TOWARDS SCALING REINFORCEMENT LEARNING TO MASSIVE POPULATIONS: LEARNING MEAN-FIELD REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern multi-agent systems are increasingly deployed at scale in settings such as ad-auctions, traffic routing, and recommendation systems. The dominant approach in such settings is to optimize each agent's policy independently, treating the other agents as part of a fixed single-agent environment rather than modeling the population dynamics. In many large-population systems, the dynamics depend on an aggregate summary of the population rather than the identity of any individual. Mean-field reinforcement learning (RL) exploits such structure, providing a principled framework that models each agent's environment as an explicit function of the population distribution. However, in large state-action spaces or high-dimensional control problems, modeling the population distribution is itself intractable. *How can we design a scalable framework for high-dimensional control problems with large populations?* This work explores this question from the perspective of representation learning. We introduce a mean-field RL framework in which the rewards and transition dynamics depend on the population only through an unknown low-dimensional aggregate statistic. We then study this framework in the offline setting and design a provable approach that learns an approximate Nash equilibrium by learning a low-dimensional representation. In a one-step routing game motivated by supply-chain optimization, we show that, under a fixed neural-network parameter count and optimization budget, learning a low-dimensional population representation improves reward prediction and the equilibrium quality of the resulting policies, compared with baselines that do not exploit this structure.

1 INTRODUCTION

Many modern decision systems involve large populations of strategic agents, from traffic routing and power-grid markets to recommendation and resource-allocation platforms. Directly modeling all agents jointly is statistically and computationally intractable, since the joint state-action space grows exponentially with the population size. Mean-field games (MFGs; Huang et al., 2006; Lasry & Lions, 2007) provide a tractable approximation for such large-population multi-agent systems by replacing the interactions among finitely many agents with the interaction of a representative agent with a limiting population law, called the mean-field distribution. However, this distribution lives in a $(|\mathcal{X}||\mathcal{A}| - 1)$-dimensional simplex, where $|\mathcal{X}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action spaces, respectively. Consequently, representing this distribution and learning how rewards and transitions vary over it becomes statistically and computationally prohibitive in large state-action spaces. The key observation motivating this work is that the rewards and transitions often depend on the mean-field distribution only through a much lower-dimensional statistic (see Figure 1). This raises the central question: *How can we exploit such structure to design statistically efficient learning algorithms that scale to high-dimensional decision problems with large numbers of agents?*

Contributions. In this work, we study this question through the lens of representation learning in MFGs. We **formalize a framework for learning low-dimensional representations** and focus on the offline setting, where the representation must be learned from logged representative-agent trajectories together with finite-population snapshots. To find an equilibrium from such data, we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

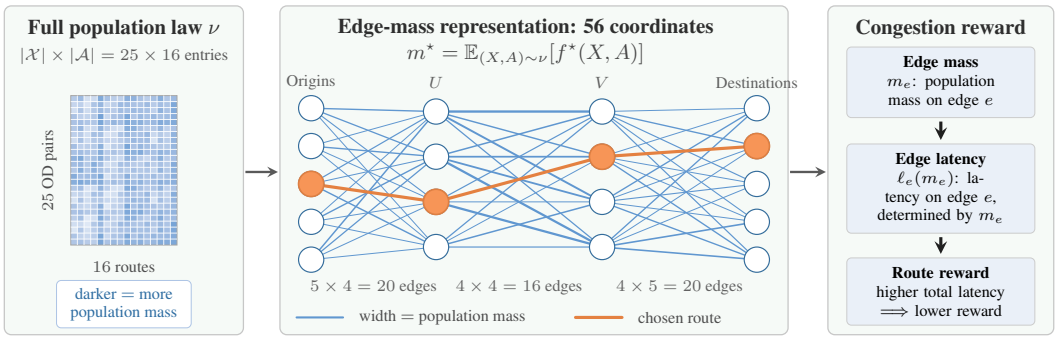

Figure 1: The one-step routing game and its mean-field representation. A state is an origin–destination pair, and an action picks one node in each of the two intermediate layers, giving a three-edge route (orange); so $|\mathcal{X}| \times |\mathcal{A}| = 25 \times 16 = 400$. The unknown feature map $f^\star(X, A) \in \{0, 1\}^{56}$ records which of the $20 + 16 + 20 = 56$ edges a route uses, so the representation has only 56 coordinates: the population mass on each edge (blue widths). The mass m_e on edge e sets its latency $\ell_e(m_e)$, and higher total route latency gives lower reward.

need to evaluate a population policy and check whether an agent can improve its return by deviating. Evaluating a candidate population policy requires predicting the population flow it induces, which may differ from every logged flow. Checking whether an agent has an incentive to deviate holds the population flow fixed, but the logged data may not cover the deviated states and actions.

Learning the representation adds two difficulties of its own: the population is observed only through finite snapshots, so the representation carries estimation error even with unlimited episodes, and an agent's values depend on the whole future path of representations, not just the current one (Proposition 1).

We develop a **model-based algorithm** (Algorithm 1) that learns representation-dependent reward and transition models from a finite class Θ, and selects a policy by minimizing an optimistic–pessimistic upper bound on its Nash gap over all models consistent with the data. We prove a finite-sample Nash-gap guarantee and show that up to coverage and horizon factors, the bound scales as $\sqrt{\log|\Theta|/N} + \sqrt{1/K}$ in the number N of logged episodes and the snapshot size K, with no explicit dependence on the number of states. The coverage requirement is adaptive: it suffices to cover the self-play of one low-gap policy and one approximate best response under the optimistic value estimates.

Finally, we evaluate the representation-learning mechanism in a **one-step congestion-routing game** motivated by fulfillment-routing problems arising in supply-chain optimization (see Figure 1). Under a fixed neural-network parameter count and optimization budget, we compare our model with baselines that either do not fully exploit the population structure or have privileged access to the true mean field. Our model outperforms the former and approaches the latter in the large-data regime.

Related Work. Our analysis is closest to Zhang et al. (2023b), who study offline equilibrium learning in Markov games by minimizing an estimated equilibrium gap based on optimistic deviation values and pessimistic policy values. We adopt their adaptive bound strategy for mean-field RL, where we need to account for shifts in representative-agent occupancies, population flows, and learned representation paths, together with estimation error from finite population snapshots.

1. Mean-field RL: Finite-state and discrete-time MFGs provide a basis for RL (Gomes et al., 2010; Saldi et al., 2018). Learning methods include adaptive control (Kizilkale & Caines, 2013), approximate dynamic programming (Yin et al., 2014), and model-free RL and mirror descent (Yang et al., 2018; Guo et al., 2019; Subramanian & Mahajan, 2019; Laurière et al., 2022a); see Laurière et al. (2022b) for a survey. Statistical guarantees have been established under known mean-field embeddings or model classes (Wang et al., 2020; Pásztor et al., 2023; Huang et al., 2024a;b). We study offline learning with an unknown low-dimensional population representation.

2. Offline RL and coverage: In single-agent offline RL, pessimism can give guarantees with coverage of a comparator policy (Jin et al., 2021; Xie et al., 2021b; Uehara & Sun, 2022; Jiang & Xie, 2025). In offline Markov games, covering the candidate policy alone is generally insufficient because equilibrium learning also requires evaluating unilateral deviations (Cui & Du, 2022b;a; Zhan et al., 2025). Mean-field coverage assumptions developed for online interaction or simulator access (Anahtarci et al., 2023; Xie et al., 2021a; Mao et al., 2022) do not directly address learning from a fixed offline dataset. Existing offline mean-field methods exploit invariance (Chen et al., 2021) or use mirror descent (Brunnbauer et al., 2025), but neither targets representation learning.

3. Representation learning: Prior work studies representation learning in low-rank and latent-structure RL (Misra et al., 2020; Agarwal et al., 2020; Uehara et al., 2022; Zhang et al., 2023a) and low-rank Markov games (Ni et al., 2023). There, features describe the current state and actions; ours are averaged over the population, so the average depends on the population policy.

Notation. For a finite set $\mathcal{X}$, write $\Delta(\mathcal{X})$ for the set of probability distributions on $\mathcal{X}$. For $p, q \in \Delta(\mathcal{X})$, write $D_{\mathrm{TV}}(p,q) := \frac{1}{2}\sum_x |p(x) - q(x)|$ and $D^2_{\mathrm{H}}(p,q) := \frac{1}{2}\sum_x (\sqrt{p(x)} - \sqrt{q(x)})^2$. We consider finite state and action spaces $\mathcal{X}$ and $\mathcal{A}$, a finite horizon H, and an initial state distribution $\mu_0 \in \Delta(\mathcal{X})$. A (non-stationary Markov) policy π is a sequence $\pi = (\pi_h)_{h=0}^{H-1}$ with $\pi_h \colon \mathcal{X} \to \Delta(\mathcal{A})$. The set of all such policies is denoted by Π. For a state law $\mu \in \Delta(\mathcal{X})$ and a decision rule $\alpha \colon \mathcal{X} \to \Delta(\mathcal{A})$, define the induced state-action law $\mu \otimes \alpha \in \Delta(\mathcal{X} \times \mathcal{A})$ as $(\mu \otimes \alpha)(x,a) := \mu(x)\alpha(a \mid x)$.

2 Mean-Field Games with Low-Dimensional Representations

Mean-field games describe the limit of a large symmetric population, in which a representative agent interacts with the others only through the population's state-action distribution, not through the identities and states of individual agents. We first define mean-field games in which this interaction passes through a low-dimensional representation, then show that value functions depend on the whole future path of representations, and finally define Nash equilibria and show that they exist.

2.1 Low-Dimensional Mean-Field Representations

We assume that the population affects rewards and transitions only through a low-dimensional mean-field representation. Specifically, there exists $d \in \mathbb{N}$ such that, for each $h \in \{0, \dots, H-1\}$, there is a feature map $f^\star_h \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, a reward map $r^\star_h \colon \mathcal{X} \times \mathcal{A} \times \mathbb{R}^d \to [0,1]$, and a transition kernel $T^\star_h(\cdot \mid x, a, m) \in \Delta(\mathcal{X})$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, $m \in \mathbb{R}^d$. If the population's state-action law at stage h is $\nu_h \in \Delta(\mathcal{X} \times \mathcal{A})$, an agent that takes action a in state x receives reward $r^\star_h(x, a, m_h)$ and transitions according to $T^\star_h(\cdot \mid x, a, m_h)$, where $m_h := \sum_{y,b} \nu_h(y,b) f^\star_h(y,b) \in \mathbb{R}^d$ is the mean-field representation of ν_h.

When the population uses policy $\eta \in \Pi$, it induces the population flow defined recursively by $\mu^{\star,\eta}_0 = \mu_0$, $\nu^{\star,\eta}_h = \mu^{\star,\eta}_h \otimes \eta_h$, and $\mu^{\star,\eta}_{h+1}(x') = \sum_{x,a} \nu^{\star,\eta}_h(x,a)\, T^\star_h(x' \mid x, a, m^{\star,\eta}_h)$, with induced representations $m^{\star,\eta}_h := \sum_{x,a} \nu^{\star,\eta}_h(x,a) f^\star_h(x,a)$. A single agent is negligible in this limit, so a representative agent who plays $\pi \in \Pi$ against a population playing η leaves the population flow unchanged. Its state-action process $(X_0, A_0, \dots, X_{H-1}, A_{H-1}, X_H)$ therefore follows

$$X_0 \sim \mu_0, \quad A_h \sim \pi_h(\cdot \mid X_h), \quad X_{h+1} \sim T^\star_h(\cdot \mid X_h, A_h, m^{\star,\eta}_h), \quad h \in \{0, \dots, H-1\}. \tag{1}$$

Let $\mathbb{P}^{\pi;\eta}_\star$ and $\mathbb{E}^{\pi;\eta}_\star$ denote the law of this process and its expectation, let $\mu^{\star,\pi;\eta}_h$ be the agent's state law under $\mathbb{P}^{\pi;\eta}_\star$, and let $\nu^{\star,\pi;\eta}_h := \mu^{\star,\pi;\eta}_h \otimes \pi_h$. The value of π against η is $J(\pi;\eta) := \mathbb{E}^{\pi;\eta}_\star \left[\sum_{h=0}^{H-1} r^\star_h(X_h, A_h, m^{\star,\eta}_h)\right]$.

In classical mean-field games, rewards and transitions may depend on the entire law ν_h, which lies in a simplex of dimension $|\mathcal{X}||\mathcal{A}| - 1$. This case is recovered by taking $d = |\mathcal{X}||\mathcal{A}|$ and letting $f^\star_h(x,a)$ be the standard basis vector indexed by (x,a), so that $m^{\star,\eta}_h = \nu^{\star,\eta}_h$. Ideally, $d \ll |\mathcal{X}||\mathcal{A}| - 1$.

2.2 Value Functions Depend on Future Representations

Against a population policy η, a representative agent faces a finite-horizon MDP whose reward and transition kernel at stage h are $r^\star_h(\cdot, \cdot, m^{\star,\eta}_h)$ and $T^\star_h(\cdot \mid \cdot, \cdot, m^{\star,\eta}_h)$. To make the dependence on the

representation path explicit, fix a suffix $z_{h:H-1} = (z_h, \dots, z_{H-1}) \in (\mathbb{R}^d)^{H-h}$, where $z_{H:H-1}$ is the empty suffix. For a policy π, set $v_H^{\star,\pi}(x, \varnothing) \coloneqq 0$ and, for $h = H-1, \dots, 0$,

$$\begin{aligned} q_h^{\star,\pi}(x, a, z_{h:H-1}) &\coloneqq r_h^\star(x, a, z_h) + \sum_{x' \in \mathcal{X}} T_h^\star(x' \mid x, a, z_h) v_{h+1}^{\star,\pi}(x', z_{h+1:H-1}), \\ v_h^{\star,\pi}(x, z_{h:H-1}) &\coloneqq \sum_{a \in \mathcal{A}} \pi_h(a \mid x) q_h^{\star,\pi}(x, a, z_{h:H-1}). \end{aligned} \tag{2}$$

The agent's values against η are these functions evaluated along $m_{h:H-1}^{\star,\eta}$. In particular, $J(\pi; \eta) = \sum_{x \in \mathcal{X}} \mu_0(x) v_0^{\star,\pi}\left(x, m_{0:H-1}^{\star,\eta}\right)$. Although the current representation z_h determines the immediate reward and next-state distribution, the values also depend on the future suffix $z_{h+1:H-1}$. Hence two population policies that induce the same current representation but different future ones can give different action values, so action values are not in general functions of $(x, a, m_h^{\star,\eta})$ alone (Proposition 1 in App. A.1.1). This is why the Bellman residuals in Section 4 evaluate value functions along complete representation suffixes.

2.3 Equilibrium and Well-Posedness

The exploitability, or Nash gap, of a population policy η is $\mathrm{Gap}(\eta) \coloneqq \sup_{\pi \in \Pi} J(\pi; \eta) - J(\eta; \eta)$. A policy is a Nash equilibrium if it is a best response to the population flow it induces, that is, if its Nash gap is zero. Let $\Pi_\star \coloneqq \{\eta \in \Pi : \mathrm{Gap}(\eta) = 0\}$ denote the set of Nash equilibria. This set is non-empty if the features are bounded and the rewards and transitions are continuous in the representation (Proposition 2 in App. A.1.2). A similar approach could be used for the social optimum, also known as mean-field control, where the objective is $\max_{\eta \in \Pi} J(\eta; \eta)$.

3 Offline Mean-Field RL Framework

We now assume that the mean-field representation of Section 2 is unknown and must be learned from offline data, using a model class that the learner can access. We describe the model class, the data, and the learning algorithm below; Section 4 states the guarantees.

Candidate mean-field models. A candidate model is a tuple $\theta = (f^\theta, r^\theta, T^\theta) \in \Theta$, where $f_h^\theta \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a feature map, $r_h^\theta \colon \mathcal{X} \times \mathcal{A} \times \mathbb{R}^d \to [0, 1]$ is a reward model, and $T_h^\theta(\cdot \mid x, a, m) \in \Delta(\mathcal{X})$ is a transition model. We assume that Θ is finite. For a population policy η, the model θ induces the population flow $(\mu_h^{\theta,\eta}, \nu_h^{\theta,\eta}, m_h^{\theta,\eta})_{h=0}^{H-1}$ just like in the true model case discussed above with $\star$ replaced with θ. For future reference, define $\bar{m}_h^{\theta,\eta} \coloneqq \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \nu_h^{\star,\eta}(x, a) f_h^\theta(x, a)$.

Against population policy η, a representative agent playing π under model θ follows the path law $\mathbb{P}_\theta^{\pi;\eta}$ defined as in (1), with transition $T_h^\theta(\cdot \mid X_h, A_h, m_h^{\theta,\eta})$. The corresponding candidate value is $J_\theta(\pi; \eta) \coloneqq \mathbb{E}_\theta^{\pi;\eta}\left[\sum_{h=0}^{H-1} r_h^\theta(X_h, A_h, m_h^{\theta,\eta})\right]$.

We work with the following assumptions.

Assumption 1 (Realizability)**.** *The true model belongs to the model class: $\theta^\star = (f^\star, r^\star, T^\star) \in \Theta$.*

Assumption 2 (Boundedness)**.** *There are constants $B_f < \infty$ and $\tau \in (0, 1]$. Let $\mathcal{B}_f \coloneqq \{m \in \mathbb{R}^d : \|m\|_2 \le B_f\}$. For every $\theta \in \Theta$, $h \in \{0, \dots, H-1\}$, $(x, a) \in \mathcal{X} \times \mathcal{A}$, $x' \in \mathcal{X}$, and $m \in \mathcal{B}_f$, we have $\|f_h^\theta(x, a)\|_2 \le B_f$ and $T_h^\theta(x' \mid x, a, m) \in \{0\} \cup [\tau, 1]$.*

Assumption 3 (Continuity)**.** *There are constants $L_{r,h} < \infty$ and $L_{T,h} < \infty$ such that, for every $\theta \in \Theta$, $h \in \{0, \dots, H-1\}$, $(x, a) \in \mathcal{X} \times \mathcal{A}$, and $m, m' \in \mathcal{B}_f$,*

$$|r_h^\theta(x, a, m) - r_h^\theta(x, a, m')| \le L_{r,h} \|m - m'\|_2, \; D_{\mathrm{TV}}\big(T_h^\theta(\cdot \mid x, a, m), T_h^\theta(\cdot \mid x, a, m')\big) \le L_{T,h} \|m - m'\|_2.$$

Offline data. We observe $N \in \mathbb{N}$ independent logged episodes with each sample containing a pair of the representative agent's trajectory and $K \in \mathbb{N}$ independent samples of population snapshots over time:

$$\mathcal{D} \coloneqq \{(X_{n,h}, A_{n,h}, R_{n,h}, X_{n,h+1}, (X_{n,h}^{(k)}, A_{n,h}^{(k)})_{k=1}^K)_{h=0}^{H-1}\}_{n=1}^N.$$

Here, for each n, a behavior population policy ρ_n is drawn independently from a distribution $\mathfrak{p}$ over Π, and conditional on ρ_n, the representative trajectory $(X_{n,0}, A_{n,0}, \dots, X_{n,H-1}, A_{n,H-1}, X_{n,H})$ is

Algorithm 1 Offline mean-field representation learning

Input: Offline data $\mathcal{D}$, model class Θ, tolerance $\beta \geq 0$, optimization tolerance $\varepsilon_{\mathrm{opt}} \geq 0$.

Output: Policy $\hat{\eta} \in \Pi$ satisfying $\widehat{\mathrm{Gap}}_\beta(\hat{\eta}) \leq \inf_{\eta \in \Pi} \widehat{\mathrm{Gap}}_\beta(\eta) + \varepsilon_{\mathrm{opt}}$.

Write $Y_{n,h} \coloneqq (X_{n,h}, A_{n,h})$, $Y_{n,h}^{(k)} \coloneqq (X_{n,h}^{(k)}, A_{n,h}^{(k)})$.

1: **for** each $\theta = (f^\theta, r^\theta, T^\theta) \in \Theta$ **do**

2: Compute $\hat{m}_{n,h}^\theta = K^{-1} \sum_{k=1}^K f_h^\theta(Y_{n,h}^{(k)})$ and the empirical risk

$$\hat{\mathcal{R}}(\theta) \coloneqq \frac{1}{N} \sum_{n=1}^{N} \sum_{h=0}^{H-1} \left\{ -\log T_h^\theta\Big(X_{n,h+1} \mid Y_{n,h}, \hat{m}_{n,h}^\theta\Big) + \Big(R_{n,h} - r_h^\theta(Y_{n,h}, \hat{m}_{n,h}^\theta)\Big)^2 \right\}. \tag{3}$$

We use the convention $-\log 0 = +\infty$.

3: Create the confidence set

$$\mathcal{C}(\beta) \coloneqq \{\theta \in \Theta : \hat{\mathcal{R}}(\theta) \leq \inf_{\theta' \in \Theta} \hat{\mathcal{R}}(\theta') + \beta\} \tag{4}$$

4: **for** each candidate population policy $\eta \in \Pi$ **do**

5: Evaluate $J_\theta(\pi; \eta)$ for all $\theta \in \mathcal{C}(\beta)$ and $\pi \in \Pi$, and set

$$\overline{J}_\beta(\pi; \eta) \coloneqq \sup_{\theta \in \mathcal{C}(\beta)} J_\theta(\pi; \eta),\ \underline{J}_\beta(\pi; \eta) \coloneqq \inf_{\theta \in \mathcal{C}(\beta)} J_\theta(\pi; \eta),\ \widehat{\mathrm{Gap}}_\beta(\eta) \coloneqq \sup_{\pi \in \Pi} \overline{J}_\beta(\pi; \eta) - \underline{J}_\beta(\eta; \eta). \tag{5}$$

6: **return** any policy satisfying the output condition.

drawn from $\mathbb{P}_\star^{\rho_n; \rho_n}$, and the observed reward is $R_{n,h} = r_h^\star(X_{n,h}, A_{n,h}, m_h^{\star,\rho_n})$. Conditional on ρ_n, the entire collection of population snapshots is independent of the representative trajectory. At each time-step h, the snapshot consists of K samples $(X_{n,h}^{(k)}, A_{n,h}^{(k)}) \overset{\text{i.i.d.}}{\sim} \nu_h^{\star,\rho_n}$.

The sample can be used to compute the empirical mean-field representation for a candidate model θ: $\hat{m}_{n,h}^\theta \coloneqq \frac{1}{K} \sum_{k=1}^K f_h^\theta(X_{n,h}^{(k)}, A_{n,h}^{(k)})$. Conditionally on ρ_n, it satisfies $\mathbb{E}[\hat{m}_{n,h}^\theta \mid \rho_n] = \bar{m}_h^{\theta,\rho_n}$. Note that $\hat{m}_{n,h}^\theta$, $\bar{m}_h^{\theta,\rho_n}$, and m_h^{θ,ρ_n} play different roles. The empirical risk can use only $\hat{m}_{n,h}^\theta$, the population prediction loss is naturally expressed at $\bar{m}_h^{\theta,\rho_n}$, and planning under model θ uses $m_h^{\theta,\eta}$.

Learning algorithm. Algorithm 1 summarizes the offline learning procedure. For each model θ, we use the finite-population samples to construct empirical representation $\hat{m}^\theta$ and fit the reward and transition components using the empirical risk in (3). We then retain the near-minimizers in the confidence set in (4). Planning is performed over this confidence set: for each candidate population policy, the algorithm evaluates deviations under all plausible models, forms optimistic deviation values and a pessimistic self-play value, and scores the policy by the interval Nash-gap objective in (5). The output is an $\varepsilon_{\mathrm{opt}}$-approximate minimizer of this objective; when the true model is contained in the confidence set, the interval objective upper bounds the true Nash gap.

Transition support. The support condition in Assumption 2 allows zero transition probabilities and includes deterministic kernels. Together with total-variation continuity, it implies that, for fixed θ, h, x, a, the support of $T_h^\theta(\cdot \mid x, a, m)$ is independent of $m \in \mathcal{B}_f$. In particular, when the kernel is deterministic, its successor may depend on h, x, a but not on m. This support stability ensures that evaluating the realizable model at a finite-population estimate of the representation cannot assign zero probability to a transition possible at the exact representation. These assumptions do not require distinct candidate models to share a transition support. We therefore measure transition prediction error using squared Hellinger distance, which remains finite under support mismatch.

4 Provable Representation Learning in Mean-Field RL

We now state the main guarantee for the offline procedure from Section 3. The likelihood-based empirical criterion yields a confidence set with small population prediction loss along the logged mean-field representation paths. This loss controls logged Bellman residuals, and the coefficient κ transfers those residuals to value errors for a target population policy. Optimistic deviation values and pessimistic self-play values then turn these value bounds into a Nash-gap guarantee.

Model confidence from data. We first bound the population prediction error of the models selected by the empirical criterion. The algorithm fits transitions by negative log-likelihood, but the analysis measures transition error in squared Hellinger distance. The reason is that a candidate model can fit every observed transition yet rule out a rare transition that never appeared in the data. Its KL error is then infinite, while its Hellinger error stays finite; this can happen even when the representation is known exactly. The likelihood criterion controls the Hellinger error, which in turn controls the total-variation error needed in the Bellman analysis. To this end, define the population prediction loss $\mathcal{L}(\theta)$ as

$$\sum_{h=0}^{H-1} \mathbb{E}_{\rho\sim\mathfrak{p},Y\sim\nu_h^{\star,\rho}}\left[2D_{\mathrm{H}}^2\Big(T_h^\star(\cdot \mid Y, m_h^{\star,\rho}), T_h^\theta(\cdot \mid Y, \bar{m}_h^{\theta,\rho})\Big) + \Big(r_h^\star(Y, m_h^{\star,\rho}) - r_h^\theta(Y, \bar{m}_h^{\theta,\rho})\Big)^2\right].$$

Under Assumption 1, $\mathcal{L}(\theta^\star) = 0$. For $s \geq 0$, define $\Theta_s \coloneqq \{\theta \in \Theta : \mathcal{L}(\theta) \leq s\}$.

Theorem 1 (Finite-class confidence event, informal constants)**.** *Suppose Assumptions 1–3 hold and Θ is finite. There are choices of β and $s_{N,K}$ with*

$$\beta,\ s_{N,K} = O\left(\frac{H\big(1+\log(1/\tau)\big)\log(|\Theta|/\delta)}{N} + \frac{B_f^2}{K}\sum_{h=0}^{H-1}\left[\frac{L_{T,h}^2}{\tau} + L_{r,h}^2\right]\right)$$

such that, with probability at least $1-\delta$, $\theta^\star \in \mathcal{C}(\beta)$ and $\mathcal{C}(\beta) \subseteq \Theta_{s_{N,K}}$.

The N^{-1} term is the usual model-selection error from N logged episodes. The K^{-1} term is specific to mean-field representation learning because each episode shows only K samples from the population, so the learner evaluates candidate models at the estimate $\hat{m}_{n,h}^\theta$ rather than at $\bar{m}_h^{\theta,\rho_n}$. If $B_f, \tau^{-1}, L_{T,h}, L_{r,h}$ are uniformly bounded in h, then $s_{N,K} = O\Big(\frac{H\log(|\Theta|/\delta)}{N} + \frac{H}{K}\Big)$. The exact statement and proof are given in Theorem 3 in App. A.2.

Value transfer through Bellman residuals. The confidence event above controls prediction loss only at logged population policies $\rho \sim \mathfrak{p}$. Planning, however, evaluates new population policies η and deviations π. We transfer this confidence through Bellman residuals of the candidate model-induced value functions.

For $\theta \in \Theta$ and $\pi \in \Pi$, define suffix-indexed candidate value functions $q_h^{\theta,\pi}$ and $v_h^{\theta,\pi}$ by the terminal condition $v_H^{\theta,\pi}(x, \varnothing) = 0$ and the recursion in (2) with $\star$ replaced by θ. For a representation $m \in \mathcal{B}_f$ and a candidate suffix $z_{h:H-1}$, define the Bellman residual $\mathfrak{b}_h^{\theta,\pi}(x, a; m, z_{h:H-1})$ by

$$r_h^\theta(x,a,z_h) - r_h^\star(x,a,m) + \sum_{x'\in\mathcal{X}}\big(T_h^\theta(x' \mid x,a,z_h) - T_h^\star(x' \mid x,a,m)\big)v_{h+1}^{\theta,\pi}(x', z_{h+1:H-1}).$$

This is the Bellman residual of the candidate suffix value under the true one-step Bellman operator.

Write $\|\mathfrak{b}^{\theta,\pi}\|_{2,\mathsf{M}}^2 \coloneqq \int \mathfrak{b}_h^{\theta,\pi}(x,a;m,z_{h:H-1})^2\,\mathrm{d}\mathsf{M}(h,m,z_{h:H-1},x,a)$ for a finite positive measure M. For each θ, define the logged and target lifted measures by

$$\mathsf{D}_{\log}^\theta \coloneqq \sum_{h=0}^{H-1}\mathbb{E}_{\rho\sim\mathfrak{p}}\left[\delta_{(h,m_h^{\star,\rho},\bar{m}_{h:H-1}^{\theta,\rho})}\otimes\nu_h^{\star,\rho}\right], \qquad \mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta} \coloneqq \sum_{h=0}^{H-1}\delta_{(h,m_h^{\star,\eta},m_{h:H-1}^{\theta,\eta})}\otimes\nu_h^{\star,\pi;\eta}.$$

The target measure weights residuals by the representative-agent occupancy under π against the population flow of η, while the logged measure weights residuals by behavior population flows.

Define the Bellman-transfer concentrability coefficient

$$\kappa(\pi;\eta;s) \coloneqq \sup_{\theta\in\Theta_s}\frac{\left(\int \mathfrak{b}^{\theta,\pi}\,\mathrm{d}\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}\right)^2}{\|\mathfrak{b}^{\theta,\pi}\|_{2,\mathsf{D}_{\log}^\theta}^2} = \sup_{\theta\in\Theta_s}\frac{\big(J_\theta(\pi;\eta) - J(\pi;\eta)\big)^2}{\|\mathfrak{b}^{\theta,\pi}\|_{2,\mathsf{D}_{\log}^\theta}^2},$$

with the conventions $0/0 = 0$ and positive-over-zero equal to $+\infty$. App. A.4 proves the last equality. A finite coefficient means that every model in Θ_s whose Bellman residual is small on logged population flows must have small signed residual along the target evaluation of π against η.

App. A.6 gives sufficient coverage conditions for finite κ. These are not necessary: κ can be finite even when different representation paths make the lifted measures mutually singular.

The target measure uses the whole suffix $m^{\theta,\eta}_{h:H-1}$ because continuation values depend on future representations (Proposition 1). The logged measure uses $\bar{m}^{\theta,\rho}_{h:H-1}$ because the offline samples identify candidate features on the true logged population flow.

Remark 1 (An L^2 Bellman coverage coefficient)**.** Another natural Bellman-error coverage ratio is

$$\mathscr{C}(\pi;\eta;s) \coloneqq \sup_{\theta\in\Theta_s} \frac{\|\mathfrak{b}^{\theta,\pi}\|^2_{2,\mathsf{D}^{\theta,\pi;\eta}_{\mathrm{tar}}}}{\|\mathfrak{b}^{\theta,\pi}\|^2_{2,\mathsf{D}^{\theta}_{\log}}},$$

with the same conventions. This is similar to the Bellman-error coverage ratios used in offline Markov games (Zhang et al., 2023b). Since $\mathsf{D}^{\theta,\pi;\eta}_{\mathrm{tar}}$ has total mass H, the Cauchy–Schwarz inequality gives $\kappa(\pi;\eta;s) \le H\mathscr{C}(\pi;\eta;s)$ (Proposition 3 in App. A.6). Thus $\mathscr{C}$ can replace κ in all bounds at the cost of an additional factor $\sqrt{H}$ in the value-transfer radius. The coefficient κ measures only the transfer needed for the signed value error, whereas the coefficient $\mathscr{C}$ controls the stronger target L^2 Bellman-residual error.

Let $a_H \coloneqq 1+(H-1)^2$. App. A.4 proves $\|\mathfrak{b}^{\theta,\pi}\|^2_{2,\mathsf{D}^{\theta}_{\log}} \le a_H\mathcal{L}(\theta)$ for all $\theta\in\Theta,\ \pi\in\Pi$. For $s>0$, define $R_s(\pi;\eta) \coloneqq \sqrt{a_H s\,\kappa(\pi;\eta;s)}$. Then, whenever $\theta\in\Theta_s$, $|J_\theta(\pi;\eta) - J(\pi;\eta)| \le R_s(\pi;\eta)$.

Nash-gap guarantee. We now combine the confidence event with the optimistic–pessimistic planner. For a population policy η, define the optimistic response sub-optimality of $\tilde\pi$ by $\mathrm{subopt}^{+}_{\beta,\eta}(\tilde\pi) \coloneqq \sup_{\pi\in\Pi} \overline{J}_\beta(\pi;\eta) - \overline{J}_\beta(\tilde\pi;\eta)$. This term is zero for an optimistic best response to η and small for an approximate maximizer of the optimistic value.

Theorem 2 (Adaptive Nash-gap bound)**.** *Assume that, for some $s>0$, the event $\theta^\star \in \mathcal{C}(\beta)\subseteq\Theta_s$ holds. Let $\hat\eta$ be an $\varepsilon_{\mathrm{opt}}$-approximate minimizer of $\widehat{\mathrm{Gap}}_\beta$. Then, for every $\eta\in\Pi$,*

$$\mathrm{Gap}(\hat\eta) \le \varepsilon_{\mathrm{opt}} + \mathrm{Gap}(\eta) + R_s(\eta;\eta) + \inf_{\tilde\pi\in\Pi}\left\{R_s(\tilde\pi;\eta) + \mathrm{subopt}^{+}_{\beta,\eta}(\tilde\pi)\right\}.$$

The proof is given in App. A.5. This theorem is the main offline equilibrium statement. Learning succeeds when, for some low-gap policy η, the data can evaluate its self-play value and one approximate best response under the optimistic value estimates. It does not require coverage of every possible unilateral deviation.

A simpler corollary is obtained by replacing the adaptive response term by a worst-case Bellman-transfer coefficient, $\kappa_{\mathrm{uni}}(\eta;s) \coloneqq \sup_{\pi\in\Pi}\kappa(\pi;\eta;s)$.

Corollary 1 (Equilibrium rate)**.** *Under the assumptions of Theorem 1, with probability at least $1-\delta$, every $\varepsilon_{\mathrm{opt}}$-approximate minimizer $\hat\eta$ of $\widehat{\mathrm{Gap}}_\beta$ satisfies, for every Nash equilibrium $\eta_\star\in\Pi_\star$,*

$$\mathrm{Gap}(\hat\eta) \le \varepsilon_{\mathrm{opt}} + 2\sqrt{a_H s_{N,K}}\sqrt{\kappa_{\mathrm{uni}}(\eta_\star;s_{N,K})}.$$

The proof is given in App. A.5, which also states the bound for a general comparison policy $\eta\in\Pi$ in (11). When $B_f, \tau^{-1}, L_{T,h}, L_{r,h}$ are bounded uniformly in h, the last display gives the rate

$$\mathrm{Gap}(\hat\eta) \le \varepsilon_{\mathrm{opt}} + O\left(\inf_{\eta_\star\in\Pi_\star}\sqrt{\kappa_{\mathrm{uni}}(\eta_\star;s_{N,K})}\left[\sqrt{\frac{a_H H\log(|\Theta|/\delta)}{N}} + \sqrt{\frac{a_H H}{K}}\right]\right).$$

The two terms have the same sources as in Theorem 1. We emphasize that the bounds have no explicit dependence on the number of states.

5 Experiments

We study a routing game motivated by fulfillment and supply-chain optimization. The game is **one-step** ($H=1$), so the learner only needs the reward, which depends on the population through an unknown low-dimensional representation. Because Algorithm 1 is not computationally tractable, this study is not an implementation of our algorithm; **it demonstrates how learning a low-dimensional mean-field representation helps** (App. B.1).

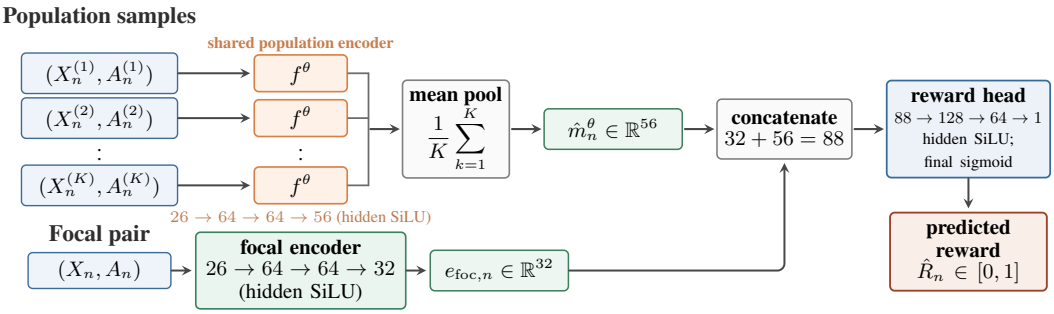

Figure 2: The learned mean-field reward model. A shared encoder f^θ maps each population sample, and the K outputs are averaged into $\hat{m}_n^\theta$. A separate focal encoder embeds (X_n, A_n), and the reward head maps the concatenated embeddings to the predicted reward.

Routing game environment. The network in Figure 1 has origins $O_1, \ldots, O_5$, destinations $D_1, \ldots, D_5$, and intermediate layers $U_1, \ldots, U_4$ and $V_1, \ldots, V_4$. A state is an origin–destination pair $x = (i, j)$, so $|\mathcal{X}| = 25$, and an action picks one node in each intermediate layer, $a = (u, v)$, so $|\mathcal{A}| = 16$; together they give the route $O_i \to U_u \to V_v \to D_j$. Each route uses three of the $5 \cdot 4 + 4 \cdot 4 + 4 \cdot 5 = 56$ edges, and the true feature map $f^\star \colon \mathcal{X} \times \mathcal{A} \to \{0, 1\}^{56}$ records which ones. The true representation $m^\star = \mathbb{E}_{(X,A)\sim\nu}[f^\star(X, A)]$ of a population law ν is therefore the vector of edge loads, with 56 coordinates instead of the 400 of ν. Each edge's latency grows quadratically with its load, and an agent's reward decreases with the total latency of its route (App. B.2). The edge parameters are asymmetric, so a route that is attractive under one population can be poor under another, and the uniform policy is not automatically close to equilibrium.

Offline data generation. Each row n has its own context: a demand law μ_n over origin–destination pairs and a behavior policy ρ_n, which induce $\nu_n = \mu_n \otimes \rho_n$ and its representation $m_n^\star$. We draw a focal pair $(X_n, A_n) \sim \nu_n$, an independent snapshot of K samples from ν_n, and the reward $R_n = r^\star(X_n, A_n, m_n^\star)$. The snapshot affects the model input but not the reward label, so N counts logged observations, while K controls how accurately each row's population is observed (details in App. B.2).

Reward models. Every model encodes a state–action pair by $z(x, a) \in \{0, 1\}^{26}$, which concatenates one-hot codes of the origin, the destination, and the action. All six models share the focal encoder, reward head, optimizer, data, and checkpoint-selection rule, and their parameter counts are approximately the same (App. B.3).

The *single-agent* model sees no population input. The *monolithic raw mean* applies a network to the average raw encoding $\frac{1}{K}\sum_k z(X_n^{(k)}, A_n^{(k)})$. This average reveals the origin and destination marginals and the loads on the 16 middle edges, but not the 40 outer-edge loads, so it is useful but structurally incomplete. The *full population law* model feeds a network the 400-bin empirical histogram $\hat{\nu}_n$ of the K samples. It has all the information but no edge-load inductive bias.

Our *learned mean field* model encodes each sample before averaging, $\hat{m}_n^\theta = \frac{1}{K}\sum_k f^\theta(X_n^{(k)}, A_n^{(k)})$ (Figure 2). Because $z(x, a)$ identifies the route, a suitable encoder can recover $f^\star$, and then $\hat{m}_n^\theta$ estimates $m_n^\star$. Two oracles receive privileged information. The *finite-K oracle* averages the true features $f^\star$ over the snapshot, which removes representation-learning error but keeps finite-K noise. The *infinite-population oracle* receives $m_n^\star$ itself, which removes both but still learns the reward from finitely many rows.

Evaluation metrics. At evaluation, every model receives its exact population input, the infinite-K limit of its training input, and we enumerate all 400 state–action pairs (App. B.4). For a context c with law ν_c and true reward $r_c^\star(x, a) = r^\star(x, a, m_c^\star)$, a fitted reward $\hat{r}_c$ has error

$$\mathrm{RMSE}_{\mathrm{pop}} = \Big[\mathbb{E}_c \sum_{x,a} \nu_c(x, a)\big(\hat{r}_c(x, a) - r_c^\star(x, a)\big)^2\Big]^{1/2}. \tag{6}$$

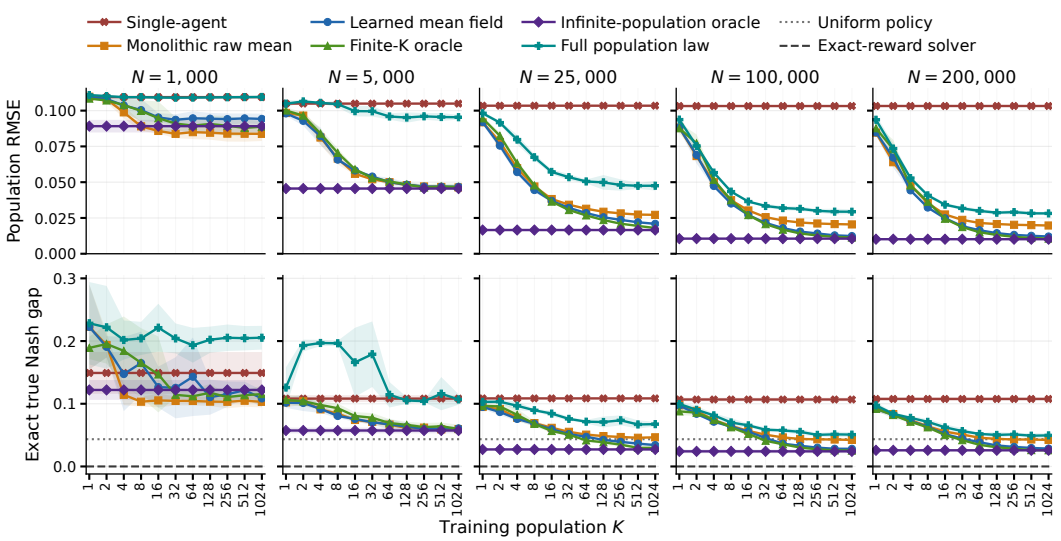

Figure 3: Population reward error (top) and exact true Nash gap of the policy obtained by solving each fitted game (bottom), as functions of the number N of offline rows and the number K of population samples per row; lower is better. Horizontal lines mark the uniform policy and the exact-reward control. Lines connect means; bands are 95% Student-t intervals over five paired seeds.

For a target demand law μ and a policy η, let $m^{\star,\eta} = \sum_{x,a} \mu(x)\eta(a \mid x) f^\star(x,a)$. The exact one-step Nash gap, the expected gain from a unilateral best response, is

$$\mathrm{Gap}_\mu(\eta) = \sum_x \mu(x) \Big[\max_a r^\star(x,a,m^{\star,\eta}) - \sum_a \eta(a \mid x)\, r^\star(x,a,m^{\star,\eta}) \Big]. \tag{7}$$

For each fitted model, we compute candidate equilibria of the fitted game using the optimization procedures described in Appendix B.5, select the candidate with the smallest Nash gap under the fitted reward, and only then evaluate it with (7). As references, we report the uniform policy and an exact-reward control that runs the same planner on the true reward.

Results. Figure 3 shows both metrics over the N-by-K grid. The single-agent and infinite-population-oracle curves are constant in K by construction. At $K = 1$ there is only one sample to pool, so encoding before averaging cannot help, as the curves confirm. As K and N grow, the models separate according to how much population information they can recover, and the finite-K oracle approaches the infinite-population oracle. At $(N, K) = (200000, 1024)$, both metrics order the models as infinite-population oracle $\approx$ finite-K oracle $<$ learned mean field $<$ monolithic raw mean $<$ single-agent. The full-population-law model does worse than the learned mean field on average. It has every population coordinate but must learn from a noisy 400-dimensional histogram without the edge-load inductive bias. So access to the full law is not by itself an effective finite-data representation at this budget, although a different architecture or a larger budget could change this.

6 Conclusion and Future Work

We showed that the population's effect on rewards and transitions can be learned offline through a low-dimensional representation, with a finite-sample Nash-gap guarantee under adaptive coverage, and that, given enough data, this improves reward prediction and equilibrium quality in a routing benchmark. There are several open directions including lower bounds that identify the weakest coverage conditions for offline equilibrium learning, scalable algorithms with comparable provable guarantees (ours is statistically motivated but computationally inefficient), and an online version, which would need new exploration criteria over population flows and unilateral deviations.

AI USE STATEMENT

LLMs were used in brainstorming ideas and proof techniques, and checking for mathematical correctness. LLMs were heavily used in writing code for the experiments. The authors take full responsibility for the content in the paper.

REPRODUCIBILITY STATEMENT

All mathematical claims in the main paper are proved in the appendix. All experiments in the paper can be reproduced from the source code provided.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107, 2020.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Learning mean-field games with discounted and average costs. *Journal of Machine Learning Research*, 24(17):1–59, 2023. URL https://www.jmlr.org/papers/v24/21-0505.html.

Axel Brunnbauer, Julian Lemmel, Zahra Babaiee, Sophie A. Neubauer, and Radu Grosu. Scalable offline reinforcement learning for mean field games. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 408–417. International Foundation for Autonomous Agents and Multiagent Systems, 2025. doi: 10.5555/3709347.3743555.

Minshuo Chen, Yan Li, Ethan Wang, Zhuoran Yang, Zhaoran Wang, and Tuo Zhao. Pessimism meets invariance: Provably efficient offline mean-field multi-agent rl. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17913–17926, 2021.

Qiwen Cui and Simon S. Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11739–11751, 2022a.

Qiwen Cui and Simon S. Du. When are offline two-player zero-sum markov games solvable? In *Advances in Neural Information Processing Systems*, volume 35, pp. 25779–25791, 2022b.

Diogo A. Gomes, Joana Mohr, and Rafael Rigão Souza. Discrete time, finite state space mean field games. *Journal de Mathématiques Pures et Appliquées*, 93(3):308–328, 2010. doi: 10.1016/j.matpur.2009.10.010.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32:4966–4976, 2019.

Jiawei Huang, Niao He, and Andreas Krause. Model-based RL for mean-field games is not statistically harder than single-agent RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19816–19870. PMLR, 2024a.

Jiawei Huang, Batuhan Yardim, and Niao He. On the statistical efficiency of mean-field reinforcement learning with general function approximation. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 289–297. PMLR, 2024b.

Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop mckean–vlasov systems and the nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252, 2006. doi: 10.4310/CIS.2006.v6.n3.a5.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 40(4):570–596, 2025. doi: 10.1214/25-STS1000. URL https://doi.org/10.1214/25-STS1000.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096. PMLR, 2021.

Arman C. Kizilkale and Peter E. Caines. Mean field stochastic adaptive control. *IEEE Transactions on Automatic Control*, 58(4):905–920, 2013. doi: 10.1109/TAC.2012.2228032.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2 (1):229–260, 2007. doi: 10.1007/s11537-007-0657-8.

Mathieu Laurière, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Théophile Cabannes, Georgios Piliouras, Julien Pérolat, Romuald Élie, Olivier Pietquin, and Matthieu Geist. Scalable deep reinforcement learning algorithms for mean field games. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12078–12095. PMLR, 2022a.

Mathieu Laurière, Sarah Perrin, Julien Pérolat, Sertan Girgin, Paul Muller, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Learning in mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022b.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravishankar K. Iyer, and Tamer Başar. A mean-field game approach to cloud resource management with function approximation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36243–36258, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/eb3c8135137c8a60425a0320869ad87e-Abstract-Conference.html.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6961–6971. PMLR, 2020.

Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Zihan Ding, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. In *International Conference on Learning Representations*, 2023.

Barna Pásztor, Ilija Bogunovic, and Andreas Krause. Efficient model-based multi-agent mean-field reinforcement learning. *Transactions on Machine Learning Research*, 2023.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018. doi: 10.1137/17M1112583.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 251–259. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10092–10103. PMLR, 2020.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11436–11447. PMLR, 2021a.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694, 2021b.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5571–5580. PMLR, 2018.

Huibing Yin, Prashant G. Mehta, Sean P. Meyn, and Uday V. Shanbhag. Learning in mean-field games. *IEEE Transactions on Automatic Control*, 59(3):629–644, 2014. doi: 10.1109/TAC.2013.2287733.

Wenhao Zhan, Scott Fujimoto, Zheqing Zhu, Jason D. Lee, Daniel R. Jiang, and Yonathan Efroni. Exploiting structure in offline multi-agent RL: The benefits of low interaction rank. In *International Conference on Learning Representations*, 2025.

Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation selection in low-rank markov decision processes: From online to offline rl. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2488–2497. PMLR, 2023a.

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline learning in Markov games with general function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40804–40829. PMLR, 2023b.

NOTATION TABLE

Conventions. Model indices $\star$ and θ distinguish true-model and candidate-model quantities. Here Θ is the finite candidate class. In $(\pi;\eta)$, π is the representative-agent policy and η is the population policy; $\rho \sim \mathfrak{p}$ denotes a logged behavior population policy, where $\mathfrak{p}$ is the logging distribution on Π. For a sequence $u = (u_0, \dots, u_{H-1})$, write $u_{h:H-1} = (u_h, \dots, u_{H-1})$.

Symbol	Meaning
Flows and representations	
$\mu_h^{\star,\eta}, \nu_h^{\star,\eta}$	True population state and state-action laws induced by η, with $\nu_h^{\star,\eta} = \mu_h^{\star,\eta} \otimes \eta_h$.
$\mu_h^{\star,\pi;\eta}, \nu_h^{\star,\pi;\eta}$	State and state-action laws of a representative agent using π against the fixed population flow induced by η.
$m_h^{\star,\eta}$	True representation $\sum_{x,a} \nu_h^{\star,\eta}(x,a) f_h^\star(x,a)$.
$m_h^{\theta,\eta}$	Candidate representation $\sum_{x,a} \nu_h^{\theta,\eta}(x,a) f_h^\theta(x,a)$, computed along the candidate population flow and used for planning.
$\bar{m}_h^{\theta,\eta}$	Candidate feature average on the true population flow, $\sum_{x,a} \nu_h^{\star,\eta}(x,a) f_h^\theta(x,a)$, used in the population prediction loss.
$\hat{m}_{n,h}^\theta$	Empirical candidate representation formed from the K population samples in episode n at stage h.
$\mathbb{P}_\star^{\pi;\eta}, \mathbb{P}_\theta^{\pi;\eta}$	True and candidate representative-agent path laws.
$J(\pi;\eta), J_\theta(\pi;\eta)$	True and candidate finite-horizon values.
$\mathrm{Gap}(\eta), \Pi_\star$	Nash gap $\sup_{\pi\in\Pi} J(\pi;\eta) - J(\eta;\eta)$; $\Pi_\star = \{\eta \in \Pi : \mathrm{Gap}(\eta) = 0\}$ is the set of Nash equilibria.
Data and confidence	
N, K	Number of logged episodes and population samples per episode-stage.
$\hat{\mathcal{R}}(\theta), \mathcal{C}(\beta)$	Empirical risk and its β-near-minimizer confidence set.
$\mathcal{L}(\theta), \Theta_s$	Population prediction loss; $\Theta_s = \{\theta \in \Theta : \mathcal{L}(\theta) \le s\}$.
$\tilde{\mathcal{L}}(\theta)$	Expected prediction loss $\mathbb{E}W_1(\theta)$ at the empirical representations $\hat{m}^\theta$, including expectation over the population snapshots.
$s_{N,K}$	High-probability upper bound on $\mathcal{L}(\theta)$ for $\theta \in \mathcal{C}(\beta)$.
$B_f, \mathcal{B}_f, \tau, L_{r,h}, L_{T,h}$	Feature bound, representation ball, lower bound on nonzero transition probabilities, and reward/transition Lipschitz constants.
Bellman transfer and interval quantities	
$q_h^{\theta,\pi}, v_h^{\theta,\pi}; \mathfrak{b}_h^{\theta,\pi}$	Candidate suffix-indexed action-value and value functions, and their scalar Bellman residual under the true one-step Bellman operator.
$\mathsf{D}_{\log}^\theta, \mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}$	Logged and target measures on tuples $(h, m, z_{h:H-1}, x, a)$, each with total mass H.
$\kappa(\pi;\eta;s), \mathscr{C}(\pi;\eta;s)$	Scalar Bellman-transfer coefficient and the stronger target-L^2 Bellman-coverage coefficient.
$\kappa_{\mathrm{uni}}(\eta;s)$	Worst-case Bellman-transfer coefficient over representative-agent policies, $\sup_{\pi\in\Pi} \kappa(\pi;\eta;s)$.
$\mathrm{rad}_{\mathrm{Bell},s}(\pi;\eta)$	Worst-case absolute value error over Θ_s.
$R_s(\pi;\eta), a_H$	Value-error bound $R_s(\pi;\eta) = \sqrt{a_H s\, \kappa(\pi;\eta;s)}$, $s > 0$, with $a_H = 1 + (H-1)^2$.
$\overline{J}_{\mathcal{C}}, \underline{J}_{\mathcal{C}}, \widehat{\mathrm{Gap}}_{\mathcal{C}}$	Optimistic and pessimistic values and the induced interval gap for a nonempty candidate-model set $\mathcal{C} \subseteq \Theta$.
$r_{\mathcal{C}}^{\pm}, \mathrm{subopt}_{\mathcal{C},\eta}^{+}, \varepsilon_{\mathrm{opt}}$	Nonnegative one-sided interval errors when $\theta^\star \in \mathcal{C}$, optimistic response suboptimality, and optimization tolerance; set $\mathcal{C} = \mathcal{C}(\beta)$ for the main-text quantities.

A Proofs and Additional Details

A.1 Elementary structural facts

A.1.1 The current representation need not determine action values

Proposition 1 (The current representation need not determine the action value)**.** *There exists a horizon-two mean-field model satisfying the representation structure of Section 2, together with policies $\eta, \tilde{\eta}, \pi \in \Pi$, a state $x \in \mathcal{X}$, and an action $a \in \mathcal{A}$, such that $m_0^{\star,\eta} = m_0^{\star,\tilde{\eta}}$, but $q_0^{\star,\pi}\left(x, a, m_{0:1}^{\star,\eta}\right) \neq q_0^{\star,\pi}\left(x, a, m_{0:1}^{\star,\tilde{\eta}}\right)$. Consequently, action value functions cannot in general be represented uniformly over population policies as functions of $(x, a, m_h^{\star,\eta})$ alone.*

Proof. Take $H = 2$, $\mathcal{X} = \{x_0, u, v\}$, $\mathcal{A} = \{L, R\}$, and $\mu_0 = \delta_{x_0}$. Let the stage-0 reward and feature map be identically zero, so $r_0^\star \equiv 0$ and $f_0^\star \equiv 0$. Let the stage-0 transition be independent of the representation and satisfy $T_0^\star(u \mid x, L, m) = 1$ and $T_0^\star(v \mid x, R, m) = 1$ for every $x \in \mathcal{X}$ and $m \in \mathbb{R}$.

At stage 1, use the scalar feature map $f_1^\star(u, a) = 1$ and $f_1^\star(v, a) = f_1^\star(x_0, a) = 0$ for every $a \in \mathcal{A}$. Let $\sigma(m) := \min\{1, \max\{0, m\}\}$, and define the terminal reward by $r_1^\star(u, a, m) = \sigma(m)$ and $r_1^\star(v, a, m) = r_1^\star(x_0, a, m) = 0$. The stage-1 transition kernel is irrelevant.

Let η choose L at x_0 with probability one, and let $\tilde{\eta}$ choose R at x_0 with probability one. Since $f_0^\star \equiv 0$, we have $m_0^{\star,\eta} = m_0^{\star,\tilde{\eta}} = 0$. At stage 1, however, the population is concentrated at u under η and at v under $\tilde{\eta}$. Hence $m_1^{\star,\eta} = 1$, whereas $m_1^{\star,\tilde{\eta}} = 0$.

Now consider action L for the representative agent at stage 0. This action sends the representative agent to u, irrespective of the population policy. Therefore, for every representative policy π,

$$q_0^{\star,\pi}\left(x_0, L, m_{0:1}^{\star,\eta}\right) = 1, \qquad q_0^{\star,\pi}\left(x_0, L, m_{0:1}^{\star,\tilde{\eta}}\right) = 0.$$

Thus the two action values differ although the current inputs (x_0, L, m_0) coincide. □

A.1.2 Existence of Nash equilibria

Proposition 2 (Existence of Nash equilibria)**.** *Suppose $\mathcal{X}$ and $\mathcal{A}$ are finite, each feature map $f_h^\star$ is bounded, and for every h, x, a, the maps $m \mapsto r_h^\star(x, a, m)$ and $m \mapsto T_h^\star(\cdot \mid x, a, m)$ are continuous. Then there exists a policy $\eta^\star \in \Pi$ such that $\mathrm{Gap}(\eta^\star) = 0$.*

The proof uses the standard occupancy-measure formulation and a Kakutani fixed-point argument applied to the mean-field representation path.

Proof. Let

$$\mathcal{K}_h := \mathrm{conv}\,\{f_h^\star(x, a) : (x, a) \in \mathcal{X} \times \mathcal{A}\} \qquad \text{and} \qquad \mathcal{K} := \prod_{h=0}^{H-1} \mathcal{K}_h.$$

The set $\mathcal{K}$ of candidate representation paths is nonempty, compact, and convex. For a frozen path $m \in \mathcal{K}$, let $\mathrm{Occ}(m)$ be the set of $\psi \in \mathcal{D} := \prod_{h=0}^{H-1} \Delta(\mathcal{X} \times \mathcal{A})$ satisfying

$$\sum_a \psi_0(x, a) = \mu_0(x), \qquad x \in \mathcal{X},$$

$$\sum_{a'} \psi_{h+1}(x', a') = \sum_{x,a} \psi_h(x, a) T_h^\star(x' \mid x, a, m_h), \quad 0 \le h \le H-2,\ x' \in \mathcal{X}.$$

These are precisely the state-action occupancy flows of Markov policies in the MDP with rewards $r_h^\star(x, a, m_h)$ and transitions $T_h^\star(\cdot \mid x, a, m_h)$. For each m, the set $\mathrm{Occ}(m)$ is nonempty, compact, and convex. Continuity of the transitions gives the correspondence $m \mapsto \mathrm{Occ}(m)$ a closed graph. Define

$$J_m(\psi) := \sum_{h,x,a} \psi_h(x, a) r_h^\star(x, a, m_h), \qquad \mathrm{OptOcc}(m) := \operatorname*{arg\,max}_{\psi \in \mathrm{Occ}(m)} J_m(\psi).$$

Linearity of J_m in ψ makes $\mathrm{OptOcc}(m)$ nonempty, compact, and convex. The optimal frozen-path value $V^*(m)$ is continuous: a finite-horizon MDP has a deterministic optimal Markov policy, so $V^*(m)$ is the maximum of the continuous values of finitely many such policies. Since $J_m(\psi)$ is jointly continuous, the characterization $\mathrm{OptOcc}(m) = \{\psi \in \mathrm{Occ}(m) : J_m(\psi) = V^*(m)\}$ then shows that OptOcc has closed graph.

Let

$$\Phi(\psi) \coloneqq \left(\sum_{x,a} \psi_h(x,a) f_h^\star(x,a)\right)_{h=0}^{H-1}, \qquad \Gamma(m) \coloneqq \Phi\big(\mathrm{OptOcc}(m)\big).$$

The linear map Φ takes $\mathcal{D}$ into $\mathcal{K}$, so Γ has nonempty compact convex values in $\mathcal{K}$. Its graph is closed: if $m^n \to m$ and $\Phi(\psi^n) \to z$ with $\psi^n \in \mathrm{OptOcc}(m^n)$, compactness of $\mathcal{D}$ gives a convergent subsequence with limit $\psi \in \mathrm{OptOcc}(m)$, and continuity of Φ gives $z = \Phi(\psi)$. Because $\mathcal{K}$ is compact, this implies upper hemicontinuity. Kakutani's theorem therefore gives $m^\star \in \Gamma(m^\star)$, witnessed by some $\psi^\star \in \mathrm{OptOcc}(m^\star)$ with $\Phi(\psi^\star) = m^\star$.

Define $\eta_h^\star(a \mid x) = \psi_h^\star(x,a) / \sum_b \psi_h^\star(x,b)$ whenever the denominator is positive, and choose any action distribution otherwise. The flow constraints and $\Phi(\psi^\star) = m^\star$ imply by forward induction that this policy induces $\nu_h^{\star,\eta^\star} = \psi_h^\star$ and $m_h^{\star,\eta^\star} = m_h^\star$ at every stage. Thus every deviation $\pi \in \Pi$ faces the frozen MDP at $m^\star$, and its occupancy $\psi^{\pi,m^\star}$ belongs to $\mathrm{Occ}(m^\star)$. Optimality of $\psi^\star$ gives

$$J(\pi;\eta^\star) = J_{m^\star}(\psi^{\pi,m^\star}) \le J_{m^\star}(\psi^\star) = J(\eta^\star;\eta^\star).$$

Taking the supremum over π and noting equality for $\pi = \eta^\star$ proves $\mathrm{Gap}(\eta^\star) = 0$. □

A.1.3 Support and distance facts

Lemma 1 (Support stability). *Under Assumptions 2 and 3, fix $\theta \in \Theta$, h, (x,a), and $x' \in \mathcal{X}$. Then either*

$$T_h^\theta(x' \mid x,a,m) = 0 \quad \textit{for every } m \in \mathcal{B}_f,$$

or

$$T_h^\theta(x' \mid x,a,m) \ge \tau \quad \textit{for every } m \in \mathcal{B}_f.$$

Consequently, $\operatorname{supp} T_h^\theta(\cdot \mid x,a,m)$ *does not depend on* $m \in \mathcal{B}_f$.

Proof. For fixed x', the coordinate map $g(m) = T_h^\theta(x' \mid x,a,m)$ is continuous because

$$|g(m) - g(m')| \le D_{\mathrm{TV}}\big(T_h^\theta(\cdot \mid x,a,m), T_h^\theta(\cdot \mid x,a,m')\big) \le L_{T,h}\|m - m'\|_2.$$

The ball $\mathcal{B}_f$ is connected, so $g(\mathcal{B}_f)$ is connected. Since $g(\mathcal{B}_f) \subseteq \{0\} \cup [\tau, 1]$, its image lies entirely in one of these two components. Applying the argument to every $x' \in \mathcal{X}$ proves the support claim. □

For $p, q \in \Delta(\mathcal{X})$, write $D_{\mathrm{KL}}(p\|q) \coloneqq \sum_x p(x) \log\big(p(x)/q(x)\big)$, with the conventions $0 \log(0/b) = 0$ for $b \ge 0$ and $a \log(a/0) = +\infty$ for $a > 0$.

Lemma 2 (Elementary distance bounds). *For any probability distributions p, q on $\mathcal{X}$,*

$$D_{\mathrm{TV}}(p,q)^2 \le 2 D_{\mathrm{H}}^2(p,q).$$

If p and q have the same support and every positive probability in both distributions is at least τ, then

$$D_{\mathrm{H}}^2(p,q) \le \frac{D_{\mathrm{TV}}(p,q)^2}{2\tau}, \qquad D_{\mathrm{KL}}(p\|q) \le \frac{4}{\tau} D_{\mathrm{TV}}(p,q)^2.$$

Proof. Factoring $|p(x) - q(x)|$ and applying Cauchy–Schwarz gives

$$\begin{aligned} 2D_{\mathrm{TV}}(p,q) &= \sum_x |\sqrt{p(x)} - \sqrt{q(x)}|(\sqrt{p(x)} + \sqrt{q(x)}) \\ &\le \big(2D_{\mathrm{H}}^2(p,q)\big)^{1/2} \left(2 + 2\sum_x \sqrt{p(x)q(x)}\right)^{1/2} \le 2\sqrt{2D_{\mathrm{H}}^2(p,q)}, \end{aligned}$$

which proves the first inequality.

For the second claim, sum over the common support. Since $(\sqrt{p(x)} + \sqrt{q(x)})^2 \geq 4\tau$,

$$D_{\mathrm{H}}^2(p,q) = \frac{1}{2}\sum_x \frac{(p(x)-q(x))^2}{(\sqrt{p(x)}+\sqrt{q(x)})^2} \leq \frac{\|p-q\|_1^2}{8\tau} = \frac{D_{\mathrm{TV}}(p,q)^2}{2\tau}.$$

Finally, $D_{\mathrm{KL}}(p\|q) \leq \chi^2(p\|q) \coloneqq \sum_x (p(x)-q(x))^2/q(x)$, and hence

$$D_{\mathrm{KL}}(p\|q) \leq \sum_x \frac{(p(x)-q(x))^2}{q(x)} \leq \frac{\|p-q\|_1^2}{\tau} = \frac{4}{\tau} D_{\mathrm{TV}}(p,q)^2.$$

□

A.2 Proof of the confidence theorem

We prove Theorem 1 by first bounding prediction error at the empirical representations $\hat{m}_{n,h}^{\theta}$, then transferring the bound to the candidate feature averages $\bar{m}_h^{\theta,\rho_n}$ along the true population flow.

A.2.1 Episode quantities

For episode n and stage h, define $P_{n,h} \coloneqq T_h^\star(\cdot \mid Y_{n,h}, m_h^{\star,\rho_n})$ and $Q_{n,h}^\theta \coloneqq T_h^\theta(\cdot \mid Y_{n,h}, \hat{m}_{n,h}^\theta)$, and let $e_{n,h}^\theta \coloneqq R_{n,h} - r_h^\theta(Y_{n,h}, \hat{m}_{n,h}^\theta)$. The one-episode likelihood difference and prediction loss are

$$A_n(\theta) \coloneqq \sum_{h=0}^{H-1} \left[\log \frac{P_{n,h}(X_{n,h+1})}{Q_{n,h}^\theta(X_{n,h+1})} + (e_{n,h}^\theta)^2 \right],$$

and

$$W_n(\theta) \coloneqq \sum_{h=0}^{H-1} \left[2D_{\mathrm{H}}^2(P_{n,h}, Q_{n,h}^\theta) + (e_{n,h}^\theta)^2 \right].$$

The log ratio is $+\infty$ when its numerator is positive and its denominator is zero. Its value when both probabilities are zero is immaterial, since $X_{n,h+1}$ is drawn from $P_{n,h}$. Write $\overline{A}_N(\theta) = N^{-1}\sum_{n=1}^N A_n(\theta)$ and $\overline{W}_N(\theta) = N^{-1}\sum_{n=1}^N W_n(\theta)$, and define $\tilde{\mathcal{L}}(\theta) \coloneqq \mathbb{E}W_1(\theta)$ as the expected prediction loss at the empirical representations $\hat{m}_{n,h}^\theta$. The population prediction loss $\mathcal{L}(\theta)$ instead evaluates predictions at the corresponding population averages $\bar{m}_h^{\theta,\rho_n}$.

The empirical criterion decomposes as

$$\hat{\mathcal{R}}(\theta) = \hat{H}_N + \overline{A}_N(\theta), \qquad \hat{H}_N \coloneqq \frac{1}{N}\sum_{n=1}^N \sum_{h=0}^{H-1} -\log P_{n,h}(X_{n,h+1}).$$

Every observed next state has positive probability under $P_{n,h}$, and that probability is at least τ. Thus $\hat{H}_N < \infty$ almost surely, and the decomposition is valid in the extended reals. The quantity $\hat{H}_N$ does not depend on θ.

A.2.2 Likelihood and prediction loss

Lemma 3 (Likelihood-to-Hellinger bound)**.** *Let $(\mathcal{F}_i)$ be a filtration, and suppose that Z_i, conditionally on $\mathcal{F}_{i-1}$, has probability mass function p_i. For each $\theta \in \Theta$, let q_i^θ be an $\mathcal{F}_{i-1}$-measurable probability mass function. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $\theta \in \Theta$,*

$$\sum_i \log \frac{p_i(Z_i)}{q_i^\theta(Z_i)} \geq 2\sum_i D_{\mathrm{H}}^2(p_i, q_i^\theta) - 2\log\frac{|\Theta|}{\delta}.$$

The log ratio is interpreted p_i-almost surely, with value $+\infty$ when $p_i(Z_i) > 0$ and $q_i^\theta(Z_i) = 0$.

Proof. Fix θ. With the convention $e^{-\infty} = 0$,

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}\log\frac{p_i(Z_i)}{q_i^\theta(Z_i)}\right)\middle|\mathcal{F}_{i-1}\right] = \sum_z \sqrt{p_i(z)q_i^\theta(z)}$$
$$= 1 - D_{\mathrm{H}}^2(p_i, q_i^\theta) \le e^{-D_{\mathrm{H}}^2(p_i,q_i^\theta)}.$$

Consequently,

$$\exp\left(\sum_{j\le i}\left[D_{\mathrm{H}}^2(p_j, q_j^\theta) - \frac{1}{2}\log\frac{p_j(Z_j)}{q_j^\theta(Z_j)}\right]\right)$$

is a nonnegative supermartingale. Markov's inequality followed by a union bound over Θ proves the claim. □

Fix $\delta \in (0,1)$ and set $u := \log(3|\Theta|/\delta)$. For each episode, reveal ρ_n and the entire population-snapshot collection before revealing the representative trajectory chronologically. The conditional independence in the offline data model ensures that this does not change the representative trajectory law. Immediately before $X_{n,h+1}$ is revealed, $Y_{n,h}$, $P_{n,h}$, and $Q_{n,h}^\theta$ are measurable, and the conditional law of $X_{n,h+1}$ is $P_{n,h}$. Applying Lemma 3 across all episode-stage pairs and adding the reward errors gives, with probability at least $1 - \delta/3$,

$$\overline{A}_N(\theta) \ge \overline{W}_N(\theta) - \frac{2u}{N} \qquad \text{for every } \theta \in \Theta. \tag{8}$$

Lemma 4 (Empirical prediction loss)**.** *With probability at least* $1 - \delta/3$*,*

$$\tilde{\mathcal{L}}(\theta) \le 2\overline{W}_N(\theta) + \frac{8Hu}{N} \qquad \textit{for every } \theta \in \Theta.$$

Proof. For every θ, we have $0 \le W_n(\theta) \le 3H$ and $W_n(\theta)^2 \le 3H\, W_n(\theta)$. Let $\mu_\theta = \mathbb{E}W_1(\theta)$. A lower-tail Bernstein inequality gives, with probability at least $1 - e^{-u}$,

$$\mu_\theta - \overline{W}_N(\theta) \le \sqrt{\frac{6H\mu_\theta u}{N}} + \frac{Hu}{N} \le \frac{\mu_\theta}{2} + \frac{4Hu}{N}.$$

Rearranging and taking a union bound over Θ proves the result. □

A.2.3 The realizable model

Set $M := \log(1/\tau)$ and

$$b_K := \frac{B_f^2}{K}\sum_{h=0}^{H-1}\left(\frac{L_{T,h}^2}{\tau} + L_{r,h}^2\right).$$

The exact representation $m_h^{\star,\rho}$, every $\hat{m}_{n,h}^\theta$, and every $\bar{m}_h^{\theta,\rho}$ lie in $\mathcal{B}_f$, since they are averages of vectors with norm at most B_f. For $\theta^\star$, realizability and Lemma 1 therefore imply that $P_{n,h}$ and $Q_{n,h}^{\theta^\star}$ have the same support. On this support, all positive probabilities are at least τ, so

$$\left|\log\frac{P_{n,h}(x')}{Q_{n,h}^{\theta^\star}(x')}\right| \le M. \tag{9}$$

In particular, the true model has finite empirical risk almost surely.

Lemma 5 (Realizable score)**.** *There is a numerical constant* C_0 *such that, with probability at least* $1 - \delta/3$*,*

$$\overline{A}_N(\theta^\star) \le C_0\left(b_K + \frac{H(1+M)u}{N}\right).$$

Proof. Conditioning on the logged state-action pair and the population snapshot gives

$$\mathbb{E}A_1(\theta^\star) = \sum_{h=0}^{H-1}\mathbb{E}\left[D_{\mathrm{KL}}(P_{1,h}\|Q_{1,h}^{\theta^\star}) + (e_{1,h}^{\theta^\star})^2\right].$$

For a generic population snapshot, conditional on ρ,

$$\mathbb{E}\left[\|\hat{m}_h^{\theta,\rho} - \bar{m}_h^{\theta,\rho}\|_2^2 \middle| \rho\right] = \frac{1}{K} \operatorname{tr} \operatorname{Cov}_{\nu_h^{\star,\rho}}(f_h^\theta(X,A)) \le \frac{B_f^2}{K}.$$

For the realizable model, $\bar{m}_h^{\theta^\star,\rho} = m_h^{\star,\rho}$. By Lemmas 1 and 2, the total-variation Lipschitz condition, and reward Lipschitzness,

$$\mathbb{E}A_1(\theta^\star) \le \frac{B_f^2}{K} \sum_{h=0}^{H-1} \left(\frac{4L_{T,h}^2}{\tau} + L_{r,h}^2 \right) \le 4b_K.$$

We next record a log-moment bound. If p, q have common support and $|\log(p/q)| \le M$, then

$$\mathbb{E}_{X\sim p}\left[\left(\log \frac{p(X)}{q(X)} \right)^2 \right] \le 16(1+M) D_{\mathrm{KL}}(p\|q). \tag{10}$$

Indeed, write $x = p(y)/q(y)$. For $0 < x \le e^M$, we have $x(\log x)^2/(x\log x - x + 1) \le 16(1+M)$, where the ratio is interpreted continuously at $x = 1$. To see this, put $x = e^t$. The ratio is at most 2 when $t \le 0$, at most $2e^2$ when $0 \le t \le 2$, and at most $2t$ when $t \ge 2$. Multiplying by $q(y)(x\log x - x + 1)$, summing over y, and using $D_{\mathrm{KL}}(p\|q) = \sum_y q(y)\big(x\log x - x + 1\big)$ proves (10).

Using (9), (10), $0 \le (e_{1,h}^{\theta^\star})^2 \le 1$, and $(\sum_h z_h)^2 \le H\sum_h z_h^2$, we obtain $\mathbb{E}A_1(\theta^\star)^2 \le 128H(1+M)b_K$. Also, $|A_1(\theta^\star)| \le H(1+M)$. Bernstein's inequality across the N independent episodes, followed by Young's inequality, now gives

$$\overline{A}_N(\theta^\star) \le C_0 \left(b_K + \frac{H(1+M)u}{N} \right)$$

with probability at least $1 - e^{-u}$, which is at least $1 - \delta/3$. □

A.2.4 Finite-population error

Lemma 6 (Finite-population error). *For every $\theta \in \Theta$,*

$$\mathcal{L}(\theta) \le 2\tilde{\mathcal{L}}(\theta) + 2b_K.$$

Proof. Fix h, ρ, x, a, and let $\hat{m}_h^{\theta,\rho}$ denote a generic size-K snapshot average. Write $P = T_h^\star(\cdot \mid x, a, m_h^{\star,\rho})$, $Q = T_h^\theta(\cdot \mid x, a, \hat{m}_h^{\theta,\rho})$, and $\bar{Q} = T_h^\theta(\cdot \mid x, a, \bar{m}_h^{\theta,\rho})$. Since Hellinger distance is a metric, $2D_{\mathrm{H}}^2(P,\bar{Q}) \le 4D_{\mathrm{H}}^2(P,Q) + 4D_{\mathrm{H}}^2(Q,\bar{Q})$. The two candidate representations lie in $\mathcal{B}_f$, so Lemma 1 gives Q and $\bar{Q}$ the same support. Lemma 2 and Assumption 3 then give

$$D_{\mathrm{H}}^2(Q,\bar{Q}) \le \frac{D_{\mathrm{TV}}(Q,\bar{Q})^2}{2\tau} \le \frac{L_{T,h}^2}{2\tau} \|\hat{m}_h^{\theta,\rho} - \bar{m}_h^{\theta,\rho}\|_2^2.$$

For rewards,

$$\begin{aligned} &\left(r_h^\star(x,a,m_h^{\star,\rho}) - r_h^\theta(x,a,\bar{m}_h^{\theta,\rho}) \right)^2 \\ &\quad \le 2\left(r_h^\star(x,a,m_h^{\star,\rho}) - r_h^\theta(x,a,\hat{m}_h^{\theta,\rho}) \right)^2 + 2L_{r,h}^2 \|\hat{m}_h^{\theta,\rho} - \bar{m}_h^{\theta,\rho}\|_2^2. \end{aligned}$$

Averaging over the snapshot, the logged state-action pair, and ρ, then summing over h, proves the claim using $\mathbb{E}[\|\hat{m}_h^{\theta,\rho} - \bar{m}_h^{\theta,\rho}\|_2^2 \mid \rho] \le B_f^2/K$. □

A.2.5 Confidence theorem

Theorem 3 (Finite-class confidence event). *Assume $|\Theta| < \infty$ and Assumptions 1–3. Fix $\delta \in (0,1)$, and let u, M, b_K be as defined above. For numerical constants C_β, C_s, set*

$$\beta \coloneqq C_\beta \left(b_K + \frac{H(1+M)u}{N} \right), \qquad s_{N,K} \coloneqq C_s \left(b_K + \frac{H(1+M)u}{N} \right).$$

For sufficiently large C_β, C_s, with probability at least $1-\delta$,

$$\theta^\star \in \mathcal{C}(\beta), \qquad \mathcal{C}(\beta) \subseteq \Theta_{s_{N,K}}.$$

In particular,

$$\beta,\ s_{N,K} = O\left(\frac{H\big(1+\log(1/\tau)\big)\log(|\Theta|/\delta)}{N} + \frac{B_f^2}{K}\sum_{h=0}^{H-1}\left[\frac{L_{T,h}^2}{\tau} + L_{r,h}^2\right]\right).$$

Proof. Intersect the events in (8) and Lemmas 4 and 5. Their intersection has probability at least $1-\delta$.

First, (8) and $\overline{W}_N(\theta) \geq 0$ imply $\inf_{\theta\in\Theta} \overline{A}_N(\theta) \geq -2u/N$. Let $r_{N,K} \coloneqq b_K + H(1+M)u/N$. The decomposition of $\hat{\mathcal{R}}$ and Lemma 5 give

$$\begin{aligned}\hat{\mathcal{R}}(\theta^\star) - \inf_{\theta\in\Theta}\hat{\mathcal{R}}(\theta) &= \overline{A}_N(\theta^\star) - \inf_{\theta\in\Theta}\overline{A}_N(\theta) \\ &\leq C_0 r_{N,K} + \frac{2u}{N} \leq (C_0+2)r_{N,K}.\end{aligned}$$

Thus $\theta^\star \in \mathcal{C}(\beta)$ when $C_\beta \geq C_0 + 2$.

Now fix $\theta \in \mathcal{C}(\beta)$. Since $\inf_{\theta'}\hat{\mathcal{R}}(\theta') \leq \hat{\mathcal{R}}(\theta^\star)$, we have $\overline{A}_N(\theta) \leq \overline{A}_N(\theta^\star) + \beta$. Using (8) once more, $\overline{W}_N(\theta) \leq \overline{A}_N(\theta) + 2u/N \leq (C_0+C_\beta+2)r_{N,K}$. Lemma 4 then yields $\tilde{\mathcal{L}}(\theta) \leq 2(C_0+C_\beta+2)r_{N,K} + 8Hu/N \leq 2(C_0+C_\beta+6)r_{N,K}$. Finally, Lemma 6 gives $\mathcal{L}(\theta) \leq [4(C_0+C_\beta+6)+2]r_{N,K}$. Choosing C_s at least this large proves $\theta \in \Theta_{s_{N,K}}$. The order bound follows from the definitions of u, M, b_K. □

A.3 Interval gaps and deterministic oracle inequalities

For a nonempty set $\mathcal{C} \subseteq \Theta$, define

$$\overline{J}_{\mathcal{C}}(\pi;\eta) \coloneqq \sup_{\theta\in\mathcal{C}} J_\theta(\pi;\eta), \qquad \underline{J}_{\mathcal{C}}(\pi;\eta) \coloneqq \inf_{\theta\in\mathcal{C}} J_\theta(\pi;\eta),$$

and

$$\widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta) \coloneqq \sup_{\pi\in\Pi} \overline{J}_{\mathcal{C}}(\pi;\eta) - \underline{J}_{\mathcal{C}}(\eta;\eta).$$

When $\mathcal{C} = \mathcal{C}(\beta)$, these agree with the main-text quantities $\overline{J}_\beta$, $\underline{J}_\beta$, and $\widehat{\mathrm{Gap}}_\beta$. For $\theta^\star \in \mathcal{C}$, define the one-sided interval errors

$$r_{\mathcal{C}}^+(\pi;\eta) \coloneqq \overline{J}_{\mathcal{C}}(\pi;\eta) - J(\pi;\eta), \qquad r_{\mathcal{C}}^-(\pi;\eta) \coloneqq J(\pi;\eta) - \underline{J}_{\mathcal{C}}(\pi;\eta).$$

They are nonnegative. Define also

$$\mathrm{subopt}^+_{\mathcal{C},\eta}(\tilde{\pi}) \coloneqq \sup_{\pi\in\Pi} \overline{J}_{\mathcal{C}}(\pi;\eta) - \overline{J}_{\mathcal{C}}(\tilde{\pi};\eta).$$

When $\mathcal{C} = \mathcal{C}(\beta)$, this is $\mathrm{subopt}^+_{\beta,\eta}$.

Lemma 7 (Interval objective upper-bounds the true gap)**.** *If $\theta^\star \in \mathcal{C}$, then for every $\eta \in \Pi$,*

$$\mathrm{Gap}(\eta) \leq \widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta).$$

Proof. The claim follows from $J(\pi;\eta) = J_{\theta^\star}(\pi;\eta) \leq \overline{J}_{\mathcal{C}}(\pi;\eta)$ and $J(\eta;\eta) = J_{\theta^\star}(\eta;\eta) \geq \underline{J}_{\mathcal{C}}(\eta;\eta)$. Taking the supremum over π gives $\mathrm{Gap}(\eta) \leq \widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta)$. □

Theorem 4 (Deterministic interval oracle inequality)**.** *Let $\mathcal{C} \subseteq \Theta$ satisfy $\theta^\star \in \mathcal{C}$. Let $\hat{\eta} \in \Pi$ obey*

$$\widehat{\mathrm{Gap}}_{\mathcal{C}}(\hat{\eta}) \leq \inf_{\eta\in\Pi} \widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta) + \varepsilon_{\mathrm{opt}}.$$

Then, for every comparator $\eta \in \Pi$,

$$\mathrm{Gap}(\hat{\eta}) \leq \varepsilon_{\mathrm{opt}} + \mathrm{Gap}(\eta) + r_{\mathcal{C}}^-(\eta;\eta) + \inf_{\tilde{\pi}\in\Pi}\left\{ r_{\mathcal{C}}^+(\tilde{\pi};\eta) + \mathrm{subopt}^+_{\mathcal{C},\eta}(\tilde{\pi}) \right\}.$$

Proof. By Lemma 7 and approximate optimality,

$$\mathrm{Gap}(\hat{\eta}) \le \widehat{\mathrm{Gap}}_{\mathcal{C}}(\hat{\eta}) \le \widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta) + \varepsilon_{\mathrm{opt}}.$$

Fix $\tilde{\pi} \in \Pi$. By definition of $\mathrm{subopt}^{+}_{\mathcal{C},\eta}$,

$$\widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta) = \overline{J}_{\mathcal{C}}(\tilde{\pi};\eta) - \underline{J}_{\mathcal{C}}(\eta;\eta) + \mathrm{subopt}^{+}_{\mathcal{C},\eta}(\tilde{\pi}).$$

Using the one-sided errors, $\overline{J}_{\mathcal{C}}(\tilde{\pi};\eta) \le J(\tilde{\pi};\eta) + r^{+}_{\mathcal{C}}(\tilde{\pi};\eta)$ and $-\underline{J}_{\mathcal{C}}(\eta;\eta) \le -J(\eta;\eta) + r^{-}_{\mathcal{C}}(\eta;\eta)$. Thus

$$\widehat{\mathrm{Gap}}_{\mathcal{C}}(\eta) \le J(\tilde{\pi};\eta) - J(\eta;\eta) + r^{+}_{\mathcal{C}}(\tilde{\pi};\eta) + r^{-}_{\mathcal{C}}(\eta;\eta) + \mathrm{subopt}^{+}_{\mathcal{C},\eta}(\tilde{\pi}).$$

Since $J(\tilde{\pi};\eta) - J(\eta;\eta) \le \mathrm{Gap}(\eta)$, taking the infimum over $\tilde{\pi}$ completes the proof. □

A.4 Bellman residuals and value-transfer radii

We use the suffix-indexed candidate values, Bellman residuals, lifted measures, and associated norms defined in Section 4. Write $\gamma_h := H - h - 1$.

Lemma 8 (Bellman-residual value identity)**.** *For every $\theta \in \Theta$ and every $\pi, \eta \in \Pi$,*

$$J_\theta(\pi;\eta) - J(\pi;\eta) = \int \mathfrak{b}^{\theta,\pi}\, d\mathsf{D}^{\theta,\pi;\eta}_{\mathrm{tar}}.$$

Proof. Fix θ, π, η, and write $z_{h:H-1} = m^{\theta,\eta}_{h:H-1}$. The candidate value can be written as

$$J_\theta(\pi;\eta) = \mathbb{E}_{X_0\sim\mu_0} v^{\theta,\pi}_0(X_0, z_{0:H-1}).$$

Under the true law $\mathbb{P}^{\pi;\eta}_\star$, the population path is $m^{\star,\eta}$. Since $v^{\theta,\pi}_H \equiv 0$, telescoping gives

$$J_\theta(\pi;\eta) - J(\pi;\eta) = \sum_{h=0}^{H-1} \mathbb{E}^{\pi;\eta}_\star\Big[v^{\theta,\pi}_h(X_h, z_{h:H-1}) - r^\star_h(X_h, A_h, m^{\star,\eta}_h) - v^{\theta,\pi}_{h+1}(X_{h+1}, z_{h+1:H-1})\Big].$$

Because $A_h \sim \pi_h(\cdot \mid X_h)$,

$$\mathbb{E}^{\pi;\eta}_\star\Big[v^{\theta,\pi}_h(X_h, z_{h:H-1})\Big] = \mathbb{E}^{\pi;\eta}_\star\Big[q^{\theta,\pi}_h(X_h, A_h, z_{h:H-1})\Big].$$

Conditioning the last term on (X_h, A_h) and using $X_{h+1} \sim T^\star_h(\cdot \mid X_h, A_h, m^{\star,\eta}_h)$ gives exactly

$$\sum_{h=0}^{H-1} \mathbb{E}_{(X,A)\sim\nu^{\star,\pi;\eta}_h}\left[\mathfrak{b}^{\theta,\pi}_h(X, A; m^{\star,\eta}_h, m^{\theta,\eta}_{h:H-1})\right],$$

which is the displayed integral. □

Lemma 9 (Logged Bellman residual is controlled by prediction loss)**.** *For every $\theta \in \Theta$ and every $\pi \in \Pi$,*

$$\|\mathfrak{b}^{\theta,\pi}\|^2_{2,\mathsf{D}^\theta_{\log}} \le a_H \mathcal{L}(\theta).$$

Proof. Fix h, ρ, x, a, and write $z_{h:H-1} = \bar{m}^{\theta,\rho}_{h:H-1}$, $m = m^{\star,\rho}_h$. Since rewards lie in $[0,1]$, backward induction gives $0 \le v^{\theta,\pi}_{h+1}(x', z_{h+1:H-1}) \le \gamma_h$ for all x'. Thus, with $P = T^\star_h(\cdot \mid x, a, m)$ and $Q = T^\theta_h(\cdot \mid x, a, z_h)$,

$$\left|\sum_{x'}(Q(x') - P(x'))v^{\theta,\pi}_{h+1}(x', z_{h+1:H-1})\right| \le \gamma_h D_{\mathrm{TV}}(P, Q).$$

Therefore,

$$\left|\mathfrak{b}^{\theta,\pi}_h(x, a; m, z_{h:H-1})\right| \le \left|r^\theta_h(x, a, z_h) - r^\star_h(x, a, m)\right| + \gamma_h D_{\mathrm{TV}}(P, Q).$$

By Cauchy–Schwarz in $\mathbb{R}^2$ and Lemma 2,

$$\begin{aligned}\mathfrak{b}_h^{\theta,\pi}(x,a;m,z_{h:H-1})^2 &\le (1+\gamma_h^2)\left[\left(r_h^\theta(x,a,z_h)-r_h^\star(x,a,m)\right)^2 + D_{\mathrm{TV}}(P,Q)^2\right]\\ &\le (1+\gamma_h^2)\left[\left(r_h^\theta(x,a,z_h)-r_h^\star(x,a,m)\right)^2 + 2D_{\mathrm{H}}^2(P,Q)\right].\end{aligned}$$

Since $1+\gamma_h^2 \le a_H$, integrating this display over $\rho\sim\mathfrak{p}$, $(x,a)\sim\nu_h^{\star,\rho}$, and summing over h gives the result by the definition of $\mathcal{L}(\theta)$. □

For $s\ge 0$, define the exact Bellman-transfer radius

$$\mathrm{rad}_{\mathrm{Bell},s}(\pi;\eta) \coloneqq \sup_{\theta\in\Theta_s}\left|\int \mathfrak{b}^{\theta,\pi}\, d\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}\right|,$$

with the convention that the supremum of an empty set is zero. For $s>0$, this radius satisfies $\mathrm{rad}_{\mathrm{Bell},s}(\pi;\eta)\le R_s(\pi;\eta)$, where R_s is defined in Section 4. If $\kappa(\pi;\eta;s)=+\infty$, the claim is immediate. Otherwise, the zero-denominator conventions imply that a zero logged residual must also have zero target integral. Hence, for every $\theta\in\Theta_s$, the definition of κ and Lemma 9 give

$$\left(\int \mathfrak{b}^{\theta,\pi}\, d\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}\right)^2 \le \kappa(\pi;\eta;s)\|\mathfrak{b}^{\theta,\pi}\|_{2,\mathsf{D}_{\log}^\theta}^2 \le a_H s\,\kappa(\pi;\eta;s).$$

Lemma 10 (One-sided interval control by Bellman-transfer radii)**.** *Suppose $\theta^\star\in\mathcal{C}\subseteq\Theta_s$. Then, for every $\pi,\eta\in\Pi$,*

$$r_{\mathcal{C}}^+(\pi;\eta)\le \mathrm{rad}_{\mathrm{Bell},s}(\pi;\eta), \qquad r_{\mathcal{C}}^-(\pi;\eta)\le \mathrm{rad}_{\mathrm{Bell},s}(\pi;\eta).$$

Consequently, for $s>0$,

$$r_{\mathcal{C}}^+(\pi;\eta)\le R_s(\pi;\eta), \qquad r_{\mathcal{C}}^-(\pi;\eta)\le R_s(\pi;\eta).$$

Proof. For the upper side,

$$\begin{aligned}r_{\mathcal{C}}^+(\pi;\eta) &= \sup_{\theta\in\mathcal{C}}\{J_\theta(\pi;\eta)-J(\pi;\eta)\}\\ &\le \sup_{\theta\in\mathcal{C}}|J_\theta(\pi;\eta)-J(\pi;\eta)|\\ &= \sup_{\theta\in\mathcal{C}}\left|\int \mathfrak{b}^{\theta,\pi}\, d\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}\right| \le \mathrm{rad}_{\mathrm{Bell},s}(\pi;\eta),\end{aligned}$$

where the equality is Lemma 8. The lower side is identical. The coefficient form follows from $\mathrm{rad}_{\mathrm{Bell},s}\le R_s$. □

A.5 Proof of the adaptive Nash-gap theorem and rate corollary

Proof of Theorem 2. Apply Theorem 4 with $\mathcal{C}=\mathcal{C}(\beta)$. Since $\theta^\star\in\mathcal{C}(\beta)\subseteq\Theta_s$, Lemma 10 gives $r_{\mathcal{C}(\beta)}^-(\eta;\eta)\le R_s(\eta;\eta)$. For every $\tilde\pi\in\Pi$, it also gives $r_{\mathcal{C}(\beta)}^+(\tilde\pi;\eta)\le R_s(\tilde\pi;\eta)$. Substituting these two inequalities into the deterministic oracle inequality proves

$$\mathrm{Gap}(\hat\eta)\le \varepsilon_{\mathrm{opt}}+\mathrm{Gap}(\eta)+R_s(\eta;\eta)+\inf_{\tilde\pi\in\Pi}\left\{R_s(\tilde\pi;\eta)+\mathrm{subopt}_{\beta,\eta}^+(\tilde\pi)\right\}.$$

□

Proof of Corollary 1. On the event of Theorem 3, $\theta^\star\in\mathcal{C}(\beta)\subseteq\Theta_{s_{N,K}}$. Fix $\eta\in\Pi$. Each candidate model induces a finite-horizon MDP against η and hence has a deterministic optimal Markov policy. Since $\mathcal{C}(\beta)$ is finite, an optimistic best response $\pi_\eta^+\in\arg\max_{\pi\in\Pi}\overline{J}_\beta(\pi;\eta)$ therefore exists and satisfies $\mathrm{subopt}_{\beta,\eta}^+(\pi_\eta^+)=0$. Apply Theorem 2 with this response and bound its transfer coefficient by $\sup_{\pi\in\Pi}\kappa(\pi;\eta;s_{N,K})$. Taking the infimum over comparison policies yields

$$\mathrm{Gap}(\hat\eta)\le \varepsilon_{\mathrm{opt}}+\inf_{\eta\in\Pi}\left\{\mathrm{Gap}(\eta)+\sqrt{a_H s_{N,K}}\left(\sqrt{\kappa(\eta;\eta;s_{N,K})}+\sup_{\pi\in\Pi}\sqrt{\kappa(\pi;\eta;s_{N,K})}\right)\right\}. \quad (11)$$

For $\eta_\star\in\Pi_\star$, choose $\eta=\eta_\star$ in the last bound. Its Nash gap is zero, and both square-root coefficients are at most $\sqrt{\kappa_{\mathrm{uni}}(\eta_\star;s_{N,K})}$, giving

$$\mathrm{Gap}(\hat\eta)\le \varepsilon_{\mathrm{opt}}+2\sqrt{a_H s_{N,K}}\sqrt{\kappa_{\mathrm{uni}}(\eta_\star;s_{N,K})}.$$

□

A.6 Scalar and L^2 Bellman coverage

We prove the comparison between the coefficients κ and $\mathscr{C}$ defined in Section 4.

Proposition 3 (L^2 Bellman coverage controls scalar Bellman transfer)**.** *For every $\pi, \eta \in \Pi$ and every $s \geq 0$,*

$$\kappa(\pi; \eta; s) \leq H\mathscr{C}(\pi; \eta; s).$$

Proof. Fix $\theta \in \Theta_s$. Since $\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}$ has total mass H, the Cauchy–Schwarz inequality gives

$$\left(\int \mathfrak{b}^{\theta,\pi}\, d\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}\right)^2 \leq H\|\mathfrak{b}^{\theta,\pi}\|^2_{2,\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta}}.$$

Dividing by $\|\mathfrak{b}^{\theta,\pi}\|^2_{2,\mathsf{D}_{\log}^{\theta}}$ and taking the supremum over $\theta \in \Theta_s$ proves the claim, with the stated zero-denominator conventions. □

Sufficient conditions for finite κ. Suppose that, for every $\theta \in \Theta_s$, $\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta} \ll \mathsf{D}_{\log}^{\theta}$ with density w_θ satisfying $\int w_\theta^2\, \mathrm{d}\mathsf{D}_{\log}^{\theta} \leq C < \infty$. Then Cauchy–Schwarz gives $\kappa(\pi; \eta; s) \leq C$. For example, suppose logging assigns probability $p_\eta := \mathfrak{p}(\{\eta\}) > 0$ to η, the representation paths satisfy $\bar{m}_{0:H-1}^{\theta,\eta} = m_{0:H-1}^{\theta,\eta}$ for every $\theta \in \Theta_s$, and $\nu_h^{\star,\pi;\eta}(x,a) \leq c\nu_h^{\star,\eta}(x,a)$ for all h, x, a, with $c < \infty$. Then $\mathsf{D}_{\mathrm{tar}}^{\theta,\pi;\eta} \leq (c/p_\eta)\mathsf{D}_{\log}^{\theta}$, so $\int w_\theta^2\, \mathrm{d}\mathsf{D}_{\log}^{\theta} \leq cH/p_\eta$ and hence $\kappa(\pi; \eta; s) \leq cH/p_\eta$. The representation equality is automatic for $H = 1$; for longer horizons, it is an extra condition on the candidate models.

B Routing Experiment Details

B.1 Scope and relation to the theory

We study representation learning in a one-step routing game with deterministic reward labels. We fit neural reward models and compute candidate equilibria of the fitted games. Unlike the theoretical analysis, the experiment involves neither transition learning nor a finite candidate class, and planning uses a single fitted model rather than a confidence set. It therefore does not implement Algorithm 1 or test the finite-sample bound of Section 4.

Population contexts are generated using scrambled Sobol sequences. The demand law varies across rows, whereas the theory fixes the initial state law. Conditional on each context, the focal state–action pair and the population snapshot are sampled independently, as in Section 3. We vary the number of rows N and the snapshot size K separately.

B.2 Environment and offline context distribution

The edge set is ordered deterministically: first the 20 origin-to-U edges, then the 16 U-to-V edges, and finally the 20 V-to-destination edges. For $x = (i, j)$ and $a = (u, v)$, the vector $f^\star(x, a)$ has ones at the coordinates of $O_i \to U_u$, $U_u \to V_v$, and $V_v \to D_j$, and zeros elsewhere. Consequently, the edge loads within each of the three layers sum to one for every population law.

Each edge e has a minimum latency $\tau_e > 0$, capacity $c_e > 0$, and positive congestion coefficients α_e, β_e. Its latency at load $z \in [0, 1]$ is

$$\ell_e(z) = \tau_e + \alpha_e \frac{z}{c_e} + \beta_e \left(\frac{z}{c_e}\right)^2. \tag{12}$$

With a small nonnegative fixed route offset $\delta_{x,a}$, the route cost and reward are

$$C(x, a, m) = \delta_{x,a} + \sum_{e=1}^{56} f_e^\star(x, a)\ell_e(m_e), \qquad r^\star(x, a, m) = 1 - \frac{C(x, a, m)}{C_{\max}}. \tag{13}$$

Since each edge load lies in $[0, 1]$, the normalization

$$C_{\max} = \max_{x,a} \left\{ \delta_{x,a} + \sum_{e=1}^{56} f_e^\star(x, a) \left(\tau_e + \frac{\alpha_e}{c_e} + \frac{\beta_e}{c_e^2} \right) \right\} \tag{14}$$

guarantees $r^\star \in [0, 1]$. We draw the asymmetric edge parameters and route offsets once and hold the resulting network fixed across all training seeds and all choices of N and K.

Demand law. A context first generates product-form origin–destination demand. Let $g^O, g^D \in \mathbb{R}^5$ have independent standard-normal coordinates, and let $T_O, T_D > 0$ be independent temperatures. With a fixed uniform floor ε_μ, set

$$q_i^O = (1 - \varepsilon_\mu) \frac{\exp(g_i^O / T_O)}{\sum_{i'} \exp(g_{i'}^O / T_O)} + \frac{\varepsilon_\mu}{5}, \tag{15}$$

$$q_j^D = (1 - \varepsilon_\mu) \frac{\exp(g_j^D / T_D)}{\sum_{j'} \exp(g_{j'}^D / T_D)} + \frac{\varepsilon_\mu}{5}, \tag{16}$$

and $\mu_c(i, j) = q_i^O q_j^D$. The product form is deliberate: the raw population mean identifies the complete demand law through its origin and destination marginals.

Behavior policy. The behavior population uses structured route preferences. Draw independent standard-normal arrays $\xi_{i,u}^{OU}$, $\xi_{u,v}^{UV}$, and $\xi_{v,j}^{VD}$, and a route-choice temperature $T_\rho > 0$. Define

$$s_{i,j,u,v} = -\kappa_{\text{ff}} \big(\tau_{i,u} + \tau_{u,v} + \tau_{v,j} \big) + \sigma_{OU} \xi_{i,u}^{OU} + \sigma_{UV} \xi_{u,v}^{UV} + \sigma_{VD} \xi_{v,j}^{VD}, \tag{17}$$

and, with exploration floor ε_ρ,

$$\rho_c(u, v \mid i, j) = (1 - \varepsilon_\rho) \frac{\exp(s_{i,j,u,v} / T_\rho)}{\sum_{u',v'} \exp(s_{i,j,u',v'} / T_\rho)} + \frac{\varepsilon_\rho}{16}. \tag{18}$$

The temperatures are drawn from fixed log-uniform ranges. These ranges, the exploration floors, and all scalar coefficients are the same for all models and all choices of N and K.

Sampling across dataset and snapshot sizes. For each context, we compute μ_c, ρ_c, $\nu_c = \mu_c \otimes \rho_c$, and $m_c^\star$ analytically. We then draw one focal pair and an independent ordered stream of population samples. We experiment with $N \in \{1000, 5000, 25000, 100000, 200000\}$ and $K \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. Smaller N-datasets are prefixes of a common row stream, and smaller K-snapshots are prefixes of the same ordered population stream within each row. Training, validation, and final evaluation use disjoint random-number streams.

B.3 Reward-model architectures and fitting

The common focal encoder uses the 26-dimensional encoding $z(x, a)$ and has shape $26 \to 64 \to 64 \to 32$, with SiLU nonlinearities. The learned model's encoder for each population sample and the monolithic population network have shape $26 \to 64 \to 64 \to 56$. The oracle adapter has shape $56 \to 64 \to 64 \to 56$. After concatenating the focal and population embeddings, the common reward head has shape $88 \to 128 \to 64 \to 1$, followed by a sigmoid. The single-agent model replaces the population branch with a second encoder of the focal state–action pair, matching the learned model's parameter count.

The full-population-law branch receives the 400-entry empirical histogram and has shape $400 \to 20 \to 20 \to 56$. With these widths, the complete full-law model has 37297 parameters, compared with 37209 for the learned mean-field model, a difference of 0.237%. It is trained independently for each (N, K) pair.

We train all models by minimizing mean squared reward error with AdamW, using learning rate 10^{-3} and weight decay 10^{-5}. The effective batch size is 256 rows, and the gradient norm is clipped at one. Training runs for at most 40,000 updates, with validation on 5,000 rows every 250 updates.

We select the checkpoint with the lowest validation MSE. Validation uses the same form of input as training: no population input for the single-agent model, finite snapshots for the monolithic, learned,

finite-K-oracle, and full-law models, and the analytic representation for the infinite-population oracle. Final population RMSE and true Nash gap are not available to the trainer or checkpoint selector.

For each N and seed, we train the single-agent and infinite-population-oracle models once and reuse their checkpoints across all K. The other four models are trained separately for each (N, K) pair and seed.

B.4 Exact reward evaluation

For an evaluation context c, all 400 focal categories are enumerated. The exact population inputs are

$$\overline{m}_c^{\text{learned}} = \sum_{x,a} \nu_c(x,a) f^\theta(x,a), \tag{19}$$

$$\overline{m}_c^{\text{mono}} = g^\theta\left(\sum_{x,a} \nu_c(x,a) z(x,a)\right), \tag{20}$$

$$\overline{m}_c^{\text{oracle-}K} = \overline{m}_c^{\text{oracle-}\infty} = m_c^\star, \tag{21}$$

$$\overline{m}_c^{\text{full}} = h^\theta(\nu_c), \tag{22}$$

where g^θ is the monolithic population network and h^θ is the full-law population network. The single-agent model receives no population input. Thus the learned model is evaluated using the expected encoding of a population sample under ν_c, while the full-law model is evaluated at the exact 400-entry law. Only the two oracle models receive the true edge-load vector $m_c^\star$ as an input.

The inner sum in (6) is exact and accumulated in float64. The outer expectation is approximated using 131072 contexts in eight independently scrambled Sobol blocks shared across models and seeds.

B.5 Fitted-game planning and exact controls

For each of 512 target demand laws μ and each fitted reward model, a candidate policy η is evaluated using that model's exact population input under $\mu \otimes \eta$. Let $\hat{r}_\theta^\eta(x,a)$ denote the resulting reward table. The fitted-game Nash gap is

$$\widehat{\text{Gap}}_{\mu,\theta}(\eta) = \sum_x \mu(x) \left[\max_a \hat{r}_\theta^\eta(x,a) - \sum_a \eta(a \mid x) \hat{r}_\theta^\eta(x,a)\right]. \tag{23}$$

This quantity can be computed without consulting the true environment.

For every fitted reward model, we compute 12 candidate policies using the following optimization procedures. Entropic mirror descent and entropic Mirror–Prox each start from the uniform policy and use initial step sizes in $\{5, 20, 30\}$, producing six candidates. The other six candidates minimize a smoothed Nash-gap objective with Adam, using learning rates in $\{0.03, 0.1\}$, temperatures $0.1, 0.03, 0.01$ in sequence, and three fixed initial policies. Each candidate optimization uses at most 5000 iterations, checks the fitted-game Nash gap every 10 iterations, and stops early when that gap is at most 10^{-5}. The entropic mirror-descent update is

$$\tilde{\eta}_{t+1}(a \mid x) = \frac{\eta_t(a \mid x) \exp(\gamma_t \hat{r}_\theta^{\eta_t}(x,a))}{\sum_b \eta_t(b \mid x) \exp(\gamma_t \hat{r}_\theta^{\eta_t}(x,b))}. \tag{24}$$

For each checkpoint and target demand, the returned candidate is the one with the smallest fitted-game Nash gap (23). The true reward is never used for candidate selection. The selected policy is then evaluated under the true reward via (7). We average the fitted-game Nash gaps of the selected policies over the 512 target demand laws. This average was below 1.4×10^{-5} for every plotted model, (N, K) pair, and training seed.

The *uniform policy* is $\eta_{\text{unif}}(a \mid x) = 1/16$. The *exact-reward control* computes 12 candidate policies using the same optimization procedures after replacing the learned reward by the analytic reward in (13). This control measures the error remaining when the same planning procedure is applied to the true reward. The exact-reward control attains an exact true Nash gap of 9.58×10^{-6}.

B.6 UNCERTAINTY ESTIMATES

We compute uncertainty across five training seeds. Within a seed, models share contexts, focal samples, ordered population streams, nested datasets and snapshots, and the fixed quasi–Monte Carlo context samples used for evaluation. For each seed, we average the population-weighted squared prediction errors over the evaluation contexts and then take the square root to obtain population RMSE. We summarize these seed-level RMSE values across the five training seeds. For each seed, we also average the true Nash gaps of the selected policies over the 512 target demand laws. If L_s is the resulting population RMSE or mean true Nash gap for seed $s \in \{0, \dots, 4\}$, we report

$$\overline{L} \pm t_{0.975,4} \frac{\mathrm{sd}(L_0, \dots, L_4)}{\sqrt{5}}.$$

Thus every "$\pm$" quantity is a 95% Student-t half-width over five paired training seeds.