MEAN-FIELD Q-LEARNING FOR HETEROGENEOUS MULTI-AGENT SYSTEMS WITH CONVERGENCE GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study cooperative reinforcement learning in large populations of heterogeneous agents subject to common noise. A non-exchangeable mean-field approximation preserves agent labels while replacing the joint state by a population law. The resulting control problem has infinite-dimensional state and kernel action spaces. We develop a finite-net Q-learning scheme with guarantees that distinguish policy restriction, discretisation, learning, and finite-population approximation. We prove that restricting relaxed policies to Lipschitz kernels is uniformly consistent, quantify the finite-net bias, and establish almost-sure convergence to the projected Bellman fixed point under a generative mean-field oracle. Under the finite-population approximation conditions, the greedy policy admits a performance bound in the original agent system. Congestion and cybersecurity experiments examine these error mechanisms through projected fixed-point diagnostics, net refinement, and finite-population transfer. At the tested resolutions, they identify state discretisation as the main practical bottleneck.

1 INTRODUCTION

We consider a cooperative MARL problem with N heterogeneous agents, each with state space $\mathcal{X}$. Direct learning on the joint state–action space becomes intractable as N grows. We therefore use a mean-field approximation, replacing the finite system by a control problem on the population law. In the classical exchangeable model, all agents are statistically identical: their dynamics and rewards do not depend on their identity, and the model is unchanged when agents are permuted. The relevant aggregate state is then the empirical state distribution

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{X_t^{i,N}} \in \mathcal{P}(\mathcal{X}).$$

This measure records how many agents are in each state, but not which agent is in which state. Under exchangeability, that loss of identity is harmless. Many systems of interest are not exchangeable. Agents may have different locations, classes, priorities, network positions, or intrinsic parameters, and these fixed features may affect both their own dynamics and how they interact with the population. A standard extension is a finite-type model with label space $I_K = \{1, \dots, K\}$. Agents of the same type remain exchangeable, whereas different types may have different rewards and transition laws. If N_k is the number of agents of type k, one tracks the empirical state distribution within each type,

$$\frac{1}{N_k}\sum_{i:u_i=k}\delta_{X_t^{i,N}} \in \mathcal{P}(\mathcal{X}), \quad k = 1, \dots, K,$$

and represents the population by these K type-specific distributions. This is appropriate when heterogeneity is well described by a small number of meaningful classes, but it is restrictive when heterogeneity varies continuously. In a spatial system, for example, position may affect dynamics and interactions smoothly, so forcing agents into finitely many classes creates artificial boundaries and can discard relevant information. This motivates a continuum of fixed labels. In our setting, each agent i carries a label $u_i \in I = [0, 1]$, which may encode location, priority, graphon position, or

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.

another persistent feature. The finite population is summarised by the empirical label–state law

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{(u_i, X_t^{i,N})} \in \mathcal{P}(I \times \mathcal{X}).$$

This label–state representation follows the formulation introduced by Lacker and Soret (2023) for stochastic graphon games. Unlike the classical empirical state distribution, this object retains the association between labels and states. It therefore preserves the heterogeneity relevant for control without returning to the full N-agent state space. Figure 1 illustrates the distinction between a finite partition of agents into types and a continuum of labels.

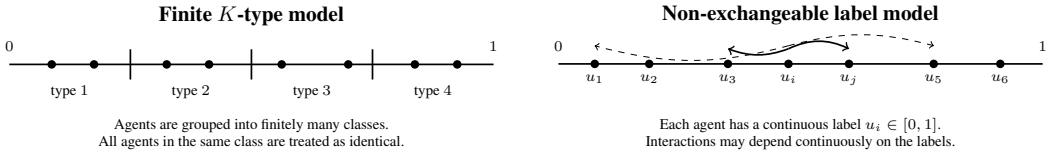

Figure 1: Finite K-type heterogeneity versus continuous non-exchangeable labels.

Contributions. Our contribution is a quantitative approximation and learning analysis for the label–state mean-field MDP. The key difficulty is to approximate its population-law states and relaxed kernel actions while controlling the resulting policy loss. We establish regularity of the integrated Bellman problem and uniform consistency of the Lipschitz kernel restriction (Theorems 6–7). We then bound the finite-net bias and prove almost-sure convergence to the projected fixed point (Theorem 9). Finally, the greedy-policy residual links these approximation errors to finite-agent performance under the finite-population comparison conditions (Corollaries 10–11). The finite-table learning step uses classical asynchronous Q-learning; the analysis quantifies how the measure and kernel approximations enter the control guarantee.

Related literature. Cooperative MARL studies agents that learn to maximise a common return. Early work considered independent learning and coordination (Tan, 1993; Claus and Boutilier, 1998). Centralised training with decentralised execution is now a standard design for cooperative problems; representative actor–critic methods include MADDPG (Lowe et al., 2017) and COMA (Foerster et al., 2018). Value-decomposition methods include VDN (Sunehag et al., 2018), QMIX (Rashid et al., 2018), QTRAN (Son et al., 2019), and QPLEX (Wang et al., 2021), while MAPPO is a widely used on-policy baseline (Yu et al., 2022). SMAC (Samvelyan et al., 2019) and comparative studies such as Papoudakis et al. (2021) provide common empirical benchmarks; broader surveys and perspectives include Zhang et al. (2021); Gronauer and Diepold (2022); Gemp et al. (2022). For large populations, mean-field MARL replaces direct many-agent interactions with a population distribution. Yang et al. (2018) develop mean-field Q-learning and actor–critic algorithms, and Gu et al. (2021; 2023) study cooperative mean-field control and integrated Q-functions with convergence guarantees. Related mean-field game learning methods include scalable online mirror descent (Pérolat et al., 2022). Heterogeneous populations require richer population states. Graphon models encode structured interactions (Hu et al., 2023), while the label–state formulation of Lacker and Soret (2023) keeps persistent agent features in the population law. Mekkaoui and Pham (2026) study non-exchangeable mean-field control with common noise and finite-population approximation, and Mekkaoui et al. (2026) study learning operators on labelled conditional distributions. In the linear–quadratic non-exchangeable setting, De Crescenzo et al. (2026) derive Riccati-based optimal controls. These works provide the control and population-approximation framework on which we build. Our focus is its finite-net learning approximation: the consistency of Lipschitz kernel restriction, the bias introduced by projecting both laws and kernels, and the resulting greedy-policy guarantee. We combine these approximation bounds with asymptotic convergence of the finite-table learner and the finite-population comparison.

2 Problem formulation

We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports all random variables introduced below.

2.1 FINITE-POPULATION MARL

Consider N cooperative heterogeneous agents with joint state–action space $\mathcal{X}^N \times A^N$. Let $I = [0, 1]$ be the label space. For a population of size N, agent i has a fixed deterministic label $u_i^N \in I$. Let $(\mathcal{X}, d)$ and (A, d_A) be compact metric state and action spaces, and let E and E^0 be standard Borel spaces for the idiosyncratic and common noises. Fix a discount factor $\beta \in (0, 1)$. For each N, let

$$f_N : I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \to \mathbb{R}$$

be a bounded Borel reward function and let

$$F_N : I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \times E \times E^0 \to \mathcal{X}$$

be a Borel transition function. Let $\mathcal{F}_t^N$ be the filtration generated by the finite-system idiosyncratic noises $(\epsilon_t^{i,N})_{i,t}$, the common noise $(\epsilon_t^0)_t$, and the randomisers $(Z_t^{i,N})_{i,t}$ up to time t. Here $(\epsilon_t^{i,N})_{i,t}$ are mutually independent with law ϱ, $(\epsilon_t^0)_t$ is an i.i.d. sequence with law ϱ^0, and the two families are independent. We also allow independent randomisers $(Z_t^{i,N})_{i,t}$ so that randomised policies can be represented explicitly. An admissible joint control is an A^N-valued process $\alpha^N = (\alpha_t^N)_{t\in\mathbb{N}}$ adapted to $(\mathcal{F}_t^N)_{t\in\mathbb{N}}$. We denote the admissible set by

$$\mathcal{A}_N := \left\{ \alpha^N = (\alpha_t^{i,N})_{1\le i\le N,\, t\in\mathbb{N}} : \alpha_t^{i,N} \in A \text{ is } \mathcal{F}_t^N\text{-measurable for every } i, t \right\}.$$

Let $X_t^N = (X_t^{1,N}, \dots, X_t^{N,N}) \in \mathcal{X}^N$ and $\alpha^N = (\alpha^{1,N}, \dots, \alpha^{N,N}) \in \mathcal{A}_N$. The interaction is summarised by the empirical label–state–action law

$$\nu_t^N := \frac{1}{N} \sum_{j=1}^{N} \delta_{(u_j^N, X_t^{j,N}, \alpha_t^{j,N})} \in \mathcal{P}(I \times \mathcal{X} \times A),$$

at time t. Given an initial state $x = (x^1, \dots, x^N) \in \mathcal{X}^N$, the controlled dynamics are

$$\begin{cases} X_0^{i,N} = x^i, \\ X_{t+1}^{i,N} = F_N\left(u_i^N, X_t^{i,N}, \alpha_t^{i,N}, \nu_t^N, \epsilon_{t+1}^{i,N}, \epsilon_{t+1}^0\right), \quad i = 1, \dots, N, \quad t \in \mathbb{N}. \end{cases}$$

For $\alpha^N \in \mathcal{A}_N$, the cooperative discounted return is

$$V_N^{\alpha^N}(x) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t f_N\left(u_i^N, X_t^{i,N}, \alpha_t^{i,N}, \nu_t^N\right) \right],$$

and the finite-population value function is

$$V_N(x) := \sup_{\alpha^N \in \mathcal{A}_N} V_N^{\alpha^N}(x), \quad x \in \mathcal{X}^N.$$

The objective is to learn a policy with value close to V_N. Because direct tabular learning on $\mathcal{X}^N \times A^N$ scales with N, we learn on the non-exchangeable mean-field approximation and transfer the policy to the finite population.

2.2 RELAXED MEAN-FIELD CONTROL

The mean-field model is specified by the Borel maps

$$f : I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \to \mathbb{R}, \quad F : I \times \mathcal{X} \times A \times \mathcal{P}(I \times \mathcal{X} \times A) \times E \times E^0 \to \mathcal{X}.$$

The maps f and F are the large-population limits of f_N and F_N under the assumptions of Section 5. We use the weak label–state formulation. Fix a label distribution $\lambda \in \mathcal{P}(I)$ and define

$$\mathcal{P}_\lambda(I \times \mathcal{X}) := \{\mu \in \mathcal{P}(I \times \mathcal{X}) : \mathrm{pr}_1 \# \mu = \lambda\},$$

where $\mathrm{pr}_1 \# \mu$ denotes the pushforward of μ under projection onto the label coordinate. Let U be an I-valued random variable with law λ, let $\epsilon = (\epsilon_t)_{t\ge1}$ be an i.i.d. sequence in E with law ϱ, and let $\epsilon^0 = (\epsilon_t^0)_{t\ge1}$ be an i.i.d. sequence in E^0 with law ϱ^0. We assume that U, ϵ, and ϵ^0 are mutually independent. Let Γ be an initial-information random variable with values in a standard Borel space G,

independent of the noise sequences, and let the initial state ξ be $\mathcal{X}$-valued and $\sigma(U,\Gamma)$-measurable. We impose the randomisation hypothesis that $\sigma(\Gamma)$ contains a uniform random variable on $[0,1]$ independent of ξ. This allows randomised controls to be represented by Borel open-loop policies. A control α is admissible if there exists a sequence of Borel maps

$$a_t : I \times G \times E^t \times (E^0)^t \longrightarrow A, \quad t \in \mathbb{N},$$

with a_0 defined on $I \times G$, such that

$$\alpha_t = a_t\left(U, \Gamma, (\epsilon_s)_{1\le s\le t}, (\epsilon^0_s)_{1\le s\le t}\right), \quad t \in \mathbb{N}, \quad \mathbb{P}\text{-a.s.}$$

We denote the set of admissible randomised open-loop controls by $\mathcal{A}^W$. For $t \in \mathbb{N}$, define the common-noise filtration $\mathcal{F}^0_t := \sigma(\epsilon^0_s : 1 \le s \le t)$, with $\mathcal{F}^0_0$ trivial, and write $\mathbb{P}^0_Y$ for a version of the conditional law of a random variable Y given the common noise. For an A-valued control process $\alpha = (\alpha_t)_{t\in\mathbb{N}}$, the limiting state evolves as

$$\begin{cases} X_0 = \xi, \\ X_{t+1} = F\left(U, X_t, \alpha_t, \mathbb{P}^0_{(U,X_t,\alpha_t)}, \epsilon_{t+1}, \epsilon^0_{t+1}\right), \quad t \in \mathbb{N}. \end{cases}$$

For $\alpha \in \mathcal{A}^W$, define

$$V^\alpha_W(\xi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t f\left(U, X_t, \alpha_t, \mathbb{P}^0_{(U,X_t,\alpha_t)}\right)\right], \quad V_W(\xi) := \sup_{\alpha\in\mathcal{A}^W} V^\alpha_W(\xi). \tag{1}$$

Remark 1. The weak formulation avoids the joint-measurability issue caused by an uncountable family of idiosyncratic noises. Under the assumptions of Mekkaoui and Pham (2026), it has the same value as the strong formulation.

We now express the weak control problem as an MDP on conditional laws. We use the product metrics

$$\rho((u,x),(u',x')) := |u-u'| + d(x,x'),$$
$$\rho_A((u,x,a),(u',x',a')) := |u-u'| + d(x,x') + d_A(a,a'),$$

and write $\Delta := \mathrm{diam}(I \times \mathcal{X})$ and $\Delta_A := \mathrm{diam}(I \times \mathcal{X} \times A)$. Every Wasserstein distance below is the 1-Wasserstein distance for the displayed cost. Set

$$\mu_t := \mathbb{P}^0_{(U,X_t)} \in \mathcal{P}_\lambda(I \times \mathcal{X}).$$

Since the label U is time-independent, the label marginal remains fixed at λ, while the conditional state distribution evolves. A relaxed mean-field action is a Borel kernel $\pi \in L^0(I \times \mathcal{X}; \mathcal{P}(A))$. For $\mu \in \mathcal{P}_\lambda(I \times \mathcal{X})$ and $\pi \in L^0(I \times \mathcal{X}; \mathcal{P}(A))$, define the induced label–state–action law

$$\nu := \mu \otimes \pi, \quad \nu(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a) := \mu(\mathrm{d}u, \mathrm{d}x)\pi(u,x)(\mathrm{d}a).$$

For $e^0 \in E^0$, set

$$\overline{f}(\mu,\pi) := \int_{I\times\mathcal{X}\times A} f(u,x,a,\nu)\, \nu(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a),$$
$$\overline{F}(\mu,\pi,e^0) := \Big[(u,x,a,e) \mapsto \big(u, F(u,x,a,\nu,e,e^0)\big)\Big] \# \big(\nu \otimes \varrho\big). \tag{2}$$

Since the first coordinate is unchanged, $\overline{F}(\mu,\pi,e^0) \in \mathcal{P}_\lambda(I \times \mathcal{X})$, and the lifted dynamics are $\mu_{t+1} = \overline{F}(\mu_t, \pi_t, \epsilon^0_{t+1})$.

Assumption 2. The reward is bounded: $\|f\|_\infty < \infty$. There are constants $L_f, L_F \ge 0$ such that, for all $u,u' \in I$, $x,x' \in \mathcal{X}$, $a,a' \in A$, $\nu,\nu' \in \mathcal{P}(I \times \mathcal{X} \times A)$, and $e^0 \in E^0$,

$$|f(u,x,a,\nu) - f(u',x',a',\nu')| \le L_f\Big(\rho((u,x),(u',x')) + d_A(a,a') + \mathcal{W}(\nu,\nu')\Big).$$
$$\mathbb{E}_e\, d\big(F(u,x,a,\nu,e,e^0), F(u',x',a',\nu',e,e^0)\big) \le L_F\Big(\rho((u,x),(u',x')) + d_A(a,a') + \mathcal{W}(\nu,\nu')\Big).$$

The following dynamic programming result follows from Theorem 3.1, together with the Bellman construction in Section 4.1.2 of Mekkaoui and Pham (2026).

Proposition 3 (Dynamic Programming Principle)**.** *Under Assumption 2, problem* (1) *has a bounded value $V^\star : \mathcal{P}_\lambda(I \times \mathcal{X}) \to \mathbb{R}$. It is the unique bounded fixed point of*

$$[\mathcal{T}V](\mu) := \sup_{\pi\in L^0(I\times\mathcal{X};\mathcal{P}(A))} \left\{\overline{f}(\mu,\pi) + \beta\mathbb{E}\big[V(\overline{F}(\mu,\pi,\epsilon^0))\big]\right\}.$$

In particular, if $\mu = \mathbb{P}_{(U,\xi)}$, then $V_W(\xi) = V^\star(\mu)$, and an optimal randomised feedback exists.

2.3 INTEGRATED Q-FUNCTION

Definition 4 (Integrated Q-function)**.** For $(\mu, \pi) \in \mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0(I \times \mathcal{X}; \mathcal{P}(A))$, define

$$Q^\star(\mu, \pi) := \overline{f}(\mu, \pi) + \beta \mathbb{E}\big[V^\star(\overline{F}(\mu, \pi, \epsilon^0))\big]. \tag{3}$$

Theorem 5 (Bellman equation and uniqueness)**.** *The function $Q^\star$ is bounded and*

$$V^\star(\mu) = \sup_{\pi \in L^0(I \times \mathcal{X}; \mathcal{P}(A))} Q^\star(\mu, \pi), \tag{4}$$

$$Q^\star(\mu, \pi) = \overline{f}(\mu, \pi) + \beta \mathbb{E}\left[\sup_{\pi' \in L^0(I \times \mathcal{X}; \mathcal{P}(A))} Q^\star(\overline{F}(\mu, \pi, \epsilon^0), \pi')\right]. \tag{5}$$

It is the unique bounded solution Q of (5) *for which $\mu \mapsto \sup_\pi Q(\mu, \pi)$ is Borel measurable.*

Proof. The proof is in Appendix A. The measurable-supremum condition is needed because the common-noise expectation must be defined. No compactness of the full kernel class is used. □

3 REGULARITY AND LIPSCHITZ POLICY RESTRICTION

To obtain a compact action class for finite-net Q-learning, we restrict the relaxed kernels to a Lipschitz class and quantify the resulting approximation error. For $L \geq 0$, let

$$L^0_L(I \times \mathcal{X}; \mathcal{P}(A)) := \left\{\pi \in L^0(I \times \mathcal{X}; \mathcal{P}(A)) : \mathcal{W}(\pi(u, x), \pi(u', x')) \leq L\rho((u, x), (u', x'))\right\},$$

and equip $L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ with

$$d_\infty(\pi, \pi') := \sup_{(u,x) \in I \times \mathcal{X}} \mathcal{W}(\pi(u, x), \pi'(u, x)).$$

On $\mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$, set

$$D_L((\mu, \pi), (\mu', \pi')) := \mathcal{W}(\mu, \mu') + d_\infty(\pi, \pi').$$

Both $\mathcal{P}_\lambda(I \times \mathcal{X})$ and $(L^0_L(I \times \mathcal{X}; \mathcal{P}(A)), d_\infty)$ are compact. Set

$$K_0 := 1 + 2L_F, \qquad c_0 := 2L_f, \qquad K_L := K_0(1 + L), \qquad c_L := 2L_f(1 + L).$$

For $c, K \geq 0$, define

$$\Omega_{c,K}(r) := c \sum_{t=0}^{\infty} \beta^t \min\{K^t r, \Delta\}. \tag{6}$$

Appendix B gives explicit Hölder pairs (α_0, C_V) and $(\alpha_L, C_{V,L})$ for Ω_{c_0,K_0} and Ω_{c_L,K_L}, respectively.

For $V \in C(\mathcal{P}_\lambda(I \times \mathcal{X}))$, define

$$[\mathcal{T}_L V](\mu) := \max_{\pi \in L^0_L(I \times \mathcal{X}; \mathcal{P}(A))} \left\{\overline{f}(\mu, \pi) + \beta \mathbb{E}[V(\overline{F}(\mu, \pi, \epsilon^0))]\right\}. \tag{7}$$

Theorem 6 (Restricted problem and regularity)**.** *The operator $\mathcal{T}_L : C(\mathcal{P}_\lambda(I \times \mathcal{X})) \to C(\mathcal{P}_\lambda(I \times \mathcal{X}))$ is a β-contraction with unique fixed point V_L. An optimal stationary feedback exists, and*

$$|V_L(\mu) - V_L(\mu')| \leq \Omega_{c_L,K_L}(\mathcal{W}(\mu, \mu')). \tag{8}$$

Define

$$Q_L(\mu, \pi) := \overline{f}(\mu, \pi) + \beta \mathbb{E}[V_L(\overline{F}(\mu, \pi, \epsilon^0))], \qquad \pi \in L^0_L(I \times \mathcal{X}; \mathcal{P}(A)).$$

Then $\max_{\pi \in L^0_L(I \times \mathcal{X}; \mathcal{P}(A))} Q_L(\mu, \pi) = V_L(\mu)$*, and*

$$|Q_L(\mu, \pi) - Q_L(\mu', \pi')| \leq C_{Q,L} D_L((\mu, \pi), (\mu', \pi'))^{\alpha_L}, \qquad C_{Q,L} := c_L \Delta_A^{1-\alpha_L} + \beta C_{V,L} K_L^{\alpha_L}. \tag{9}$$

Theorem 7 (Consistency of the policy restriction)**.** *As $L \to \infty$,*

$$V_L \uparrow V^\star \qquad \textit{uniformly on } \mathcal{P}_\lambda(I \times \mathcal{X}).$$

Hence

$$g_L := \|V^\star - V_L\|_\infty \longrightarrow 0, \tag{10}$$

and, on $\mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$,

$$0 \leq Q^\star - Q_L \leq \beta g_L. \tag{11}$$

The coupling estimates, regularity of the unrestricted problem, and an explicit bound on g_L are given in Appendices C–D. A rate in L requires additional regularity of near-optimal kernels.

Algorithm 1 Finite-net non-exchangeable mean-field Q-learning

1: Initialise a bounded table Q_0 and a visit count $N_0(c) = 0$ for each $c \in \mathcal{R}_\mu \times \mathcal{R}_\pi$.
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: Choose $c_n = (\mu_n, \pi_n)$ using information available before the fresh oracle draw.
4: Draw $\mu'_n = \overline{F}(\mu_n, \pi_n, \epsilon^0_{n+1})$ and set $Y_n = \overline{f}(\mu_n, \pi_n) + \beta \max_{\pi' \in \mathcal{R}_\pi} Q_n(\Pi_\mu \mu'_n, \pi')$.
5: Set $k = N_n(c_n)$ and update $Q_{n+1}(c_n) = (1 - \eta_k)Q_n(c_n) + \eta_k Y_n$; leave all other entries unchanged.
6: Increase only the count of c_n.
7: **end for**

4 Finite-Net Mean-Field Q-Learning

Fix L. Let $\mathcal{R}_\mu \subset \mathcal{P}_\lambda(I \times \mathcal{X})$ and $\mathcal{R}_\pi \subset L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ be finite internal nets of radii ϵ_μ for $\mathcal{W}$ and ϵ_π for d_∞. Let Π_μ and Π_π be Borel nearest-neighbour maps with fixed tie-breaking. Existence follows from compactness; Appendix E gives covering bounds.

Assumption 8 (Generative mean-field oracle)**.** For any $(\mu, \pi) \in \mathcal{R}_\mu \times \mathcal{R}_\pi$, the learner can obtain $\overline{f}(\mu, \pi)$ and an independent draw $\overline{F}(\mu, \pi, \epsilon^0)$ with $\epsilon^0 \sim \varrho^0$.

Define the projected Bellman operator on the finite table by

$$[\mathcal{B}_\mathcal{R} q](\mu, \pi) := \overline{f}(\mu, \pi) + \beta \mathbb{E}\left[\max_{\pi' \in \mathcal{R}_\pi} q(\Pi_\mu \overline{F}(\mu, \pi, \epsilon^0), \pi')\right]. \tag{12}$$

It is a β-contraction and has a unique fixed point $Q_\mathcal{R}$.

For a table q, define its nearest-neighbour extension $[\Gamma q](\mu, \pi) := q(\Pi_\mu \mu, \Pi_\pi \pi)$.

Theorem 9 (Convergence and discretisation error)**.** *Assume every table entry is visited infinitely often almost surely, the oracle draw is conditionally independent with the stated law, and*

$$\eta_k \in [0, 1], \quad \sum_k \eta_k = \infty, \quad \sum_k \eta_k^2 < \infty.$$

Then $Q_n \to Q_\mathcal{R}$ almost surely. Define

$$e_\mathcal{R} := \frac{\beta\big(C_{V,L}\epsilon_\mu^{\alpha_L} + C_{Q,L}\epsilon_\pi^{\alpha_L}\big)}{1 - \beta}, \quad E_\mathcal{R} := e_\mathcal{R} + C_{Q,L}(\epsilon_\mu + \epsilon_\pi)^{\alpha_L}.$$

Then

$$\max_{\mathcal{R}_\mu \times \mathcal{R}_\pi} |Q_\mathcal{R} - Q_L| \le e_\mathcal{R}, \tag{13}$$

$$\|\Gamma Q_\mathcal{R} - Q_L\|_\infty \le E_\mathcal{R}, \tag{14}$$

$$\limsup_{n \to \infty} \|\Gamma Q_n - Q^\star\|_{\mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0_L(I \times \mathcal{X}; \mathcal{P}(A))} \le E_\mathcal{R} + \beta g_L \quad \textit{almost surely}. \tag{15}$$

A deterministic round-robin schedule satisfies the visit condition. The convergence argument follows standard asynchronous Q-learning (Watkins and Dayan, 1992; Jaakkola et al., 1994; Tsitsiklis, 1994). The theorem distinguishes convergence to $Q_\mathcal{R}$ from approximation of Q_L and $Q^\star$: learning removes the table error, while projection and policy restriction leave the stated bias. It does not give a finite-sample learning rate.

Corollary 10 (Greedy policy)**.** *Let*

$$\pi_\mathcal{R}(\mu) \in \arg\max_{\pi \in \mathcal{R}_\pi} Q_\mathcal{R}(\Pi_\mu \mu, \pi)$$

with fixed tie-breaking, and set

$$\eta_\mathcal{R} := 2e_\mathcal{R} + C_{Q,L}(\epsilon_\mu^{\alpha_L} + \epsilon_\pi^{\alpha_L}) + C_{V,L}\epsilon_\mu^{\alpha_L}.$$

Then $\pi_\mathcal{R}$ is Borel and

$$[\mathcal{T}^{\pi_\mathcal{R}} V_L](\mu) \ge V_L(\mu) - \eta_\mathcal{R}, \tag{16}$$

$$V^\star(\mu) - V^{\pi_\mathcal{R}}(\mu) \le g_L + \frac{\eta_\mathcal{R}}{1 - \beta}, \tag{17}$$

$$[\mathcal{T}^{\pi_\mathcal{R}} V^\star](\mu) \ge V^\star(\mu) - (\eta_\mathcal{R} + g_L). \tag{18}$$

5 TRANSFER TO FINITE-POPULATION MARL

For $x = (x^1, \dots, x^N) \in \mathcal{X}^N$ and labels $\mathbf{u}^N = (u_i^N)_{i=1}^N$, let $\iota_N = N^{-1} \sum_{i=1}^N \delta_i$ on $\{1, \dots, N\}$ and choose a coupling $\Gamma_N \in \mathcal{P}(I \times \{1, \dots, N\})$ with first marginal λ and second marginal ι_N. Define the λ-preserving lift

$$\mu_N^\lambda[\mathbf{u}^N, x](\mathrm{d}u, \mathrm{d}x') := \sum_{i=1}^N \Gamma_N(\mathrm{d}u, \{i\})\, \delta_{x^i}(\mathrm{d}x'). \tag{19}$$

The coupling Γ_N is chosen according to the label approximation in the finite-population assumptions (for instance, to couple u_i^N closely to λ). By construction $\mu_N^\lambda[\mathbf{u}^N, x] \in \mathcal{P}_\lambda(I \times \mathcal{X})$. The raw empirical measure $N^{-1} \sum_i \delta_{(u_i^N, x^i)}$ generally has label marginal $N^{-1} \sum_i \delta_{u_i^N}$ rather than λ, and therefore need not belong to $\mathcal{P}_\lambda(I \times \mathcal{X})$.

Assume first that $A = [a_-, a_+] \subset \mathbb{R}$. For a probability measure ϑ on A, let F_ϑ^{-1} be its quantile function. Define a randomised action representation by

$$a(\mu, u, x, r) := F^{-1}_{\pi_\mathcal{R}(\mu)(u,x)}(r), \quad r \sim \mathrm{Unif}[0, 1]. \tag{20}$$

Because the same r is used on both sides,

$$\mathbb{E}_r |a(\mu, u, x, r) - a(\mu, u', x', r)| = \mathcal{W}(\pi_\mathcal{R}(\mu)(u, x), \pi_\mathcal{R}(\mu)(u', x')) \le L\rho((u, x), (u', x')).$$

This is the regularity required for finite-N transfer.

Corollary 11 (Finite-N policy bound)**.** *Assume the finite-population approximation conditions used for the propagation-of-chaos comparison: M_N controls the empirical Wasserstein error, ϵ_N^f the one-step reward mismatch, ϵ_N^F the transition mismatch, and $\gamma \in (0, 1]$ is the regularity exponent entering the value comparison. Let each agent use*

$$\alpha_t^{i,N} = a\big(\mu_N^\lambda[\mathbf{u}^N, X_t^N], u_i^N, X_t^{i,N}, Z_t^i\big), \tag{21}$$

where the Z_t^i are independent uniform random variables on $[0, 1]$. There is a data-dependent constant C such that

$$V_N(x) - V_N^{\pi_\mathcal{R}, N}(x) \le C(1 + L)\left[\eta_\mathcal{R} + g_L + M_N^\gamma + \epsilon_N^f + (\epsilon_N^F)^\gamma\right]. \tag{22}$$

For a general compact action space, choose an action vector whose lifted empirical label–state–action law minimises the Wasserstein distance to the mean-field feedback law. This gives the same approximation order but not the explicit policy (21)*. The comparison follows the quantitative label–state argument of Mekkaoui and Pham (2026); Appendix G gives the proof in the present notation.*

Corollary 11 combines the mean-field verification residual with the finite-population comparison. Its conclusion is conditional on the population-approximation conditions invoked there, in addition to the assumptions used for mean-field learning.

6 NUMERICAL STUDY

The experiments examine three questions suggested by the error decomposition: does the learner approach its projected fixed point, does net refinement improve the policy, and does a frozen policy transfer consistently as the population grows? These diagnostics assess the approximation mechanisms; they do not establish empirical convergence rates or finite-population optimality.

Benchmarks and protocol. The primary method is nearest-neighbour Q-learning (NN), as analysed in Algorithm 1 and Theorem 9. Positive interpolation (HAT) is an exploratory comparator that changes the projected operator and lies outside this theorem. We evaluate both on two models satisfying Assumption 2. The first is a congestion model with a smaller calibration problem whose finite-MDP fixed points can be computed exactly. The second model is a cybersecurity model with continuous labels, heterogeneous interactions, and common noise. We choose parameters for which the optimal policy depends on the current state, so that feedback control is needed. We use value iteration on a fine grid as a reference and evaluate the resulting policy by simulation. Without common noise, we also compare this reference with a Pontryagin solution. Full parameters, training budgets, and evaluation details are given in Appendix I. Calibration uses nets with proven radii; the main experiments use structured kernel subclasses and reachable-set nets with measured radii (Appendix H).

Learning and discretisation. On the calibration problem, both methods approach their own projected fixed points (Figure 2). Increasing the oracle budget makes the learning error small relative to the finite-net error, separating incomplete learning from the bias of the chosen representation.

Refining the state net improves the exact finite-table policies, especially for NN, while refining the action grid or the kernel class has a smaller effect (Figure 3). On this calibration problem, HAT remains above NN across the tested refinements. These results show that the state discretisation has an important effect on the quality of the final policy. The theoretical error bounds are conservative at the tested resolutions, and the holdout certificate covers the observed table error while remaining loose.

On the full congestion model, HAT closely matches its own exact projected fixed point, indicating that the remaining gap mainly comes from the finite representation. Its broad interpolation weights reduce locality. NN is more stable in this setting and remains the main implementation. Neural policy-gradient baselines achieve higher returns at the tested resolutions. Together, these comparisons identify finite representation as a practical limitation even when the projected learning problem is accurately solved.

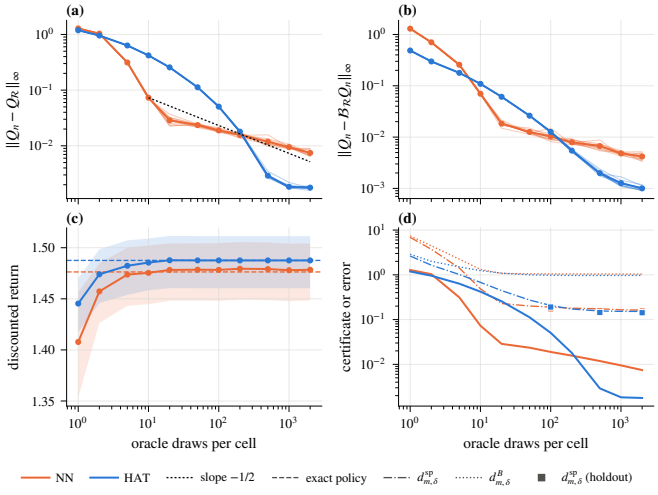

Figure 2: Congestion calibration: error relative to $Q_{\mathcal{R}}$, exact Bellman residual, and test return of learned and exact finite-table policies. Orange: NN; blue: HAT.

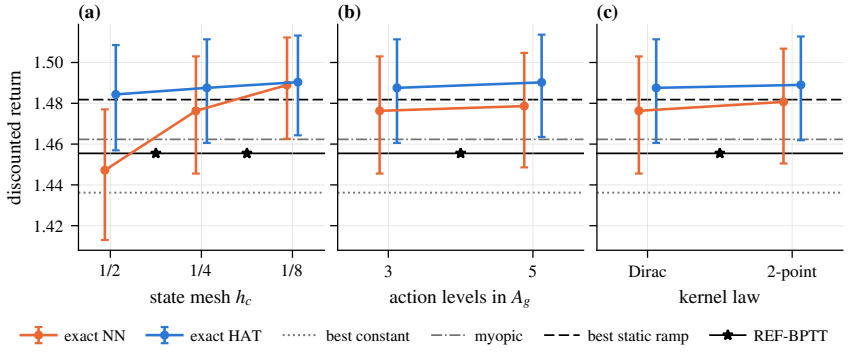

Figure 3: Calibration refinement: test return of the exact finite-table policies when (a) the state mesh, (b) the action grid, and (c) the kernel class are refined separately. Orange: NN; blue: HAT; horizontal lines: reference controls.

Cybersecurity model. The cybersecurity model gives a second example of the effect of discretisation. The learner again approaches $Q_{\mathcal{R}}$, while most of the remaining control error comes from the state projection (Figure 12 in Appendix I). A single decision step can move the population by a

distance close to the state-net radius. As a result, several actions can be mapped to the same small set of next states, and the finite Bellman operator cannot clearly separate them.

Refining the state net reduces this effect, although the improvement is not smooth at every tested resolution. Refining the action grid has a smaller effect, which suggests that the state projection is the main source of error in this example. We also test a local interpolation rule in Wasserstein distance, following Gu et al. (2021). This reduces the projection error, but it changes the Bellman operator and is therefore not covered by Theorem 9. In the tested setting, the optimal policy does not depend strongly on the label.

Finite-population transfer. Corollary 11 separates the mean-field error from the terms that decrease with the population size. In the congestion model, where $\epsilon_N^f = \epsilon_N^F = 0$, the one-step Wasserstein difference between the lifted finite-population next law and the mean-field next law decreases as N grows (Figure 4).

In the cybersecurity model, we evaluate the same fixed policy in the finite-agent system and in the mean-field model using the same common-noise paths. The difference between the two also decreases as the population grows (Figure 13 in Appendix I). These results are consistent with stable transfer of the learned mean-field policies to large finite populations. They also suggest that, at the tested resolutions, the main practical error comes from the finite approximation of the mean-field problem rather than from the finite population itself.

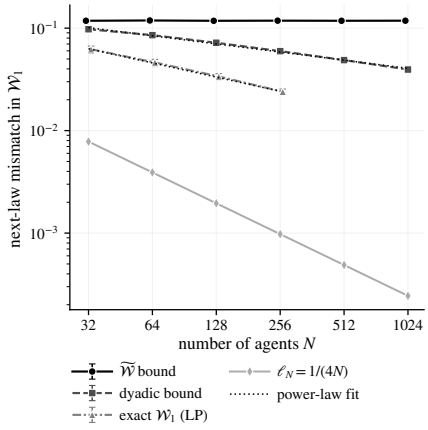

Figure 4: Congestion model: one-step Wasserstein difference between the lifted finite-population and mean-field next laws as N increases. The plot also shows upper bounds and the label discretisation error.

7 Conclusion and limitations

We develop a finite-net Q-learning analysis for heterogeneous mean-field control with common noise. Uniform consistency of Lipschitz kernel restriction and explicit projection bounds connect the finite-table learner to the mean-field control problem. Under the finite-population comparison conditions, the greedy-policy residual also controls deployment loss. Numerical diagnostics document the approach to the projected fixed points, better policy values under state-net refinement, and decreasing finite-population transfer error in the tested settings.

The main computational limitation is the metric complexity of the population-law and kernel spaces. Table storage is independent of the population size N, but constructing and refining the nets can be costly, and deployment requires evaluating the current population state. The experiments identify state projection as a major source of control error and motivate adaptive nets and richer policy classes.

The learning guarantee assumes a generative oracle for the mean-field MDP and is asymptotic. Direct learning from a single finite-agent trajectory requires further analysis. The interpolation variants and the empirical radii of reachable-set nets provide useful diagnostics, but are outside the global nearest-neighbour discretisation guarantee.

AI USE STATEMENT

Generative AI tools were used to review mathematical arguments, identify possible errors, suggest proof repairs, and edit the writing. The authors checked all AI-assisted claims and proofs and take responsibility for the final content.

ETHICS STATEMENT

This work is methodological and uses synthetic benchmark environments. It does not involve human subjects, personal data, or deployed decision systems. Potential downstream risks are those of reinforcement learning and control methods more broadly, especially if applied without domain validation in high-stakes settings.

REPRODUCIBILITY STATEMENT

All numerical comparisons use paired seeds and matched simulator-budget accounting as described in the paper. Code and run scripts are available at the project code repository.

REFERENCES

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning. *The Annals of Applied Probability*, 33(6B):5334–5381, 2023.

Anna De Crescenzo, Filippo de Feo, and Huyên Pham. Linear-quadratic optimal control for non-exchangeable mean-field SDEs and applications to systemic risk. *ESAIM: Control, Optimisation and Calculus of Variations*, 32:34, 2026. doi:10.1051/cocv/2026017.

James Dugundji. An extension of Tietze's theorem. *Pacific Journal of Mathematics*, 1(3):353–367, 1951.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with Q-learning for cooperative MARL: Convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021. doi:10.1137/20M1360700.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 71(4):1040–1054, 2023. doi:10.1287/opre.2022.2395.

Yuanquan Hu, Xiaoli Wei, Junji Yan, and Hengxi Zhang. Graphon mean-field control for cooperative multi-agent reinforcement learning. *Journal of the Franklin Institute*, 360(18):14783–14805, 2023. doi:10.1016/j.jfranklin.2023.09.002.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

László Lovász. *Large Networks and Graph Limits*. American Mathematical Society, 2012.

Samy Mekkaoui and Huyên Pham. Non-exchangeable mean field Markov decision processes with common noise: From Bellman equation to quantitative propagation of chaos. arXiv:2603.00009, 2026.

Samy Mekkaoui, Huyên Pham, and Xavier Warin. Learning operators on labelled conditional distributions with applications to mean field control of non exchangeable systems. arXiv:2603.21683, 2026.

Médéric Motte and Huyên Pham. Mean-field Markov decision processes with common noise and open-loop controls. *The Annals of Applied Probability*, 32(2):1421–1458, 2022. doi:10.1214/21-AAP1713.

Médéric Motte and Huyên Pham. Quantitative propagation of chaos for mean field Markov decision process with common noise. *Electronic Journal of Probability*, 28:1–24, 2023. doi:10.1214/23-EJP978.

John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

Daniel Lacker and Agathe Soret. A label-state formulation of stochastic graphon games and approximate equilibria on large networks. *Mathematics of Operations Research*, 48(4):1987–2018, 2023.

Ming Tan. Multi-agent reinforcement learning: Independent versus cooperative agents. In *Proceedings of the 10th International Conference on Machine Learning*, pages 330–337, 1993.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 746–752, 1998.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pages 6379–6390, 2017.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2974–2982, 2018.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087, 2018.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4295–4304, 2018.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5887–5896, 2019.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex dueling multi-agent Q-learning. In *International Conference on Learning Representations*, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pages 24611–24624, 2022.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2186–2188, 2019.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

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*, pages 5571–5580, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*, volume 325 of *Studies in Systems, Decision and Control*, pages 321–384. Springer, 2021.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

Ian Gemp, Thomas Anthony, Yoram Bachrach, Avishkar Bhoopchand, Kalesha Bullard, Jerome Connor, Vibhavari Dasagi, Bart De Vylder, Edgar A. Duéñez-Guzmán, Romuald Elie, Richard Everett, Daniel Hennes, Edward Hughes, Mina Khan, Marc Lanctot, Kate Larson, Guy Lever, Siqi Liu, Luke Marris, Kevin R. McKee, Paul Muller, Julien Pérolat, Florian Strub, Andrea Tacchetti, Eugene Tarassov, Zhe Wang, and Karl Tuyls. Developing, evaluating and scaling learning agents in multi-agent environments. *AI Communications*, 35(4):271–284, 2022. doi:10.3233/AIC-220113.

Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1028–1037, 2022.

Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

Laëtitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learning in cooperative Markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, volume 27, pages 1–31, 2012.

Qiaomin Cui and Heinz Koeppl. Learning decentralized partially observable mean field control for artificial collective behavior. *arXiv preprint*, 2022.

Landon Kraemer and Bikramjit Banerjee. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, volume 190, pages 82–94, 2016.

A Bellman and measurability details

A.1 The fixed marginal and measurable transition

For $(\mu, \pi) \in \mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0(I \times \mathcal{X}; \mathcal{P}(A))$, set $\nu := \mu \otimes \pi$. The pushforward definition (2) gives, for every bounded Borel g on I,

$$\int g(u)\, \overline{F}(\mu, \pi, e^0)(\mathrm{d}u, \mathrm{d}x) = \int g(u)\, \nu(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a) = \int g \, \mathrm{d}\lambda.$$

Hence $\overline{F}(\mu, \pi, e^0) \in \mathcal{P}_\lambda(I \times \mathcal{X})$.

Fix (μ, π) and write $\nu := \mu \otimes \pi$. For every $\phi \in C(I \times \mathcal{X})$, the map

$$e^0 \longmapsto \int \phi\big(u, F(u, x, a, \nu, e, e^0)\big)\nu(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a)\varrho(\mathrm{d}e)$$

is Borel by joint measurability and Fubini's theorem. Since $I \times \mathcal{X}$ is a compact metric space, a countable dense subset of $C(I \times \mathcal{X})$ generates the Borel sigma-field of $(\mathcal{P}(I \times \mathcal{X}), \mathcal{W})$. Thus $e^0 \mapsto \overline{F}(\mu, \pi, e^0)$ is Borel. This proves that (3) is well defined once $V^\star$ is Borel. The regularity proof in Appendix C gives continuity of $V^\star$.

A.2 Proof of Theorem 5

Proof. By Definition 4 and Proposition 3,

$$\sup_{\pi \in L^0(I\times\mathcal{X};\mathcal{P}(A))} Q^\star(\mu,\pi) = [\mathcal{T}V^\star](\mu) = V^\star(\mu),$$

which proves (4). Substituting this identity at the random next state in (3) gives (5).

For uniqueness, let q be a bounded solution of (5), and suppose that $v_q(\mu) := \sup_\pi q(\mu,\pi)$ is Borel. Taking the supremum over π in its Bellman equation gives $v_q = \mathcal{T}v_q$. Proposition 3 yields $v_q = V^\star$. Substitution back into the equation for q gives $q = Q^\star$. Finally, $|Q^\star| \le \|f\|_\infty + \beta\|V^\star\|_\infty \le \|f\|_\infty/(1-\beta)$. □

Remark 12 (Why the label law is fixed here). Throughout, $\lambda \in \mathcal{P}(I)$ is an arbitrary fixed Borel probability measure. The fixed-marginal condition is structural: labels are persistent and the lifted dynamics preserve their distribution. No assumption $\lambda(\mathrm{d}u) = \mathrm{d}u$ is imposed. We use the uniform metric on kernels, which is independent of the support of λ and therefore separates distinct continuous kernels.

B The capped geometric modulus

Lemma 13 (Explicit Hölder constants)**.** *Let $c \ge 0$, $K \ge 0$, $\Delta > 0$, and $\Omega_{c,K}(r) = c\sum_{t\ge0}\beta^t \min\{K^t r, \Delta\}$. Then $\Omega_{c,K}$ is nondecreasing, concave, and vanishes at zero. The following bounds hold.*

(i) If $\beta K < 1$, then

$$\Omega_{c,K}(r) \le \frac{c}{1-\beta K} r.$$

(ii) If $\beta K > 1$, let $\alpha = \log(1/\beta)/\log K \in (0,1)$. Then

$$\Omega_{c,K}(r) \le c\left(\frac{K}{\beta K - 1} + \frac{1}{1-\beta}\right)\Delta^{1-\alpha} r^\alpha. \tag{23}$$

(iii) If $\beta K = 1$, then

$$\Omega_{c,K}(r) \le cr\left(1 + \frac{1}{1-\beta} + \frac{\log_+(\Delta/r)}{\log K}\right). \tag{24}$$

Hence, for every fixed $\alpha \in (0,1)$,

$$\Omega_{c,K}(r) \le c\left(1 + \frac{1}{1-\beta} + \frac{1}{e(1-\alpha)\log K}\right)\Delta^{1-\alpha} r^\alpha. \tag{25}$$

In general, no Lipschitz bound holds at this boundary.

Proof. Concavity and monotonicity hold term by term. Part (i) follows by replacing the minimum by $K^t r$ and summing the geometric series.

For $0 < r \le \Delta$ and $K > 1$, let

$$T = \left\lceil \frac{\log(\Delta/r)}{\log K} \right\rceil.$$

If $\beta K > 1$, split the sum before T:

$$\Omega_{c,K}(r) \le cr\frac{(\beta K)^T}{\beta K - 1} + c\Delta\frac{\beta^T}{1-\beta}.$$

The definition of T gives $K^T \le K\Delta/r$ and $\beta^T \le (r/\Delta)^\alpha$, which proves (23). The case $r > \Delta$ follows from $\Omega_{c,K}(r) = c\Delta/(1-\beta)$.

If $\beta K = 1$, the head has at most T terms, each at most cr, and the tail is at most $c\Delta\beta^T/(1-\beta) \le cr/(1-\beta)$. This gives (24). The inequality $s^{1-\alpha}\log(1/s) \le 1/(e(1-\alpha))$ for $s \in (0,1]$ yields (25). Finally, at $r_n = \Delta K^{-n}$ the first n terms all equal cr_n, so $\Omega_{c,K}(r_n)/r_n \ge cn \to \infty$ when $c > 0$. □

For later use, we choose (α_0, C_V) from Lemma 13 with $(c, K, \Delta) = (c_0, K_0, \Delta)$, and $(\alpha_L, C_{V,L})$ with $(c, K, \Delta) = (c_L, K_L, \Delta)$. If $c = 0$, all corresponding moduli are zero and any admissible exponent may be used.

C Proofs of the regularity results

C.1 Coupling estimates

Lemma 14 (Lifted Lipschitz estimates)**.** *For $(\mu, \pi), (\mu', \pi') \in \mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0(I \times \mathcal{X}; \mathcal{P}(A))$, set $\nu := \mu \otimes \pi$ and $\nu' := \mu' \otimes \pi'$. Then, for $e^0 \in E^0$,*

$$|\overline{f}(\mu, \pi) - \overline{f}(\mu', \pi')| \le 2L_f \mathcal{W}(\nu, \nu'), \tag{26}$$

$$\mathcal{W}(\overline{F}(\mu, \pi, e^0), \overline{F}(\mu', \pi', e^0)) \le (1 + 2L_F)\mathcal{W}(\nu, \nu'). \tag{27}$$

For $(\mu, \pi), (\mu', \pi') \in \mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$, with the same notation,

$$\mathcal{W}(\nu, \nu') \le (1 + L)\mathcal{W}(\mu, \mu') + d_\infty(\pi, \pi') \le (1 + L)D_L((\mu, \pi), (\mu', \pi')). \tag{28}$$

Finally, for every $\mu, \mu' \in \mathcal{P}_\lambda(I \times \mathcal{X})$ and $\pi \in L^0(I \times \mathcal{X}; \mathcal{P}(A))$, there is a Borel kernel $\pi' \in L^0(I \times \mathcal{X}; \mathcal{P}(A))$ such that, with $\nu := \mu \otimes \pi$ and $\nu' := \mu' \otimes \pi'$,

$$\mathcal{W}(\nu, \nu') \le \mathcal{W}(\mu, \mu'). \tag{29}$$

The 1 in (27) is essential: the next-state measure contains the same label coordinate as the current measure. Equation (29) is obtained by transporting the label–state pair and carrying the same action, then disintegrating the resulting joint law at the target state.

Proof of Lemma 14. Set $\nu = \mu \otimes \pi$ and $\nu' = \mu' \otimes \pi'$, and let γ be an optimal coupling of ν and ν'. For the reward,

$$\begin{aligned}
|\overline{f}(\mu, \pi) - \overline{f}(\mu', \pi')| &\le \int |f(u, x, a, \nu) - f(u', x', a', \nu')| \, \gamma(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a, \mathrm{d}u', \mathrm{d}x', \mathrm{d}a') \\
&\le L_f \left(\int \rho_A((u, x, a), (u', x', a')) \, \mathrm{d}\gamma + \mathcal{W}(\nu, \nu') \right) \\
&= 2L_f \mathcal{W}(\nu, \nu').
\end{aligned}$$

For the transition, draw a coupled pair from γ and use the same idiosyncratic noise e. By the pushforward definition (2), this constructs a coupling of $\overline{F}(\mu, \pi, e^0)$ and $\overline{F}(\mu', \pi', e^0)$. Its expected cost is at most

$$\begin{aligned}
&\int |u - u'| \, \mathrm{d}\gamma + L_F \int (\rho((u, x), (u', x')) + d_A(a, a') + \mathcal{W}(\nu, \nu')) \, \mathrm{d}\gamma \\
&\quad \le (1 + 2L_F)\mathcal{W}(\nu, \nu').
\end{aligned}$$

This proves (26)–(27).

For (28), set $\nu_0 := \mu' \otimes \pi$. First compare ν with ν_0. Let ξ be an optimal coupling of μ and μ'. Conditional on $((u, x), (u', x'))$, couple $\pi(u, x)$ and $\pi(u', x')$ optimally. A measurable choice of optimal couplings exists on compact Polish spaces; alternatively, one may use measurable δ-optimal couplings and send $\delta \downarrow 0$. The resulting cost is at most

$$\int (\rho((u, x), (u', x')) + \mathcal{W}(\pi(u, x), \pi(u', x'))) \, \xi(\mathrm{d}u, \mathrm{d}x, \mathrm{d}u', \mathrm{d}x') \le (1 + L)\mathcal{W}(\mu, \mu').$$

Next compare ν_0 and ν' while keeping (u, x) fixed. Their cost is at most

$$\int \mathcal{W}(\pi(u, x), \pi'(u, x))\mu'(\mathrm{d}u, \mathrm{d}x) \le d_\infty(\pi, \pi').$$

The triangle inequality gives (28).

For (29), let ξ be an optimal coupling of μ and μ'. On $(I \times \mathcal{X}) \times (I \times \mathcal{X}) \times A$ define

$$\chi(\mathrm{d}u, \mathrm{d}x, \mathrm{d}u', \mathrm{d}x', \mathrm{d}a) := \xi(\mathrm{d}u, \mathrm{d}x, \mathrm{d}u', \mathrm{d}x')\pi(u, x)(\mathrm{d}a).$$

Let ν' be the (u', x', a)-marginal of χ. Its (u', x')-marginal is μ', so standard disintegration gives a Borel kernel π' with $\nu'(\mathrm{d}u', \mathrm{d}x', \mathrm{d}a) = \mu'(\mathrm{d}u', \mathrm{d}x')\pi'(u', x')(\mathrm{d}a)$. Under χ, couple (u, x, a) with (u', x', a), carrying exactly the same action. The cost is $\int \rho((u, x), (u', x'))\mathrm{d}\xi = \mathcal{W}(\mu, \mu')$, which proves the claim. □

C.2 Full value and Q regularity

Theorem 15 (Regularity of the full value and Q-function)**.** *For all $\mu, \mu' \in \mathcal{P}_\lambda(I \times \mathcal{X})$,*

$$|V^\star(\mu) - V^\star(\mu')| \le \Omega_{c_0,K_0}(\mathcal{W}(\mu,\mu')). \tag{30}$$

Let (α_0, C_V) be the explicit Hölder pair from Appendix B for (c_0, K_0). Then, on $\mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$,

$$|Q^\star(\mu,\pi) - Q^\star(\mu',\pi')| \le C^\star_Q(L) D_L((\mu,\pi),(\mu',\pi'))^{\alpha_0}, \tag{31}$$

where

$$C^\star_Q(L) := c_L \Delta_A^{1-\alpha_0} + \beta C_V K_L^{\alpha_0}.$$

Proof of Theorem 15. Let V have modulus Ω_{c_0,K_0}. Fix μ, μ' and $\pi \in L^0(I \times \mathcal{X}; \mathcal{P}(A))$. Choose π' by (29), and write $r_0 = \mathcal{W}(\mu,\mu')$. Lemma 14 gives

$$\begin{aligned}&\overline{f}(\mu,\pi) + \beta\mathbb{E}[V(\overline{F}(\mu,\pi,\epsilon^0))] - \overline{f}(\mu',\pi') - \beta\mathbb{E}[V(\overline{F}(\mu',\pi',\epsilon^0))] \\ &\quad \le c_0 r_0 + \beta\Omega_{c_0,K_0}(K_0 r_0) \le \Omega_{c_0,K_0}(r_0).\end{aligned}$$

The last step follows from the shifted series in (6); the cap by Δ is understood. Taking the supremum over π, then reversing μ, μ', shows that $\mathcal{T}$ preserves this modulus.

Start value iteration at $V_0 = 0$. Proposition 3 and the contraction of $\mathcal{T}$ imply $\mathcal{T}^n V_0 \to V^\star$ uniformly. Every iterate has the same modulus, so the limit satisfies (30). Lemma 13 gives the Hölder pair (α_0, C_V).

For the Q-function, put $R = D_L((\mu,\pi),(\mu',\pi'))$. Equations (26), (27), and (28) give

$$\begin{aligned}|Q^\star(\mu,\pi) - Q^\star(\mu',\pi')| &\le c_L R + \beta C_V (K_L R)^{\alpha_0} \\ &\le \left(c_L \Delta_A^{1-\alpha_0} + \beta C_V K_L^{\alpha_0}\right) R^{\alpha_0},\end{aligned}$$

because $R \le \Delta_A$. This is (31). □

C.3 Compactness and the restricted Bellman operator

Lemma 16 (Compact restricted spaces)**.** *The spaces $(\mathcal{P}_\lambda(I \times \mathcal{X}), \mathcal{W})$ and $(L^0_L(I \times \mathcal{X}; \mathcal{P}(A)), d_\infty)$ are compact.*

Proof. The space $\mathcal{P}(I \times \mathcal{X})$ is compact because $I \times \mathcal{X}$ is compact. The marginal map is continuous, so $\mathcal{P}_\lambda(I \times \mathcal{X})$ is closed and compact.

The codomain $(\mathcal{P}(A), \mathcal{W})$ is compact. The family $L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ is equicontinuous and has pointwise relatively compact range. Arzelà–Ascoli therefore gives relative compactness in the uniform metric. A uniform limit of L-Lipschitz maps is L-Lipschitz, so the class is closed and hence compact. □

Proof of Theorem 6. For fixed e^0, Lemma 14 shows that $(\mu,\pi) \mapsto \overline{F}(\mu,\pi,e^0)$ is Lipschitz on $\mathcal{P}_\lambda(I \times \mathcal{X}) \times L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$. The bound is uniform in e^0. The reward is also Lipschitz. Thus, for $V \in C(\mathcal{P}_\lambda(I \times \mathcal{X}))$, the objective in (7) is jointly continuous in (μ,π): use uniform continuity of V and bounded convergence for the common-noise expectation. Compactness of $L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ gives a maximum, and the maximum theorem gives $\mathcal{T}_L V \in C(\mathcal{P}_\lambda(I \times \mathcal{X}))$. Also,

$$\|\mathcal{T}_L V - \mathcal{T}_L V'\|_\infty \le \beta \|V - V'\|_\infty.$$

Hence $\mathcal{T}_L$ has a unique fixed point $V_L \in C(\mathcal{P}_\lambda(I \times \mathcal{X}))$.

The argmax correspondence is nonempty and compact-valued with a Borel graph. The measurable maximum theorem gives a Borel selector π_L. Standard finite-horizon iteration, followed by the contraction limit, shows that V_L is the restricted control value and that π_L is optimal.

To prove (8), use the same kernel $\pi \in L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ at μ and μ'. Equation (28) with $\pi' = \pi$ gives an input-law distance at most $(1 + L)\mathcal{W}(\mu,\mu')$. Therefore $\mathcal{T}_L$ preserves the modulus Ω_{c_L,K_L}. Value iteration from zero gives the modulus for V_L, and Lemma 13 gives $(\alpha_L, C_{V,L})$.

The identities for Q_L follow from $\mathcal{T}_L V_L = V_L$. Its regularity is the same calculation as for $Q^\star$, replacing (α_0, C_V) by $(\alpha_L, C_{V,L})$, which yields (9). □

D PROOF OF CONSISTENCY

For the explicit restriction bound, define

$$\mathcal{C}(\mu) := \{\nu \in \mathcal{P}(I \times \mathcal{X} \times A) : \mathrm{pr}_{12}\#\nu = \mu\},$$
$$\mathcal{C}_L(\mu) := \{\nu \in \mathcal{C}(\mu) : \nu = \mu \otimes \pi \text{ for some } \pi \in L^0_L(I \times \mathcal{X}; \mathcal{P}(A))\},$$
$$h_L := \sup_{\mu \in \mathcal{P}_\lambda(I\times\mathcal{X})} \sup_{\nu \in \mathcal{C}(\mu)} \mathrm{dist}_{\mathcal{W}}(\nu, \mathcal{C}_L(\mu)),$$

and

$$\omega_G(r) := 2L_f r + \beta C_V (K_0 r)^{\alpha_0}.$$

The proof below shows that $h_L \downarrow 0$ and

$$g_L \le \frac{\omega_G(h_L)}{1-\beta}. \tag{32}$$

Lemma 17 (Lipschitz kernels are dense under a fixed state law)**.** *Let $\mu \in \mathcal{P}_\lambda(I \times \mathcal{X})$ and let $\kappa : I \times \mathcal{X} \to \mathcal{P}(A)$ be Borel. For every $\delta > 0$, there is an $L < \infty$ and $\pi \in L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ such that*

$$\int_{I\times\mathcal{X}} \mathcal{W}(\kappa(u,x), \pi(u,x))\, \mu(\mathrm{d}u, \mathrm{d}x) < \delta.$$

Proof. The target $\mathcal{P}(A)$ is a compact metric space and therefore separable. Lusin's theorem gives a compact set $K \subset I \times \mathcal{X}$ such that $\mu(K^c)$ is as small as desired and $\kappa|_K$ is continuous. The set $\mathcal{P}(A)$ is convex in the locally convex vector space of finite signed measures with the weak topology. Because A is compact, that topology agrees with the $\mathcal{W}$ topology on $\mathcal{P}(A)$. Dugundji's extension theorem (Dugundji, 1951) gives a continuous extension $\widetilde{\kappa} : I \times \mathcal{X} \to \mathcal{P}(A)$ with $\widetilde{\kappa} = \kappa$ on K.

It remains to approximate $\widetilde{\kappa}$ uniformly by a Lipschitz map. Fix $r > 0$ and let $(u_1, x_1), \ldots, (u_p, x_p)$ be an $r/2$-net of $I \times \mathcal{X}$. Define

$$b_j(u,x) := (r - \rho((u,x),(u_j,x_j)))_+, \quad \psi_j(u,x) := \frac{b_j(u,x)}{\sum_{k=1}^p b_k(u,x)},$$

$$\pi_r(u,x) := \sum_{j=1}^p \psi_j(u,x)\widetilde{\kappa}(u_j,x_j).$$

The denominator is at least $r/2$. Each ψ_j is Lipschitz, so π_r is Lipschitz when $\mathcal{P}(A)$ is equipped with $\mathcal{W}$. Indeed, by Kantorovich–Rubinstein duality,

$$\mathcal{W}(\pi_r(u,x), \pi_r(u',x')) \le \Delta_A \sum_{j=1}^p |\psi_j(u,x) - \psi_j(u',x')|.$$

Only centres within distance r of (u,x) receive positive weight. If $\widetilde{\omega}$ is a modulus of continuity for $\widetilde{\kappa}$, convexity of $\mathcal{W}$ in either argument gives

$$\sup_{(u,x)\in I\times\mathcal{X}} \mathcal{W}(\pi_r(u,x), \widetilde{\kappa}(u,x)) \le \widetilde{\omega}(r).$$

Choose r so that this is less than $\delta/2$. Also choose K so that $\Delta_A \mu(K^c) < \delta/2$. Since $\kappa = \widetilde{\kappa}$ on K,

$$\int_{I\times\mathcal{X}} \mathcal{W}(\kappa(u,x), \pi_r(u,x))\mu(\mathrm{d}u,\mathrm{d}x) \le \Delta_A \mu(K^c) + \sup_{(u,x)\in I\times\mathcal{X}} \mathcal{W}(\widetilde{\kappa}(u,x), \pi_r(u,x)) < \delta.$$

Thus $\pi = \pi_r$ proves the claim. □

Proof of Theorem 7. For $\nu \in \mathcal{P}(I \times \mathcal{X} \times A)$ with label marginal λ, define the proof-local objective directly by

$$\mathcal{G}(\nu) := \int_{I\times\mathcal{X}\times A} f(u,x,a,\nu)\, \nu(\mathrm{d}u, \mathrm{d}x, \mathrm{d}a)$$
$$+ \beta \mathbb{E}\left[V^\star\left(\left[(u,x,a,e) \mapsto \left(u, F(u,x,a,\nu,e,\epsilon^0)\right)\right]\#(\nu \otimes \varrho)\right)\right].$$

Every $\nu \in \mathcal{C}(\mu)$ disintegrates as $\nu = \mu \otimes \kappa$ for a Borel kernel after changing it on a μ-null set. Hence

$$[\mathcal{T}V^\star](\mu) = \sup_{\nu \in \mathcal{C}(\mu)} \mathcal{G}(\nu), \quad [\mathcal{T}_L V^\star](\mu) = \max_{\nu \in \mathcal{C}_L(\mu)} \mathcal{G}(\nu).$$

By Lemma 14 and the Hölder bound for $V^\star$,

$$|\mathcal{G}(\nu) - \mathcal{G}(\nu')| \le 2L_f \mathcal{W}(\nu, \nu') + \beta C_V (K_0 \mathcal{W}(\nu, \nu'))^{\alpha_0} = \omega_G(\mathcal{W}(\nu, \nu')).$$

It follows that

$$0 \le [\mathcal{T}V^\star](\mu) - [\mathcal{T}_L V^\star](\mu) \le \omega_G(h_L).$$

Since $\mathcal{T}V^\star = V^\star$ and $\mathcal{T}_L$ is a contraction,

$$\|V^\star - V_L\|_\infty \le \omega_G(h_L) + \beta \|V^\star - V_L\|_\infty,$$

which proves (32), once $h_L \to 0$ is shown.

Let

$$D := \{(\mu, \nu) : \mu \in \mathcal{P}_\lambda(I \times \mathcal{X}),\ \nu \in \mathcal{C}(\mu)\}.$$

The set D is closed in the compact product $\mathcal{P}_\lambda(I \times \mathcal{X}) \times \mathcal{P}(I \times \mathcal{X} \times A)$, hence compact. For each integer $n \ge 0$, define

$$g_n(\mu, \nu) := \mathrm{dist}_{\mathcal{W}}(\nu, \mathcal{C}_n(\mu)).$$

The set $\mathcal{C}_n(\mu)$ is compact. Also,

$$|g_n(\mu, \nu) - g_n(\mu', \nu')| \le \mathcal{W}(\nu, \nu') + (1 + n)\mathcal{W}(\mu, \mu'),$$

because the same n-Lipschitz kernel sends μ and μ' to laws at distance at most $(1 + n)\mathcal{W}(\mu, \mu')$. Thus g_n is continuous on D, and g_n decreases with n.

Fix $(\mu, \nu) \in D$ and disintegrate $\nu = \mu \otimes \kappa$. By Lemma 17, for every $\delta > 0$ there is $\pi \in L_L^0(I \times \mathcal{X}; \mathcal{P}(A))$. Set $\nu_\pi := \mu \otimes \pi$. Then $\int_{I \times \mathcal{X}} \mathcal{W}(\kappa(u, x), \pi(u, x))\mu(\mathrm{d}u, \mathrm{d}x) < \delta$, and coupling at the same (u, x) gives

$$\mathcal{W}(\nu, \nu_\pi) \le \int_{I \times \mathcal{X}} \mathcal{W}(\kappa(u, x), \pi(u, x))\mu(\mathrm{d}u, \mathrm{d}x) < \delta.$$

Therefore $g_n(\mu, \nu) \downarrow 0$ pointwise. Dini's theorem gives uniform convergence on D, so $h_n = \sup_D g_n \downarrow 0$. Monotonicity extends this to real $L \to \infty$.

The inclusion $L_L^0(I \times \mathcal{X}; \mathcal{P}(A)) \subset L^0(I \times \mathcal{X}; \mathcal{P}(A))_{L'}$ for $L \le L'$ implies $V_L \le V_{L'} \le V^\star$ by Bellman comparison. Together with (10), this proves uniform monotone convergence. Finally,

$$Q^\star(\mu, \pi) - Q_L(\mu, \pi) = \beta \mathbb{E}[(V^\star - V_L)(\overline{F}(\mu, \pi, \epsilon^0))],$$

which proves (11). □

Proposition 18 (A useful finite-L special case)**.** *Suppose that for some L_0 and $\delta \ge 0$ there is a Borel feedback $\pi : \mathcal{P}_\lambda(I \times \mathcal{X}) \to L^0(I \times \mathcal{X}; \mathcal{P}(A))_{L_0}$ such that*

$$\overline{f}(\mu, \pi(\mu)) + \beta \mathbb{E}[V^\star(\overline{F}(\mu, \pi(\mu), \epsilon^0))] \ge V^\star(\mu) - \delta$$

for every μ. Then

$$\|V^\star - V_L\|_\infty \le \frac{\delta}{1 - \beta} \quad \textit{for every } L \ge L_0.$$

Proof. The assumption gives $V^\star \le \mathcal{T}_{L_0} V^\star + \delta$. Compare with the fixed point V_{L_0} and use the contraction:

$$\|V^\star - V_{L_0}\|_\infty \le \delta + \beta \|V^\star - V_{L_0}\|_\infty.$$

Monotonicity in L completes the proof. □

E FINITE NETS AND COVERING BOUNDS

For a compact metric space (Y, d), write $N_Y(r)$ for the smallest size of an internal r-net of Y. The next lemma gives a simple bound for probability measures. It is not sharp, but it is enough to show that all tables used in Section 4 are finite.

Lemma 19 (Covering probability measures)**.** *Let (Y, d) be compact with diameter $\Delta_A^0 > 0$. For $0 < r \le \Delta_A^0$, set $m := N_Y(r/4)$. Then*

$$\log N_{\mathcal{P}(Y)}(r) \le m \log\left(1 + \frac{8m\Delta_A^0}{r}\right),$$

where $\mathcal{P}(Y)$ is equipped with $\mathcal{W}$. For $r > \Delta_A^0$, one point is enough.

Proof. Choose an internal $r/4$-net $\{y_1, \dots, y_m\}$ of Y and a Borel nearest-point map $p : Y \to \{y_1, \dots, y_m\}$. For every $\mu \in \mathcal{P}(Y)$, $\mathcal{W}(\mu, p\#\mu) \le r/4$.

It remains to quantise the vector of masses of $p\#\mu$. Let $K := \lceil 4m\Delta_A^0/r \rceil$. Every probability vector in the m-simplex has a vector $q = (q_1, \dots, q_m)$ with $q_i = k_i/K$, $k_i \in \mathbb{N} \cup \{0\}$, $\sum_i k_i = K$, and

$$\|q - p\|_1 \le \frac{2m}{K}.$$

One construction rounds the first $m-1$ cumulative sums to multiples of $1/K$ and takes differences. Two measures supported on the same m points satisfy $\mathcal{W} \le (\Delta_A^0/2)\|q - p\|_1$. Hence this second rounding costs at most $\Delta_A^0 m/K \le r/4$. The total error is at most $r/2$, and therefore also at most r.

There are $\binom{K+m-1}{m-1} \le (K+1)^m$ rounded vectors. Since $r \le \Delta_A^0$ and $m \ge 1$, $K+1 \le 1+8m\Delta_A^0/r$. Taking logarithms proves the result. □

Proposition 20 (State and kernel nets)**.** *For every $\epsilon_\mu, \epsilon_\pi > 0$, finite internal nets $\mathcal{R}_\mu \subset \mathcal{P}_\lambda(I \times \mathcal{X})$ and $\mathcal{R}_\pi \subset L_L^0(I \times \mathcal{X}; \mathcal{P}(A))$ exist. Their sizes may be chosen so that*

$$N_{\mathcal{P}_\lambda(I\times\mathcal{X})}(\epsilon_\mu) \le N_{\mathcal{P}(I\times\mathcal{X})}(\epsilon_\mu/2), \tag{33}$$

$$\log N_{L_L^0(I\times\mathcal{X};\mathcal{P}(A))}(\epsilon_\pi) \le N_{I\times\mathcal{X}}\left(\frac{\epsilon_\pi}{4L}\right) \log N_{\mathcal{P}(A)}(\epsilon_\pi/4), \quad L > 0. \tag{34}$$

For $L = 0$, $N_{L^0(I\times\mathcal{X};\mathcal{P}(A))_0}(\epsilon_\pi) \le N_{\mathcal{P}(A)}(\epsilon_\pi)$. Each finite net admits a Borel nearest-neighbour map with deterministic tie-breaking.

Proof. For (33), start with an external $\epsilon_\mu/2$-net of $\mathcal{P}(I \times \mathcal{X})$. For every ball that intersects $\mathcal{P}_\lambda(I \times \mathcal{X})$, choose one point of the intersection. If $\mu \in \mathcal{P}_\lambda(I \times \mathcal{X})$ lies in the original ball, it is within ϵ_μ of the chosen point. This makes the net internal. Lemma 19, with $Y = I \times \mathcal{X}$, gives an explicit finite upper bound.

Assume $L > 0$. Let $\{(u_1, x_1), \dots, (u_M, x_M)\}$ be an internal $\epsilon_\pi/(4L)$-net of $I \times \mathcal{X}$ and let $\mathcal{Q}$ be an internal $\epsilon_\pi/4$-net of $\mathcal{P}(A)$. Take a maximal subset $\{\pi_1, \dots, \pi_J\} \subset L_L^0(I \times \mathcal{X}; \mathcal{P}(A))$ whose distinct elements have d_∞-distance greater than ϵ_π. Compactness makes J finite, and maximality makes this set an internal ϵ_π-net.

Encode π_j by rounding each value $\pi_j(u_i, x_i)$ to a nearest point of $\mathcal{Q}$. Two kernels with the same code satisfy, for every (u, x) and a nearest (u_i, x_i),

$$\begin{aligned}\mathcal{W}(\pi(u,x), \pi'(u,x)) \le{}& \mathcal{W}(\pi(u,x), \pi(u_i,x_i)) + \mathcal{W}(\pi(u_i,x_i), q_i)\\ &+ \mathcal{W}(q_i, \pi'(u_i,x_i)) + \mathcal{W}(\pi'(u_i,x_i), \pi'(u,x)) \le \epsilon_\pi.\end{aligned}$$

They therefore cannot be two distinct members of the strictly separated family. The code is injective, so $J \le |\mathcal{Q}|^M$, which is (34). If $L = 0$, every kernel is constant on the nonempty space $I \times \mathcal{X}$, so $L^0(I \times \mathcal{X}; \mathcal{P}(A))_0$ is isometric to $\mathcal{P}(A)$.

For a finite ordered net $\{y_1, \dots, y_J\}$, assign a point to the smallest-index nearest centre. Its cells are finite intersections of sets of the form $\{d(\cdot, y_i) \le d(\cdot, y_j)\}$ and their complements, hence are Borel. This gives the claimed projection maps. □

The bounds above are worst-case bounds. They prove existence, not computational tractability. In particular, they do not imply that an arbitrary net point has a closed-form reward or transition oracle. Appendix H gives structured, explicitly representable quantisers for one benchmark.

F PROOFS FOR FINITE-NET Q-LEARNING

F.1 REDUCTION TO A FINITE MARKOV DECISION PROCESS

The table in Algorithm 1 is exactly the Q-table of a finite discounted MDP. Its state set is $\mathcal{R}_\mu$, its action set is $\mathcal{R}_\pi$, its reward at (μ, π) is $\overline{f}(\mu, \pi)$, and its next state is $\Pi_\mu \overline{F}(\mu, \pi, \epsilon^0)$. Its optimality operator is $\mathcal{B}_\mathcal{R}$ in (12). Since rewards are bounded, the fixed point $Q_\mathcal{R}$ is bounded by $\|f\|_\infty/(1-\beta)$.

Proof of the convergence part of Theorem 9. Let $\mathcal{F}_n$ contain the table, visit counts, and all choices and oracle samples through time n. The pair c_n is $\mathcal{F}_n$-measurable, while the fresh common-noise draw has its stated conditional law. Therefore

$$\mathbb{E}[Y_n \mid \mathcal{F}_n, c_n = (\mu, \pi)] = [\mathcal{B}_\mathcal{R} Q_n](\mu, \pi).$$

The noise $Y_n - [\mathcal{B}_\mathcal{R} Q_n](c_n)$ has conditional mean zero. The iterates remain bounded: with

$$B := \max\left\{\|Q_0\|_\infty, \frac{\|f\|_\infty}{1-\beta}\right\},$$

induction gives $\|Q_n\|_\infty \le B$, because every update is a convex combination and $|Y_n| \le \|f\|_\infty + \beta B \le B$. Thus the conditional second moments of the noise are uniformly bounded.

The operator $\mathcal{B}_\mathcal{R}$ is a monotone β-contraction in the sup norm. The per-visit steps satisfy the Robbins–Monro conditions, and every component is updated infinitely often. The standard asynchronous stochastic-approximation theorem for discounted Q-learning (Jaakkola et al., 1994; Tsitsiklis, 1994) now gives $Q_n \to Q_\mathcal{R}$ almost surely. Notice that this argument uses one fresh sample for the single component updated at time n; it does not analyse a simultaneous sweep. □

F.2 PROJECTION BIAS

Proof of the error bounds in Theorem 9. Set

$$\delta := \max_{(\mu,\pi)\in\mathcal{R}_\mu\times\mathcal{R}_\pi} |Q_\mathcal{R}(\mu, \pi) - Q_L(\mu, \pi)|.$$

For a realised next state $s \in \mathcal{P}_\lambda(I \times \mathcal{X})$, first compare the two maximisations at $\Pi_\mu s$. The action net and the Hölder continuity of Q_L give

$$V_L(\Pi_\mu s) - C_{Q,L}\epsilon_\pi^{\alpha_L} \le \max_{g\in\mathcal{R}_\pi} Q_L(\Pi_\mu s, g) \le V_L(\Pi_\mu s). \tag{35}$$

The state modulus gives

$$|V_L(\Pi_\mu s) - V_L(s)| \le C_{V,L}\epsilon_\mu^{\alpha_L}. \tag{36}$$

Finally, replacing $Q_\mathcal{R}$ by Q_L on the finite table changes a maximum by at most δ. Using the two fixed-point equations, (35), and (36), we obtain

$$\delta \le \beta\left(\delta + C_{V,L}\epsilon_\mu^{\alpha_L} + C_{Q,L}\epsilon_\pi^{\alpha_L}\right).$$

Solving this inequality proves (13).

For arbitrary $(\mu, \pi) \in \mathcal{P}_\lambda(I \times \mathcal{X}) \times L_L^0(I \times \mathcal{X}; \mathcal{P}(A))$,

$$\begin{aligned}|[\Gamma Q_\mathcal{R}](\mu, \pi) - Q_L(\mu, \pi)| &\le e_\mathcal{R} + |Q_L(\Pi_\mu\mu, \Pi_\pi\pi) - Q_L(\mu, \pi)| \\ &\le e_\mathcal{R} + C_{Q,L}(\epsilon_\mu + \epsilon_\pi)^{\alpha_L} = E_\mathcal{R}.\end{aligned}$$

This is (14). Almost-sure table convergence makes the same bound hold for the limiting extensions of Q_n. By (11), $\|Q^\star - Q_L\|_{\mathcal{P}_\lambda(I\times\mathcal{X})\times L_L^0(I\times\mathcal{X};\mathcal{P}(A))} \le \beta g_L$, and the triangle inequality proves (15). □

F.3 THE GREEDY POLICY

Proof of Corollary 10. Fix $\mu \in \mathcal{P}_\lambda(I \times \mathcal{X})$, put $\bar{\mu} = \Pi_\mu \mu$, and let $\pi_\mathcal{R} = \pi_\mathcal{R}(\mu)$. Starting from the exact restricted maximum at $\bar{\mu}$, use the action net, the grid error twice, and then move the state back from $\bar{\mu}$ to μ:

$$\begin{aligned} V_L(\mu) &\le V_L(\bar{\mu}) + C_{V,L}\epsilon_\mu^{\alpha_L} \\ &\le \max_{g \in \mathcal{R}_\pi} Q_L(\bar{\mu}, g) + C_{Q,L}\epsilon_\pi^{\alpha_L} + C_{V,L}\epsilon_\mu^{\alpha_L} \\ &\le \max_{g \in \mathcal{R}_\pi} Q_\mathcal{R}(\bar{\mu}, g) + e_\mathcal{R} + C_{Q,L}\epsilon_\pi^{\alpha_L} + C_{V,L}\epsilon_\mu^{\alpha_L} \\ &= Q_\mathcal{R}(\bar{\mu}, \pi_\mathcal{R}) + e_\mathcal{R} + C_{Q,L}\epsilon_\pi^{\alpha_L} + C_{V,L}\epsilon_\mu^{\alpha_L} \\ &\le Q_L(\bar{\mu}, \pi_\mathcal{R}) + 2e_\mathcal{R} + C_{Q,L}\epsilon_\pi^{\alpha_L} + C_{V,L}\epsilon_\mu^{\alpha_L} \\ &\le Q_L(\mu, \pi_\mathcal{R}) + \eta_\mathcal{R}. \end{aligned}$$

Because $Q_L(\mu, \pi_\mathcal{R}) = [\mathcal{T}^{\pi_\mathcal{R}} V_L](\mu)$, this is (16).

Let $V^{\pi_\mathcal{R}}$ be the fixed point of the policy operator $\mathcal{T}^{\pi_\mathcal{R}}$. Monotonicity and contraction give

$$0 \le V_L - V^{\pi_\mathcal{R}} \le \frac{\eta_\mathcal{R}}{1-\beta}.$$

Adding $0 \le V^\star - V_L \le g_L$ proves (17). Also, $\mathcal{T}^{\pi_\mathcal{R}} V^\star \ge \mathcal{T}^{\pi_\mathcal{R}} V_L \ge V_L - \eta_\mathcal{R} \ge V^\star - (g_L + \eta_\mathcal{R})$, which is (18). The policy is Borel because it is constant on the finitely many Borel cells of Π_μ and tie-breaking is fixed. □

G FINITE-POPULATION INTERFACE

This section checks the two details needed to apply the finite-population theorem: the randomised feedback must be measurable, and it must satisfy the pointwise regularity condition in the label–state variable.

Lemma 21 (Quantile representation)**.** *Let $A = [a_-, a_+]$ with the Euclidean metric. There is a Borel map*

$$q : \mathcal{P}(A) \times [0,1] \to A, \quad q(\vartheta, r) = F_\vartheta^{-1}(r),$$

such that $q(\vartheta, Z)$ has law ϑ for $Z \sim \mathrm{Unif}[0,1]$. Moreover,

$$\int_0^1 |q(\vartheta, r) - q(\vartheta', r)| \, \mathrm{d}r = \mathcal{W}(\vartheta, \vartheta'). \tag{37}$$

Proof. Use the left-continuous generalised inverse $F_\vartheta^{-1}(r) = \inf\{a \in A : F_\vartheta(a) \ge r\}$. Its joint Borel measurability follows, for example, because $\{(\vartheta, r) : F_\vartheta^{-1}(r) < t\} = \bigcup_{s<t,\, s\in\mathbb{Q}} \{(\vartheta, r) : F_\vartheta(s) \ge r\}$ and $\vartheta \mapsto F_\vartheta(s) = \vartheta([a_-, s])$ is Borel. The probability integral transform gives the first claim. Identity (37) is the one-dimensional monotone-coupling formula for $\mathcal{W}$. □

Proof of Corollary 11. The map $\pi_\mathcal{R} : \mathcal{P}_\lambda(I \times \mathcal{X}) \to L_L^0(I \times \mathcal{X}; \mathcal{P}(A))$ is Borel. Evaluation $(\pi, u, x) \mapsto \pi(u, x)$ is continuous from $L_L^0(I \times \mathcal{X}; \mathcal{P}(A)) \times I \times \mathcal{X}$ to $\mathcal{P}(A)$, and the quantile map in Lemma 21 is Borel. Hence a in (20) is a Borel randomised feedback.

For fixed μ, $u, u' \in I$, and $x, x' \in \mathcal{X}$, use the same uniform random variable in the two actions. Equations (37) and the definition of $L_L^0(I \times \mathcal{X}; \mathcal{P}(A))$ give

$$\mathbb{E}_Z |a(\mu, u, x, Z) - a(\mu, u', x', Z)| = \mathcal{W}(\pi_\mathcal{R}(\mu)(u, x), \pi_\mathcal{R}(\mu)(u', x')) \le L\rho((u, x), (u', x')).$$

Thus the transferred randomised feedback is pointwise Lipschitz in the label–state variable with constant L.

Corollary 10 gives the one-step verification inequality

$$V^\star(\mu) \le [\mathcal{T}^{\pi_\mathcal{R}} V^\star](\mu) + \epsilon_B, \quad \epsilon_B := \eta_\mathcal{R} + g_L.$$

We now compare, under the same common-noise realisation, the finite system controlled by (21) with a mean-field system controlled by $\pi_\mathcal{R}$. Couple the idiosyncratic noises and use the block lift (19).

At one step, the mismatch between the two joint label–state–action laws is the sum of three effects: empirical sampling (M_N), the difference between F_N and F (ϵ_N^F), and the discrepancy already present in the current state law. Because the same uniform randomiser is used in the two actions, the latter discrepancy is amplified by at most $1 + L$. Iterating this one-step coupling and using the Hölder regularity exponent γ for the mean-field value turns the transition error into $M_N^\gamma + (\epsilon_N^F)^\gamma$.

The reward comparison is handled in the same coupling: the direct reward mismatch contributes ϵ_N^f, while the state–action mismatch is controlled by the same Wasserstein error. Summing the discounted one-step inequalities and inserting the verification residual ϵ_B gives

$$V_N(x) - V_N^{\pi_{\mathcal{R}},N}(x) \le C(1+L)\left[\epsilon_B + M_N^\gamma + \epsilon_N^f + (\epsilon_N^F)^\gamma\right],$$

which is (22).

For a general compact metric action space, a measurable randomisation still realises each action kernel, but there is no one-dimensional quantile coupling guaranteeing the same-noise Lipschitz estimate. Instead, at each finite state, choose an action vector minimising the Wasserstein distance between its lifted empirical label–state–action law and the joint law prescribed by the mean-field feedback. The same one-step comparison and discounted iteration then give the same asymptotic order, although the resulting policy is no longer the explicit formula (21). □

H A BENCHMARK SATISFYING THE ASSUMPTIONS

We specify the nominal congestion model and the window-throughput variant used in Section 6. Both satisfy Assumption 2.

H.1 ENVIRONMENT

Agents sit on a line and share a resource whose capacity is changed by common noise. Let $I = [0,1]$, $\lambda \in \mathcal{P}(I)$ be the fixed label distribution, and

$$w(u,v) := e^{-|u-v|}.$$

Let $\mathcal{X} = \{0,1,\dots,\bar{S}\}$ with $d(x,x') = |x-x'|/\bar{S}$, and let $A = [0,1]$. Both idiosyncratic and common noise are uniform on $[0,1]$. For $\nu \in \mathcal{P}(I \times \mathcal{X} \times A)$, define

$$\begin{aligned} J(u;\nu) &:= \int w(u,v)\frac{x'}{\bar{S}}a'\,\nu(\mathrm{d}v,\mathrm{d}x',\mathrm{d}a'), \qquad (38)\\ \kappa(e^0) &:= \kappa_{\min} + (\kappa_{\max} - \kappa_{\min})e^0, \quad c(u) := c_0 + c_1 u,\\ \ell(u,\nu,e^0) &:= \mathrm{clip}\left(J(u;\nu) - \kappa(e^0)c(u)\right),\\ p(u,a,\nu,e^0) &:= \mathrm{clip}\left(\ell(u,\nu,e^0) + a/2\right), \end{aligned}$$

where $\mathrm{clip}(s) = \min\{1, \max\{0,s\}\}$. The state follows an increase/backoff rule:

$$F(u,x,a,\nu,e,e^0) = \begin{cases} \min\{x+1,\bar{S}\}, & e > p(u,a,\nu,e^0),\\ \lceil x/2 \rceil, & e \le p(u,a,\nu,e^0). \end{cases} \qquad (39)$$

Let $\bar{\ell}(u,\nu) = \int_0^1 \ell(u,\nu,e^0)\,\mathrm{d}e^0$ and set

$$f(u,x,a,\nu) = r_1 a(1-\bar{\ell}(u,\nu)) - r_2\left(\frac{x}{\bar{S}} - \tau\right)^2 - r_3\bar{\ell}(u,\nu).$$

A concrete default is

$$\bar{S} = 8, \quad \kappa_{\min} = 0.4, \quad \kappa_{\max} = 1, \quad c_0 = c_1 = \tfrac{1}{2}, \quad r_1 = 1,\ r_2 = \tfrac{1}{2},\ r_3 = 1,\ \tau = \tfrac{1}{2}.$$

The graphon makes interaction depend on the label, and e^0 makes the limiting state law random. A higher action value gives more immediate throughput but also raises the chance of a backoff.

Variant used in the experiments. In the nominal experiments, the constant action $a = 1$ was best among the evaluated constant and 625 ramp controls and matched the model-gradient reference. The main comparisons therefore use window throughput: replace $r_1 a(1-\bar{\ell})$ by $r_1(x/\bar{S})a(1-\bar{\ell})$ and set $\kappa_{\min} = 0.2$ and $\kappa_{\max} = 0.5$, leaving the other defaults unchanged. Equivalently, $\kappa(e^0) = \bar{\kappa} + \sigma(2e^0 - 1)$ with $\bar{\kappa} = 0.35$ and $\sigma = 0.15$. The discount is $\beta = 0.8$, with $\beta = 0.9$ used for sensitivity analysis. This variant couples immediate throughput to the current window state; the nominal regime is retained as a baseline diagnostic.

H.2 Correct constants

The reward is bounded by

$$\|f\|_\infty \le r_1 + r_2 \max\{\tau, 1-\tau\}^2 + r_3.$$

The map w is 1-Lipschitz in each argument, and all multiplicative factors in (38) lie in $[0,1]$. Kantorovich–Rubinstein duality therefore gives

$$|J(u;\nu) - J(u';\nu')| \le |u-u'| + \mathcal{W}(\nu,\nu').$$

Since clipping is 1-Lipschitz,

$$\begin{aligned} |\ell(u,\nu,e^0) - \ell(u',\nu',e^0)| &\le (1+\kappa_{\max}c_1)|u-u'| + \mathcal{W}(\nu,\nu'), \\ |p(u,a,\nu,e^0) - p(u',a',\nu',e^0)| &\le (1+\kappa_{\max}c_1)|u-u'| + \tfrac{1}{2}|a-a'| + \mathcal{W}(\nu,\nu'). \end{aligned} \tag{40}$$

The same bound as (40) holds for $\bar{\ell}$.

Couple two transitions with the same uniform e. When both systems choose the same branch of (39), the normalised state distance does not increase. The branches differ with probability $|p-p'|$, and their distance is at most one. Hence

$$\mathbb{E}_e d(F,F') \le d(x,x') + |p-p'|.$$

Thus Assumption 2 holds for the transition with

$$L_F = 1 + \kappa_{\max}c_1.$$

For the reward, use

$$|a(1-\bar{\ell}) - a'(1-\bar{\ell}')| \le |a-a'| + |\bar{\ell}-\bar{\ell}'|$$

and $|(x/\bar{S}-\tau)^2 - (x'/\bar{S}-\tau)^2| \le 2d(x,x')$. A single constant that multiplies the sum of all four input increments is

$$L_f = \max\left\{r_1,\ 2r_2,\ (r_1+r_3)(1+\kappa_{\max}c_1)\right\}. \tag{41}$$

At the nominal defaults, (41) gives $L_f = 3$, with $L_F = 3/2$ and $K_0 = 1 + 2L_F = 4$. For window throughput, the additional factor satisfies

$$\left|\frac{x}{\bar{S}}a(1-\bar{\ell}) - \frac{x'}{\bar{S}}a'(1-\bar{\ell}')\right| \le d(x,x') + |a-a'| + |\bar{\ell}-\bar{\ell}'|.$$

The same valid constants $L_F = 3/2$, $L_f = 3$, $K_0 = 4$ and the same reward bound therefore apply to the experimental variant. These constants need not be sharp.

H.3 Structured quantisers

The abstract covering proof in Appendix E does not by itself give an exact oracle. For this benchmark, useful net points can instead be represented by label blocks and finite probability vectors.

For $\mathcal{X} = \{0,\dots,\bar{S}\}$, one-dimensional transport gives

$$\mathcal{W}(\vartheta,\vartheta') = \frac{1}{\bar{S}}\sum_{j=0}^{\bar{S}-1} |F_\vartheta(j) - F_{\vartheta'}(j)|. \tag{42}$$

Rounding a CDF to a monotone grid of mesh h gives an internal fibre quantiser of radius at most $h/2$. Averaging the conditional state laws over n_b equal label blocks and then rounding their CDFs gives a state quantiser with radius at most

$$\frac{1}{n_b} + \frac{h}{2}.$$

The first term comes from moving each label to a representative point in its block; the second comes from (42).

For actions on $[0, 1]$, approximate a quantile function by its *average* on each of k equal quantile blocks, then round those averages to a monotone value grid of mesh g. Because a monotone $[0, 1]$-valued function has total variation at most one, the block-average error in L^1 is at most $1/(2k)$, and value rounding adds at most $g/2$. Thus the action-law radius is

$$\frac{1}{2k} + \frac{g}{2}.$$

Using arbitrary point samples inside the quantile blocks would not justify the factor $1/(2k)$; the block averages are needed.

A full net of $L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$ must index values by both a label grid and the state. Constant or label-only kernels cover only smaller compact subclasses. Learning on such a subclass is mathematically valid, but its value has an extra restriction gap relative to V_L unless a separate density argument is supplied. The main experiments use a structured ramp-kernel subclass and reachable-set nets with measured covering radii. Calibration instead uses CDF-grid product nets with a proven radius. These distinctions are retained when interpreting Theorems 6–7 and Corollary 11.

I Details of the numerical illustration

I.1 Representation and protocol

The physical label quadrature uses 16 midpoint atoms, with uniform masses or masses corresponding to the density $2u$. Main experiments use two label blocks and nested farthest-point nets on the reachable set. Their radii are measured over node successors and validation laws, not certified as global radii on $\mathcal{P}_\lambda(I \times \mathcal{X})$. The 625 ramp kernels have the form $a(u, x) = \mathrm{clip}(\alpha(u) + s(u)(x/\bar{S} - \tau))$ with two label anchors. Calibration uses $\bar{S} = 2$, label atoms $\{0, 1\}$, 289 Dirac kernels at $L = 1$, and CDF-grid product nets of 36, 225, and 2025 nodes with proven radius $h_c/2$.

HAT uses positive weights proportional to $(2\hat{\epsilon}_\mu - \mathcal{W}(\mu, \mu_j))_+$, normalised to sum to one, and interpolates grid-state maxima in the target and the table in deployment. Here $\hat{\epsilon}_\mu$ is the measured radius. At 256 nodes its bandwidth is 0.229 and covers 73% to 79% of nodes. This removes the measured successor-node action aliasing (12% of distinct kernel pairs for NN, 0% for HAT), but does not improve main-benchmark returns. HAT is an experimental variant; the theoretical Algorithm 1 in this version remains the nearest-neighbour scheme.

Training, validation, and test initial laws are disjoint. The test set contains 20 laws with independent labels and states and 20 correlated laws, with 16 paired common-noise paths per law. Tuning uses validation only, with four configurations and three seeds per method; ties within 0.01 select the smallest step-size exponent. There are 10 training seeds, increased to 20 for confirmatory tests. Intervals use a two-way bootstrap over training runs and test laws. All four confirmatory tests use Holm correction and a predeclared practical threshold of 0.01; their adjusted p-values are $4 \cdot 10^{-4}$.

I.2 Confirmatory comparisons and diagnostic scope

Table 1 retains the paired comparisons and their uncertainty. C2 compares representations: the label-blind version uses 25 kernels and a bandwidth covering 98% of its nodes, so the difference does not isolate label information. The C3 comparator uses the number of blocks selected using the validation set. CE denotes HAT trained at $\sigma = 0$. C4 is below the practical threshold. A post hoc, exploratory decomposition attributes about three quarters of its difference to certainty-equivalent reward bias rather than conditioning on the realised law. The four tests reuse the same 20 HAT runs; they are not independent replications, and their adjusted p-values lie at the bootstrap floor.

In calibration (10 training seeds), the holdout residual diagnostic covers the table error in 100/100 HAT batches (Clopper–Pearson interval $[0.964, 1]$) and 20/20 NN batches. At 1600 fresh draws per cell, the bound/error ratio ranges from 3.8 to 83 and one batch costs $1.04 \cdot 10^8$ oracle calls. This is observed coverage, not a finite experimental proof of a probability guarantee. Exact HAT policy returns under calibration refinement are 1.4844, 1.4876, and 1.4904.

Table 1: Predeclared confirmatory tests. Reported differences are in discounted test returns, with 95% two-way bootstrap intervals. CE uses one trained table, not 20 independent tables.

Comparison	Training seeds	Difference [95% CI]
C1: HAT − NN	20 vs 20	-0.0307 $[-0.0361, -0.0248]$
C2: HAT − label-blind HAT	20 vs 20	$+0.0154$ $[0.0115, 0.0192]$
C3: HAT − selected block PPO	20 vs 20	-0.0794 $[-0.0835, -0.0751]$
C4: HAT − CE, $\sigma = 0.15$	20 vs 1 table	$+0.0047$ $[0.0025, 0.0070]$

The transfer study uses frozen tables and labels $(i - \frac{1}{2})/N$. Its exact linear-programming next-law diagnostic has empirical slope -0.46 $[-0.51, -0.41]$; this is not a universal rate for the continuous-label metric. The label approximation error is $\ell_N = 1/(4N)$ and the model-mismatch terms vanish by construction. Thus the study isolates sampling and policy-transfer effects. A supplementary bounded-attack-noise botnet study is a baseline check. The best constant control matches the numerically converged but uncertified value-iteration reference wherever it was run, and no learned policy beats that constant in the tested cells. This does not establish global optimality or a benefit from feedback or label information.

I.3 Calibration and finite-population measurements

On the calibration problem, the sup-norm error of each method relative to its own projected fixed point decreases steadily with the number of draws per cell and is ultimately smaller for HAT than for NN, with little variation across seeds. These measurements do not establish a convergence rate.

Frozen HAT tables deployed in the finite population return the same value as their own exact projected reference, up to sampling noise, at every tested N. This comparison differs from the same-policy finite-N minus mean-field return, which is small but not zero at the smallest population. Figures 9 and 10 distinguish these comparisons. Neither measures finite-N optimality gaps. Here $\epsilon_N^f = \epsilon_N^F = 0$, so model mismatch is not tested.

I.4 Heterogeneity and common-noise comparisons

Figure 5 compares HAT with label-blind HAT (LB), validation-selected block PPO, and label-aware PPO (LA-PPO). The gain over LB is a representation ablation, not an isolated label-information effect. Figure 6 compares conditional-law HAT with certainty-equivalent HAT (CE, trained without common noise) and the unconditional-law ablation (UL). These comparisons do not establish a material advantage from label information or from conditioning on the realised common-noise law.

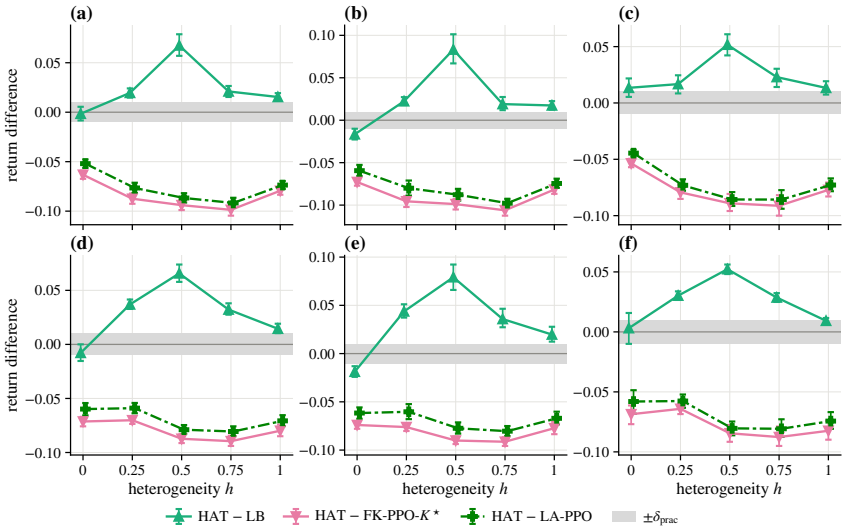

Figure 5: HAT minus comparator return as the heterogeneity h varies, for a uniform (top) and a linear (bottom) label law, on all test laws and on the independent and correlated label–state laws separately. The grey band marks the practical-significance threshold; only the comparisons at full heterogeneity under the uniform law are confirmatory.

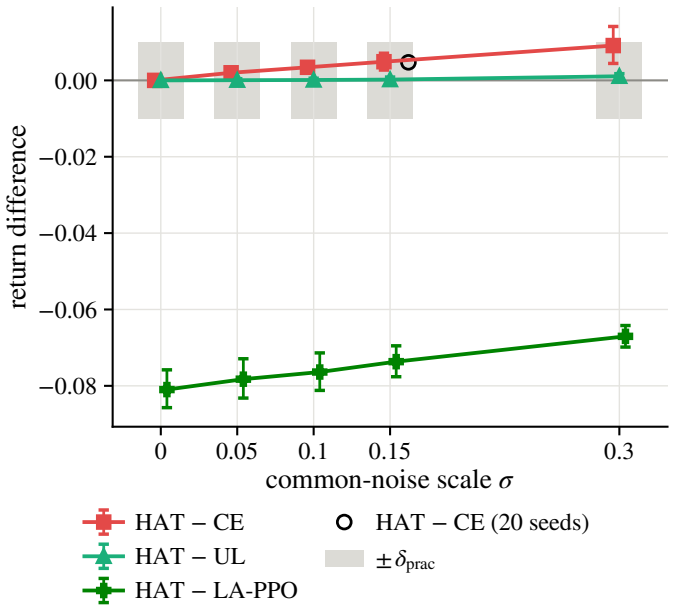

Figure 6: HAT minus CE, UL, and LA-PPO returns against the common-noise scale σ, with paired bootstrap intervals; the open circle is the confirmatory comparison with CE. The grey band marks the practical-significance threshold.

I.5 Additional calibration and projection figures

Figure 7 reports the residual diagnostics recorded by the experimental suite. Here $d^{\mathrm{sp}}_{m,\delta}$ and $d^{B}_{m,\delta}$ denote its span-based and range-based diagnostics; they are not additional guarantees of the nearest-neighbour theorem in this manuscript. Figure 8 reports exact-policy returns over net size and action scale in the main model.

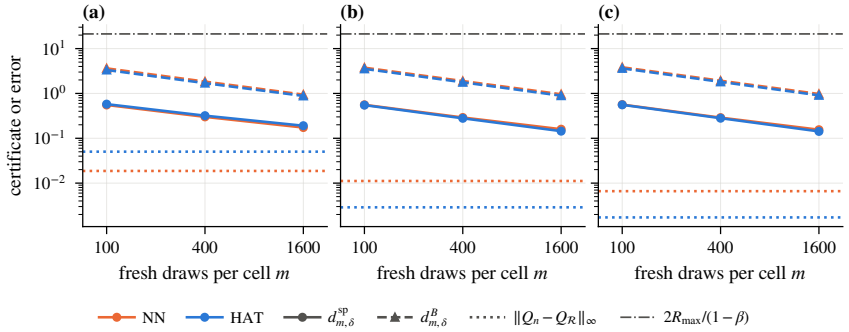

Figure 7: Holdout residual certificate on the calibration problem at three learning checkpoints, against the number m of fresh draws per cell. Solid: $d^{\mathrm{sp}}_{m,\delta}$; dashed: $d^{B}_{m,\delta}$; dotted: observed table error $\|Q_n - Q_{\mathcal{R}}\|_\infty$; dash-dot: vacuity level. Blue: HAT; orange: NN. The certificate covers the error in every batch but is loose, the span-based version being markedly sharper.

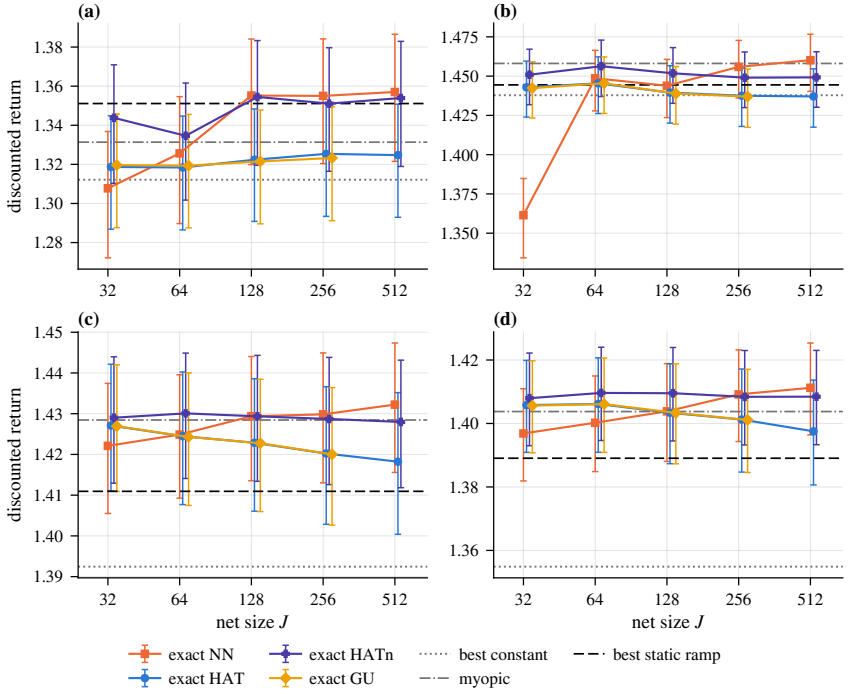

Figure 8: Test return of the exact finite-MDP policies against net size J, for decreasing action scales ζ (a)–(d): exact NN (orange), HAT (blue), HATn (purple), and GU (gold). Horizontal lines: best constant, myopic, and best static ramp controls. Exact HAT and GU are flat in J, while exact NN and HATn improve up to intermediate net sizes.

I.6 Transfer and cybersecurity figures

Figure 9 compares each frozen table with finite-population reference policies. Figure 10 instead compares the same policy in the finite and mean-field systems. These are different comparisons;

neither is a certificate of finite-population optimality. Figure 11 gives a bounded-noise variant of the botnet model in which a constant defence is optimal.

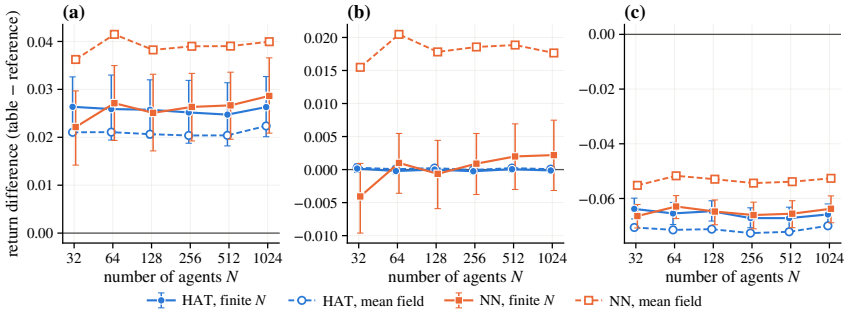

Figure 9: Return difference $V_N(\text{table}) - V_N(\text{reference})$ between the frozen HAT (blue) and NN (orange) tables and three finite-N reference policies, against N: (a) the validated constant action, (b) the exact HAT table, (c) REF-BPTT. Solid: finite-N gaps; dashed: gaps predicted by the mean-field model. Positive values favour the table; the references are comparators, not finite-N optima.

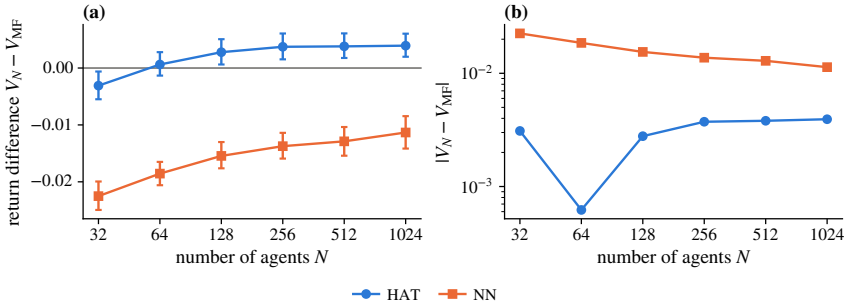

Figure 10: Same-policy transfer: return V_N of a frozen HAT (blue) or NN (orange) table with N agents minus its mean-field return on the same common-noise path, against N: (a) mean difference; (b) magnitude, for which no rate is claimed. The HAT difference changes sign and remains small; the NN difference shrinks but does not vanish over the tested range.

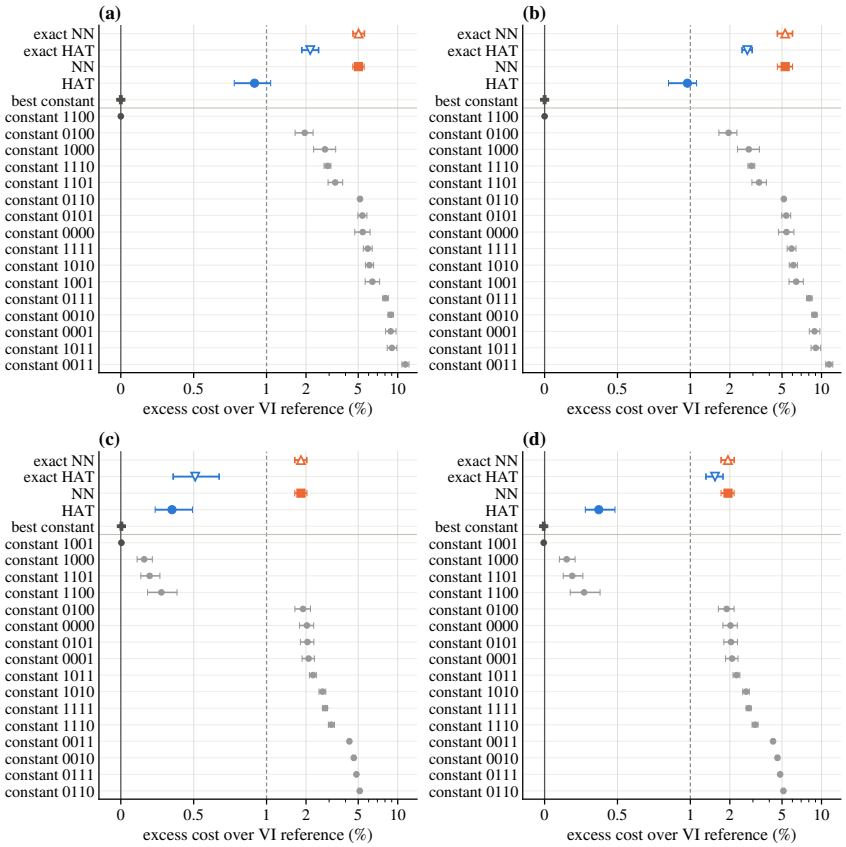

Figure 11: Bounded-noise botnet variant: relative excess cost over the fine-lattice value iteration reference (lower is better), in the nominal (a, b) and stress (c, d) regimes with one or two label blocks. Exact and learned NN (orange) and HAT (blue) tables, the validated best constant (dark grey) and all constant controls. The best constant has the lowest cost in every panel and matches the reference.

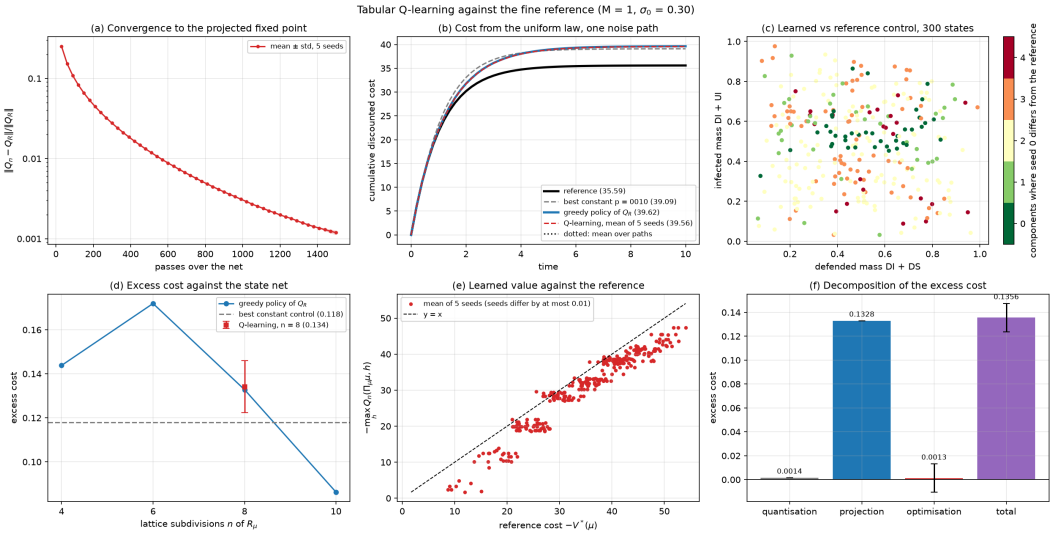

Figure 12: Botnet defence: (a) convergence to $Q_{\mathcal{R}}$; (b) cost on one noise path; (c) policy disagreement; (d) excess cost against the state net; (e) learned value against $-V^\star$; (f) decomposition of the excess cost.

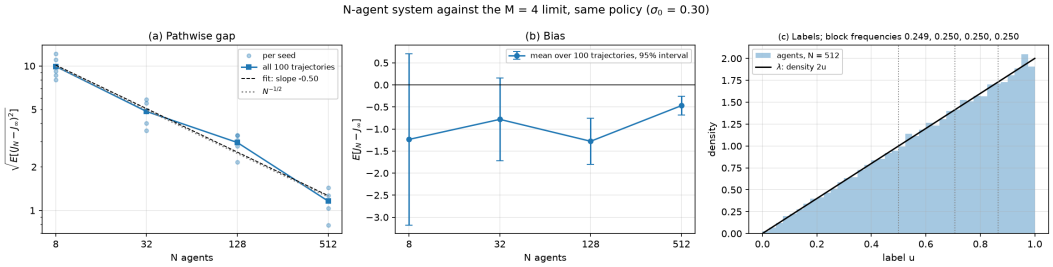

Figure 13: Botnet defence, finite-N transfer of the same frozen policy on paired common-noise paths: (a) pathwise gap $\mathbb{E}[(J_N - J_\infty)^2]^{1/2}$; (b) bias $\mathbb{E}[J_N - J_\infty]$; (c) empirical label marginal against λ.

I.7 Policy-class and deployment diagnostics

Figure 14 distinguishes the projected table value from the realised return as the structured kernel class grows. Figure 15 separates population-independent table storage from the cost of executing the policy for N agents. The sweep concerns finite projected problems, not a numerical proof of the unrestricted consistency theorem.

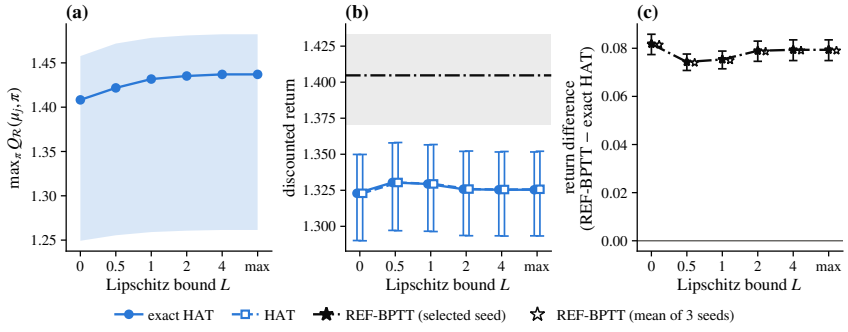

Figure 14: Lipschitz sweep for HAT over nested ramp subclasses of increasing L: (a) exact net value; (b) test return of exact and learned HAT against REF-BPTT; (c) gap to REF-BPTT, an upper estimate of the restriction plus projection error. The projected value increases with L, whereas the realised return does not and stays below REF-BPTT.

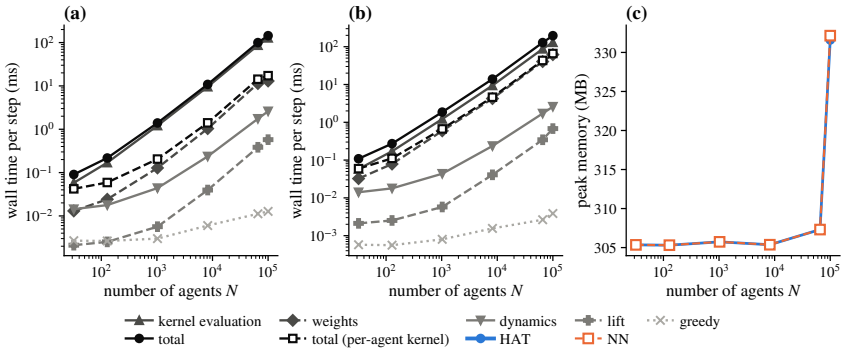

Figure 15: Deployment of the frozen HAT (a) and NN (b) tables: wall time per finite-N step, by component and in total, and (c) peak memory, against N. Execution cost grows about linearly in N, while table storage does not depend on N.

I.8 Computational record

All predeclared tasks were completed on a single host from a single code version; run configurations and seed-level outputs are archived. Invariants hold to machine precision, and the Bellman residuals and quadrature errors of the references are negligible. Truncated check runs of label-aware PPO are excluded from all conclusions. Table storage is independent of the population size, whereas execution cost grows linearly with it.

J Notation and dependency map

Symbol	Meaning
$\mathcal{P}_\lambda(I \times \mathcal{X})$	lifted state laws with fixed label marginal λ
$L^0(I \times \mathcal{X}; \mathcal{P}(A))$	
$L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$	Borel relaxed kernels; L-Lipschitz subclass
$\nu = \mu \otimes \pi$	induced label–state–action law $\mu(\mathrm{d}u, \mathrm{d}x)\pi(u, x)(\mathrm{d}a)$
K_0, c_0	$1 + 2L_F$ and $2L_f$ for the unrestricted state modulus
K_L, c_L	$(1 + 2L_F)(1 + L)$ and $2L_f(1 + L)$
g_L	uniform value loss caused by restricting to $L^0_L(I \times \mathcal{X}; \mathcal{P}(A))$
$e_\mathcal{R}$	error at finite state–action net points
$E_\mathcal{R}$	error of the nearest-neighbour extension
$\eta_\mathcal{R}$	one-step loss of the finite-net greedy policy for V_L
$\eta_\mathcal{R} + g_L$	one-step loss relative to $V^\star$ used for finite-N transfer
τ	target normalised state of the congestion benchmark

The proof dependencies are: Lemma 14 gives the direct constants; Lemma 13 turns them into Hölder bounds; these bounds justify the net bias; Theorem 7 adds the restriction error; and Corollary 10 supplies the verification residual used by the finite-population theorem.