{|\mathcal{A}|}$  such that  $\sum_a \delta(a) = 0$ .

**Proof.** See the Bellman optimality equation (e.g., [Agarwal et al. 2022](#), Theorem 1.9). ■

**Lemma E.2.** *Under [Assumption 2.4](#), it holds that, for all  $\pi, \tilde{\pi} \in (\Delta(\mathcal{A})^{\mathcal{S}})^H$ ,*

$$J^{\lambda, \sigma}(m[\pi], \pi) + J^{\lambda, \sigma}(m[\tilde{\pi}], \tilde{\pi}) - J^{\lambda, \sigma}(m[\pi], \tilde{\pi}) - J^{\lambda, \sigma}(m[\tilde{\pi}], \pi) \leq 0,$$

where  $m$  is defined in [\(2.1\)](#).

**Proof of Lemma E.2.** The proof is similar to ([F. Zhang et al. 2023](#), §H). Set  $\mu = m[\pi]$  and  $\tilde{\mu} = m[\tilde{\pi}]$ . One can obtain that

$$\begin{aligned}
&J^{\lambda, \sigma}(m[\pi], \pi) + J^{\lambda, \sigma}(m[\tilde{\pi}], \tilde{\pi}) - J^{\lambda, \sigma}(m[\pi], \tilde{\pi}) - J^{\lambda, \sigma}(m[\tilde{\pi}], \pi) \\
&= (J^{\lambda, \sigma}(\mu, \pi) - J^{\lambda, \sigma}(\tilde{\mu}, \pi)) + (J^{\lambda, \sigma}(\tilde{\mu}, \tilde{\pi}) - J^{\lambda, \sigma}(\mu, \tilde{\pi})) \\
&= \sum_{h=1}^H \sum_{s_h \in \mathcal{S}} m[\pi]_h(s_h) \sum_{a_h \in \mathcal{A}} \pi_h(a_h | s_h) (r_h(s_h, a_h, \mu_h) - r_h(s_h, a_h, \tilde{\mu}_h)) \\
&\quad + \sum_{h=1}^H \sum_{s_h \in \mathcal{S}} m[\tilde{\pi}]_h(s_h) \sum_{a_h \in \mathcal{A}} \tilde{\pi}_h(a_h | s_h) (r_h(s_h, a_h, \tilde{\mu}_h) - r_h(s_h, a_h, \mu_h))
\end{aligned}$$

$$= \sum_{h,s,a} (\pi_h(a | s) \mu_h(s) - \tilde{\pi}_h(a | s) \tilde{\mu}_h(s)) (r_h(s_h, a_h, \mu_h) - r_h(s_h, a_h, \tilde{\mu}_h)),$$

and the right-hand side of the above inequality is less than 0 by [Assumption 2.4](#). ■

**Lemma E.3.** *Let  $V_h^{\lambda,\sigma}$  be the state value function defined in (4.2) and  $Q_h^{\lambda,\sigma}$  be the state action value function defined in (4.3). For any  $s \in \mathcal{A}$ ,  $a \in \mathcal{A}$ , and  $h \in [H]$ , it holds that*

$$\begin{aligned} \lambda(H - h + 1) \log \sigma_{\min} &\leq V_h^{\lambda,\sigma}(s, \mu, \pi) \leq H - h + 1, \\ \lambda(H - h + 1) \log \sigma_{\min} &\leq Q_h^{\lambda,\sigma}(s, a, \mu, \pi) \leq H - h + 2. \end{aligned}$$

**Proof.** We prove the inequalities by backward induction on  $h$ . By definition, we have

$$\begin{aligned} &V_h^{\lambda,\sigma}(s, \mu, \pi) \\ &= \mathbb{E} \left[ \sum_{l=h}^H (r_l(s_l, a_l, \mu_l) - \lambda D_{\text{KL}}(\pi_l(s_l), \sigma_l(s_l))) \mid s_h = s \right] \\ &= \langle r_h(s, \bullet, \mu_h), \pi_h(s) \rangle - \lambda D_{\text{KL}}(\pi_h(s_h), \sigma_h(s_h)) \\ &\quad + \sum_{s_{h+1} \in \mathcal{S}} V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \sum_{a_h \in \mathcal{A}} P_h(s_{h+1} | s, a_h) \pi_h(a_h | s) \\ &\leq 1 + \max_{s_{h+1} \in \mathcal{S}} V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi), \end{aligned}$$

and

$$\begin{aligned} &V_h^{\lambda,\sigma}(s, \mu, \pi) \\ &= \langle r_h(s, \bullet, \mu_h), \pi_h(s) \rangle - \lambda D_{\text{KL}}(\pi_h(s_h), \sigma_h(s_h)) \\ &\quad + \sum_{s_{h+1} \in \mathcal{S}} V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \sum_{a_h \in \mathcal{A}} P_h(s_{h+1} | s, a_h) \pi_h(a_h | s) \\ &\geq \lambda \log \sigma_{\min} + \max_{s_{h+1} \in \mathcal{S}} V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi). \end{aligned}$$

Then, we have

$$V_h^{\lambda,\sigma}(s, \mu, \pi) \in [\lambda(H - h + 1) \log \sigma_{\min}, H - h + 1],$$

by the induction. The definition of  $Q_h^{\lambda,\sigma}$  in (4.3) immediately yields the bound. ■

**Lemma E.4.** *For all  $\pi, \tilde{\pi} \in (\Delta(\mathcal{A})^{\mathcal{S}})^H$ , it holds that*

$$\begin{aligned} &\sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle (\pi_h - \tilde{\pi}_h)(s), Q_h^{\lambda,\sigma}(s, \bullet, \pi, \mu) \right\rangle \right] \\ &= J^{\lambda,\sigma}(\mu, \pi) - J^{\lambda,\sigma}(\mu, \tilde{\pi}) - \lambda D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma), \end{aligned}$$

where we set  $\mu = m[\pi]$ .

**Proof.** From the definition of  $V^{\lambda,\sigma}$  and  $Q^{\lambda,\sigma}$  in (4.2) and (4.3), we have

$$\begin{aligned}
& \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle \pi_h(s), Q_h^{\lambda,\sigma}(s, \bullet, \pi, \mu) \right\rangle \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle \pi_h(s), r_h(s, \bullet, \mu_h) + \mathbb{E} \left[ V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \mid s_{h+1} \sim P(s, \bullet, \mu_h) \right] \right\rangle \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s_h \sim m[\tilde{\pi}]_h} \left[ \mathbb{E}_{a_h \sim \pi_h(s)} [r_h(s_h, a_h, \mu_h) - \lambda D_{\text{KL}}(\pi(s_h), \sigma(s_h))] \right] + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma) \\
&\quad + \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \mathbb{E} \left[ V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \mid s_{h+1} \sim P(s, a_h, \mu_h), a_h \sim \pi_h(s) \right] \right] \tag{E.2} \\
&= \sum_{h=1}^H \mathbb{E}_{s_h \sim m[\tilde{\pi}]_h} \left[ V_h^{\lambda,\sigma}(s_h, \mu, \pi) - \mathbb{E} \left[ V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \mid \begin{matrix} s_{h+1} \sim P(s, a_h, \mu_h), \\ a_h \sim \pi_h(s) \end{matrix} \right] \right] \\
&\quad + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma) + \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \mathbb{E} \left[ V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \mid \begin{matrix} s_{h+1} \sim P(s, a_h, \mu_h), \\ a_h \sim \pi_h(s) \end{matrix} \right] \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ V_h^{\lambda,\sigma}(s, \mu, \pi) \right] + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma).
\end{aligned}$$

Similarly, (4.1) and (2.1) gives us

$$\begin{aligned}
& \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle \tilde{\pi}_h(s), Q_h^{\lambda,\sigma}(s, \bullet, \pi, \mu) \right\rangle \right] \\
&= \sum_{h=1}^H \mathbb{E}_{s_h \sim m[\tilde{\pi}]_h} \left[ \mathbb{E}_{a_h \sim \tilde{\pi}_h(s)} [r_h(s_h, a_h, \mu_h) - \lambda D_{\text{KL}}(\tilde{\pi}(s_h), \sigma(s_h))] \right] + \lambda D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) \\
&\quad + \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \mathbb{E} \left[ V_{h+1}^{\lambda,\sigma}(s_{h+1}, \mu, \pi) \mid s_{h+1} \sim P(s, a_h, \mu_h), a_h \sim \tilde{\pi}_h(s) \right] \right] \tag{E.3} \\
&= J^{\lambda,\sigma}(\mu, \tilde{\pi}) + \lambda D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) + \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_{h+1}} \left[ V_{h+1}^{\lambda,\sigma}(s, \mu, \pi) \right].
\end{aligned}$$

Combining (E.2) and (E.3) yields

$$\begin{aligned}
& \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\mu}]_h} \left[ \left\langle (\pi_h - \tilde{\pi}_h)(s), Q_h^{\lambda,\sigma}(s, \bullet, \pi, \mu) \right\rangle \right] \\
&= \left( \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ V_h^{\lambda,\sigma}(s, \mu, \pi) \right] + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma) \right) \\
&\quad - \left( J^{\lambda,\sigma}(\mu, \tilde{\pi}) + \lambda D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) + \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_{h+1}} \left[ V_{h+1}^{\lambda,\sigma}(s, \mu, \pi) \right] \right) \\
&= \left( \mathbb{E}_{s \sim m[\tilde{\pi}]_1} \left[ V_1^{\lambda,\sigma}(s, \mu, \pi) \right] + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma) \right) - \left( J^{\lambda,\sigma}(\mu, \tilde{\pi}) + \lambda D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) \right) \\
&= \mathbb{E}_{s \sim \mu_1} \left[ V_1^{\lambda,\sigma}(s, \mu, \pi) \right] - J^{\lambda,\sigma}(\mu, \tilde{\pi}) + \lambda D_{m[\tilde{\pi}]}(\pi, \sigma) - \lambda D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma),
\end{aligned}$$

which concludes the proof. ■

**Lemma E.5.** For all  $\pi, \tilde{\pi} \in (\Delta(\mathcal{A})^S)^H$ , it holds that

$$\sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle (\pi_h - \tilde{\pi}_h)(s), \log \frac{\pi_h(s)}{\sigma_h(s)} \right\rangle \right] = D_{m[\tilde{\pi}]}(\pi, \sigma) - D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) + D_{\tilde{\pi}}(\tilde{\pi}, \pi).$$

**Proof.** A direct computation yields

$$\begin{aligned}
& \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle (\pi_h - \tilde{\pi}_h)(s), \log \frac{\pi_h(s)}{\sigma_h(s)} \right\rangle \right] \\
&= D_{m[\tilde{\pi}]}(\pi, \sigma) - \sum_{h=1}^H \mathbb{E}_{s \sim m[\tilde{\pi}]_h} \left[ \left\langle \tilde{\pi}_h(s), \log \frac{\tilde{\pi}_h(s)}{\sigma_h(s)} - \log \frac{\tilde{\pi}(s)}{\pi(s)} \right\rangle \right] \\
&= D_{m[\tilde{\pi}]}(\pi, \sigma) - D_{m[\tilde{\pi}]}(\tilde{\pi}, \sigma) + D_{m[\tilde{\pi}]}(\tilde{\pi}, \pi).
\end{aligned}$$

■

**Lemma E.6.** The operator  $m$  defined in (2.1) is 1-Lipschitz, namely, it holds that

$$\|m[\pi]_{h+1} - m[\pi']_{h+1}\| \leq \sum_{l=0}^h \mathbb{E}_{s_l \sim m[\pi]_l} [\|\pi_l(s_l) - \pi'_l(s_l)\|], \quad (\text{E.4})$$

for  $\pi, \pi' \in (\Delta(\mathcal{A})^{\mathcal{S}})^H$  and all  $h \in \{0, \dots, H\}$ . Here, we set  $\pi_0(s) = \pi'_0(s) = \mathbf{U}_{\mathcal{A}}$  for all  $s \in \mathcal{S}$ .

**Proof.** Fix  $\pi, \pi' \in (\Delta(\mathcal{A})^{\mathcal{S}})^H$ . We prove the inequality by induction on  $h$ .

**(I) Base step  $h = 0$ :** It is obvious because  $\|m[\pi]_1 - m[\pi']_1\| = \|\mu_1 - \mu_1\| = 0$ .

**(II) Inductive step:** Suppose that there exists  $h \in [H]$  satisfying the inequality (E.4). By (2.1), we obtain

$$\begin{aligned}
& \|m[\pi]_{h+2} - m[\pi']_{h+2}\| \\
& \leq \sum_{\substack{s_{h+2} \in \mathcal{S}, \\ (s_{h+1}, a_{h+1}) \in \mathcal{S} \times \mathcal{A}}} P_{h+1}(s_{h+2} \mid s_{h+1}, a_{h+1}) m[\pi]_{h+1}(s_{h+1}) |\pi_{h+1}(a_{h+1} \mid s_{h+1}) - \pi'_{h+1}(a_{h+1} \mid s_{h+1})| \\
& \quad + \sum_{\substack{s_{h+2} \in \mathcal{S}, \\ (s_{h+1}, a_{h+1}) \in \mathcal{S} \times \mathcal{A}}} P_{h+1}(s_{h+2} \mid s_{h+1}, a_{h+1}) \pi'_{h+1}(a_{h+1} \mid s_{h+1}) |m[\pi]_{h+1}(s_{h+1}) - m[\pi']_{h+1}(s_{h+1})| \\
& \leq \sum_{(s_{h+1}, a_{h+1}) \in \mathcal{S} \times \mathcal{A}} m[\pi]_{h+1}(s_{h+1}) |\pi_{h+1}(a_{h+1} \mid s_{h+1}) - \pi'_{h+1}(a_{h+1} \mid s_{h+1})| \\
& \quad + \sum_{s_{h+1} \in \mathcal{S}} |m[\pi]_{h+1}(s_{h+1}) - m[\pi']_{h+1}(s_{h+1})| \\
& = \mathbb{E}_{s_{h+1} \sim m[\pi]_{h+1}} [\|\pi_{h+1}(s_{h+1}) - \pi'_{h+1}(s_{h+1})\|] + \|m[\pi]_{h+1} - m[\pi']_{h+1}\|.
\end{aligned}$$

By the hypothesis of the induction, we finally obtain

$$\begin{aligned}
& \|m[\pi]_{h+2} - m[\pi']_{h+2}\| \\
& \leq \mathbb{E}_{s \sim m[\pi]_{h+1}} [\|\pi_{h+1}(s) - \pi'_{h+1}(s)\|] + \sum_{l=1}^h \mathbb{E}_{s \sim m[\pi]_l} \|\pi_l(s) - \pi'_l(s)\| \\
& \leq \sum_{l=1}^{h+1} \mathbb{E}_{s \sim m[\pi]_l} \|\pi_l(s) - \pi'_l(s)\|.
\end{aligned}$$

■

**Lemma E.7.** Let  $\pi, \pi' \in (\Delta(\mathcal{A})^{\mathcal{S}})^H$ ,  $\mu, \mu' \in \Delta(\mathcal{S})^H$ ,  $s \in \mathcal{S}$ , and  $h \in \{1, \dots, H+1\}$ . Assume

$$\min_{(h,a,s) \in [H] \times \mathcal{A} \times \mathcal{S}} \min\{\pi_h(a \mid s), \pi'_h(a \mid s)\} > 0,$$

and set  $\mu_{H+1} = \mu'_{H+1} = \mathbf{U}_{\mathcal{S}}$ ,  $\pi_{H+1}(s) = \pi'_{H+1}(s) = \mathbf{U}_{\mathcal{A}}$  for all  $s \in \mathcal{S}$ .

$$\begin{aligned}
& |V_h^{\lambda, \sigma}(s, \pi, \mu) - V_h^{\lambda, \sigma}(s, \pi', \mu')| \\
& \leq \mathbb{E} \left[ \sum_{l=h}^{H+1} (C^{\lambda, \sigma}(\pi, \pi') \|\pi_l(s_l) - \pi'_l(s_l)\|_1 + L \|\mu_l - \mu'_l\|_1) \right] \left[ \begin{array}{l} s_h = s, \\ s_{l+1} \sim P_l(s_l, a_l), \\ a_l \sim \pi_l(s_l) \\ \text{for each } l \in \{h, \dots, H+1\} \end{array} \right]
\end{aligned}$$

for Here,  $C^{\lambda,\sigma}(\pi, \pi') > 0$  is defined in [Proposition E.8](#), and the discrete time stochastic process  $(s_l)_{l=h}^H$  is induced recursively as  $s_{l+1} \sim P_l(s_l, a_l)$ ,  $a_l \sim \pi_l(s_l)$  for each  $l \in \{h, \dots, H-1\}$ .

**Proof.** Fix  $\pi, \pi', \mu$  and  $\mu'$ . We prove the inequality by backward induction on  $h$ .

**(I) Base step  $h = H + 1$ :** It is obvious because  $|V_{H+1}^{\lambda,\sigma}(s, \pi, \mu) - V_{H+1}^{\lambda,\sigma}(s, \pi', \mu')| = |0 - 0| = 0$ .

**(II) Inductive step:** Suppose that there exists  $h \in [H]$  satisfying

$$\begin{aligned} & \left| V_{h+1}^{\lambda,\sigma}(s, \pi, \mu) - V_{h+1}^{\lambda,\sigma}(s, \pi', \mu') \right| \\ & \leq \mathbb{E} \left[ \sum_{l=h+1}^{H+1} (C^{\lambda,\sigma}(\pi, \pi') \|\pi_l(s_l) - \pi'_l(s_l)\|_1 + L \|\mu_h - \mu'_h\|_1) \left| \begin{array}{l} s_{h+1} = s, \\ s_{l+1} \sim P_l(s_l, a_l), \\ a_l \sim \pi_l(s_l) \\ \text{for each } l \in \{h+1, \dots, H+1\} \end{array} \right. \right], \end{aligned} \quad (\text{E.5})$$

for all  $s \in \mathcal{S}$ . By the definition of the value function in (4.2) and [Assumption 2.5](#), we have

$$\begin{aligned} & \left| V_h^{\lambda,\sigma}(s, \pi, \mu) - V_h^{\lambda,\sigma}(s, \pi', \mu') \right| \\ & \leq \left| \sum_{a_h \in \mathcal{A}} (\pi_h(a_h | s) r_h(s, a_h, \mu_h) - \pi'_h(a_h | s) r_h(s, a_h, \mu'_h)) \right| \\ & \quad + \lambda |D_{\text{KL}}(\pi_h(s), \sigma_h(s)) - D_{\text{KL}}(\pi'_h(s), \sigma_h(s))| \\ & \quad + \left| \sum_{\substack{a_h \in \mathcal{A}, \\ s_{h+1} \in \mathcal{S}}} P_h(s_{h+1} | s, a_h) \left( \pi_h(a_h | s) V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi, \mu) - \pi'_h(a_h | s) V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi', \mu') \right) \right| \\ & \leq \|\pi_h(s) - \pi'_h(s)\|_1 + \sum_{a_h \in \mathcal{A}} \pi_h(a_h | s) |r_h(s, a_h, \mu_h) - r_h(s, a_h, \mu'_h)| \\ & \quad + \lambda \left| \sum_{a_h \in \mathcal{A}} \left( \pi_h(a_h | s) \left( \log \frac{\pi_h(a_h | s)}{\sigma_h(a_h | s)} - 1 \right) - \pi'_h(a_h | s) \left( \log \frac{\pi'_h(a_h | s)}{\sigma_h(a_h | s)} - 1 \right) \right) \right| \\ & \quad + \|\pi_h(s) - \pi'_h(s)\|_1 \\ & \quad + \sum_{\substack{a_h \in \mathcal{A}, \\ s_{h+1} \in \mathcal{S}}} P_h(s_{h+1} | s, a_h) \pi_h(a_h | s) \left| V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi, \mu) - V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi', \mu') \right| \\ & \leq 2\|\pi_h(s) - \pi'_h(s)\|_1 + L\|\mu_h - \mu'_h\|_1 \\ & \quad + \lambda \max_{(h,a,s)} \log \frac{1}{(\sigma\pi\pi')_h(a | s)} \|\pi_h(s) - \pi'_h(s)\|_1 \\ & \quad + \sum_{\substack{a_h \in \mathcal{A}, \\ s_{h+1} \in \mathcal{S}}} P_h(s_{h+1} | s, a_h) \pi_h(a_h | s) \left| V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi, \mu) - V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi', \mu') \right| \\ & \leq C^{\lambda,\sigma}(\pi, \pi') \|\pi_h(s) - \pi'_h(s)\|_1 + L\|\mu_h - \mu'_h\|_1 \\ & \quad + \mathbb{E} \left[ \left| V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi, \mu) - V_{h+1}^{\lambda,\sigma}(s_{h+1}, \pi', \mu') \right| \left| \begin{array}{l} s_h = s, \\ s_{h+1} \sim P_h(s_h, a_h), \\ a_h \sim \pi_h(s_h) \end{array} \right. \right]. \end{aligned}$$

Combining the above inequality and the hypothesis of the induction completes the proof. ■



**Proposition E.8.** Let  $Q^{\lambda,\sigma}$  be the function defined by (4.3), and  $(\pi, \pi') \in ((\Delta(\mathcal{A})^S)^H)^2$  be policies with full supports. Under Assumptions 2.5 and 4.1, it holds that

$$\begin{aligned} & \left| Q_h^{\lambda,\sigma}(s, a, \pi, \mu) - Q_h^{\lambda,\sigma}(s, a, \pi', \mu') \right| \\ & \leq L \sum_{l=h}^H \|\mu_l - \mu'_l\| + C^{\lambda,\sigma}(\pi, \pi') \mathbb{E}_{(s_l)_{l=h+1}^H} \left[ \sum_{l=h+1}^H \|\pi_l(s_l) - \pi'_l(s_l)\| \mid s_h = s \right], \end{aligned}$$

for  $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$  and  $\mu, \mu' \in \Delta(\mathcal{S})^H$ . Here, the random variables  $(s_l)_{l=h+1}^H$  follows the stochastic process starting from state  $s$  at time  $h$ , induced from  $P$  and  $\pi$ , and the function  $C^{\lambda,\sigma}: ((\Delta(\mathcal{A})^S)^H)^2 \rightarrow \mathbb{R}$  is given by  $C^{\lambda,\sigma}(\pi, \pi') = 2 - \lambda \inf_{(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}} \log(\sigma\pi\pi')_h(a \mid s)$ .

**Proof of Proposition E.8.** Let  $h$  be larger than 2. By the definition of  $Q_h^{\lambda,\sigma}$  given in (4.3) and Lemma E.7, we have

$$\begin{aligned} & \left| Q_{h-1}^{\lambda,\sigma}(s, a, \pi, \mu) - Q_{h-1}^{\lambda,\sigma}(s, a, \pi', \mu') \right| \\ & \leq \left| r_{h-1}(s, a, \mu_{h-1}) - r_{h-1}(s, a, \mu'_{h-1}) \right| + \mathbb{E}_{s_h \sim P_{h-1}(s, a)} \left[ \left| V_h^{\lambda,\sigma}(s_h, \pi, \mu) - V_h^{\lambda,\sigma}(s_h, \pi', \mu') \right| \right] \\ & \leq L \|\mu_{h-1} - \mu'_{h-1}\| + \mathbb{E}_{s_h \sim P_{h-1}(s, a)} \left[ \left| V_h^{\lambda,\sigma}(s_h, \pi, \mu) - V_h^{\lambda,\sigma}(s_h, \pi', \mu') \right| \right]. \end{aligned}$$

Combining the above inequality and Lemma E.7 completes the proof.  $\blacksquare$

## F Experiment Details

We ran experiments on a laptop with an 11th Gen Intel Core i7-1165G7 8-core CPU, 16GB RAM, running Windows 11 Pro with WSL. As is clear from Algorithm 1, APP is deterministic. Thus, we ran the algorithm only once for each experimental setting. We implemented APP using Python. The computation of  $Q^{\lambda,\sigma}$  and  $\mu$  in Algorithm 1 was based on the implementation provided by Fabian et al. (2023).

**Algorithms.** In this experiment, we implement APP in Algorithm 1. For comparison, we also implement RMD (i.e., Algorithm 1 without the update of  $\sigma_k$ ) in (4.4). For both algorithms, the learning rate is fixed at  $\eta = 0.1$ , and we vary the regularization parameter  $\lambda$  and update time  $T$  to run the experiments.

We show further details for the Beach Bar Process. We set  $H = 10, |\mathcal{S}| = 10, \mathcal{A} = \{-1, \pm 0, +1\}, \lambda = 0.1, \eta = 0.1$ , and

$$P_h(s' \mid s, a) = \begin{cases} 1 - \varepsilon & \text{if } a = \pm 0 \text{ \& } s' = s, \\ \frac{\varepsilon}{2} & \text{if } a = \pm 1 \text{ \& } s' = s \pm 1, \\ 0 & \text{otherwise,} \end{cases}$$

where we choose  $\varepsilon = 0.1$ . In addition, we initialize  $\sigma^0$  and  $\pi^0$  in Algorithm 1 as the uniform distributions on  $\mathcal{A}$ .

**Remark F.1.** When the contraction factor  $1 - \lambda\eta$  is close to 1, we can observe a small  $\tau$  can lead to instability in the outer PP loop.

## NeurIPS Paper Checklist

### (i) Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The introduction clearly states the two main contributions—LIC for the proximal-point update ([Theorem 3.1](#)) and exponential convergence of RMD ([Theorem 4.5](#))—and these match the theoretical and experimental results presented.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### (ii) Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper includes § 6 on limitations of the applicability of our theory.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### (iii) Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All assumptions are stated alongside [Theorems 3.1](#) and [4.5](#), and detailed proofs are provided in the appendices.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

(iv) **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: [Appendix F](#) gives full details of the benchmark (Beach Bar Process), hyperparameters, and implementation environment (Python, deterministic runs).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

(v) **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides the code for reproducing the results in § 5 as supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

(vi) **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: § 5 and Appendix F fully specify the algorithm parameters, benchmark setup, and evaluation metric (exploitability)

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

(vii) **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our proposed algorithms (Algorithms 1 and 2) are deterministic, thus error bars or confidence intervals are less meaningful.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

(viii) **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: [Appendix F](#) reports execution on an 11th Gen Intel Core i7-1165G7 CPU with 16 GB RAM under Windows 11 Pro (WSL)

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

(ix) **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The work is foundational and does not involve sensitive data, human subjects, or applications raising ethical issues.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

(x) **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: No broader impacts discussion appears in the main text or appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

(xi) **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release high-risk models or datasets needing special safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

(xii) **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: No license or terms of use are specified for the external implementation by (Fabian et al. 2024)

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.



- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

(xiii) **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new datasets or code assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

(xiv) **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human-subject research or crowdsourcing is involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

(xv) **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The work does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

(xvi) **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core methods do not involve any use of large language models.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.