# Maximum Causal Entropy IRL in Mean-Field Games and GNEP Framework for Forward RL

**Berkay Anahtarci**                                    BERKAY.ANAHTARCI@OZYEGIN.EDU.TR
*Department of Natural and Mathematical Sciences*
*Özyeğin University*
*İstanbul, Turkey*

**Can Deha Kariksiz**                                   DEHA.KARIKSIZ@OZYEGIN.EDU.TR
*Department of Natural and Mathematical Sciences*
*Özyeğin University*
*İstanbul, Turkey*

**Naci Saldi**                                          NACI.SALDI@BILKENT.EDU.TR
*Department of Mathematics*
*Bilkent University*
*Ankara, Turkey*

## Abstract

This paper explores the use of Maximum Causal Entropy Inverse Reinforcement Learning (IRL) within the context of discrete-time stationary Mean-Field Games (MFGs) characterized by finite state spaces and an infinite-horizon, discounted-reward setting. Although the resulting optimization problem is non-convex with respect to policies, we reformulate it as a convex optimization problem in terms of state-action occupation measures by leveraging the linear programming framework of Markov Decision Processes. Based on this convex reformulation, we introduce a gradient descent algorithm with a guaranteed convergence rate to efficiently compute the optimal solution. Moreover, we develop a new method that conceptualizes the MFG problem as a Generalized Nash Equilibrium Problem (GNEP), enabling effective computation of the mean-field equilibrium for forward reinforcement learning (RL) problems and marking an advancement in MFG solution techniques. We further illustrate the practical applicability of our GNEP approach by employing this algorithm to generate data for numerical MFG examples.

**Keywords:** Mean-field games, inverse reinforcement learning, maximum causal entropy, discounted reward.

## 1. Introduction

Mean-field games (MFGs) were introduced to analyze continuous-time differential games with a vast number of agents. These agents strategically interact through a mean-field term, capturing the average distribution of the population's states. This concept was pioneered in the works of Caines et al. (2006) and Lasry and Lions (2007).

In stationary MFGs, a typical agent characterizes the collective behavior of other agents (Weintraub et al., 2005) through a time-invariant distribution, which leads to a Markov Decision Process (MDP) constrained by the state's stationary distribution. In this case, the

equilibrium, referred to as the "stationary mean-field equilibrium", involves a policy and a distribution satisfying the Nash Certainty Equivalence (NCE) principle (Caines et al., 2006). According to this principle, the policy should be optimal under a specified distribution, assumed to be the stationary infinite population limit of the mean-field term. Additionally, when the generic agent applies this policy, the resulting stationary distribution of the agent's state should align with this distribution. Under relatively mild assumptions, the existence of a stationary MFE can be proven using Kakutani's fixed-point theorem. Furthermore, it can be established that with a sufficiently large number of agents, the policy in a stationary MFE approximates a Nash equilibrium for a finite-agent scenario (Adlakha et al., 2015).

Classical MFG theory excels at computing equilibria when well-defined reward functions are provided, often employing forward reinforcement learning (RL) techniques (see Laurière et al., 2024). However, specifying the reward function in MFG problems can be difficult in practical applications. Inverse Reinforcement Learning (IRL) addresses this issue by inferring the reward function from expert demonstrations, aiming to develop policies that allow agents to effectively imitate the expert behavior. This method has significant real-world applications in diverse fields such as finance (Bernasconi et al., 2023), autonomous vehicles (Chen et al., 2023a), and epidemic control (Doncel et al., 2022).

The potential benefits of IRL are multifaceted. Firstly, by learning from demonstrations, IRL facilitates the incorporation of human knowledge and guidance into the learning process. This can be particularly valuable in situations where reward functions are complex or difficult to define. Secondly, IRL promotes interpretability and transparency in agent behavior. By inferring the reward function from demonstrations, we gain insights into the agent's motivations and decision-making process, which is useful in safety-critical applications where human-agent collaboration is essential. Finally, IRL has the potential to improve agent generalization to unseen scenarios. By focusing on the underlying objectives, IRL might lead to agents that can adapt and perform well in situations not explicitly experienced during training (see Adams et al., 2022).

Several recent papers have tackled the IRL problem in the context of MFGs. Yang et al. (2018) reduce a specific MFG to an MDP and employs the principle of maximum entropy to solve the corresponding IRL problem. However, this reduction from MFG to MDP is only applicable in fully cooperative settings, where all agents share the same societal reward. In contrast, typical MFGs involve decentralized information structures and misaligned objectives between agents. To address this, Chen et al. (2022) formulate the IRL problem for MFGs in a decentralized and non-cooperative setting, then tackle it using a maximum margin approach. Compared to these works, the most relevant one to ours is Chen et al. (2023b), making it important to highlight the differences. A key difference lies in the problem setting: Chen et al. (2023b) focus on a finite-horizon cost framework, whereas our work addresses the infinite-horizon case. As we discuss in the appendix, this distinction is crucial because, in the infinite-horizon setting, the problem cannot be formulated over the set of path probabilities, which makes a maximum likelihood approach infeasible. Another fundamental difference concerns the assumptions on the expert's trajectory. Chen et al. (2023b) assume that the expert follows an entropy-regularized mean-field equilibrium. This assumption enables them to recover the expert's policy using a likelihood-based optimization framework. In contrast, our approach assumes that the expert trajectory is drawn from a standard (non-regularized) mean-field equilibrium. Given this trajectory, we con-

struct the mean-field term and the expected feature vector, then seek policies that explain the observed behavior while maximizing entropy. Hence, the solution to our optimization problem–together with the mean-field term–yields a standard mean-field equilibrium rather than an entropy-regularized one. A further distinction lies in the uniqueness of the recovered solution. In Chen et al. (2023b), the authors ensure the uniqueness of the solution upfront by leveraging the variational formula. This guarantees that their approach recovers a unique policy that explains the observed behavior. In contrast, uniqueness is not required in our framework, multiple policies may be consistent with the observed behavior, and our goal is to select the one with the highest entropy among them. Overall, while our approaches differ in their technical formulations and underlying assumptions, they are complementary in nature. Exploring deeper connections between these formulations presents an interesting avenue for future research.

Closely related to the IRL framework is imitation learning (IL). It offers a complementary approach to the IRL framework. While IRL infers the reward function from demonstrations, IL directly learns a policy that mimics the desired behavior. This can be particularly beneficial when expert demonstrations are readily available and the reward function is difficult to define explicitly. Recent work by Ramponi et al. (2023) showcases the potential of IL applications in the context of MFGs.

It is worth mentioning that the aforementioned papers focus solely on finite-horizon cost structures, leading to convex optimization problems that employ classical maximum entropy principle and maximum margin approach. However, the classical maximum entropy principle used in Chen et al. (2023b) generally does not work for infinite-horizon problems due to the fact that the distribution induced by the state-action process on the path space becomes ill-defined in such cases. To address this limitation, Zhou et al. (2018) introduce the maximum causal entropy principle for infinite-horizon problems within MDPs. In this paper, we address the maximum entropy IRL problem in the context of infinite-horizon MFGs, where the expert trajectory is assumed to arise from a standard (non-regularized) MFE. Unlike the aforementioned approaches, our framework applies the maximum entropy principle without relying on entropy regularization. Specifically, we estimate the mean-field term and expected feature vector from this trajectory, identifying policies that maximize entropy to explain the observed behavior. Unlike approaches that enforce uniqueness of the MFE, we select the highest entropy solution from among potentially multiple MFEs that align with the behavior.

## 1.1 Contributions

1. We introduce the maximum causal entropy IRL problem for discrete-time stationary MFGs. This is an extension of the framework initially established for MDPs in Zhou et al. (2018). We address scenarios in both MDPs and MFGs where the optimality criterion is characterized by an unknown infinite-horizon discounted reward.

2. The Maximum Causal Entropy IRL problem for MFGs is inherently non-convex. To address this, we reformulate it as a convex optimization problem using a linear programming framework based on state-action occupation measures. Within this dual

context, we demonstrate the smoothness and strong convexity of the objective function. Then, we apply a gradient descent algorithm with a constant learning rate to find the optimal solution to the dual problem, ensuring a guaranteed convergence rate (Section 2).

3. We propose an alternative approach that operates under more relaxed assumptions to overcome restrictive conditions typically required for contraction-based methods in computing MFE. By reframing the MFG problem as a Generalized Nash Equilibrium Problem (GNEP) through the linear programming (LP) framework of MDPs, we adapt an algorithm from the GNEP literature to develop a novel MFE computation algorithm. This approach reduces dependency on stringent conditions such as Lipschitz continuity and strong convexity. Hence, the GNEP formulation provides a more flexible framework that can handle a broader class of MFGs. We demonstrate the method's practical applicability by using this algorithm to generate data for numerical examples (Section 3).

**Notation.** For a finite set $\mathsf{E}$, we let $\mathcal{P}(\mathsf{E})$ denote the set of all probability distributions on $\mathsf{E}$ endowed with the $l_2$-norm $\|\cdot\|$. For any $e \in \mathsf{E}$, $\delta_e$ is the Dirac delta measure. For any $a, b \in \mathbb{R}^d$, $\langle a, b \rangle$ denotes the inner product. The notation $v \sim \nu$ means that the random element $v$ has distribution $\nu$.

## 2. Maximum Causal Entropy Principle in MFGs

Drawing on the MDP fundamentals established in the appendix, this section extends the maximum causal entropy IRL framework to the realm of MFGs. We adapt the existing formulation for infinite-horizon MDPs to accommodate the unique characteristics of MFGs, specifically the interplay between individual agent behavior and the aggregate behavior of the entire population. This adaptation ensures that the principles of maximum causal entropy remain both valid and insightful within the MFG context.

A discrete-time stationary MFG is specified by $(\mathsf{X}, \mathsf{A}, p, r, \beta)$, where $\mathsf{X}$ is the finite state space and $\mathsf{A}$ is the finite action space. The components $p : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{X})$ and $r : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to [0, \infty)$ are the transition probability and the one-stage reward function, respectively. Therefore, given current state $x(t)$, action $a(t)$, and state-measure $\mu$, the reward $r(x(t), a(t), \mu)$ is received immediately, and the next state $x(t+1)$ evolves to a new state probabilistically according to the following distribution:

$$x(t+1) \sim p(\cdot | x(t), a(t), \mu).$$

In MFGs, a state-measure $\mu \in \mathcal{P}(\mathsf{X})$ represents the collective behaviour of the other agents, that is, $\mu$ can be considered as the infinite population limit of the empirical distribution of the states of other agents.[1]

To complete the description of the model dynamics, we should also specify how the agent selects its action. To that end, a policy $\pi$ is a conditional distribution on $\mathsf{A}$ given $\mathsf{X}$, that is, $\pi : \mathsf{X} \to \mathcal{P}(\mathsf{A})$. Let $\Pi$ denote the set of all policies.

---

1. In classical mean-field game literature, the exogenous behavior of other agents is typically modeled by a time-varying state measure-flow $\{\mu(t)\}$, where $\mu(t) \in \mathcal{P}(\mathsf{X})$ for all $t$. This implies a non-stationary total population behavior. However, our work only focuses on the stationary case, assuming a constant state measure-flow $\mu(t) = \mu$ for all $t$.

Before presenting the optimality notion adapted to MFGs, we first introduce the discounted reward of any policy given any state measure. In discounted MFGs, for a fixed $\mu$, the infinite-horizon discounted-reward function of any policy $\pi$ is given by

$$J_\mu(\pi, \mu_0) = E^\pi \left[ \sum_{t=0}^\infty \beta^t r(x(t), a(t), \mu) \right],$$

where $\beta \in (0, 1)$ is the discount factor and $x(0) \sim \mu_0$ is the initial state distribution. In this model, we define the set-valued mapping $\Psi : \mathcal{P}(\mathsf{X}) \to 2^\Pi$ as follows (here, $2^\Pi$ is the collection of all subsets of $\Pi$):

$$\Psi(\mu) = \{\hat\pi \in \Pi : J_\mu(\hat\pi, \mu) = \sup_\pi J_\mu(\pi, \mu)\}.$$

The set $\Psi(\mu)$ indeed represents the optimal policies for a specified $\mu$. Similarly, we define the set-valued mapping $\Lambda : \Pi \to 2^{\mathcal{P}(\mathsf{X})}$ as follows: for any $\pi \in \Pi$, the state-measure $\mu_\pi \in \Lambda(\pi)$ is an invariant distribution of the transition probability $p(\cdot | x, \pi, \mu_\pi)$, that is,

$$\mu_\pi(\cdot) = \sum_{x \in \mathsf{X}} p(\cdot | x, a, \mu_\pi) \, \pi(a|x) \, \mu_\pi(x).$$

Then, the notion of equilibrium for the MFG is defined as follows.

**Definition 1** *A pair $(\pi_*, \mu_*) \in \Pi \times \mathcal{P}(\mathsf{X})$ is called a mean-field equilibrium if $\pi_* \in \Psi(\mu_*)$ and $\mu_* \in \Lambda(\pi_*)$.*

Since IRL seeks to infer the reward function from expert demonstrations, it is important to impose some structure on the set of possible rewards to keep the problem manageable. As with MDPs, a common assumption in IRL for MFGs is that the reward can be represented as a linear combination of feature vectors, which depend on the state, action, and mean-field term:

$$\mathcal{R} := \left\{ r(x, a, \mu) = \langle \theta, f(x, a, \mu) \rangle : \theta \in \mathbb{R}^k, \; f : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}^k \right\}.$$

In this context, $f(x, a, \mu) \in \mathbb{R}^k$ is the feature vector that captures information about the corresponding state $x$, action $a$, and the mean-field term $\mu$.

In the IRL setting, we suppose that an expert generates trajectories

$$\mathcal{D} = \left\{ (x_i(t), a_i(t))_{t \geq 0} \right\}_{i=1}^d$$

under some mean-field equilibrium $(\pi_E, \mu_E)$. Since $\mu_E$ is the stationary distribution of the transition probability under policy $\pi_E$ when the mean-field term in state dynamics is $\mu_E$, the ergodic theorem implies (as $\mu_E$ is the invariant distribution of the transition probability $p(\cdot | x, \pi_E, \mu_E)$) that

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^T \left( \frac{1}{d} \sum_{i=1}^d \mathbb{1}_{\{x_i(t)=x\}} \right) = \mu_E(x)$$

for all $x \in \mathsf{X}$. In the above limit, it is sufficient to consider only one sample path in $\mathcal{D}$. To obtain a more robust estimate of $\mu_E$, one can use all sample paths in $\mathcal{D}$. Furthermore, when $d$ is sufficiently large, we can approximate the feature expectation vector as

$$\frac{1}{d} \sum_{i=1}^{d} \left( \sum_{t=0}^{\infty} \beta^t \, f(x_i(t), a_i(t), \mu_E) \right) \simeq \langle f \rangle_{\pi_E, \mu_E}, \tag{1}$$

where $\langle f \rangle_{\pi_E, \mu_E} := E^{\pi_E, \mu_E} \left[ \sum_{t=0}^{\infty} \beta^t \, f(x(t), a(t), \mu_E) \right]$ represents the expected cumulative discounted feature value under the MFE $(\pi_E, \mu_E)$, with the initial distribution also given by $\mu_E$.

Therefore, in the remainder of this paper, we suppose that the discounted feature expectation vector $\langle f \rangle_{\pi_E, \mu_E}$ under $(\pi_E, \mu_E)$ and the mean-field term $\mu_E$ are given.

**Remark 2** *In the literature on IRL for MDPs, the feature expectation vector is commonly assumed to be known. Extending this to MFGs, recent studies (Chen et al., 2022, 2023b) demonstrate that an unbiased estimate of the mean-field term can be obtained from agent trajectories. By the ergodic theorem, this estimate converges to the true mean-field term with probability 1. Therefore, the assumption that the true mean-field term $\mu_E$ is known does not impose a significant practical restriction.*[2]

We define the maximum discounted causal entropy IRL problem as follows:

$$(\textbf{OPT}_\textbf{1}) \quad \text{maximize}_\pi \quad H(\pi)$$

$$\text{subject to} \quad \pi(a|x) \geq 0 \; \forall (x,a) \in \mathsf{X} \times \mathsf{A}$$

$$\sum_{a \in \mathsf{A}} \pi(a|x) = 1 \; \forall x \in \mathsf{X}$$

$$\mu_E(x) = \sum_{(a,y) \in \mathsf{A} \times \mathsf{X}} p(x|y, a, \mu_E) \, \pi(a|y) \, \mu_E(y) \; \forall x \in \mathsf{X}$$

$$\sum_{t=0}^{\infty} \beta^t \, E^{\pi, \mu_E} [f(x(t), a(t), \mu_E)] = \langle f \rangle_{\pi_E, \mu_E},$$

where

$$H(\pi) = \sum_{t=0}^{\infty} \beta^t E^{\pi, \mu_E} \left[ -\log \pi(a(t)|x(t)) \right]$$

is the discounted causal entropy of the policy $\pi$. In this problem, the expert behaves according to some mean-field equilibrium $(\pi_E, \mu_E)$ under some unknown reward function $r_E(x, a, \mu) = \langle \theta_E, f(x, a, \mu) \rangle$. Therefore, $\pi_E$ is the optimal policy for $\mu_E$ under $r_E$. On the other hand, $\mu_E$ is the stationary distribution of the state under policy $\pi_E$ and the initial distribution $\mu_E$, when the mean-field term in state dynamics is $\mu_E$. Typically, there can be many $\theta$ values that can explain this behavior, much like in the setting of MDPs. To address this inherent ambiguity, we employ the maximum causal entropy principle, which dictates that when confronted with multiple candidates explaining the behavior, one should select

---

2. Using estimates instead of true quantities produces an approximate MFE without fundamentally changing the problem's structure, although estimation errors may occur. Since the error analysis is straightforward and adds little insight, we do not explore it further.

the one with the highest causal entropy. This allows us to avoid any bias except for the bias introduced by the feature expectation constraint.

Let $\pi^*$ be the solution of ($\mathbf{OPT_1}$). Then, we have

$$E^{\pi^*, \mu_E}[f(x(t), a(t), \mu)] = \langle f \rangle_{\pi_E, \mu_E},$$

and from (1), it follows that

$$E^{\pi^*, \mu_E}[r_E(x(t), a(t), \mu)] = E^{\pi_E, \mu_E}[r_E(x(t), a(t), \mu_E)].$$

Therefore, $\pi^*$ is also an optimal policy for $\mu_E$ likewise $\pi_E$ (i.e., $\pi^* \in \Psi(\mu_E)$). From

$$\mu_E(x) = \sum_{(a,y) \in \mathsf{A} \times \mathsf{X}} p(x|y, a, \mu_E)\, \pi^*(a|y)\, \mu_E(y) \ \forall x \in \mathsf{X},$$

it follows that $\mu_E \in \Lambda(\pi^*)$; and therefore, $(\pi^*, \mu_E)$ is a mean-field equilibrium as well. Hence, solving ($\mathbf{OPT_1}$) also leads to a MFE, similar to the MDP setting.

**Remark 3** *In ($\mathbf{OPT_1}$), we can include the mean-field term $\mu$ as a variable in addition to $\pi$ if we suppose that $\mu_E$ is not available to us and obtain*

$$
\begin{aligned}
(\mathbf{OPT_o}) \ \ maximize_{\pi, \mu} \quad & H(\pi, \mu) \\
subject \ to \quad & \pi(a|x) \geq 0 \ \forall (x, a) \in \mathsf{X} \times \mathsf{A} \\
& \sum_{a \in \mathsf{A}} \pi(a|x) = 1 \ \forall x \in \mathsf{X} \\
& \sum_{x \in \mathsf{X}} \mu(x) = 1 \\
& \mu(x) \geq 0 \ \forall x \in \mathsf{X} \\
& \mu(x) = \sum_{(a,y) \in \mathsf{A} \times \mathsf{X}} p(x|y, a, \mu)\, \pi(a|y)\, \mu(y) \ \forall x \in \mathsf{X} \\
& \sum_{t=0}^{\infty} \beta^t\, E^{\pi, \mu}[f(x(t), a(t), \mu)] = \langle f \rangle_{\pi_E, \mu_E},
\end{aligned}
$$

*where $H(\pi, \mu)$ is defined as*

$$H(\pi, \mu) := \sum_{t=0}^{\infty} \beta^t E^{\pi, \mu}\left[ -\log \pi(a(t)|x(t)) \right].$$

*If $(\pi^*, \mu^*)$ is the optimal solution to ($\mathbf{OPT_o}$), a potential issue arises when $\mu^* \neq \mu_E$. In this scenario, the policy $\pi^*$ no longer qualifies as an optimal policy for either $\mu^*$ or $\mu_E$ because the condition $\sum_{t=0}^{\infty} \beta^t E^{\pi^*, \mu^*}[f(x(t), a(t), \mu^*)] = \langle f \rangle_{\pi_E, \mu_E}$ does not imply either of these optimality results. Consequently, $(\pi^*, \mu^*)$ and $(\pi^*, \mu_E)$ cannot be considered a mean-field equilibrium, which is an undesirable outcome.*

**Remark 4** *Entropy-regularized MFGs, as developed in Cui and Koeppl (2021); Anahtarci et al. (2023), ensure the uniqueness of the MFE by incorporating an entropy term into the*

*reward structure. This regularization smooths the optimization landscape, improves convergence, and ensures the uniqueness of equilibrium solutions by favoring balanced policies. However, it leads to an approximate MFE that may deviate from the true equilibrium of the original game. An alternative approach to achieving a unique MFE is the use of ($\mathbf{OPT_1}$) which selects the MFE that maximizes entropy from among all possible equilibria corresponding to $\mu_E$. In contrast to entropy-regularized MFGs, which yield an approximate MFE, the proposed method delivers an exact MFE that is characterized by the highest entropy. This results in a more stable and balanced solution akin to that provided by entropy regularization. Nevertheless, effective implementation of this approach necessitates the computation of the MFE and the generation of the feature expectation vector, or alternatively, the direct construction of this vector, which can be computationally demanding.*

### 2.1 Convexification of ($\mathbf{OPT_1}$)

Note that ($\mathbf{OPT_1}$) is not a convex optimization problem as its last constraint, required for discounted feature expectation matching, is not convex in $\pi$. To convexify the problem, we employ normalized occupation measures induced by policies, a technique similarly used in Zhou et al. (2018) to address the IRL problem in the infinite-horizon MDPs. For any policy $\pi$, we define the state-action normalized occupation measure as

$$\nu_\pi(x, a) := (1 - \beta) \sum_{t=0}^{\infty} \beta^t E^{\pi, \mu_E} \left[ 1_{\{(x(t), a(t)) = (x,a)\}} \right].$$

The constant factor $(1 - \beta)$ in the definition makes $\nu_\pi$ a probability measure. Let us also define the corresponding marginal occupation measure as $\nu_\pi^{\mathsf{X}}(x) := \sum_{a \in \mathsf{A}} \nu_\pi(x, a)$. The following result establishes a flow condition satisfied by state-action normalized occupation measure under feasible policies of ($\mathbf{OPT_1}$).

**Lemma 5** *If $\pi$ is a feasible point for ($\mathbf{OPT_1}$), then $\nu_\pi^{\mathsf{X}}(x) = \mu_E(x)$ for every $x \in \mathsf{X}$. Moreover, $\nu_\pi$ satisfies the following flow condition:*

$$\nu_\pi^{\mathsf{X}}(x) = \sum_{(y,a) \in \mathsf{X} \times \mathsf{A}} p(x|y, a, \mu_E) \, \nu_\pi(y, a) \ \forall x \in \mathsf{X}.$$

**Proof** Without any constraint on $\pi$, this occupation measure satisfies the Bellman flow condition

$$\nu_\pi^{\mathsf{X}}(x) = (1 - \beta) \mu_E(x) + \beta \sum_{(y,a) \in \mathsf{X} \times \mathsf{A}} p(x|y, a, \mu_E) \, \nu_\pi(y, a)$$

for all $x \in \mathsf{X}$. Note that $\nu_\pi$ can be disintegrated as

$$\nu_\pi(x, a) = \pi(a|x) \, \nu_\pi^{\mathsf{X}}(x),$$

where

$$\nu_\pi^{\mathsf{X}}(x) = (1 - \beta) \sum_{t=0}^{\infty} \beta^t E^{\pi, \mu_E} \left[ 1_{\{x(t) = x\}} \right] = (1 - \beta) \sum_{t=0}^{\infty} \beta^t \mathrm{Law}^{\pi, \mu_E} \{x(t)\}(x).$$

If $\pi$ is a feasible point for $(\mathbf{OPT_1})$, then it satisfies the following additional constraint

$$\mu_E(x) = \sum_{(a,y)\in\mathsf{A}\times\mathsf{X}} p(x|y,a,\mu_E)\,\pi(a|y)\,\mu_E(y)\ \forall x\in\mathsf{X}.$$

Hence, we have $\mathrm{Law}^{\pi,\mu_E}\{x(t)\} = \mu_E$ for all $t\geq 0$, as $x(0)\sim\mu_E$. This implies $\nu_\pi^\mathsf{X} = \mu_E$, and the Bellman flow condition in this case can be written as

$$\nu_\pi^\mathsf{X}(x) = \sum_{(y,a)\in\mathsf{X}\times\mathsf{A}} p(x|y,a,\mu_E)\,\nu_\pi(y,a)\ \forall x\in\mathsf{X}.$$

$\blacksquare$

Using Lemma 5, we can also prove the following.

**Lemma 6** *If $\pi$ is a feasible point for $(\mathbf{OPT_1})$, then $H(\pi)$ can be written as*

$$H(\pi) = \frac{1}{1-\beta}\sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} -\log\left(\frac{\nu_\pi(x,a)}{\mu_E(x)}\right)\nu_\pi(x,a).$$

**Proof**   Recall that, for any policy $\pi$, we have

$$H(\pi) = \sum_{t=0}^{\infty}\beta^t E^{\pi,\mu_E}\left[-\log\,\pi(a(t)|x(t))\right]$$

$$= \sum_{t=0}^{\infty}\beta^t\sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} E^{\pi,\mu_E}\left[-\log\,\pi(a|x)\,1_{\{(x(t),a(t))=(x,a)\}}\right]$$

$$= \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}}\sum_{t=0}^{\infty}\beta^t\,E^{\pi,\mu_E}\left[-\log\,\pi(a|x)\,1_{\{(x(t),a(t))=(x,a)\}}\right]$$

$$= \frac{1}{1-\beta}\sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} -\log\left(\pi(a|x)\right)\nu_\pi(x,a).$$

Namely, one can write the discounted causal entropy of any policy $\pi$ as

$$H(\pi) = \frac{1}{1-\beta}\sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} -\log\left(\pi(a|x)\right)\nu_\pi(x,a)$$

using the occupation measure $\nu_\pi$. Note that we can disintegrate $\nu_\pi$ as

$$\nu_\pi(x,a) = \pi(a|x)\,\nu_\pi^\mathsf{X}(x).$$

Hence, we can alternatively write the discounted causal entropy of the policy $\pi$ as

$$H(\pi) = \frac{1}{1-\beta}\sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} -\log\left(\frac{\nu_\pi(x,a)}{\nu_\pi^\mathsf{X}(x)}\right)\nu_\pi(x,a).$$

9

This is true for any $\pi \in \Pi$. If $\pi$ is feasible for $(\mathbf{OPT_1})$, then by Lemma 5, $\nu_\pi^{\mathsf{X}} = \mu_E$. Hence, the discounted causal entropy of $\pi$ can also be expressed in the form of

$$H(\pi) = \frac{1}{1-\beta} \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} -\log \left( \frac{\nu_\pi(x,a)}{\mu_E(x)} \right) \nu_\pi(x,a).$$

∎

In addition to Lemma 6, the discounted feature expectation vector can also be written as

$$\sum_{t=0}^{\infty} \beta^t E^{\pi,\mu_E}[f(x(t), a(t), \mu_E)] := \langle f \rangle_{\pi,\mu_E} = \frac{1}{1-\beta} \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} f(x, a, \mu_E) \nu_\pi(x,a). \quad (2)$$

Consequently, we can define the following (convex) optimization problem, which will be proven to be equivalent to $(\mathbf{OPT_1})$:

$$
\begin{aligned}
(\mathbf{OPT_2}) \ \text{maximize}_\nu \quad & \frac{1}{1-\beta} \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} -\log \left( \frac{\nu(x,a)}{\mu_E(x)} \right) \nu(x,a) \\
\text{subject to} \quad & \frac{1}{1-\beta} \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} f(x, a, \mu_E) \nu(x,a) = \langle f \rangle_{\pi_E, \mu_E} \\
& \mu_E(x) = \sum_{(y,a) \in \mathsf{X} \times \mathsf{A}} p(x|y, a, \mu_E) \nu(y,a) \ \forall x \in \mathsf{X} \\
& \nu^{\mathsf{X}}(x) = \mu_E(x) \ \forall x \in \mathsf{X} \\
& \nu(x,a) \geq 0 \ \forall (x,a) \in \mathsf{X} \times \mathsf{A}.
\end{aligned}
$$

Here, $\nu^{\mathsf{X}}(x) := \sum_{a \in \mathsf{A}} \nu(x,a)$. Instead of formulating the problem over the set of policies, we express it with respect to the set of occupation measures induced by policies. This reformulation convexifies the problem, as demonstrated below.

**Lemma 7** *The optimization problem $(\mathbf{OPT_2})$ is convex.*

**Proof** Firstly, the equality and inequality constraints in $(\mathbf{OPT_2})$ are all linear in $\nu$. Moreover, the objective function is strongly concave in $\nu$. Indeed, rewriting the objective function as

$$\frac{1}{1-\beta} \left( \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} -\log (\nu(x,a)) \, \nu(x,a) + \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} \log (\mu_E(x)) \, \nu(x,a) \right),$$

the first term in above sum is the entropy of the distribution $\nu$, which is known to be strongly concave in $\nu$, and the second term is linear in $\nu$. Hence, the objective function is strongly concave overall. ∎

Now it is time to prove the equivalence of $(\mathbf{OPT_1})$ and $(\mathbf{OPT_2})$.

**Theorem 8** *The optimization problems $(\mathbf{OPT_1})$ and $(\mathbf{OPT_2})$ are equivalent; that is, there are mappings between feasible points and two equivalent feasible points under these mappings lead to the same objective value.*

**Proof** Let $\pi$ be a feasible point for ($\mathbf{OPT_1}$). Then, consider the corresponding occupation measure $\nu_\pi$. By Lemma 5, it follows that

$$\mu_E(x) = \sum_{(y,a)\in\mathsf{X}\times\mathsf{A}} p(x|y,a,\mu_E)\,\nu_\pi(y,a), \ \forall x \in \mathsf{X}.$$

Moreover, from (2), we also have

$$\frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E)\,\nu_\pi(x,a) = \langle f \rangle_{\pi_E,\mu_E}.$$

Therefore, $\nu_\pi$ is feasible for ($\mathbf{OPT_2}$). On the other hand, from Lemma 6 we obtain

$$H(\pi) = \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} -\log\left(\frac{\nu_\pi(x,a)}{\mu_E(x)}\right)\nu_\pi(x,a),$$

indicating that the objectives of $\pi$ and $\nu_\pi$ are equivalent. This completes the proof of the "only if" part.

Conversely, let $\nu$ be a feasible point for ($\mathbf{OPT_2}$). Then, define

$$\pi_\nu(a|x) = \begin{cases} \frac{\nu(x,a)}{\nu^\mathsf{X}(x)} & \text{if } \nu^\mathsf{X}(x) > 0, \\ \pi(a) & \text{if } \nu^\mathsf{X}(x) = 0, \end{cases}$$

where $\pi \in \mathcal{P}(\mathsf{X})$ is arbitrary. Consider the occupation measure $\nu_{\pi_\nu}$ induced by $\pi_\nu$. Since

$$\nu^\mathsf{X}(x) = \sum_{(y,a)\in\mathsf{X}\times\mathsf{A}} p(x|y,a,\mu_E)\,\pi_\nu(a|y)\,\nu^\mathsf{X}(y) \ \forall x \in \mathsf{X}$$

$$\nu^\mathsf{X}(x) = \mu_E(x) \ \forall x \in \mathsf{X}, \ \ x(0) \sim \mu_E,$$

we have $\mathrm{Law}^{\pi_\nu,\mu_E}\{x(t)\} = \nu^\mathsf{X}$ for all $t \geq 0$. Also, since

$$\nu^\mathsf{X}_{\pi_\nu} = (1-\beta) \sum_{t=0}^{\infty} \beta^t \, \mathrm{Law}^{\pi_\nu,\mu_E}\{x(t)\},$$

we have $\nu^\mathsf{X}_{\pi_\nu} = \nu^\mathsf{X} = \mu_E$. This implies that

$$\nu_{\pi_\nu}(x,a) = \pi_\nu(a|x)\,\nu^\mathsf{X}_{\pi_\nu}(x) = \pi_\nu(a|x)\,\nu^\mathsf{X}(x) = \nu(x,a) \ \forall (x,a) \in \mathsf{X}\times\mathsf{A}.$$

This completes the proof of the "if" part in view of the following:

$$H(\pi_\nu) = \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} -\log\left(\frac{\nu_{\pi_\nu}(x,a)}{\mu_E(x)}\right)\nu_{\pi_\nu}(x,a) \quad \text{(Lemma 6)}$$

$$\sum_{t=0}^{\infty} \beta^t \, E^{\pi_\nu,\mu_E}[f(x(t),a(t),\mu_E)] = \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E)\,\nu_{\pi_\nu}(x,a) \quad \text{(Eq. (2))}.$$

$\blacksquare$

**Remark 9** *The relation between feasible points of* $(\mathbf{OPT_1})$ *and* $(\mathbf{OPT_2})$ *is bijective if* $\mu_E(x) > 0$ *for all* $x \in \mathsf{X}$. *Indeed, we have*

$$\Pi \ni \pi \mapsto \nu_\pi =: (1 - \beta) \sum_{t=0}^{\infty} \beta^t \, E^{\pi,\mu_E} \left[ 1_{\{(x(t),a(t))=:(x,a)\}} \right] \in \mathcal{P}(\mathsf{X} \times \mathsf{A})$$

$$\mathcal{P}(\mathsf{X} \times \mathsf{A}) \ni \nu \mapsto \pi_\nu = \frac{\nu(x,a)}{\nu^{\mathsf{X}}(x)} \in \Pi.$$

*In both cases, we have* $\nu_\pi^{\mathsf{X}}(x) = \nu^{\mathsf{X}}(x) = \mu_E(x) > 0$ *for all* $x \in \mathsf{X}$ *by assumption. This implies that the second mapping above is injective; and so, above mappings define a bijective relation between feasible points of* $(\mathbf{OPT_1})$ *and* $(\mathbf{OPT_2})$.

## 2.2 Dual of $(\mathbf{OPT_2})$ and Its Solution

In the remainder of this section, our aim is to establish an algorithm to compute the optimal solution of $(\mathbf{OPT_2})$, as it is equivalent to $(\mathbf{OPT_1})$. To this end, we first prove the following result.

**Proposition 10** *The dual of* $(\mathbf{OPT_2})$ *is*

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{\mathsf{X}}} \left\{ \frac{1}{1-\beta} \log \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)} - \langle \boldsymbol{\theta}, \langle f \rangle_{\pi_E,\mu_E} \rangle - \sum_{x \in \mathsf{X}} \boldsymbol{\lambda}_x \, \mu_E(x) \right\},$$

*where*

$$k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) := \log \mu_E(x) + \langle \boldsymbol{\theta}, f(x,a,\mu_E) \rangle + (1-\beta) \left[ \boldsymbol{\lambda}_x + \sum_{z \in \mathsf{X}} \boldsymbol{\xi}_z \, (p(z|x,a,\mu_E) - \mu_E(z)) \right].$$

*Moreover, the optimal value of the dual problem is equal to the optimal value of* $(\mathbf{OPT_2})$; *that is, there is no duality gap. Finally, if* $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*)$ *is the optimal solution of the dual problem, then the optimal solution of* $(\mathbf{OPT_2})$ *is the following Boltzman distribution*

$$\nu^*_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}(x,a) := \frac{e^{k_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}(x,a)}}{\sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}(x,a)}}.$$

**Proof** To obtain the dual form of $(\mathbf{OPT_2})$, let us introduce the problem $(\mathbf{OPT_2})$ in a min-max formulation:

$$(\mathbf{OPT_2}) = \max_{\nu \in \mathcal{P}(\mathsf{X} \times \mathsf{A})} \min_{\boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{\mathsf{X}}} \frac{1}{1-\beta} \left[ H(\nu) + \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) \, \nu(x,a) \right]$$
$$- \langle \boldsymbol{\theta}, \langle f \rangle_{\pi_E,\mu_E} \rangle - \sum_{x \in \mathsf{X}} \boldsymbol{\lambda}_x \, \mu_E(x).$$

Then, according to Sion's minimax theorem (see Sion, 1958), we can interchange the minimum and maximum in the expression above, leading to the dual form:

$$(\mathbf{OPT_2}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{\mathsf{X}}} \max_{\nu \in \mathcal{P}(\mathsf{X} \times \mathsf{A})} \frac{1}{1-\beta} \left[ H(\nu) + \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a) \, \nu(x,a) \right]$$
$$- \langle \boldsymbol{\theta}, \langle f \rangle_{\pi_E, \mu_E} \rangle - \sum_{x \in \mathsf{X}} \boldsymbol{\lambda}_x \, \mu_E(x)$$

$$= \min_{\boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{\mathsf{X}}} \left\{ \frac{1}{1-\beta} \max_{\nu \in \mathcal{P}(\mathsf{X} \times \mathsf{A})} \left[ H(\nu) + \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a) \, \nu(x,a) \right] \right.$$
$$\left. - \langle \boldsymbol{\theta}, \langle f \rangle_{\pi_E, \mu_E} \rangle - \sum_{x \in \mathsf{X}} \boldsymbol{\lambda}_x \, \mu_E(x) \right\}$$

$$= \min_{\boldsymbol{\theta} \in \mathbb{R}^k, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{\mathsf{X}}} \left\{ \frac{1}{1-\beta} \log \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a)} - \langle \boldsymbol{\theta}, \langle f \rangle_{\pi_E, \mu_E} \rangle - \sum_{x \in \mathsf{X}} \boldsymbol{\lambda}_x \, \mu_E(x) \right\}.$$

Here, the last equality follows from the variational formula

$$\log \sum_{z \in \mathsf{Z}} e^{k(z)} = \max_{\nu \in \mathcal{P}(\mathsf{Z})} \left[ H(\nu) + \sum_{z \in \mathsf{Z}} k(z) \, \nu(z) \right]$$

(see Dupuis and Ellis, 1997, Proposition 1.4.2).[3] This implies that above minimization problem is the dual of $(\mathbf{OPT_2})$ and by Sion's minimax theorem, there is no duality gap. Moreover, for a given $\boldsymbol{\theta} \in \mathbb{R}^k$, $\boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{\mathsf{X}}$, the probability measure $\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}$ that maximizes

$$H(\nu) + \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a) \, \nu(x,a)$$

is the Boltzman distribution defined by

$$\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a) := \frac{e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a)}}{Z_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}},$$

where $Z_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}} = \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a)}$ (Dupuis and Ellis, 1997, Proposition 1.4.2). This completes the proof of the last assertion. ∎

We denote the objective function in dual formulation of $(\mathbf{OPT_2})$ as

$$g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}) := \frac{1}{1-\beta} \log \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a)} - \langle \boldsymbol{\theta}, \langle f \rangle_{\pi_E, \mu_E} \rangle - \sum_{x \in \mathsf{X}} \boldsymbol{\lambda}_x \, \mu_E(x).$$

Note that since $H(\nu) + \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a) \, \nu(x,a)$ is linear in $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$, its maximum over $\nu$, which is given by

$$\log \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x,a)},$$

---

3. Normally, in large deviation theory, the variational formula is formulated using relative entropy. However, for finite spaces, the above result can be obtained by considering the relationship between the entropy of a distribution and the relative entropy of that distribution with respect to the uniform distribution.

is a convex function of $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$. Consequently, $g$ is a convex function as the remaining terms in $g$ are linear in $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$. Hence, the dual formulation of $(\mathbf{OPT_2})$ is an unconstrained convex optimization problem in the variables $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$. Therefore, it can be solved using a gradient descent algorithm, which we will proceed to describe. Before introducing the algorithm, we first establish the smoothness and strong convexity of $g$. This guarantees convergence of the gradient descent algorithm with an explicit rate under a constant step-size. To obtain the strong convexity of $g$, we need to make the following assumption.

**Assumption 1**

$$\mathrm{span}\left\{ \left( f(x, a, \mu_E), p(\cdot|x, a, \mu_E), e(\cdot|x, a) \right) : (x, a) \in \mathsf{X} \times \mathsf{A} \right\} = \mathbb{R}^k \times \mathbb{R}^\mathsf{X} \times \mathbb{R}^\mathsf{X},$$

where

$$e(y|x, a) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise.} \end{cases}$$

**Remark 11** *Assumption 1 is not particularly restrictive, and it is easy to construct an example that satisfies this condition. To keep the discussion simple, let us consider the following scenario. Define the transition probability function as*

$$p(y|x, a, \mu_E) = \begin{cases} 1, & \text{if } l(x) = y, \\ 0, & \text{otherwise,} \end{cases}$$

*where $l : \mathsf{X} \to \mathsf{X}$ is a bijective mapping. Additionally, define the function*

$$f_{y,b}(x, a, \mu_E) = \begin{cases} 1, & \text{if } h(x, a) = (y, b), \\ 0, & \text{otherwise,} \end{cases}$$

*where $f(x, a, \mu_E) = (f_{y,b}(x, a, \mu_E))_{(y,b) \in \mathsf{X} \times \mathsf{A}}$ and $h : \mathsf{X} \times \mathsf{A} \to \mathsf{X} \times \mathsf{A}$ is also bijective. Thus, in this particular example, $k = |\mathsf{X}| + |\mathsf{A}|$. Now, let $u = (u_1, u_2, u_3) \in \mathbb{R}^k \times \mathbb{R}^\mathsf{X} \times \mathbb{R}^\mathsf{X}$ and consider the inner product of $(u_1, u_2, u_3)$ with $\left( f(x, a, \mu_E), p(\cdot|x, a, \mu_E), e(\cdot|x, a) \right)$:*

$$\sum_{(y,b) \in \mathsf{X} \times \mathsf{A}} f_{y,b}(x, a, \mu_E)\, u_1(y, b) + \sum_{y \in \mathsf{X}} p(y|x, a, \mu_E)\, u_2(y) + \sum_{y \in \mathsf{X}} e(y|x, a)\, u_3(y)$$
$$= u_1(h(x, a)) + u_2(l(x)) + u_3(x).$$

*If this inner product equals zero for all $(x, a) \in \mathsf{X} \times \mathsf{A}$, then, since $l$ and $h$ are bijections, the vector $u$ must be zero. Therefore, Assumption 1 holds for this example.*

Now, we can state the following theorem about smoothness and strong convexity of the objective function $g$.

**Theorem 12** *The objective function g is L-smooth where*

$$L := 2M \left( \frac{M_1}{1 - \beta} + 2 \sqrt{|\mathsf{X}| \, |\mathsf{A}|} \right)$$

*and the constants $M_1$ and $M$ can be determined explicitly. Moreover, under Assumption 1, for any convex compact subset $D$ in $\mathbb{R}^m$, $g$ is also $\rho(D)$-strongly convex on $D$ for some $\rho(D)$ that depends on $D$.*

**Proof** Note that the partial gradients of $g$ with respect to the vectors $\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}$ are given as follows:

$$\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}) = \frac{1}{1 - \beta} \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} f(x, a, \mu_E) \, \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) - \langle f \rangle_{\pi_E, \mu_E},$$

$$\nabla_{\boldsymbol{\xi}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}) = \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} p(\cdot | x, a, \mu_E) \, \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) - \mu_E(\cdot),$$

$$\nabla_{\boldsymbol{\lambda}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}) = \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} e(\cdot | x, a) \, \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) - \mu_E(\cdot) = \nu^{*, \mathsf{X}}_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(\cdot) - \mu_E(\cdot).$$

Therefore, to establish the Lipschitz continuity of $\nabla g$ (or, equivalently smoothness of $g$), we need to first prove Lipschitz continuity of $\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}$ with respect to $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$. The partial gradients of $\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)$ with respect to the vectors $\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}$ are given by

$$\nabla_{\boldsymbol{\epsilon}} \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)$$

$$= \frac{e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}}(x, a) \, Z_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}} \, \nabla_{\boldsymbol{\epsilon}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) - e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}}(x, a) \sum_{(y,b) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}}(y, b) \, \nabla_{\boldsymbol{\epsilon}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(y, b)}{(Z_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}})^2}$$

$$= \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) \, \nabla_{\boldsymbol{\epsilon}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) - \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a) \, \langle \nabla_{\boldsymbol{\epsilon}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}} \rangle_{\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}},$$

where $\boldsymbol{\epsilon} \in \{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}\}$. Note that we have

$$\sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} \|\nabla_{\boldsymbol{\theta}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)\| = \sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} \|f(x, a, \mu_E)\| =: M_1 < \infty,$$

$$\sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} \|\nabla_{\boldsymbol{\lambda}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)\| = \sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} (1 - \beta) \|\mathrm{e}_x\| =: M_2 < \infty,$$

$$\sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} \|\nabla_{\boldsymbol{\xi}} k_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)\| = \sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} (1 - \beta) \|p(\cdot | x, a, \mu_E) - \mu_E(\cdot)\| =: M_3 < \infty,$$

where $\mathrm{e}_x \in \mathbb{R}^{|\mathsf{X}|}$ is the vector whose $x^{th}$ term is 1 and the rest are 0. This implies that for every $\boldsymbol{\epsilon} \in \{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}\}$ we have

$$\sup_{\substack{(x,a) \in \mathsf{X} \times \mathsf{A} \\ \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}} \|\nabla_{\boldsymbol{\epsilon}} \nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)\| \leq 2 \max\{M_1, M_2, M_3\} =: 2M.$$

Hence, by the mean-value theorem, $\nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)$ is $2M$-Lipschitz continuous with respect to $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$ for all $(x,a) \in \mathsf{X} \times \mathsf{A}$. This implies that for any $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$ and $(\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')$, we have

$$\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')\| \leq \frac{M_1}{1-\beta}\, 2M \, \|(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) - (\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')\|,$$

$$\|\nabla_{\boldsymbol{\xi}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) - \nabla_{\boldsymbol{\xi}} g(\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')\| \leq 2M \, \sqrt{|\mathsf{X}|\,|\mathsf{A}|} \, \|(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) - (\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')\|,$$

$$\|\nabla_{\boldsymbol{\lambda}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) - \nabla_{\boldsymbol{\lambda}} g(\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')\| \leq 2M \, \sqrt{|\mathsf{X}|\,|\mathsf{A}|} \, \|(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) - (\boldsymbol{\theta}',\boldsymbol{\lambda}',\boldsymbol{\xi}')\|.$$

Hence $g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$ is $L$-smooth with respect to $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$, where

$$L := 2M \left( \frac{M_1}{1-\beta} + 2\,\sqrt{|\mathsf{X}|\,|\mathsf{A}|} \right).$$

Now it is time to prove the $\rho$-strong convexity of $g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$ over compact subsets. To this end, let us compute the partial Hessian of this function with respect to variables $\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}$, building upon the previously derived results. Denoting the outer (or tensor) product of two vectors by $\otimes$, we have

$$\nabla_{\boldsymbol{\theta},\boldsymbol{\theta}}^2 g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E) \otimes \nabla_{\boldsymbol{\theta}} \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)$$

$$= \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E) \otimes \left[ \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)\, \nabla_{\boldsymbol{\theta}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) - \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)\, \langle \nabla_{\boldsymbol{\theta}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}} \rangle_{\nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*} \right]$$

$$= \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E) \otimes \left[ \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)\, f(x,a,\mu_E) - \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)\, \langle f(x,a,\mu_E) \rangle_{\nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*} \right]$$

$$= \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E) \otimes f(x,a,\mu_E) \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)$$

$$- \left( \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E) \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a) \right) \otimes \langle f(x,a,\mu_E) \rangle_{\nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*}$$

$$= \frac{1}{1-\beta} \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} f(x,a,\mu_E) \otimes f(x,a,\mu_E) \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)$$

$$- \langle f(x,a,\mu_E) \rangle_{\nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*} \otimes \langle f(x,a,\mu_E) \rangle_{\nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*}.$$

Define the random vector $\mathcal{X}_f$ on the discrete probability space $(\mathsf{X} \times \mathsf{A}, \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*)$ as

$$\mathcal{X}_f(x,a) := \frac{1}{1-\beta}\, f(x,a,\mu_E) \in \mathbb{R}^k.$$

Then, the computations above imply that

$$\nabla_{\boldsymbol{\theta},\boldsymbol{\theta}}^2 g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = E[\mathcal{X}_f \otimes \mathcal{X}_f] - E[\mathcal{X}_f] \otimes E[\mathcal{X}_f] = \operatorname{Cov}(\mathcal{X}_f).$$

Similarly, we have

$$\nabla_{\boldsymbol{\xi},\boldsymbol{\xi}}^2 g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} p(\cdot|x,a,\mu_E) \otimes \nabla_{\boldsymbol{\xi}} \nu_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}^*(x,a)$$

16

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} p(\cdot|x,a,\mu_E) \otimes \left[ \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, \nabla_{\boldsymbol{\xi}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) - \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, \langle \nabla_{\boldsymbol{\xi}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}} \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}} \right]$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} p(\cdot|x,a,\mu_E) \otimes \left[ \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, p(\cdot|x,a,\mu_E) - \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, \langle p(\cdot|x,a,\mu_E) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}} \right]$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} p(\cdot|x,a,\mu_E) \otimes p(\cdot|x,a,\mu_E) \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)$$

$$- \left( \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} p(\cdot|x,a,\mu_E) \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) \right) \otimes \langle p(\cdot|x,a,\mu_E) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}}$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} p(\cdot|x,a,\mu_E) \otimes p(\cdot|x,a,\mu_E) \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)$$

$$- \langle p(\cdot|x,a,\mu_E) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}} \otimes \langle p(\cdot|x,a,\mu_E) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}}.$$

Defining the random vector $\mathcal{X}_p$ on the discrete probability space $(\mathsf{X}\times\mathsf{A}, \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}})$ as

$$\mathcal{X}_p(x,a) := p(\cdot|x,a,\mu_E) \in \mathbb{R}^{\mathsf{X}},$$

we obtain

$$\nabla^2_{\boldsymbol{\xi},\boldsymbol{\xi}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = E[\mathcal{X}_p \otimes \mathcal{X}_p] - E[\mathcal{X}_p] \otimes E[\mathcal{X}_p] = \mathrm{Cov}(\mathcal{X}_p).$$

Finally, we have

$$\nabla^2_{\boldsymbol{\lambda},\boldsymbol{\lambda}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} e(\cdot|x,a) \otimes \nabla_{\boldsymbol{\lambda}} \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} e(\cdot|x,a) \otimes \left[ \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, \nabla_{\boldsymbol{\lambda}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) - \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, \langle \nabla_{\boldsymbol{\lambda}} k_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}} \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}} \right]$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} e(\cdot|x,a) \otimes \left[ \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, e(\cdot|x,a) - \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)\, \langle e(\cdot|x,a) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}} \right]$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} e(\cdot|x,a) \otimes e(\cdot|x,a) \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)$$

$$- \left( \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} e(\cdot|x,a) \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) \right) \otimes \langle e(\cdot|x,a) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}}$$

$$= \sum_{(x,a)\in \mathsf{X}\times\mathsf{A}} e(\cdot|x,a) \otimes e(\cdot|x,a) \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a) - \langle e(\cdot|x,a) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}} \otimes \langle e(\cdot|x,a) \rangle_{\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}}.$$

Defining another random vector $\mathcal{X}_e$ on the discrete probability space $(\mathsf{X}\times\mathsf{A}, \nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}})$ as

$$\mathcal{X}_e(x,a) := e(\cdot|x,a) \in \mathbb{R}^{\mathsf{X}},$$

we similarly obtain

$$\nabla^2_{\boldsymbol{\lambda},\boldsymbol{\lambda}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = E[\mathcal{X}_e \otimes \mathcal{X}_e] - E[\mathcal{X}_e] \otimes E[\mathcal{X}_e] = \mathrm{Cov}(\mathcal{X}_e).$$

Note that one can also compute the cross terms similarly and obtain

$$\nabla^2_{\boldsymbol{\xi},\boldsymbol{\theta}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = E[\mathcal{X}_f \otimes \mathcal{X}_p] - E[\mathcal{X}_f] \otimes E[\mathcal{X}_p],$$
$$\nabla^2_{\boldsymbol{\lambda},\boldsymbol{\theta}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = E[\mathcal{X}_f \otimes \mathcal{X}_e] - E[\mathcal{X}_f] \otimes E[\mathcal{X}_e],$$
$$\nabla^2_{\boldsymbol{\xi},\boldsymbol{\lambda}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = E[\mathcal{X}_p \otimes \mathcal{X}_e] - E[\mathcal{X}_p] \otimes E[\mathcal{X}_e].$$

Hence, if we define the random vector $\boldsymbol{\mathcal{X}} := (\mathcal{X}_f, \mathcal{X}_p, \mathcal{X}_e)$, then the Hessian of $g(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$ can be written as

$$\mathrm{Hes}(g)(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) = \mathrm{Cov}(\boldsymbol{\mathcal{X}}).$$

Clearly, $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ is dependent on the parameters $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$. Moreover, each element of $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ represents an expectation of a random variable with respect to the Boltzmann distribution $\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}$. Since $\nu^*_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}}(x,a)$ has been demonstrated to be $2M$-Lipschitz continuous with respect to $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$ for all $(x,a) \in \mathsf{X} \times \mathsf{A}$, it is evident that $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ is continuous in terms of $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$.

Furthermore, as covariance matrices are inherently symmetric and positive semi-definite, it follows that $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ is positive semi-definite for any given $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$. However, to establish its positive definiteness, Assumption 1 is required.

Under this assumption, suppose in contrary that, $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ is not positive definite. Then, there exists a vector $\mathbf{a} \in \mathbb{R}^k \times \mathbb{R}^\mathsf{X} \times \mathbb{R}^\mathsf{X} =: \mathbb{R}^m$ such that $\langle \mathbf{a}, \mathrm{Cov}(\boldsymbol{\mathcal{X}})\,\mathbf{a}\rangle = 0$; that is, if $\boldsymbol{\mathcal{X}} = (\mathcal{X}_i)_{i=1}^m$, then

$$0 = \sum_{i,j=1}^m a_j \,\mathrm{Cov}(\mathcal{X}_j, \mathcal{X}_i) a_i = \mathrm{Var}\left(\sum_{i=1}^m a_i\, \mathcal{X}_i\right).$$

This implies that the random variable $\sum_{i=1}^m a_i\, \mathcal{X}_i$ is almost surely deterministic, concentrated at a point $\alpha \in \mathbb{R}$. This means that the support of the distribution of the random vector $\boldsymbol{\mathcal{X}}$ is a subset of the hyperplane $\{\boldsymbol{d} : \langle \mathbf{a}, \boldsymbol{d}\rangle = \alpha\}$; that is,

$$\mathrm{supp}\,\{\mathrm{Law}(\boldsymbol{\mathcal{X}})\} \subset \{\boldsymbol{d} : \langle \mathbf{a}, \boldsymbol{d}\rangle = \alpha\}.$$

However, since $\boldsymbol{\mathcal{X}}$ is defined as the image of the vector-valued function

$$\big(f(x, a, \mu_E), p(\cdot|x, a, \mu_E), e(\cdot|x)\big)$$

from $\mathsf{X} \times \mathsf{A}$ to $\mathbb{R}^m$, and as the probabilities of all image vectors are positive (since they are derived from the push-forward of the Boltzmann distribution $\nu^*_{\boldsymbol{\theta},\boldsymbol{\xi}}$), we must have

$$\mathrm{supp}\,\{\mathrm{Law}(\boldsymbol{\mathcal{X}})\} \not\subset \{\boldsymbol{d} : \langle \mathbf{a}, \boldsymbol{d}\rangle = \alpha\}$$

as $\mathrm{span}\,\big\{\big(f(x, a, \mu_E), p(\cdot|x, a, \mu_E), e(\cdot|x)\big) : (x,a) \in \mathsf{X} \times \mathsf{A}\big\} = \mathbb{R}^m$ by Assumption 1, which contradicts with the above conclusion. Hence, $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ is positive definite. Let $\lambda_{\min}(\boldsymbol{\mathcal{X}})$ be the minimum eigenvalue of $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$, and the positive definiteness of $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ ensures that $\lambda_{\min}(\boldsymbol{\mathcal{X}}) > 0$. Since $\mathrm{Cov}(\boldsymbol{\mathcal{X}})$ varies continuously with respect to $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$, the minimum eigenvalue $\lambda_{\min}(\boldsymbol{\mathcal{X}})$ also changes continuously concerning $(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi})$. This implies that if $D \subset \mathbb{R}^m$ is a compact subset, then

$$\min_{(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\xi}) \in D} \lambda_{\min}(\boldsymbol{\mathcal{X}}) =: \lambda_{\min}(D) > 0$$

by uniform continuity. This means that $\mathrm{Hes}(g) = \mathrm{Cov}(\boldsymbol{\mathcal{X}}) \succcurlyeq \lambda_{\min}(D)\,\mathrm{Id}$ for all $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}) \in D$, and so, $g$ is $\rho(D)$-strongly convex on $D$, where $\rho(D) := \lambda_{\min}(D)$. ∎

We may now introduce the gradient descent algorithm for finding the minimizer of $g$.

---

**Algorithm 1** Gradient Descent

---

Inputs $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0), \gamma > 0$
Start with $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0)$
**for** $k = 0, \ldots, K - 1$ **do**

$$(\boldsymbol{\theta}_{k+1}, \boldsymbol{\lambda}_{k+1}, \boldsymbol{\xi}_{k+1}) = (\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) - \gamma\,\nabla g(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k)$$

**end for**
**return** $(\boldsymbol{\theta}_K, \boldsymbol{\lambda}_K, \boldsymbol{\xi}_K)$ and $\nu^*_{\boldsymbol{\theta}_K, \boldsymbol{\lambda}_K, \boldsymbol{\xi}_K}$

---

**Remark 13** *In gradient descent for convex functions, L-smoothness, characterized by the Lipschitz continuity of the gradient, guarantees convergence to a minimum when an appropriate learning rate $\gamma$ is selected (i.e., $\gamma \leq \frac{1}{L}$). A learning rate that is too large can cause divergence or oscillations, while one that is too small can hinder convergence. The optimal learning rate typically falls within the range $(0, \frac{1}{L}]$. Unlike adaptive learning rates, which vary based on the optimization process, a constant learning rate remains unchanged throughout this algorithm. This stability simplifies the tuning process, making implementation easier in practice.*

**Theorem 14** *Suppose that the step-size in gradient descent algorithm satisfies $0 < \gamma \leq \frac{1}{L}$. Then, we have two results of increasing strength, which depend on whether Assumption 1 is imposed or not:*

*(a) For any $k$, we have*

$$g(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) - \min_{\boldsymbol{\theta} \in \mathbb{R}^k,\, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{|\mathsf{X}|}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}) \leq \frac{\|(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\|^2}{2\gamma k},$$

*where $(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k,\, \boldsymbol{\lambda}, \boldsymbol{\xi} \in \mathbb{R}^{|\mathsf{X}|}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$. Moreover,*

$$\lim_{k \to \infty} \|(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\| = 0$$

*if we run the algorithm indefinitely. Therefore, since $\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)$ is $2M$-Lipschitz continuous with respect to $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$ for all $(x, a) \in \mathsf{X} \times \mathsf{A}$, we also have*

$$\lim_{k \to \infty} \|\nu^*_{\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k} - \nu^*_{\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*}\| = 0.$$

*(b) Suppose that Assumption 1 holds. Define the following compact subset of $\mathbb{R}^m$:*

$$D := \left\{ \boldsymbol{d} \in \mathbb{R}^m : \|\boldsymbol{d} - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\| \leq \|(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\| \right\}.$$

*Then, for any $k$, we have*

$$\|(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\| \leq (1 - \gamma\, \rho(D))^k \|(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\|.$$

*Therefore, since $\nu^*_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi}}(x, a)$ is $2M$-Lipschitz continuous with respect to $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\xi})$ for all $(x, a) \in \mathsf{X} \times \mathsf{A}$, we also have*

$$\|\nu^*_{\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k} - \nu^*_{\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*}\| \leq \sqrt{|\mathsf{X}||\mathsf{A}|}\, 2M\, (1 - \gamma\, \rho(D))^k \|(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\|.$$

**Proof** Since $g$ is convex and $L$-smooth, the part (a) follows from Garrigos and Gower (2023, Theorem 3.4). In the case of part (b), by examining the proof of Garrigos and Gower (2023, Theorem 3.4), it becomes evident that for any $k$,

$$\|(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\| \leq \|(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, \boldsymbol{\xi}_0) - (\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*, \boldsymbol{\xi}_*)\|.$$

Hence $(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k, \boldsymbol{\xi}_k) \in D$ for any $k$. Then, part (b) follows from Garrigos and Gower (2023, Theorem 3.6) and the fact that $g$ is $\rho(D)$-strongly convex on $D$. ∎

Let us elaborate on the connection between the optimizer of $g$ and the policy that resolves the maximum entropy IRL problem. If $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*)$ is the minimizer of $g$, which can be computed via above gradient descent algorithm, then by Proposition 10, the optimal solution of $(\mathbf{OPT_2})$ is

$$\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x, a) = \frac{e^{k_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x, a)}}{\sum_{(x, a) \in \mathsf{X} \times \mathsf{A}} e^{k_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x, a)}}.$$

Then, in view of the proof of Theorem 8, the policy

$$\pi_{\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}}(a|x) = \frac{\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x, a)}{\nu^{*, \mathsf{X}}_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x)}$$

solves the maximum causal entropy problem $(\mathbf{OPT_1})$.

## 3. Mean-Field Game as a GNEP

This section diverges from the preceding one by shifting focus from IRL to forward RL in the context of MFGs. Here, we reformulate the MFG problem as a generalized Nash equilibrium problem (GNEP). We derive the MFE by leveraging established algorithms for GNEPs (see Facchinei and Kanzow, 2010; Dreves et al., 2011). This alternative solution framework complements our prior analysis of the IRL problem for MFGs, providing a comprehensive perspective on MFG solution techniques. In the subsequent discussion, we provide a detailed rationale for our GNEP formulation, underscoring its key advantages and its broader applicability to a range of problems that may not be readily addressed by conventional MFE computation techniques (see Laurière et al., 2024).

The standard approach to compute the MFE in MFGs involves iteratively solving an optimal control problem, which reduces to finding an optimal stationary policy for an MDP

parameterized by a given mean-field term $\mu \in \mathcal{P}(\mathsf{X})$. This can be achieved using established algorithms like value iteration or $Q$-iteration. Alternatively, a policy iteration-like approach can be used, deriving an improved policy greedily from the $Q$-function induced by the previous policy under $\mu$, without affecting overall convergence properties.

Once the optimal or improved policy $\pi$ is determined, the invariant distribution $\mu^{\mathsf{inv}}$ associated with the transition probability $p(\cdot\,|x, \pi, \mu)$ can be computed. The new mean-field term is then defined as a convex combination of the current mean-field term and this invariant distribution, defining an operator $H$ mapping $\mathcal{P}(\mathsf{X})$ to $\mathcal{P}(\mathsf{X})$ as

$$\mu^+ = H(\mu) = (1 - \alpha)\mu + \alpha\mu^{\mathsf{inv}}.$$

Convergence to an MFE relies on showing $H$ is a contraction mapping ensuring that iterates converge to a unique fixed point $\mu^* = H(\mu^*)$. Consequently, $\mu^*$ and the corresponding policy $\pi^*$ constitute an MFE, as applied in various studies (see Laurière et al., 2024, Section 3.4).

Establishing the contraction of $H$ requires restrictive conditions on the MFG model, such as Lipschitz continuity, strong convexity, and smoothness, with additional constraints on the relevant constants. While some studies establish convergence without a contraction condition, these typically apply to continuous-time settings or specifically structured MFGs such as potential or linear-quadratic MFGs (see Laurière et al., 2024, Section 3.4). The GNEP approach offers a distinct advantage over standard methods by imposing more relaxed and readily verifiable conditions on the game structure. Notably, it circumvents the need for stringent conditions, such as Lipschitz continuity of system components or the contraction property of $H$, that are often required by conventional techniques.

Our GNEP-based approach, while requiring twice continuous differentiability and convexity conditions on the model components, provides a new methodology for computing MFEs. While convergence relies on the invertibility of a specific Jacobian, this can be addressed practically by employing the Moore-Penrose pseudoinverse and monitoring convergence behavior. Although initialization sensitivity arises (similar to neural networks), this framework allows leveraging a wide range of existing GNEP solution algorithms (see Facchinei and Kanzow, 2010), significantly expanding the tools available for MFE computation.

Extending prior work on linear MFGs by the third author (see Saldi, 2023), this paper proposes a novel approach for computing the MFE in classical non-linear MFGs. We take advantage of the flexibility of our GNEP formulation, adapting an interior-point method from Dreves et al. (2011) originally designed for solving the Karush-Kuhn-Tucker (KKT) conditions of GNEPs. Our main approach is to formulate the MDP associated with a given mean-field term, $\mu \in \mathcal{P}(\mathsf{X})$, as an LP problem utilizing occupation measures. This is a well-established technique in stochastic control, and we refer the reader to Hernández-Lerma and González-Hernández (2000); Hernández-Lerma and Lasserre (1996) for an in-depth discussion of the LP formulation of MDPs. Incorporating the mean-field consistency condition directly into this LP leads to a natural representation of the MFG problem as a GNEP.

### 3.1 GNEP Formulation

For a finite set $\mathsf{E}$, let $\mathcal{M}(\mathsf{E})$ denote the set of finite signed measures on $\mathsf{E}$ and $\mathcal{F}(\mathsf{E})$ denote the set of real functions on $\mathsf{E}$ (i.e., $\mathcal{M}(\mathsf{E}) = \mathcal{F}(\mathsf{E}) = \mathbb{R}^{\mathsf{E}}$). We define bilinear forms on

$\big(\mathcal{M}(\mathsf{X} \times \mathsf{A}), \mathcal{F}(\mathsf{X} \times \mathsf{A})\big)$ and on $\big(\mathcal{M}(\mathsf{X}), \mathcal{F}(\mathsf{X})\big)$ as inner products

$$\langle \nu, v \rangle := \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} v(x,a)\, \nu(x,a),$$

$$\langle \mu, u \rangle := \sum_{x \in \mathsf{X}} u(x)\, \mu(x),$$

where $\nu \in \mathcal{M}(\mathsf{X} \times \mathsf{A})$, $v \in \mathcal{F}(\mathsf{X} \times \mathsf{A})$, $\mu \in \mathcal{M}(\mathsf{X})$, and $u \in \mathcal{F}(\mathsf{X})$. We define the linear map $\mathrm{T}_\mu : \mathcal{M}(\mathsf{X} \times \mathsf{A}) \to \mathcal{M}(\mathsf{X})$ by

$$\mathrm{T}_\mu \nu(\,\cdot\,) = \nu^{\mathsf{X}}(\,\cdot\,) - \beta \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} p_\mu(\,\cdot\,|x,a)\, \nu(x,a) =: \nu^{\mathsf{X}} - \beta\, \nu\, p_\mu,$$

which depends on $\mu$. Recall that, for a given $\mu$, the corresponding MDP, denoted as $\mathrm{MDP}_\mu$, has the following components

$$\{\mathsf{X}, \mathsf{A}, r_\mu, p_\mu, \mu\},$$

where

$$r_\mu(x,a) := r(x,a,\mu), \;\; p_\mu(\,\cdot\,|\,x,a) := p(\,\cdot\,|\,x,a,\mu), \;\; x(0) \sim \mu.$$

Then, the optimal control problem associated to $\mathrm{MDP}_\mu$ is equivalent to the following equality constrained linear program (see Hernández-Lerma and González-Hernández, 2000, Lemma 3.3 and Section 4):

$$\mathrm{maximize}_{\nu \in \mathcal{M}_+(\mathsf{X} \times \mathsf{A})} \; \langle \nu, r_\mu \rangle$$
$$\text{subject to } \mathrm{T}_\mu(\nu) = (1 - \beta)\mu,$$

where $\mathcal{M}_+(\mathsf{E})$ denotes the set of positive measures on the finite set $\mathsf{E}$. Here, the feasible points $\nu$ of above linear program indeed corresponds to the set of (normalized) state-action occupation measures of policies. Using this LP formulation, we first establish the following result.

**Lemma 15** *Let $(\nu^*, \mu^*) \in \mathcal{M}(\mathsf{X} \times \mathsf{A})_+ \times \mathcal{M}(\mathsf{X})_+$ be a pair with the following properties:*

*(a) $\nu^*$ is the optimal solution to the above LP formulation of $\mathrm{MDP}_{\mu^*}$.*

*(b) $\mu^*$ satisfies the following equation*

$$\mu^*(\,\cdot\,) = \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} p_{\mu^*}(\,\cdot\,|x,a)\, \nu^*(x,a).$$

*Disintegrating $\nu^*$ as $\nu^*(x,a) = \pi^*(a|x)\, \nu^{*,\mathsf{X}}(x)$, the pair $(\pi^*, \mu^*)$ is an MFE for the related MFG.*

**Remark 16** *In part (b) of the above lemma, if we had the expression*

$$\mu^*(\,\cdot\,) = \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} p_{\mu^*}(\,\cdot\,|x,a)\,\pi^*(a|x)\,\mu^*(x),$$

*where $\nu^*(x,a) = \pi^*(a|x)\,\mu^*(x)$, the conclusion would be straightforward. However, in the given statement, $\pi^*(a|x)\,\mu^*(x)$ is replaced by $\nu^*(x,a)$, allowing us to avoid an additional non-linearity introduced by the product of $p_{\mu^*}(\,\cdot\,|x,a)$ and $\mu^*(\,\cdot\,)$, since both terms depend on $\mu^*$. As we will see, this expression forms a constraint for the second player in the GNEP formulation. Therefore, avoiding unnecessary non-linearities is crucial, and this is the significance of this result.*

**Proof** The proof is analogous to the linear MFG case (see Saldi, 2023, Lemma 5.1). For the sake of completeness, we provide a detailed proof.

Note that $\mu^*$ and $\nu^*$ are initially not assumed to be probability measures, and so, we establish this first. Since

$$\nu^{*,\mathsf{X}} = (1-\beta)\,\mu^* + \beta\,\nu^*\,p_{\mu^*}, \tag{3}$$

we have $\nu^*(\mathsf{X}\times\mathsf{A}) = (1-\beta)\,\mu^*(\mathsf{X}) + \beta\,\nu^*(\mathsf{X}\times\mathsf{A})\,\mu^*(\mathsf{X})$. Similarly, since

$$\mu^* = \nu^*\,p_{\mu^*},$$

we have $\mu^*(\mathsf{X}) = \nu^*(\mathsf{X}\times\mathsf{A})\,\mu^*(\mathsf{X})$, which implies that $\nu^*$ is a probability measure. In view of this and using (3), we obtain the following:

$$1 = (1-\beta)\,\mu^*(\mathsf{X}) + \beta\,\mu^*(\mathsf{X}) = \mu^*(\mathsf{X}).$$

Therefore, $\mu^*$ is also a probability measure.

Note that $\nu^*$ is the optimal occupation measure of the LP formulation of $\mathrm{MDP}_{\mu^*}$, and so, $\pi^*$ is the optimal policy. Hence, $\pi^* \in \Psi(\mu^*)$. Furthermore, since

$$\nu^*(x,a) := (1-\beta)\sum_{t=0}^{\infty}\beta^t\,E^{\pi^*,\mu^*}\left[1_{\{(x(t),a(t))=(x,a)\}}\right]$$

$$= (1-\beta)\sum_{t=0}^{\infty}\beta^t\,\mathsf{Pr}^{\pi^*,\mu^*}\left[(x(t),a(t))=(x,a)\right],$$

we have

$$\nu^*\,p_{\mu^*}(\,\cdot\,) = \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} p_{\mu^*}(\,\cdot\,|x,a)\,\nu^*(x,a)$$

$$= \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} p_{\mu^*}(\,\cdot\,|x,a)\left\{(1-\beta)\sum_{t=0}^{\infty}\beta^t\,\mathsf{Pr}^{\pi^*,\mu^*}\left[(x(t),a(t))=(x,a)\right]\right\}$$

$$= (1-\beta)\sum_{t=0}^{\infty}\beta^t\left\{\sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} p_{\mu^*}(\,\cdot\,|x,a)\,\mathsf{Pr}^{\pi^*,\mu^*}\left[(x(t),a(t))=(x,a)\right]\right\}$$

23

$$= (1 - \beta) \sum_{t=0}^{\infty} \beta^t \, \mathrm{Pr}^{\pi^*, \mu^*} \left[ x(t+1) \in \cdot \right]$$

$$= \frac{1 - \beta}{\beta} \sum_{t=1}^{\infty} \beta^t \, \mathrm{Pr}^{\pi^*, \mu^*} \left[ x(t) \in \cdot \right] + \frac{1 - \beta}{\beta} \, \mathrm{Pr}^{\pi^*, \mu^*} \left[ x(0) \in \cdot \right] - \frac{1 - \beta}{\beta} \, \mathrm{Pr}^{\pi^*, \mu^*} \left[ x(0) \in \cdot \right]$$

$$= \frac{1 - \beta}{\beta} \sum_{t=0}^{\infty} \beta^t \, \mathrm{Pr}^{\pi^*, \mu^*} \left[ x(t) \in \cdot \right] - \frac{1 - \beta}{\beta} \mu^*(\cdot) \ (\text{as } x(0) \sim \mu^*)$$

$$= \frac{\nu^{*, \mathsf{X}}(\cdot)}{\beta} - \frac{\mu^*(\cdot)}{\beta} + \mu^*(\cdot).$$

As $\mu^* = \nu^* p_{\mu^*}$, the last expression implies that $\nu^{*, \mathsf{X}} = \mu^*$. Hence, $\mu^*$ satisfies (see remark above)

$$\mu^*(\cdot) = \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} p_{\mu^*}(\cdot \,|x, a) \, \pi^*(a|x) \, \mu^*(x),$$

that is, $\mu^* \in \Lambda(\pi^*)$. This means that $(\pi^*, \mu^*)$ is an MFE. ∎

In order to determine the MFE, it is now sufficient to compute a pair $(\nu^*, \mu^*)$, which satisfies the conditions outlined in Lemma 15. To achieve this, we formulate an artificial game involving two players, which naturally leads to a GNEP. Solving for the Nash equilibrium in this artificial game yields the desired pair.

In the two-player game, the first player represents the generic agent in the MFG, while the second player represents the entire population. In this setup, an additional reward function alongside $\langle \nu, r_\mu \rangle$ is required, where it serves as the reward for the second player. It is worth mentioning that we possess complete freedom in selecting this reward function. Hence, one can regard this extra reward as a design parameter, which can be tailored to fulfill specific objectives.

Let $h : \mathcal{M}(\mathsf{X} \times \mathsf{A}) \times \mathcal{M}(\mathsf{X}) \to [0, \infty)$ be a continuous artificial objective function for the second player, which depends on both $(\nu, \mu)$. With this, we formulate the following GNEP.

| Player 1 | Player 2 |
|---|---|
| Given $\mu$: maximize$_{\nu \in \mathcal{M}_+(\mathsf{X} \times \mathsf{A})} \langle \nu, r_\mu \rangle$ | Given $\nu$: maximize$_{\mu \in \mathcal{M}_+(\mathsf{X})} h(\nu, \mu)$ |
| subject to $\nu^{\mathsf{X}} = (1 - \beta)\mu + \beta \, \nu \, p_\mu$ | subject to $\mu = \nu \, p_\mu$ |

It is worth noting that in this formulation, there is an interdependency between the two reward functions and the admissible strategy sets. As a result, this situation precisely fits the definition of a GNEP. This allows us to employ techniques and methods that have been developed for solving such games in our pursuit of computing the MFE. For a more comprehensive introduction to GNEPs, we refer the reader to the survey by Facchinei and Kanzow (2010).

The following result immediately follows from Lemma 15.

**Lemma 17** *If $(\nu^*, \mu^*)$ is a Nash equilibrium of the GNEP above, then $(\pi^*, \mu^*)$ is an MFE for the related MFG, where $\nu^*(x, a) = \pi^*(a|x) \, \nu^{*, \mathsf{X}}(x)$.*

Typically, GNEPs are expressed with inequality constraints rather than equality constraints. While it is possible to transform equality constraints into inequality constraints by duplicating them, we can instead use an alternative approach that keeps the number of constraints manageable, as shown below. Additionally, GNEPs are usually framed as a minimization problem for each agent. To align with this, we replace the terms $\langle \nu, r_\mu \rangle$ and $h(\nu, \mu)$ with their negative counterparts, $\langle \nu, -r_\mu \rangle$ (denoted as $\langle \nu, c_\mu \rangle$) and $-h(\nu, \mu)$ (denoted as $g(\nu, \mu)$). These changes allow us to reformulate each problem as a minimization.

|  Player 1 | Player 2 |
|---|---|
| Given $\mu$: minimize$_{\nu \in \mathcal{M}_+(\mathsf{X} \times \mathsf{A})}$ $\langle \nu, c_\mu \rangle$ | Given $\nu$: minimize$_{\mu \in \mathcal{M}_+(\mathsf{X})}$ $g(\nu, \mu)$ |
| subject to $\nu^{\mathsf{X}} \geq (1 - \beta)\mu + \beta \, \nu \, p_\mu$ | subject to $\mu \geq \nu \, p_\mu$, $\langle \mu, \mathbf{1} \rangle \geq 1$ |

Here, $\mathbf{1}$ denotes the constant function equal to 1. To represent the GNEP with inequality constraints, we introduce the extra constraint $\langle \mu, \mathbf{1} \rangle \geq 1$, which does not substantially expand the number of constraints.

**Lemma 18** *If $(\nu^*, \mu^*)$ is an equilibrium solution of the GNEP with inequality constraint, then $(\pi^*, \mu^*)$ is an MFE for the related MFG, where $\nu^*(x, a) = \pi^*(a|x) \, \nu^{*,\mathsf{X}}(x)$.*

**Proof** We initially prove that both $\nu^*$ and $\mu^*$ are probability measures. Since

$$\nu^{*,\mathsf{X}} \geq (1 - \beta) \, \mu^* + \beta \, \nu^* \, p_{\mu^*},$$

we have $\nu^*(\mathsf{X} \times \mathsf{A}) \geq (1 - \beta) \, \mu^*(\mathsf{X}) + \beta \, \nu^*(\mathsf{X} \times \mathsf{A}) \, \mu^*(\mathsf{X})$. Similarly, since

$$\mu^* \geq \nu^* \, p_{\mu^*},$$

we have $\mu^*(\mathsf{X}) \geq \nu^*(\mathsf{X} \times \mathsf{A}) \, \mu^*(\mathsf{X})$. Hence,

$$\nu^*(\mathsf{X} \times \mathsf{A}) \geq (1 - \beta)\nu^*(\mathsf{X} \times \mathsf{A}) \, \mu^*(\mathsf{X}) + \beta \, \nu^*(\mathsf{X} \times \mathsf{A}) \, \mu^*(\mathsf{X}) = \nu^*(\mathsf{X} \times \mathsf{A}) \, \mu^*(\mathsf{X}).$$

Therefore, $\mu^*(\mathsf{X}) \leq 1$ and $\nu^*(\mathsf{X} \times \mathsf{A}) \leq 1$. Since $\langle \mu^*, \mathbf{1} \rangle = \mu^*(\mathsf{X}) \geq 1$, we also have $\mu^*(\mathsf{X}) = 1$, and so,

$$\nu^*(\mathsf{X} \times \mathsf{A}) \geq (1 - \beta) + \beta \, \nu^*(\mathsf{X} \times \mathsf{A}).$$

Hence, $\nu^*(\mathsf{X} \times \mathsf{A}) \geq 1$. This implies that $\nu^*(\mathsf{X} \times \mathsf{A}) = 1$. That is, both $\nu^*$ and $\mu^*$ are probability measures.

In view of the above result, $(1 - \beta) \, \mu^* + \beta \, \nu^* \, p_{\mu^*}$ and $\nu^* \, p_{\mu^*}$ are also probability measures. But it is known that if two probability measures $\nu$ and $\theta$ satisfy $\nu \leq \theta$ for any event, then $\nu = \theta$. Hence, $\nu^{*,\mathsf{X}} = (1 - \beta) \, \mu^* + \beta \, \nu^* \, p_{\mu^*}$, $\mu^* = \nu^* \, p_{\mu^*}$ Note that given $\mu^*$, the following optimization problems are equivalent

|  Problem 1 | Problem 2 |
|---|---|
| minimize$_{\nu \in \mathcal{M}_+(\mathsf{X} \times \mathsf{A})}$ $\langle \nu, c_{\mu^*} \rangle$ | minimize$_{\nu \in \mathcal{M}_+(\mathsf{X} \times \mathsf{A})}$ $\langle \nu, c_{\mu^*} \rangle$ |
| subject to $\nu^{\mathsf{X}} \geq (1 - \beta)\mu^* + \beta \, \nu \, p_{\mu^*}$ | subject to $\nu^{\mathsf{X}} = (1 - \beta)\mu^* + \beta \, \nu \, p_{\mu^*}$ |

The second problem is an LP formulation of $\mathrm{MDP}_{\mu^*}$ and so $\nu^*$ is the optimal occupation measure. Hence, $\pi^* \in \Lambda(\mu^*)$. With the same analysis as in the proof of Lemma 15, we can also prove that $\nu^{*,\mathsf{X}} = \mu^*$ using $\mu^* = \nu^* p_{\mu^*}$. Hence, by $\nu^{*,\mathsf{X}} = (1-\beta)\,\mu^* + \beta\,\nu^*\,p_{\mu^*}$, we have

$$\mu^*(\,\cdot\,) = \sum_{(x,a) \in \mathsf{X} \times \mathsf{A}} p_{\mu^*}(\,\cdot\,|x,a)\,\pi^*(a|x)\,\mu^*(x);$$

that is, $\mu^* \in \Phi(\pi^*)$. This implies that $(\mu^*, \pi^*)$ is MFE. ∎

The primary novelty of our approach lies in the GNEP formulation above and its connection to the MFE. This framework enriches the MFG analysis while offering flexibility, as existing GNEP-solving algorithms (see Facchinei and Kanzow, 2010) can now be applied to compute the MFE.

In the upcoming section, we will adapt an algorithm proposed by Dreves et al. (2011) to compute an MFE. This algorithm employs an interior-point method specifically designed to solve the Karush-Kuhn-Tucker (KKT) conditions associated with the GNEP formulation.

### 3.2 Computing Equilibrium of GNEP

We now give a more explicit formulation of the inequality constrained GNEP that is introduced in the previous section.

| Player 1 | Player 2 |
|---|---|
| Given $\mu$: $\text{minimize}_{\nu \in \mathbb{R}^{\mathsf{X} \times \mathsf{A}}}\ \langle \nu, c_\mu \rangle$ | Given $\nu$: $\text{minimize}_{\mu \in \mathbb{R}^{\mathsf{X}}}\ g(\nu, \mu)$ |
| subject to $\nu^{\mathsf{X}} \geq (1-\beta)\mu + \beta\,\nu\,p_\mu$ | subject to $\mu \geq \nu\,p_\mu,\ \langle \mu, \mathbf{1} \rangle \geq 1$ |
| $\mathrm{Id} \cdot \nu \geq 0$ | $\mathrm{Id} \cdot \mu \geq 0$ |

To solve this problem, we adapt an algorithm introduced by Dreves et al. (2011) that employs an interior-point method leading to a solution for KKT conditions. Since we have the flexibility in selecting the auxiliary cost function $g$ for the second player, we assume that $g$ is twice-continuously differentiable and convex with respect to $\mu$ for any given $\nu$.

**Assumption 2** *We suppose that $p_\mu$ and $c_\mu$ are twice-continuously differentiable with respect to $\mu$. Moreover, $p_\mu(y|x,a)$ is convex with respect to $\mu$, for all $(y,x,a) \in \mathsf{X} \times \mathsf{X} \times \mathsf{X}$.*

In linear MFGs (see Saldi, 2023), the state transition probability $p_\mu$ and the one-stage reward function $r_\mu$ are linear in $\mu$, meaning the above assumption is unnecessary for the linear case. However, in the GNEP formulation, both the objectives and constraints become nonlinear. As a result, when the policy of one agent is fixed, the optimization problem faced by the other agent turns into a nonlinear programming problem, unlike in linear MFGs.

Under these conditions, the inequality constrained GNEP meets the requirements in assumptions A1 and A2 in Dreves et al. (2011).

Let us define the functions $h_1 : \mathbb{R}^{\mathsf{X} \times \mathsf{A}} \times \mathbb{R}^{\mathsf{X}} \to \mathbb{R}^{\mathsf{X} \times \mathsf{A}} \times \mathbb{R}^{\mathsf{X}}$ and $h_2 : \mathbb{R}^{\mathsf{X} \times \mathsf{A}} \times \mathbb{R}^{\mathsf{X}} \to \mathbb{R}^{\mathsf{X}} \times \mathbb{R} \times \mathbb{R}^{\mathsf{X}}$ as

$$h_1(\nu, \mu) := \begin{pmatrix} -\operatorname{Id} \cdot \nu \\ -\nu^{\mathsf{X}} + (1 - \beta)\mu + \beta \, \nu \, p_\mu \end{pmatrix}, \quad h_2(\nu, \mu) := \begin{pmatrix} -\operatorname{Id} \cdot \mu \\ -\langle \mu, \mathbf{1} \rangle + 1 \\ -\mu + \nu \, p_\mu \end{pmatrix}.$$

Then we can write the GNEP above in the following form:

| Player 1 | Player 2 |
|---|---|
| Given $\mu$: $\text{minimize}_{\nu \in \mathbb{R}^{\mathsf{X} \times \mathsf{A}}} \ \langle \nu, c_\mu \rangle$ | Given $\nu$: $\text{minimize}_{\mu \in \mathbb{R}^{\mathsf{X}}} \ g(\nu, \mu)$ |
| subject to $h_1(\nu, \mu) \leq 0$ | subject to $h_2(\nu, \mu) \leq 0$ |

Now, let us derive the joint KKT conditions for player 1 and player 2, whose solution gives a Nash equilibrium for GNEP. To this end, we start by defining the functions

$$L_1(\nu, \mu, \lambda) := \langle \nu, c_\mu \rangle + \langle h_1(\nu, \mu), \lambda \rangle,$$
$$L_2(\nu, \mu, \gamma) := g(\nu, \mu) + \langle h_2(\nu, \mu), \gamma \rangle,$$

where $\lambda$ and $\gamma$ are Lagrange multipliers of player 1 and player 2, respectively. Setting $\mathbf{F}(\nu, \mu, \lambda, \gamma) := (\nabla_\nu L_1(\nu, \mu, \lambda), \nabla_\mu L_2(\nu, \mu, \gamma))$ and $\mathbf{h}(\nu, \mu) := (h_1(\nu, \mu), h_2(\nu, \mu))$, the joint KKT conditions for player 1 and player 2 can be written as

$$\mathbf{F}(\nu, \mu, \lambda, \gamma) = 0, \ \lambda, \gamma \geq 0, \ \mathbf{h}(\nu, \mu) \leq 0, \ \langle \mathbf{h}(\nu, \mu), (\lambda, \gamma) \rangle = 0.$$

**Root Finding Problem and Algorithm** To transform the joint KKT conditions into a root finding problem, we introduce slack variables $(\bar{\lambda}, \bar{\gamma})$, where $\bar{\lambda} \in \mathbb{R}^{\mathsf{X} \times \mathsf{A}} \times \mathbb{R}^{\mathsf{X}}$ and $\bar{\gamma} \in \mathbb{R}^{\mathsf{X}} \times \mathbb{R} \times \mathbb{R}^{\mathsf{X}}$, and define

$$H(z) := H(\nu, \mu, \lambda, \gamma, \bar{\lambda}, \bar{\gamma}) := \begin{pmatrix} \mathbf{F}(\nu, \mu, \lambda, \gamma) \\ \mathbf{h}(\nu, \mu) + (\bar{\lambda}, \bar{\gamma}) \\ (\lambda, \gamma) \circ (\bar{\lambda}, \bar{\gamma}) \end{pmatrix}$$

and $Z := \{ z = (\nu, \mu, \lambda, \gamma, \bar{\lambda}, \bar{\gamma}) : (\lambda, \gamma), (\bar{\lambda}, \bar{\gamma}) \geq 0 \}$, where $(\lambda, \gamma) \circ (\bar{\lambda}, \bar{\gamma})$ is the vector formed by diagonal elements of the outer product of the vectors $(\lambda, \gamma)$ and $(\bar{\lambda}, \bar{\gamma})$. Then it is straightforward to show that $(\nu, \mu, \lambda, \gamma)$ satisfy joint KKT conditions if and only if $z := (\nu, \mu, \lambda, \gamma, \bar{\lambda}, \bar{\gamma})$ satisfies the constrained root finding problem

$$H(z) = 0, \ z \in Z$$

for some $(\bar{\lambda}, \bar{\gamma})$.

To find a solution to the constrained root finding problem, an interior-point algorithm is developed in Dreves et al. (2011). In the remainder of this section, we explain this algorithm, which depends on the potential reduction method from Monteiro and Pang (1999). Let $n = |\mathsf{X} \times \mathsf{A}| + |\mathsf{X}|$ (total number of variables in GNEP) and $m := |\mathsf{X} \times \mathsf{A}| + 3\,|\mathsf{X}| + 1$ (total number of constraints in the GNEP). Hence $H : \mathbb{R}^n \times \mathbb{R}^{2m} \to \mathbb{R}^n \times \mathbb{R}^{2m}$ and $Z = \mathbb{R}^n \times \mathbb{R}^{2m}_+$. We take a potential function on the interior of $Z$ as

$$p(u,v) = K \, \log\left(\|u\|^2 + \|v\|^2\right) - \sum_{i=1}^{2m} \log(v_i),$$

where $K > m$, which penalizes points that are close to the boundary of $Z$ that are far from the origin. Composing $p$ and $H$, we get a potential function for the constrained root finding problem as

$$\psi(z) := p(H(z)),$$

where $z \in (\mathrm{int}\, Z) \cap H^{-1}(\mathrm{int}\, Z) =: Z_I$.

Let $\nabla H$ denote the Jacobian of the function $H$. Before describing the algorithm, we need to impose the following condition to establish its convergence.

**Assumption 3** *For any $z \in Z_I$, the Jacobian $\nabla H(z)$ is invertible.*

In general, providing a sufficient condition for the invertibility of the Jacobian of $H$ at any point $z \in Z_I$ in terms of system components is quite challenging. Therefore, for practical applications of the below algorithm, it is preferable to substitute the inverse of $\nabla H(z)$ with its Moore-Penrose pseudo-inverse in order to be on the safe side. Let

$$a := \begin{pmatrix} \mathbf{0}_n^T & \mathbf{1}_{2m}^T \end{pmatrix} \Big/ \left\| \begin{pmatrix} \mathbf{0}_n^T & \mathbf{1}_{2m}^T \end{pmatrix} \right\|.$$

Then, the algorithm is given below.

---
**Algorithm 2**

---
Inputs: $\kappa \in (0,1)$ and $a$
Start with $z_0$
**for** $k = 0, 1, 2 \ldots$ **do**
  (a)    Choose $\sigma_k \in [0,1)$ and set

$$d_k = (\nabla H(z_k))^{-1} \left(\sigma_k \langle a, H(z_k)\rangle\, a - H(z_k)\right)$$

  (b)    Compute a stepsize $t_k := \max\{\kappa^l : l = 0, 1, 2, \ldots\}$ such that

$$z_k + t_k\, d_k \in Z_I, \quad \psi(z_k + t_k\, d_k) \le \psi(z_k) + t_k \, \langle \nabla \psi(z_k), d_k\rangle \tag{4}$$

  (c)    Set $z_{k+1} := z_k + t_k\, d_k$
**end for**

---

Note that by Facchinei and Pang (2003, Lemma 11.3.3), we have $\langle \nabla \psi(z_k), d_k\rangle < 0$ for

$$d_k = (\nabla H(z_k))^{-1} \left(\sigma_k \langle a, H(z_k)\rangle\, a - H(z_k)\right).$$

Hence, one can always find $t_k$ that satisfies (4). The following convergence result follows from Dreves et al. (2011, Theorems 4.3 and 4.10).

**Theorem 19** *Under Assumption 3, pick $\sigma_k$ so that*

$$\limsup_{k \to \infty} \sigma_k < 1.$$

*Then, the sequence $\{z_k\} := \{(\nu_k, \mu_k, \lambda_k, \gamma_k, \bar{\lambda}_k, \bar{\gamma}_k)\}$ is bounded and any accumulation point $z^* = (\nu^*, \mu^*, \lambda^*, \gamma^*, \bar{\lambda}^*, \bar{\gamma}^*)$ of $\{z_k\}$ is a solution to the constrained root finding problem $H(z^*) = 0$; that is, $H(z_k) \to 0$ as $k \to \infty$. Hence, $(\pi^*, \mu^*)$ is an MFE for the related MFG, where $\nu^*(x, a) = \pi^*(a|x)\, \nu^{*,\mathsf{X}}(x)$.*

This algorithm provides a notable improvement compared to existing methods for MFGs. Through tracking the changes in the components of the vector function $H$ during each step of the algorithm, we can confidently confirm the convergence to the MFE when these components approach zero.

## 4. Numerical Examples

In this section, we consider two numerical examples based on the malware spread model studied in Subramanian and Mahajan (2019).

### 4.1 A Malware Spread Model With Two States

We start with a malware spread model with two states, where we suppose that there are large number of agents, and each agent has a local state $x_i(t) \in \{0, 1\}$, where $x_i(t) = 0$ represents the "healthy" state and $x_i(t) = 1$ represents the "infected" state. Each agent can take an action $a_i(t) \in \{0, 1\}$, where $a_i(t) = 0$ represents "do nothing" and $a_i(t) = 1$ represents "repair". The dynamics is given by

$$x_i(t+1) = \begin{cases} x_i(t) + (1 - x_i(t))\, w_i(t) & \text{if } a_i(t) = 0, \\ 0 & \text{if } a_i(t) = 1, \end{cases}$$

where $w_i(t) \in \{0, 1\}$ is a Bernoulli random variable with success probability $q$, which gives the probability of an agent getting infected. In this setting, if an agent chooses not to take any action, they may be infected with probability $q$, but if they choose to take a repair action, they return to the healthy state. In the infinite population limit, each agent pays a cost

$$c(x, a, \mu) = \theta_1\, \mu(0)\, x + (\theta_1 + \theta_2)\, \mu(1)\, x + \theta_3\, a = (\theta_1 + \theta_2\, \mu(1))\, x + \theta_3\, a,$$

where $\mu$ is the mean-field term. Here, $\theta_3$ is the cost of repair, and $(\theta_1 + \theta_2\, \mu(1))$ represents the risk of being infected. In the IRL problem, it is assumed that the variables $\theta := (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ are unknown. Therefore, we suppose that the cost is an element of the function class

$$\mathcal{R} := \left\{ c(x, a, \mu) = \langle \theta, f(x, a, \mu) \rangle : \theta \in \mathbb{R}^3,\ f : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}^3 \right\},$$

where $f(x, a, \mu) := (x, x \cdot \mu(1), a)$.

29

In the first step, we consider the forward RL problem with known parameters $(\theta_1, \theta_2, \theta_3)$ and compute a corresponding mean-field equilibrium $(\pi_E, \mu_E)$ by using the GNEP formulation. Subsequently, we use this computed $(\pi_E, \mu_E)$ to generate the feature expectation vector

$$\langle f \rangle_{\pi_E, \mu_E} = E^{\pi_E, \mu_E} \left[ \sum_{t=0}^{\infty} \beta^t \, f(x(t), a(t), \mu_E) \right]$$

in the maximum causal entropy IRL problem to determine the policy that maximizes the causal entropy under the feature expectation constraint.

### 4.1.1 STEP 1: FINDING MFE

In the infinite population limit, the stationary version of the problem is studied and the model is formulated as a GNEP, where the cost function for player 2 is taken to be the same as that of player 1. For numerical experiments, we use the following parameters $\theta_1 = 0.2$, $\theta_2 = 1$, $\theta_3 = 0.4$, $\beta = 0.8$, $q = 0.9$. We use MATLAB to perform the numerical experiments. The algorithm runs for 10000 iterations and uses the following parameters $\sigma_k = 0.1$, $\kappa = 0.001$. To perform (4) in Algorithm 2, we use Armijo line search.

Now let us look at the behavior of the mean-field term. Mean-field term converges to the distribution $[0.65, 0.35]$. Hence, at the equilibrium, 65% of the states are healthy. Moreover, equilibrium policy converges to the conditional distributions $\pi(\,\cdot\,|0) = [0.61, 0.39]$ and $\pi(\,\cdot\,|1) = [0, 1]$. Hence, once an agent is infected, then it should apply repair action with probability 1. However, if the agent is healthy, then it should do nothing with probability 0.61. Let $\nu_E(x, a) := \pi_E(a|x)\,\mu_E(x)$ denote the joint distribution on $\mathsf{X} \times \mathsf{A}$ induced by mean-field equilibrium $(\pi_E, \mu_E)$. Then, we have

$$\nu_E = \begin{bmatrix} 0.3965 & 0.2535 \\ 0 & 0.35 \end{bmatrix}.$$

Using this joint distribution we can recover both $\mu_E$ and $\pi_E$.

### 4.1.2 STEP 2: SOLVING MAXIMUM ENTROPY IRL GIVEN MFE

We now feed the MFE found in the previous section into the IRL problem to generate the feature expectation vector and find the policy that solves the corresponding maximum causal entropy problem. We use MATLAB for the numerical computations. The gradient descent algorithm uses the following rate $\gamma = 0.5$. We stop the iteration when the each component of the gradient of $g$ becomes less than $O(10^{-2})$.

Note that the precise value of the minimum of $g$ holds less significance in this context. The key point here is that the Boltzmann distribution $\nu_{\theta^*, \lambda^*, \xi^*}^*$ computed at the minimizer $(\theta^*, \lambda^*, \xi^*)$ of $g$ is the optimal solution for $(\mathbf{OPT_2})$. In view of the proof of Theorem 8, the policy

$$\pi_{\nu_{\theta^*, \lambda^*, \xi^*}^*}(a|x) = \frac{\nu_{\theta^*, \lambda^*, \xi^*}^*(x, a)}{\nu_{\theta^*, \lambda^*, \xi^*}^{*, \mathsf{X}}(x)}$$

solves the maximum causal entropy IRL problem.

It turns out that the gradient descent algorithm outputs the following Boltzman distribution

$$\nu^*_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*} = \begin{bmatrix} 0.3960 & 0.2540 \\ 0 & 0.35 \end{bmatrix}$$

at the minimizer $(\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*)$ of $g$. Note that this is very close to $\nu_E$. As a result, the corresponding policy

$$\pi_{\nu^*_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}} = \begin{bmatrix} 0.6093 & 0.3907 \\ 0 & 1 \end{bmatrix}$$

is, as expected, very close to the equilibrium policy $\pi_E$, which is unknown to the player. Recall that only the feature expectation vector is available to the player in the IRL setting. Although the equilibrium policy $\pi_E$ and the maximum causal entropy policy $\pi_{\nu^*_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}}$ might yield the same feature expectation vector under $\mu_E$, their behavior can differ significantly. In this numerical example, the resemblance between the policies $\pi_{\nu^*_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}}$ and $\pi_E$ occur due to the feature vector structure $f(x,a,\mu) = (x, x \cdot \mu(1), a)$. Specifically, the numerical example's feature expectation matching constraint specifies that $\nu^{*,\mathsf{X}}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*} = \nu^{\mathsf{X}}_E$ and $\nu^{*,\mathsf{A}}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*} = \nu^{\mathsf{A}}_E$. With the additional constraint

$$\mu_E(z) = \sum_{(x,a)\in\mathsf{X}\times\mathsf{A}} p(z|y,a,\mu_E) \, \nu^{*,\mathsf{X}}_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}(y,a)$$

in the maximum causal entropy problem, the equivalence of $\nu^*_{\boldsymbol{\theta}^*,\boldsymbol{\lambda}^*,\boldsymbol{\xi}^*}$ and $\nu_E$ can be established. Changing the feature vector structure could potentially lead to different solutions for the maximum causal entropy problem compared to $\pi_E$, but this still leads to a mean-field equilibrium with $\mu_E$.

## 4.2 A Malware Spread Model With Ten States

We now consider a malware spread model with ten states, i.e., $\mathsf{X} = \{0, 0.1, 0.2, \ldots, 0.9\}$. The key difference between this model and the previous one lies in the transition probabilities. Specifically, the transitions are given by

$$p(y \mid x, a) = \begin{cases} \mathrm{U}(\{x, x+1, \ldots, 0.9\}) & \text{if } a = 0, \\ 1_{\{y=0\}} & \text{if } a = 1. \end{cases}$$

Here, choosing action $a = 1$ resets the system to the healthy state 0, whereas selecting $a = 0$ leads to a worsening state, transitioning uniformly among larger states.

In the infinite population limit, each agent incurs a cost

$$c(x, a, \mu) = \theta_1 x + \theta_2 \mu_{\mathrm{av}} x + \theta_3 a,$$

where $\mu_{\mathrm{av}}$ represents the mean of the mean-field term. The parameter $\theta_3$ corresponds to the cost of repair, while $\theta_1 + \theta_2 \mu_{\mathrm{av}}$ captures the infection risk.

In the IRL problem, the parameters $\theta := (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ are unknown. Thus, we assume that the cost function belongs to the class

$$\mathcal{R} := \left\{ c(x, a, \mu) = \langle \theta, f(x, a, \mu) \rangle : \theta \in \mathbb{R}^3, \, f : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}^3 \right\},$$

where the feature mapping is given by $f(x, a, \mu) := (x, x\mu_{\mathrm{av}}, a)$.

4.2.1 Step 1: Finding MFE

We consider the same setup as in the previous example but now with ten states. Consequently, the details of the MATLAB code remain unchanged. For numerical experiments, we use the parameter values $\theta_1 = 0.1$, $\theta_2 = 1$, $\theta_3 = 0.4$, and $\beta = 0.8$.

Next, we examine the behavior of the mean-field term. The mean-field distribution converges to

$$\mu_E = [0.3338, 0.0975, 0.0429, 0.0501, 0.0601, 0.0751, 0.1001, 0.1001, 0.1001, 0.1001].$$

Moreover, the equilibrium policy converges to the following deterministic policy:

$$\pi_E = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Let $\nu_E(x, a) := \pi_E(a|x)\,\mu_E(x)$ denote the joint distribution on $\mathsf{X} \times \mathsf{A}$ induced by the mean-field equilibrium $(\pi_E, \mu_E)$. Then, we obtain

$\nu_E =$
$$\begin{bmatrix} 0.3338 & 0.0375 & 0.0429 & 0.0501 & 0.0601 & 0.0751 & 0.1001 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1001 & 0.1001 & 0.1001 \end{bmatrix}.$$

From this joint distribution, we can recover both $\mu_E$ and $\pi_E$.

4.2.2 Step 2: Solving Maximum Entropy IRL Given MFE

We now use the MFE found in the previous section as input to the IRL problem to compute the feature expectation vector and determine the policy that solves the corresponding maximum causal entropy problem. The numerical computations are carried out in MATLAB using a gradient descent algorithm with step size $\gamma = 0.0025$.

As discussed in the previous example, the precise minimum value of $g$ is not of primary importance. The key result is that the Boltzmann distribution $\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}$, computed at the minimizer $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*)$ of $g$, serves as the optimal solution for $(\mathbf{OPT_2})$. By Theorem 8, the policy

$$\pi_{\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}}(a|x) = \frac{\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x, a)}{\nu^{*,\mathsf{X}}_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*}(x)}$$

solves the maximum causal entropy IRL problem.

The gradient descent algorithm outputs the following Boltzmann distribution at the minimizer $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*)$:

$\nu^*_{\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\xi}^*} =$
$$\begin{bmatrix} 0.3308 & 0.0371 & 0.0425 & 0.0496 & 0.0591 & 0.0724 & 0.0871 & 0.0183 & 0.0024 & 0.0004 \\ 0.0027 & 0.0003 & 0.0005 & 0.0007 & 0.0013 & 0.0031 & 0.0131 & 0.0813 & 0.0976 & 0.0998 \end{bmatrix}.$$

This distribution closely resembles $\nu_E$, with only minor differences in columns 7 and 8. Consequently, the corresponding policy

$$\pi_{\nu^*_{\theta^*,\lambda^*,\xi^*}} =$$

$$\begin{bmatrix} 0.9919 & 0.9907 & 0.9890 & 0.9856 & 0.9785 & 0.9593 & 0.8696 & 0.1834 & 0.0239 & 0.0037 \\ 0.0081 & 0.0093 & 0.0110 & 0.0144 & 0.0215 & 0.0407 & 0.1305 & 0.8166 & 0.9761 & 0.9963 \end{bmatrix}$$

is also close to the equilibrium policy $\pi_E$, again differing slightly in columns 7 and 8.

Recall that in the IRL setting, only the feature expectation vector is observable. Although $\pi_E$ and $\pi_{\nu^*_{\theta^*,\lambda^*,\xi^*}}$ may yield the same feature expectation vector under $\mu_E$, their actual behavior can vary significantly. In this case, the near equivalence of these policies stems from the feature structure $f(x, a, \mu) = (x, x \cdot \mu_{\text{av}}, a)$ and the additional invariance constraint, as observed in the previous example.

## 5. Conclusion

This paper addresses the maximum causal entropy IRL problem within the context of discrete-time MFGs under an infinite-horizon discounted-reward optimality criterion. We first formalize the maximum causal entropy IRL problem specifically for infinite-horizon MFGs, which initially presents as a non-convex optimization problem over policies. By leveraging the LP framework commonly used for MDPs, we transform this IRL problem into a convex optimization problem over state-action occupation measures. We then introduce a gradient descent algorithm, providing a guaranteed convergence rate, to compute the optimal solution. Furthermore, we introduce a novel algorithm that recasts the MFG problem as a GNEP. This contribution is significant as it is particularly capable of computing the MFE for the forward RL problem. The practical effectiveness of our algorithm is demonstrated through numerical examples.

This paper focuses on the theoretical analysis of the IRL problem under the simplifying assumption of linear rewards, which may not fully capture the complexity of real-world applications. The broader IRL field increasingly incorporates non-linear reward functions to better address practical scenarios. A natural extension would be to explore non-linear reward structures, potentially by leveraging the reproducing kernel Hilbert space (RKHS) framework for their representation.

Another further research direction involves refining the method for solving the GNEP formulation of MFGs. While we adapt an existing interior-point method, its convergence assumptions—though more relaxed than typical methods—remain somewhat restrictive. By exploiting the specific structure of the MFG-induced GNEP, future work could aim to develop less restrictive algorithms, potentially through refinements of existing methods. This presents a challenging but impactful path for advancing MFE computation.

## Acknowledgments

# Appendix A. Maximum Causal Entropy IRL in MDPs

In this section, we review the maximum entropy principle developed for IRL in MDPs. The literature on this topic is quite fragmented, and we believe that summarizing the existing approaches could be valuable for readers. A recent survey by Gleave and Toyer (2022) on maximum entropy IRL for MDPs, which focuses primarily on finite-horizon problems, also provides a helpful overview. Here, we not only present the methods developed for MDPs, but we also arrange them in chronological order to illustrate the motivation behind the evolution of different types of maximum entropy principles. This structure highlights the rationale for the introduction of the maximum causal entropy principle, particularly for handling infinite-horizon problems.

A discrete-time stochastic MDP is specified by $(\mathsf{X}, \mathsf{A}, p, r)$, where $\mathsf{X}$ is a finite state space and $\mathsf{A}$ is a finite action space. The components $p : \mathsf{X} \times \mathsf{A} \to \mathsf{X}$ and $r : \mathsf{X} \times \mathsf{A} \to [0, \infty)$ are the system dynamics and the one-stage reward function, respectively. Therefore, given the current state $x(t)$ and action $a(t)$, the reward $r(x(t), a(t))$ is received immediately, and the next state $x(t+1)$ evolves to a new state stochastically according to the following dynamics: $x(t+1) \sim p(\cdot|x(t), a(t))$. In this model, a randomized policy $\pi = \{\pi_t\}_{t=0}^{M}$, where $M$ is either finite or infinite, is a sequence of functions of the form $\pi_t : \mathsf{X} \to \mathcal{P}(\mathsf{A})$.

## A.1 Finite-Horizon MDPs

In forward RL problems, the typical goal is to maximize a finite-horizon reward defined as

$$J(\pi, \mu_0) = E^\pi \left[ \sum_{t=0}^{T-1} r(x(t), a(t)) \right],$$

where $T \in \mathbb{N}$ denotes the finite horizon (i.e., $M = T - 1$) and $x(0) \sim \mu_0$. In this context, the one-stage reward function $r$ is known. In contrast to RL where the agent learns a policy to maximize a pre-defined reward function, IRL takes a different approach. Here, the agent is presented with a collection of trajectories generated by an expert. By analyzing these expert demonstrations, the IRL algorithm aims to deduce the reward function that the expert was implicitly trying to optimize. In essence, the expert's behavior serves as a substitute for the reward function, allowing the IRL agent to learn the underlying objective and potentially develop similar or even improved policies.

Since IRL aims to recover the reward function from expert demonstrations, the set of possible rewards needs some structure to make the problem tractable. A common assumption in IRL is that the reward can be expressed as a linear combination of feature vectors corresponding to state and action pairs:

$$\mathcal{R} := \left\{ r_\theta(x, a) = \langle \theta, f(x, a) \rangle : \theta \in \mathbb{R}^k, \ f : \mathsf{X} \times \mathsf{A} \to \mathbb{R}^k \right\}.$$

In this setting, we suppose that an expert generates trajectories $\mathcal{D} = \{(x_i(t), a_i(t))_{t=0}^{T-1}\}_{i=1}^{d}$ under some optimal policy $\pi_{\text{opt}}$. Therefore, if $d$ is large enough, by the law of large numbers, we have

$$\frac{1}{d} \sum_{i=1}^{d} \left( \sum_{t=0}^{T-1} f(x_i(t), a_i(t)) \right) \simeq E^{\pi_{\text{opt}}} \left[ \sum_{t=0}^{T-1} f(x(t), a(t)) \right] =: \langle f \rangle_{\pi_{\text{opt}}},$$

where $E^{\pi_{\mathrm{opt}}}$ is the expectation under $\pi_{\mathrm{opt}}$. This suggests that the feature expectation vector, denoted by $\langle f \rangle_{\pi_{\mathrm{opt}}}$ under the optimal policy $\pi_{\mathrm{opt}}$, is readily available.

The maximum entropy principle was introduced by Ziebart et al. (2008) to address deterministic MDPs, where the state dynamics are generated by $x(t+1) = p(x(t), a(t))$ (i.e., no uncertainty in state dynamics).[4] Defining the entropy of a probability distribution $P$ on a finite set $\mathsf{E}$ as $H(P) := -\sum_{e \in \mathsf{E}} P(e) \log P(e)$, the maximum entropy IRL problem in this setting can be formulated as follows:

$$(\mathbf{OPT_d}) \ \operatorname*{maximize}_P \quad H(P)$$

$$\text{subject to} \quad P(\tau) \geq 0 \ \forall \tau \in \mathsf{Z}_{\mathrm{path}}, \ \sum_{\tau \in \mathsf{Z}_{\mathrm{path}}} P(\tau) = 1$$

$$\sum_{\tau \in \mathsf{Z}_{\mathrm{path}}} F(\tau) \, P(\tau) = \langle f \rangle_{\pi_{\mathrm{opt}}},$$

where $F(\tau) := \sum_{(x,a) \in \tau} f(x, a)$ and $\mathsf{Z}_{\mathrm{path}}$ is the path space defined as

$$\mathsf{Z}_{\mathrm{path}} := \left\{ \tau \in (\mathsf{X} \times \mathsf{A})^T : x(t+1) = p(x(t), a(t)), \ t = 0, \cdots, T-2 \right\}.$$

Here, the expert behaves according to some optimal policy $\pi_{\mathrm{opt}}$ under some unknown reward function $r_{\mathrm{opt}}(x, a) = \langle \theta_{\mathrm{opt}}, f(x, a) \rangle$. However, a key challenge in standard IRL problems arises because there can be multiple values for $\theta$ (besides the optimal $\theta_{\mathrm{opt}}$) that can explain the observed expert trajectories. The principle of maximum entropy suggests that among all these candidates, we should favor the one with the highest entropy. Hence, a solution of $(\mathbf{OPT_d})$ leads to an optimal policy with minimum bias.

In stochastic MDPs, since there is independent randomness due to noise in the state dynamics, it is not possible to formulate the maximum entropy principle over the path space as before. To see why, suppose that the problem can be formulated over the probability distributions on the path space similar to $(\mathbf{OPT_d})$. Then, the optimal solution $P^*$ must adhere to the state dynamics, meaning that $P^*$ must be factorized in the following form: [5]

$$P^*(\tau) = p_0(x(0)) \prod_{t=0}^{T-2} p(x(t+1)|x(t), a(t)) \prod_{t=0}^{T-1} \pi(a(t)|x(t)).$$

However, the optimal solution to the maximum entropy problem over the path space might not exhibit a factored form as above. This challenge can be addressed by replacing the maximum entropy principle with the maximum causal entropy principle (see Ziebart et al., 2010, 2013). In this approach, we consider the causally conditioned probability distribution of actions given states $P(\mathbf{a}||\mathbf{x}) := \prod_{t=0}^{T-1} \pi_t(a(t)|x(t))$ and maximize its entropy

$$H(P(\cdot||\cdot)) := E^\pi[-\log P(\mathbf{a}||\mathbf{x})] = \sum_{t=0}^{T-1} E^\pi[-\log \pi_t(a(t)|x(t))]$$

instead. Let $\mathcal{C}$ denote the set of causally conditioned probability distributions. Then, the maximum causal entropy IRL problem is defined as

$$(\mathbf{OPT_s}) \ \operatorname*{maximize}_{P(\cdot||\cdot) \in \mathcal{C}} \quad H(P(\cdot||\cdot))$$

$$\text{subject to} \quad \sum_{\tau \in (\mathsf{X} \times \mathsf{A})^T} F(\tau) \, \mathcal{T}_{P(\cdot||\cdot)}(\tau) = \langle f \rangle_{\pi_{\mathrm{opt}}},$$

---

4. Even though the system dynamics are deterministic, agents can still employ randomized policies.
5. Here, the first part is static, and only the second part in the product can be manipulated.

where $\mathcal{T}_{P(\cdot||\cdot)}(\tau) := p_0(x(0)) \prod_{t=0}^{T-2} p(x(t+1)|x(t), a(t)) \, P(\mathbf{a}||\mathbf{x})$. One can prove that $(\mathbf{OPT_s})$ is convex in $P(\cdot||\cdot)$ using the fact that each constraint is linear and the objective function is concave in $P(\cdot||\cdot)$.

To solve the maximum entropy problem $(\mathbf{OPT_s})$, we can introduce a Lagrange multiplier $\theta \in \mathbb{R}^k$ to penalize the deviations from feature expectation matching constraint. The solution of the Lagrangian relaxation of $(\mathbf{OPT_s})$ for a given $\theta$ can then be expressed by the following soft Bellman optimality equations

$$Q_t^\theta(x,a) = r_\theta(x,a) + \sum_{y \in \mathsf{X}} V_{t+1}^\theta(y) \, p(y|x,a), \quad V_t^\theta(x) = \log \sum_{a \in \mathsf{A}} e^{Q_t^\theta(x,a)} =: \operatorname*{softmax}_{a \in \mathsf{A}} Q_t^\theta(x,a),$$

and it can be written in the form of $P_\theta(\mathbf{a}||\mathbf{x}) = \prod_{t=0}^{T-1} \pi_t^\theta(a(t)|x(t))$, where $\pi_t^\theta(a|x) = e^{Q_t^\theta(x,a) - V_t^\theta(x)}$ (see Ziebart et al., 2010, 2013). This implies that the optimal solution of $(\mathbf{OPT_s})$ leads to the probability distribution on the path space of the following form

$$P(\tau) \propto p_0(x(0)) \prod_{t=0}^{T-2} p(x(t+1)|x(t), a(t)) \, e^{Q_t^\theta(x(t),a(t))}.$$

There are instances in the literature, for example in Snoswell et al. (2020, eq. 1), Chen et al. (2023b, eq. 3) or Fu et al. (2018, eq. 1), where it is asserted that the solution to the maximum causal entropy problem for stochastic MDPs is of the form

$$P(\tau) \propto p_0(x(0)) \prod_{t=0}^{T-2} p(x(t+1)|x(t), a(t)) \, e^{r_\theta(x(t),a(t))}.$$

However, based on the preceding calculations, it becomes evident that $r_\theta$ needs to be replaced with the soft Q-functions $Q_t^\theta$. To find an optimal $\theta^*$, we can either use the feature expectation matching constraint in $(\mathbf{OPT_s})$ or we can introduce an alternative optimization problem, whose solution satisfies the feature expectation matching constraint:

$$(\widehat{\mathbf{OPT_s}}) \quad \max_{\theta \in \mathbb{R}^k} \sum_{\tau \in (\mathsf{X} \times \mathsf{A})^T} \log P_\theta(\mathbf{a}||\mathbf{x}) \, \mathcal{T}_{P_{\mathrm{opt}}(\cdot||\cdot)}(\tau).$$

This problem in the literature is referred to as the maximum log-likelihood estimation problem. It serves as the commonly used formulation of the maximum entropy problem within the domain of IRL when dealing with stochastic MDPs.

## A.2 Infinite-Horizon MDPs

In the infinite-horizon scenario (i.e., $M = \infty$), we consider the discounted reward

$$J(\pi, \mu_0) = E^\pi \left[ \sum_{t=0}^\infty \beta^t \, r(x(t), a(t)) \right],$$

where $\beta \in (0,1)$ is the discount factor and $x(0) \sim \mu_0$. Within the field of MDPs, it is well-established that stationary Markovian policies suffice for achieving optimality with

36

discounted rewards. Consequently, we restrict our focus to such policies, where the policy remains constant for all time steps; that is, $\pi_s = \pi_t = \pi$ for all $s, t \geq 0$.

Extending the maximum causal entropy principle to an infinite horizon poses a challenge. The difficulty arises because defining the causally conditioned probability distribution becomes ill-defined in this scenario as it involves multiplying an infinite number of terms, each being less than one. Therefore, we use the policies instead of the causally conditioned probability distribution of actions given states to formulate the problem in the infinite-horizon case (see Zhou et al., 2018). Indeed, this is the appropriate variant of the maximum entropy principle that we use for the IRL problem in the infinite-horizon setting for MFGs. Defining the discounted causal entropy of the policy $\pi$ as

$$H(\pi) := \sum_{t=0}^{\infty} \beta^t E^\pi \left[ -\log \pi(a(t)|x(t)) \right],$$

the maximum discounted causal entropy IRL problem can be formulated by

$$(\mathbf{OPT}_\infty) \; \text{maximize}_\pi \quad H(\pi)$$

$$\text{subject to} \quad \pi(a|x) \geq 0 \; \forall (x, a) \in \mathsf{X} \times \mathsf{A}$$

$$\sum_{a \in \mathsf{A}} \pi(a|x) = 1 \; \forall x \in \mathsf{X}$$

$$\sum_{t=0}^{\infty} \beta^t E^\pi[f(x(t), a(t))] = \langle f \rangle_{\pi_{\mathrm{opt}}},$$

where $\langle f \rangle_{\pi_{\mathrm{opt}}} := \sum_{t=0}^{\infty} \beta^t E^{\pi_{\mathrm{opt}}}[f(x(t), a(t))]$. The challenge here is the problem's lack of convexity. This arises from the last constraint, which is non-convex with respect to policies $\pi$. Zhou et al. (2018) convert this non-convex problem with respect to the policies into a convex one by expressing the optimization problem using state-action occupation measures. This is indeed the approach we adapt to deal with maximum causal entropy IRL problem for MFGs.

To solve $(\mathbf{OPT}_\infty)$, a different approach can also be taken by considering the Lagrangian relaxation of the problem. This particular formulation was not explored by Zhou et al. (2018). Let us introduce Lagrange multiplier $\theta \in \mathbb{R}^k$ to penalize the deviations from feature expectation matching constraint. Then, the solution of the Lagrangian relaxation of $(\mathbf{OPT}_\infty)$ for a given $\theta$ can be given by the following soft Bellman optimality equations

$$Q^\theta(x, a) = r_\theta(x, a) + \sum_{y \in \mathsf{X}} V^\theta(y) \, p(y|x, a), \quad V^\theta(x) = \log \sum_{a \in \mathsf{A}} e^{Q^\theta(x,a)},$$

as the Lagrangian relaxation is indeed equivalent to an entropy regularized MDP with the reward function $r_\theta$ (see Neu et al., 2017), and it can be shown that the solution should take the form of

$$\pi^\theta(a|x) = e^{Q^\theta(x,a) - V^\theta(x)}.$$

To find an optimal $\theta^*$, we can then either use the feature expectation matching constraint in $(\mathbf{OPT}_\infty)$ or introduce an alternative optimization problem, whose solution satisfies the feature expectation matching constraint:

$$(\widehat{\mathbf{OPT}_\infty}) \; \max_{\theta \in \mathbb{R}^k} \sum_{(x,a) \in (\mathsf{X} \times \mathsf{A})} \log \pi^\theta(a|x) \, \gamma_{\pi_{\mathrm{opt}}}(x, a),$$

where $\gamma_{\pi_{\mathrm{opt}}}$ is the state-action occupation measure under the expert's optimal policy $\pi_{\mathrm{opt}}$. Similar to the finite-horizon scenario, this problem can be conceptualized as an instance of the maximum log-likelihood estimation problem.

# References

Stephen Adams, Tyler Cody, and Peter A. Beling. A survey of inverse reinforcement learning. *Artif. Intell. Rev.*, 55(6):4307–4346, Aug. 2022.

Sachin Adlakha, Ramesh Johari, and Gabriel Y. Weintraub. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory*, 156:269–316, 2015.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-Learning in Regularized Mean-field Games. *Dynamic Games and Applications*, 13(1):89–117, March 2023.

Martino Bernasconi, E. Vittori, F. Trovò, and M. Restelli. Dealer markets: A reinforcement learning mean field game approach. *The North American Journal of Economics and Finance*, 68:101974, 2023.

Peter E. Caines, Minyi Huang, and Roland P. Malhamé. Large population stochastic dynamic games: Closed loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information Systems*, 6:221–252, 2006.

Xu Chen, Shuo Liu, and Xuan Di. A hybrid framework of reinforcement learning and physics-informed deep learning for spatiotemporal mean field games. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1079–1087, 2023a.

Yang Chen, Libo Zhang, Jiamou Liu, and Shuyue Hu. Individual-level inverse reinforcement learning for mean field games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 253–262, 2022.

Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. Adversarial inverse reinforcement learning for mean field games. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1088–1096, 2023b.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1909–1917, 13–15 Apr 2021.

Josu Doncel, Nicolas Gast, and Bruno Gaujal. A mean field game analysis for sir dynamics with vaccination. *Probability in the Engineering and Informational Sciences*, 36 (2):482–499, 2022.

Axel Dreves, Francisco Facchinei, Christian Kanzow, and Simone Sagratella. On the solution of the KKT conditions of generalized nash equilibrium problems. *SIAM Journal on Optimization*, 21(3):1082–1108, 2011.

Paul Dupuis and Richard S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley-Interscience, 1997.

Francisco Facchinei and Cristian Kanzow. Generalized Nash equilibrium problems. *Annals of Operations Research*, 175:177–211, 2010.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Adam Gleave and Sam Toyer. A primer on maximum causal entropy inverse reinforcement learning. *arXiv preprint arXiv:2203.11409*, 2022.

Onésimo Hernández-Lerma and Juan González-Hernández. Constrained Markov control processes in Borel spaces: The discounted case. *Math. Meth. Oper. Res.*, 52:271–285, 2000.

Onésimo Hernández-Lerma and Jean B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2:229–260, 2007.

Mathieu Laurière, Sarah Perrin, Julien Perolat, Sertan Girgin, Paul Muller, Romuald Elie, Matthieu Geist, and Olivier Pietquin. Learning in mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2024.

Renato D. C. Monteiro and Jong-Shi Pang. A potential reduction Newton method for constrained equations. *SIAM Journal on Optimization*, 9(3):729–754, 1999.

Gergely Neu, Vicenç Gómez, and Anders Jonsson. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Laurière, and Matthieu Geist. On imitation in mean-field games. *arXiv preprint arXiv:2306.14799*, 2023.

Naci Saldi. Linear mean-field games with discounted cost. *arXiv preprint arXiv:2301.06074*, 2023.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

Aaron J. Snoswell, Surya P. N. Singh, and Nan Ye. Revisiting maximum entropy inverse reinforcement learning: New perspectives and algorithms. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 241–249, 2020.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '19, pages 251–259, 2019.

Gabriel Y. Weintraub, Lanier Benkard, and Benjamin Van Roy. Oblivious equilibrium: A mean field approximation for large-scale dynamic games. In *Advances in Neural Information Processing Systems*, 2005.

Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*, 2018.

Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Transactions on Automatic Control*, 63(9):2787–2802, 2018.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1255–1262, 2010.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. The principle of maximum causal entropy for estimating interacting processes. *IEEE Transactions on Information Theory*, 59(4):1966–1980, 2013.