A Separation Principle for Cooperative Multi-Agent Reinforcement Learning

Lucia Pezzetti
ETH Zürich
lucia.pezzetti@ai.ethz.ch

Nicolas Lanzetti
Caltech
lnicolas@caltech.edu

Antonio Terpin
ETH Zürich
aterpin@ethz.ch

Florian Dörfler
ETH Zürich
dorfler@ethz.ch

Giorgia Ramponi
University of Zürich
ramponi@ifi.uzh.ch

Abstract

We study cooperative multi-agent reinforcement learning (MARL) problems in which agents evolve under decoupled noisy dynamics but share a population objective. A canonical example is a mobility operator that orchestrates a fleet of vehicles to meet customer demand. The stochastic and time-varying nature of these problems calls for learning approaches, yet existing MARL algorithms do not scale: learning assignment and routing jointly over the full state-action space becomes prohibitive as the number of agents grows. Formulating the problem over the space of probability measures, we prove a separation principle: the population cost-to-go function is upper bounded by an optimal transport problem whose transportation cost is the target-conditioned single-agent cost-to-go function. This simplifies the multi-agent problem to learning a single-agent policy and coordinating the agents via optimal transport. Building on this principle, we introduce SALT (Separation-based Assignment and Learning via optimal Transport), a family of algorithms that learn a target-conditioned single-agent policy using any standard reinforcement learning algorithm, and then coordinate the agents via optimal transport. Since learning occurs at the single-agent level, SALT scales to fleet sizes where MARL baselines become intractable and transfers without retraining to fleet sizes unseen during training. In a vehicle-delivery experiment on the southern Manhattan road network, where MARL baselines are computationally prohibitive, SALT improves travel time by 15–21% over a shortest-path routing heuristic. In stochastic grid worlds, SALT reduces average travel times by 23–61% relative to MARL baselines and scales to population sizes where these methods fail.

1 Introduction

Consider a fleet of autonomous vehicles serving delivery requests in a city. At each decision time, the operator must simultaneously decide which vehicle should serve which request and how each vehicle should navigate to its assigned destination. These decisions are tightly coupled. The best assignment depends on how long each vehicle will actually take to reach each target, while routing itself depends on the assignment. Assignments based only on nominal shortest paths can therefore be poor when travel times depend on traffic lights, congestion, and random delays. Yet solving this full joint assignment-routing problem over the fleet quickly becomes computationally prohibitive. Beyond delivery fleets, the same structure arises across a wide range of cooperative systems: inspection drones achieving coverage patterns [1], warehouse robots repositioning to incoming orders [2], emergency response units relocating to optimal positions [3], ride-hailing platforms balancing supply and demand [4], and electric-vehicle fleets sharing charging infrastructure [5]. In all these settings,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

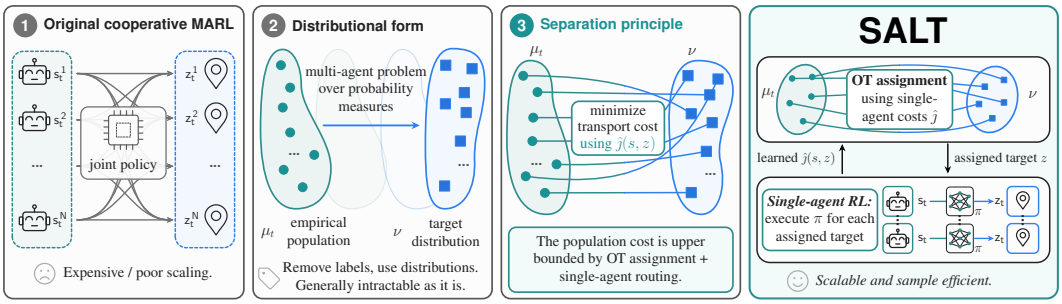


Figure 1: Overview of SALT. (1) Cooperative MARL learns a joint fleet policy and scales poorly with N. (2) Formulating the problem over probability measures replaces agent labels with an empirical population μ_t and target distribution ν, but remains intractable. (3) The separation principle decouples assignment from routing, upper bounding the population cost-to-go by an optimal transport assignment with cost $\hat{j}(s,z)$. SALT assigns targets through this optimal transport problem and routes agents with a shared single-agent policy π, yielding scalable and sample-efficient learning.

the goal is to optimally coordinate a fleet of decoupled agents so that their collective terminal state matches or approximates a target configuration, such as uniform coverage, a set of delivery locations, or a demand-weighted positioning pattern. These problems are typically solved in a receding-horizon fashion to account for changing conditions and request patterns and, thus, require modular approaches that generalize to different fleet sizes or target distributions.

Fleet assignment and routing are classically addressed via combinatorial solvers [6, 7] and, more recently, learned routing heuristics [8]. Since the joint assignment-routing problem is combinatorial in the fleet size, these methods decouple the two steps using nominal travel costs, such as shortest-path distances, to solve the assignment problem. These heuristics, however, ignore the stochasticity of the costs and may favor targets that "look close" but are expensive to reach in expectation. A natural learning-based alternative is multi-agent reinforcement learning (MARL) [9–12], often with centralized training and decentralized execution. Policy-gradient methods such as MAPPO [13] and MADDPG [14], and value-decomposition methods such as VDN [15], QMIX [16], and QPLEX [17] are effective in many settings [18, 19], but learn over a joint state-action space whose size grows exponentially with the number of agents [20], limiting scalability and generalization to unseen fleet sizes. Value-decomposition methods gain tractability by imposing architectural assumptions on the joint value function, such as additivity or monotonicity, which are provably too restrictive even for simple coordination tasks [21, 16]. Mean-field reinforcement learning [22, 23] instead replaces pairwise interactions with an aggregate representation of the population. Although practical in large systems [4, 24], its approximations can lose fidelity when modeling finite populations [25]. Closer to our motivation, some methods learn latent subtasks and dynamically assign them among agents [26], since in general cooperative MARL no such decomposition is available a priori. We instead study a structured subclass, with decoupled dynamics and a terminal optimal transport cost, in which the assignment-routing separation can be exploited directly.

In this work, we seek a principled separation of population-level assignment and single-agent control (Figure 1). Inspired by the "deterministic" separation principle for multi-agent control in [27], we formulate the cooperative MARL problem as an optimal control problem over the space of probability measures where terminal deviation from a desired target measure is quantified via an optimal transport cost. By construction, this representation is permutation invariant and captures both finite populations (as empirical measures) and a continuum of agents (as continuous measures). In the deterministic setting, Terpin et al. [27] show that the population cost-to-go is exactly an optimal transport cost whose transportation cost is the single-agent cost-to-go, thus separating assignment from single-agent control. This equality does not extend directly to the stochastic setting, where the population measure is random and future assignments can adapt to realized noise. Nevertheless, we show that a carefully designed separation principle, where noise is "averaged out", continues to upper bound the population cost-to-go. This analysis is non-trivial: the randomness of the population measure requires handling probability measures over probability measures, a modeling challenge in itself, compounded by delicate measurability issues. Finally, Terpin et al. [27] provide a purely theoretical analysis, whereas we leverage our separation principle to design scalable algorithms for MARL.

Contributions. Our contributions are as follows:

1. We cast the cooperative multi-agent control problem with homogeneous agents, decoupled stochastic dynamics, and a terminal optimal transport formation cost as an optimal control problem over probability measures in the spirit of the work of Terpin et al. [27]. This removes agent labels, makes the formulation permutation invariant, and applies to both finite populations and a continuum of agents.
2. We prove a separation principle for this optimal control problem (Theorem 3.1, Part I). The population cost-to-go is upper bounded by an optimal transport problem whose transportation cost is the target-conditioned single-agent cost-to-go. This separates target allocation from routing, which can be solved independently of the fleet size. The upper bound becomes exact in the deterministic setting, recovering [27, Corollary 4.2]. We then bound the gap between the two sides of Theorem 3.1, Part I in terms of the noise of the single-agent dynamics (Theorem D.4), and the loss of optimality under weak population interactions (Theorem E.6). Importantly, our proof is constructive and yields an implementable policy.
3. This separation principle yields SALT, a family of MARL algorithms that learn a shared target-conditioned single-agent policy and coordinate agents via optimal transport. SALT is agnostic to the single-agent algorithm and generalizes to population sizes not seen during training.
4. We validate SALT on a time-varying southern Manhattan road network with 2087 intersections and 4255 directed roads. SALT trained with $N = 20$ vehicles transfers without retraining to fleets of 1–100 vehicles and reduces travel time by 15–21% relative to a nominal shortest-path heuristic. In controlled stochastic grid worlds, the SALT variants reduce the average travel time by 23–61% relative to the standard MARL baselines, consistently across different single-agent reinforcement learning (RL) algorithms. Additionally, SALT improves sample efficiency: with PPO [28], it reaches the same reach-rate milestones as MAPPO with up to $480\times$ fewer environment interactions, and scales to $N = 128$, while MAPPO fails at $N = 8$ within a 2×10^9-interaction budget.

Notation. We work with Polish spaces, Borel functions, and Borel probability measures. For a Polish space X, $\mathcal{P}(X)$ denotes the space of probability measures on X. We write δ_x for the Dirac measure at $x \in X$. The projection from $X \times Y$ onto X is denoted by $\mathrm{proj}_X^{X\times Y}$. For $\mu \in \mathcal{P}(X)$ and $F\colon X \to Y$, the pushforward of μ through F, $F_\#\mu \in \mathcal{P}(Y)$, is defined by $F_\#\mu(B) = \mu(F^{-1}(B))$ for all Borel sets $B \subseteq Y$. A coupling between the probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ is a joint probability measure with marginals μ and ν; the set of couplings between μ and ν is $\Gamma(\mu, \nu)$. For a nonnegative cost $c\colon X \times Y \to \bar{\mathbb{R}}_{\geq 0}$, the optimal transport cost between μ and ν is

$$\mathcal{K}[c](\mu, \nu) \coloneqq \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c(x, y)\mathrm{d}\gamma(x, y).$$

If $k\colon X \to \mathcal{P}(Y)$ is a Markov kernel, $\mu \otimes k$ denotes the induced joint law on $X \times Y$. In particular, if $\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)k(x; \mathrm{d}y)$ is a disintegration, we write $\pi = \mu \otimes k^x$.

2 Problem setting

We consider cooperative MARL problems in which N homogeneous agents coordinate to reach a desired configuration over a finite horizon T, while incurring agent-level stage costs and a population-level terminal cost. Let S, A, W, and Z denote the state, action, noise, and target spaces. While $Z = S$ for canonical target matching, we keep Z distinct to account for lower-dimensional targets or task-level descriptors, such as destinations when the state also includes velocity or battery level.

2.1 Agent-based formulation

Each agent i has state $s_t^i \in S$ and chooses action $a_t^i \in A$. Its noisy dynamics are $s_{t+1}^i = f_t(s_t^i, a_t^i, w_t^i)$, where $w_t^i \sim \xi_t(\cdot \mid s_t^i, a_t^i)$ and $\xi_t(\cdot \mid s_t^i, a_t^i) \in \mathcal{P}(W)$ gives the conditional law of the noise affecting an agent in state s taking action a. We consider three noise models, which differ in how noise is correlated across agents: *individual noise*, where noise is sampled independently across agents; *local noise*, where agents with the same state-action pair share the same noise realization; and *global noise*, where all agents share a common noise realization (see Appendix A). Let $\boldsymbol{s}_t \coloneqq (s_t^1, \ldots, s_t^N)$ and $\boldsymbol{a}_t \coloneqq (a_t^1, \ldots, a_t^N)$ denote the ordered states and actions of the agents. We take the control cost to be

the sum of agent-level stage costs $g_t\colon S \times A \times W \to \bar{\mathbb{R}}_{\geq 0}$ and a population-level terminal formation cost. Given a transportation cost $g_T : S \times Z \to \bar{\mathbb{R}}_{\geq 0}$, the terminal cost compares the final population distribution $\mu_T = \frac{1}{N}\sum_{i=1}^{N} \delta_{s_T^i}$ to a target configuration $\nu \in \mathcal{P}(Z)$ through the optimal transport cost

$$\mathcal{K}[g_T](\mu_T, \nu) = \inf_{\gamma \in \Gamma(\mu_T, \nu)} \int_{S \times Z} g_T(s, z) \mathrm{d}\gamma(s, z). \tag{1}$$

For $g_T(s, z) = \|s - z\|^p$ with $p \geq 1$, $\mathcal{K}[g_T](\mu_T, \nu)$ is the p-th power of the type-p Wasserstein distance. Since this is a fully observed cooperative problem, it can be cast as a multi-agent Markov decision process (MDP) [29], which reduces to a single-agent MDP over the joint action space. By standard MDP theory [30, 31], we can restrict attention, without loss of optimality, to deterministic Markov policies $\pi = (\pi_t)_{t=0}^{T-1}$, universally measurable on general Borel state and action spaces [31, Ch. 8]. We denote this class by Π. Throughout, we assume that, given the current states and actions, the noise is independent of the past, so the law of the noise sequence $(\boldsymbol{w}_t)_{t=0}^{T-1}$, $\boldsymbol{w}_t \coloneqq (w_t^1, \ldots, w_t^N)$, under which expectations are taken, is determined by the dynamics, the policy, and the noise model.

Problem 1. *Given initial states $\bar{\boldsymbol{s}}_0 = (\bar{s}_0^1, \ldots, \bar{s}_0^N)$ and a target distribution ν, learn a policy $\pi \in \Pi$ minimizing the expected cumulative cost*

$$J^N(\bar{\boldsymbol{s}}_0, \nu) = \inf_{\pi \in \Pi} \mathbb{E}_{(\boldsymbol{w}_t)_{t=0}^{T-1}} \left[\frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} g_t(s_t^i, a_t^i, w_t^i) + \mathcal{K}[g_T](\mu_T, \nu) \right]$$
$$\text{s.t.} \quad s_{t+1}^i = f_t(s_t^i, a_t^i, w_t^i), \qquad \boldsymbol{s}_0 = \bar{\boldsymbol{s}}_0.$$

2.2 Population-based formulation

We generalize Problem 1 at the *population level* by replacing the ordered state $\boldsymbol{s}_t$ with the empirical measure $\mu_t = \frac{1}{N}\sum_{i=1}^{N} \delta_{s_t^i} \in \mathcal{P}(S)$, which discards agent labels and makes the permutation invariance in the problem explicit. We initially allow general population states $\mu_t \in \mathcal{P}(S)$ and we will later specialize this formulation back to the finite-agent setting.

Population control. A population-level control must specify how the mass at each state is split among actions. Such a control is a *state-action distribution* $\lambda_t \equiv \lambda_t(\mu_t) \in \mathcal{P}(S \times A)$ whose S-marginal is the current population state: $\left(\mathrm{proj}_S^{S \times A}\right)_{\#} \lambda_t = \mu_t$. For example, if all agents follow a common deterministic policy $\pi_t\colon S \to A$, then $\lambda_t = (\mathrm{Id}_S, \pi_t)_{\#}\mu_t$, that is, all the mass that μ_t places at state s is assigned to the state-action pair $(s, \pi_t(s))$.

Noise and population dynamics. When dynamics are deterministic, the next population state is the pushforward of λ_t under the dynamics, i.e., $\mu_{t+1} = f_{t\#}\lambda_t$. The stochastic case instead requires more care, as the transition also depends on the specific noise realizations across the population. In the agent-based formulation, each agent i draws $w_t^i \sim \xi_t(\cdot \mid s_t^i, a_t^i)$ and transitions to $f_t(s_t^i, a_t^i, w_t^i)$. At the population level, these individual state-action dependent noise draws cannot simply be replaced by their underlying distribution $\xi_t(\cdot \mid s, a)$. Indeed, the next population state depends on the *collection of noise draws actually realized* across all agents, which is itself a random object. We therefore represent the realized noise draws at each state-action pair by a state-action dependent *empirical* noise distribution $\hat{\xi}_t^{s,a} \in \mathcal{P}(W)$. The distinction between $\xi_t(\cdot \mid s, a)$ and $\hat{\xi}_t^{s,a}$ is crucial: the former is the single-agent noise law, while the latter collects the noise realizations that drive the population transition. Collecting the realized empirical kernels $\hat{\xi}_t^{s,a}$ across all state-action pairs, we define the joint state-action-noise distribution ρ_t and the population dynamics as

$$\rho_t = \lambda_t \otimes \hat{\xi}_t^{s,a} \in \mathcal{P}(S \times A \times W), \qquad \mu_{t+1} = f_{t\#}\rho_t,$$

where ρ_t records the fraction of the population experiencing each triple (s, a, w).

Example 2.1. *Consider three agents with state-action distribution $\lambda_t = \frac{2}{3}\delta_{(0,a)} + \frac{1}{3}\delta_{(1,a)}$, i.e., two agents are in state 0 and use action a, and one agent is in state 1 and uses action a. Assume that the agents are subject to individual noise and that the noise realizations for the two agents at $(0, a)$ are 0 and 1, while the noise realization for the agent at $(1, a)$ is 1. Then, $\hat{\xi}_t^{0,a} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $\hat{\xi}_t^{1,a} = \delta_1$. Accordingly, the state-action-noise distribution is $\rho_t \coloneqq \lambda_t \otimes \hat{\xi}_t^{s,a} = \frac{1}{3}\delta_{(0,a,0)} + \frac{1}{3}\delta_{(0,a,1)} + \frac{1}{3}\delta_{(1,a,1)}$ and the next population distribution is $\mu_{t+1} = (f_t)_{\#}\rho_t = \frac{1}{3}\delta_{f_t(0,a,0)} + \frac{1}{3}\delta_{f_t(0,a,1)} + \frac{1}{3}\delta_{f_t(1,a,1)}$.*

Cost. Since $\frac{1}{N}\sum_{i=1}^{N} g_t(s_t^i, a_t^i, w_t^i) = \int_{S\times A\times W} g_t(s,a,w)\mathrm{d}\rho_t(s,a,w)$, the cost of a given control $\lambda = (\lambda_t)_{t=0}^{T-1}$ is

$$J^{\lambda}(\bar{\mu}_0, \nu) = \mathbb{E}_{(\hat{\xi}_t)_{t=0}^{T-1}}\left[\sum_{t=0}^{T-1}\int_{S\times A\times W} g_t(s,a,w)\mathrm{d}\rho_t(s,a,w) + \mathcal{K}[g_T](\mu_T,\nu)\right],$$

where $\hat{\xi}_t = (\hat{\xi}_t^{s,a})_{(s,a)\in S\times A}$ collects all realized noise kernels at time t. The expectation is then taken over the randomness of these profiles. This keeps finite-population fluctuations and noise correlations explicit while reducing the population-level randomness to the single object $\hat{\xi}_t$. We refer the reader to Appendix A for additional details and explicit construction for canonical noise models. Since μ_t is random for $t \geq 1$, an *admissible population control* is a sequence of (universally measurable) feedback maps $\mu \mapsto \lambda_t(\mu)$ with S-marginal μ, applied before $\hat{\xi}_t$ is realized. We write λ_t for $\lambda_t(\mu_t)$.

Problem 2. *Given an initial population distribution $\bar{\mu}_0$ and a target distribution ν, find an admissible population control $(\lambda_t)_{t=0}^{T-1}$ that minimizes the expected cumulative cost*

$$J(\bar{\mu}_0,\nu) = \inf_{(\lambda_t)_{t=0}^{T-1}} \mathbb{E}_{(\hat{\xi}_t)_{t=0}^{T-1}}\left[\sum_{t=0}^{T-1}\int_{S\times A\times W} g_t(s,a,w)\mathrm{d}\rho_t(s,a,w) + \mathcal{K}[g_T](\mu_T,\nu)\right]$$
$$\text{s.t.}\quad \mu_0 = \bar{\mu}_0, \qquad \mu_{t+1} = f_{t\#}\rho_t, \qquad (\mathrm{proj}_S^{S\times A})_{\#}\lambda_t = \mu_t, \qquad \rho_t = \lambda_t \otimes \hat{\xi}_t^{s,a}.$$

The population formulation removes the artificial labels from the state description and generalizes Problem 1. Indeed, any finite-agent policy induces an empirical state-action distribution $\lambda_t = \frac{1}{N}\sum_{i=1}^{N}\delta_{(s_t^i,a_t^i)}$ with S-marginal μ_t, thus every policy admissible for Problem 1 is feasible in Problem 2 and $J(\bar{\mu}_0,\nu) \leq J^N(\bar{s}_0,\nu)$. The converse inequality does not hold in general. Problem 2 allows for arbitrary population state-action distributions, including those with more than N atoms, which are not realizable by N agents. However, we show in Corollary 3.3 that, for empirical target distributions ν^N (i.e., N targets with equal mass), the separation bound derived for Problem 2 directly upper bounds the agent-based cost-to-go $J^N(\bar{s}_0,\nu^N)$, yielding a deployable policy for Problem 1.

3 Separation principle

In this section, we prove a separation principle for Problem 2, showing that the population cost-to-go function separates into optimal single-agent control toward a target and optimal target assignment across the population; see Theorem 3.1 below. We then specialize the result to the finite-agent value J^N of Problem 1 and obtain a deployable policy that we use in Section 4 for algorithm design.

Target-conditioned single-agent problem. Fix a target $z \in Z$. For an agent in state $s \in S$ at time t, define $j_t(s,z)$ as the optimal cost-to-go for reaching z over the remaining horizon:

$$j_t(s,z) = \inf_{(\pi_\tau)_{\tau=t}^{T-1}} \mathbb{E}_{(w_\tau)_{\tau=t}^{T-1}}\left[\sum_{\tau=t}^{T-1} g_\tau(s_\tau,a_\tau,w_\tau) + g_T(s_T,z)\right], \qquad (2)$$
$$\text{s.t.}\quad s_t = s, \quad s_{\tau+1} = f_\tau(s_\tau,a_\tau,w_\tau), \quad w_\tau \sim \xi_\tau(\cdot \,|\, s_\tau,a_\tau).$$

This is a standard single-agent stochastic optimal control problem, whose complexity is independent of the population size N. Its Bellman recursion is

$$j_t(s_t,z) = \inf_{a_t\in A} \mathbb{E}_{w_t}\left[g_t(s_t,a_t,w_t) + j_{t+1}(f_t(s_t,a_t,w_t),z)\right], \quad j_T(s_T,z) = g_T(s_T,z). \qquad (3)$$

A separation principle. For $\mu_t \in \mathcal{P}(S)$, let $J_t(\mu_t,\nu)$ denote the optimal expected cost at time t, when the population μ_t aims to match the target distribution ν. At $t = T$, $J_T(\mu,\nu) = \mathcal{K}[g_T](\mu,\nu)$. Fix $t < T$. Since $\mathcal{P}(S)$ is a Polish space whenever S is Polish [32, Theorem 15.15], the cost is additive over time, and the population dynamics are Markovian, standard dynamic programming yields the population-level Bellman recursion [31, Ch. 8]

$$J_t(\mu_t,\nu) = \inf_{\lambda_t:\,(\mathrm{proj}_S^{S\times A})_{\#}\lambda_t=\mu_t} \mathbb{E}_{(\hat{\xi}_t)_{t=0}^{T-1}}\left[\int_{S\times A\times W} g_t(s,a,w)\,\mathrm{d}\rho_t(s,a,w) + J_{t+1}(f_{t\#}\rho_t,\nu)\right]. \qquad (4)$$

Theorem 3.1 (Separation principle, Part I: upper bound for the population cost-to-go)**.** *Consider Problem 2. For all $t \in \{0, \dots, T\}$, the population cost-to-go function* (4) *is upper bounded by an optimal transport problem whose transportation cost is the single-agent cost-to-go function j_t in* (2)*:*

$$J_t(\mu_t, \nu) \leq \mathcal{K}[j_t](\mu_t, \nu) \; = \inf_{\gamma \in \Gamma(\mu_t, \nu)} \int_{S \times Z} j_t(s, z) \mathrm{d}\gamma(s, z). \tag{5}$$

In particular, combined with the admissible control constructed in Part II below, the same upper bound holds for the optimal cost $J(\bar{\mu}_0, \nu)$ of Problem 2.

Discussion. The upper bound (5) separates the stochastic population control problem into target-conditioned single-agent control and population-level assignment, in which j_t serves as the transport cost. This separation generally comes at the price of an inequality because the two problems account for future noise differently. In the optimal transport cost, $\mathcal{K}[j_t](\mu_t, \nu)$, mass is assigned to targets based on $j_t(s, z)$, which *averages* over future noise realizations. By contrast, the Bellman recursion (4) evaluates the remaining cost $J_{t+1}(f_{t\#}\rho_t, \nu)$ at the realized future population $f_{t\#}\rho_t$, allowing the population-level assignment to *adapt* to the observed transition. The resulting minimization and expectation do not commute, which is the cause of the inequality (5). At the same time, averaging over the noise is necessary for a meaningful separation, since it preserves the structure of a standard stochastic optimal control problem in (2). Appendix D quantifies the tightness of this inequality: if $j_{\tau+1}$ is $L_{\tau+1}$-Lipschitz in the state, the gap $\mathcal{K}[j_t](\mu_t, \nu) - J_t(\mu_t, \nu)$ in (5) is bounded by $\sum_{\tau=t}^{T-1} L_{\tau+1} D_\tau$, where D_τ is the worst-case expected distance between two (conditionally independent) next states generated from the same state-action pair. The gap therefore depends linearly on the dispersion, and vanishes whenever dynamics are deterministic. In particular, under deterministic dynamics, the next population state is fully determined by (μ_t, λ_t), the expectation in (4) disappears, and the upper bound in Theorem 3.1, Part I becomes exact, recovering the deterministic separation of [27].

Corollary 3.2 (Exact separation in the deterministic setting [27, Corollary 4.2])**.** *When the dynamics are deterministic, the population cost-to-go function* (4) *is an optimal transport cost with the single-agent cost-to-go function as its transportation cost:*

$$J_t(\mu_t, \nu) = \inf_{\gamma \in \Gamma(\mu_t, \nu)} \int_{S \times Z} j_t(s, z) \mathrm{d}\gamma(s, z).$$

From a bound to a policy. For any $\varepsilon > 0$, Theorem 3.1, Part II constructs an explicit population-level control $(\lambda_t^\varepsilon)_{t=0}^{T-1}$ whose cost exceeds the bound by at most ε. Here, $J^\lambda(\bar{\mu}_0, \nu)$ denotes the expected cumulative cost of Problem 2 when the population follows the control $(\lambda_t)_{t=0}^{T-1}$.

Theorem 3.1 continued (Separation principle, Part II: constructive separation-based control). *For every $\varepsilon > 0$ and each time step t, let $\gamma_t^\varepsilon \in \Gamma(\mu_t, \nu)$ be an $\frac{\varepsilon}{2T}$-optimal coupling for the transport problem $\mathcal{K}[j_t](\mu_t, \nu)$ and let $a_t^\varepsilon : S \times Z \to A$ be an $\frac{\varepsilon}{2T}$-optimal single-agent action for the Bellman recursion* (3)*. Define the separation-based state-action distribution $\lambda_t^\varepsilon \coloneqq (\mathrm{proj}_S^{S \times Z}, a_t^\varepsilon)_\# \gamma_t^\varepsilon$. Then,*

$$J^{\lambda^\varepsilon}(\bar{\mu}_0, \nu) \leq \inf_{\gamma \in \Gamma(\bar{\mu}_0, \nu)} \int_{S \times Z} j_0(s, z) \mathrm{d}\gamma(s, z) + \varepsilon.$$

In particular, if optimal couplings $\gamma_t^\star$ and optimal single-agent actions $a_t^\star$ exist and can be selected measurably, we can choose $\varepsilon = 0$.

This construction has a natural interpretation. Assuming, for simplicity, that optimal couplings and actions exist (e.g., all spaces compact and all functions continuous; see the remark in Appendix C), the coupling $\gamma_t^\star \in \Gamma(\mu_t, \nu)$ optimally assigns the population mass at each state s to targets, while $a_t^\star$ selects the optimal action for each state–target pair. This is the core of our separation principle: the coupling $\gamma_t^\star$ indicates how to dispatch agents to targets, while $a_t^\star$ indicates how to route the single agent to the assigned target. Part II of Theorem 3.1 guarantees that the resulting population control has expected cost no greater than the transport upper bound established in Part I.

Algorithm 1 SALT: Separation-based Assignment and Learning via optimal Transport

Require: Initial target-conditioned policy $\pi^{(0)}(a \mid s, z)$, cost-to-go function estimates $\hat{j}_t^{(0)}(s, z)$, horizon T, exploration schedule $(\alpha_k)_{k \geq 0}$

1: **for** episodes $k = 0, 1, 2, \ldots$ **do**
2: Sample the episode initial population $\mu_0^{(k)}$ and target distribution ν_k.
3: **Initial population-level assignment.** With probability $1 - \alpha_k$, assign agents to targets via optimal transport with cost $\hat{j}_0^{(k)}$. Otherwise assign randomly.
4: **Single-agent learning.** Roll out the shared target-conditioned policy $\pi^{(k)}$ in parallel for all agents over the horizon T, with the assigned targets kept fixed.
5: Update $\pi^{(k+1)}$ and $\hat{j}_t^{(k+1)}(s, z)$ with the chosen single-agent RL algorithm from the N trajectories collected in Step 4.
6: **end for**
7: **Deployment (online-assignment variant).**
8: **for** $t = 0, \ldots, T-1$ **do**
9: Assign agents to targets via optimal transport with cost $\hat{j}_t$.
10: Execute one step of π_t with the current target assignments.
11: **end for**

Robustness to weakly coupled dynamics. While Theorem 3.1 assumes decoupled dynamics and stage costs, Appendix E shows that the separation-based policy remains near-optimal when both depend weakly on the population. In particular, we construct the separation-based policy of Theorem 3.1, Part II using a decoupled surrogate and deploy it in the coupled system. Let ε_τ^f bound the distance between the next states the surrogate and the coupled system produce from the same state, action, and noise, let δ_τ the expected difference between their stage costs, and let L_τ as above. The loss of optimality is then at most $\sum_{\tau=0}^{T-1} L_{\tau+1} D_\tau + 2 \sum_{\tau=0}^{T-1} (\delta_\tau + L_{\tau+1} \varepsilon_\tau^f)$. The first term is the bound on the gap in (5) derived above, while the second isolates the loss caused by population interactions, and vanishes with them.

Implications for agent-based formulation J^N. Theorem 3.1 is stated for the population formulation in Problem 2. Corollary 3.3 bridges this to the original agent-based setting: for N agents and N target masses, the same transport surrogate $\mathcal{K}[j_t]$ upper bounds J_t^N.

Corollary 3.3 (Agent-based upper bound)**.** *Let $\boldsymbol{s}_t = (s_t^1, \ldots, s_t^N) \in S^N$ be the agent states at time t, with empirical distribution $\mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{s_t^i}$. Let $\boldsymbol{z} = (z^1, \ldots, z^N) \in Z^N$ be fixed targets with empirical distribution $\nu^N = \frac{1}{N} \sum_{i=1}^N \delta_{z^i}$. Then, the cost-to-go function for the optimal control problem in Problem 1, denoted by $J_t^N(\boldsymbol{s}_t, \nu^N)$, satisfies*

$$J_t^N(\boldsymbol{s}_t, \nu^N) \;\leq\; \mathcal{K}[j_t](\mu_t^N, \nu^N). \tag{6}$$

In particular, the inequality holds true for the cost-to-go $J^N(\bar{\boldsymbol{s}}_0, \nu^N)$ in Problem 1.

We defer the proofs of Theorem 3.1, Part I, Theorem 3.1, Part II, and Corollary 3.3 to Appendix C.

4 SALT: Separation-based Assignment and Learning via optimal Transport

Corollary 3.3 suggests using the upper bound

$$\mathcal{K}[j_t](\mu_t, \nu) \;=\; \inf_{\gamma \in \Gamma(\mu_t, \nu)} \int j_t(s, z) \mathrm{d}\gamma(s, z)$$

as a tractable surrogate for the agent-based cost-to-go $J_t^N(\boldsymbol{s}_t, \nu^N)$. The resulting learning paradigm, summarized in Algorithm 1, has two phases: (i) learn the single-agent target-conditioned cost-to-go function $\hat{j}_t$, and (ii) use it as the transport cost to assign agents to targets. The choice of single-agent algorithm is entirely open: any method capable of learning a target-conditioned cost-to-go estimate $\hat{j}_t$ (Q-learning [33], DDQN [34], PPO [28], A2C [35], etc.) can be plugged in as the base learner. We refer to this family of methods as SALT, for Separation-based Assignment and Learning via optimal Transport. An instance using a single-agent RL algorithm is denoted SALT-ALG; e.g., SALT-PPO.

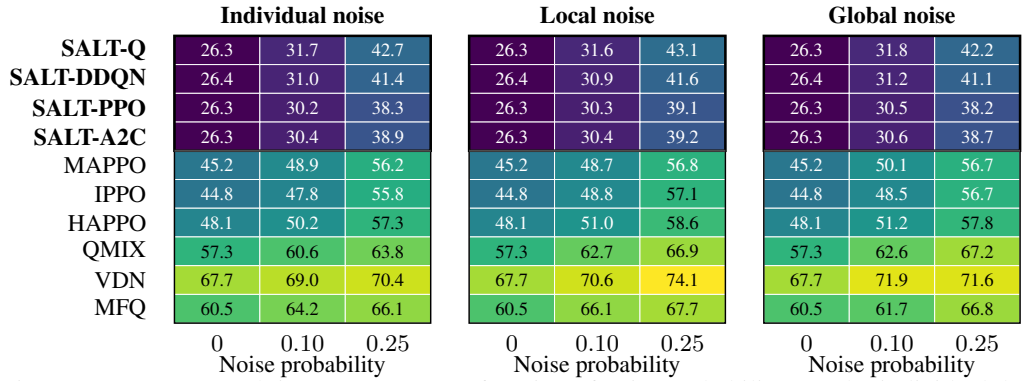


Figure 2: Average travel time to targets as a function of noise probability p under individual, local, and global noise. Cells show the pooled arithmetic mean over 50 evaluation rollouts from each of 5 independently trained policies. Figure 7 shows the corresponding distributions. Brighter cells indicate worse relative performance. SALT variants consistently outperform standard MARL baselines.

Discussion. SALT replaces learning over the joint population state-action space with a target-conditioned single-agent policy $\pi(\cdot \mid s, z)$ and cost-to-go estimate $\hat{j}(s, z)$ shared across all agents. Thus, population-level coordination is handled only through the optimal transport assignment, while both learned functions depend only on a single agent's state–target pair. This has three main benefits. (i) Scalability: learning occurs entirely at the target-conditioned single-agent level, avoiding the curse of dimensionality of the joint state-action space and generalizing naturally across population sizes. (ii) Trajectory pooling: the same single-agent model is updated from every agent's trajectory simultaneously. (iii) Modularity: the coordination layer requires only $\hat{j}(s, z)$ as its transport cost, making SALT compatible with any single-agent RL algorithm.

The time-indexed bound in Theorem 3.1, Part I makes the assignment frequency a deployment choice. The transport assignment can be recomputed at every time step with $\hat{j}_t$, as in Algorithm 1, or solved once at $t = 0$ and kept fixed for the entire episode. Both variants use the same target-conditioned learner and differ only in when the transport assignment is updated.

Training details. At each episode k, we sample the initial population $\mu_0^{(k)}$ and target distribution ν_k from the environment, and assign agents to targets by solving the optimal transport problem $\mathcal{K}[\hat{j}_0^{(k)}](\mu_0^{(k)}, \nu_k)$ with the current initial-time estimate $\hat{j}_0^{(k)}(s, z)$ as cost. For empirical population and target measures with N equally weighted samples, a coupling $\gamma \in \Gamma(\mu_t, \nu)$ can be represented as a doubly stochastic matrix (up to the factor $1/N$). By the Birkhoff–von Neumann theorem an optimal coupling can be taken to be a permutation (see the proof of Corollary 3.3), and thus the optimal transport problem reduces to a *one-to-one assignment* problem. In particular, given a cost matrix $C_{ij} = \hat{j}_t^{(k)}(s^i, z^j)$, we solve $\min_{\sigma \in S_N} \sum_{i=1}^{N} \hat{j}_t^{(k)}(s^i, z^{\sigma(i)})$, where S_N is the set of permutations of $\{1, \ldots, N\}$, and assign target $z^{\sigma(i)}$ to agent i. Such an assignment is computed exactly in $O(N^3)$ with the Hungarian algorithm, or approximately using entropic regularization [36]. During training we apply this construction at $t = 0$ with probability $1 - \alpha_k$, else we assign targets uniformly at random to ensure broad coverage of (s, z) pairs and prevent the cost-to-go estimate from overfitting to its own assignments. We then keep the assignments fixed for the entire episode. Each agent therefore faces a single-agent target-conditioned task with terminal cost $g_T(s_T, z^i)$, ensuring that the collected transitions are consistent with the cost-to-go $j_t(s, z^i)$ being learned. Once targets are assigned, agents roll out the shared target-conditioned policy in parallel, producing N trajectories that are pooled to update the shared policy and cost-to-go estimate using the chosen single-agent algorithm.

5 Experimental results

We evaluate SALT against MARL baselines in stochastic grid worlds, and test scalability and fleet-size transfer in a Manhattan urban logistics task. More ablations and experiment details are in Appendix F. The code for our experiments is available at `https://github.com/lucia-pezzetti/SALT`.

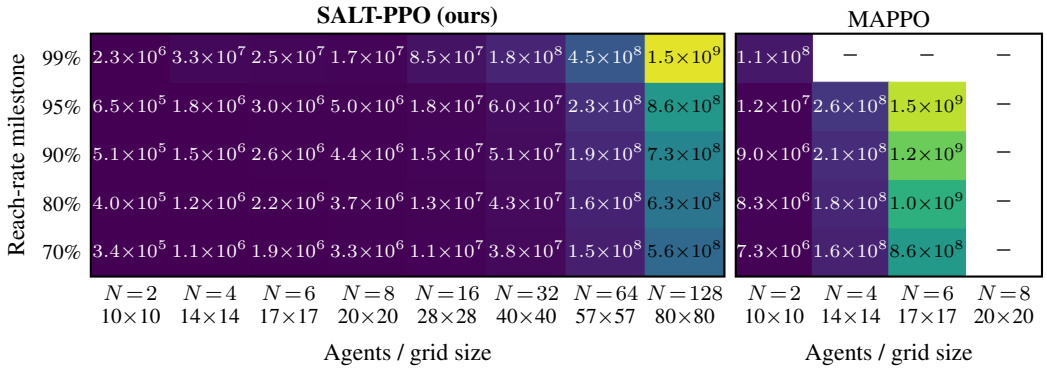


Figure 3: Training environment interactions required by SALT-PPO and MAPPO to attain each reach-rate milestone (70–99%) for 10 consecutive batches as population size N increases (grid area scaled proportionally to maintain target density). Each cell reports the mean over 5 training seeds, a dash (–) indicates that no seed attained the milestone within the training budget.

5.1 Stochastic grid-world experiments

Separation leads to better performance. We first evaluate SALT on a stochastic grid-world environment. A population of $N = 7$ homogeneous agents must reach N target locations within a horizon $T = 100$ on an 11×30 grid. Agents start in the first column, while targets are uniformly sampled from the last ten columns. At each step, the intended action is replaced uniformly by one of the other four actions with probability p. We evaluate the *individual*, *local*, and *global* noise models of Section 2, according to which replacement draws are independent across agents, shared among agents with the same state-action pair, or shared across the population. The objective is to minimize average travel time and terminal mismatch to the target distribution. We instantiate SALT with Q-learning [33] (SALT-Q), DDQN [34] (SALT-DDQN), PPO [28] (SALT-PPO), and A2C [35] (SALT-A2C), and compare them against six MARL baselines: MAPPO [13], IPPO [37], HAPPO [38], VDN [15], QMIX [16], and mean-field Q-learning (MFQ) [22]. At evaluation, SALT uses the online-assignment variant in Algorithm 1, recomputing the transport assignment at every time step. All methods are trained with 5 seeds and evaluated on 50 evaluation seeds per trained policy. Figure 2 shows that SALT consistently outperforms the MARL baselines across all noise levels and noise models. The four SALT variants perform similarly, suggesting that the gain comes from the separation structure rather than from the particular single-agent RL algorithm. Complementary metrics, as well as the complete training, evaluation, and uncertainty details, are reported in Appendix F.1.

Separation leads to better scalability. We next study sample complexity as the number of agents increases from $N = 2$ to $N = 128$. We use the same grid-world dynamics with individual noise, and increase grid area to keep the target density roughly constant and isolate the effects of scaling from the complexity of the task. During training, agents start from independently sampled rows of the first column, and targets are sampled uniformly across the grid (see Appendix F.2). We select PPO as the single-agent learner in SALT, yielding SALT-PPO, and compare it against MAPPO. We report the number of environment interactions required to attain reach-rate milestones of 70%, 80%, 90%, 95%, and 99%, up to a maximum training budget of 2×10^9 interactions. Figure 3 shows that SALT-PPO attains the same milestones with up to $480\times$ fewer interactions than MAPPO, and reaches the 99% milestone for populations of up to $N = 128$ agents, while MAPPO fails to reach the 70% milestone at $N = 8$ within the configured budget. This gap directly reflects their different scalability properties. MAPPO learns over a joint multi-agent state-action space, whose size grows exponentially with N, whereas SALT-PPO learns a target-conditioned policy that is independent of N.

5.2 Towards a realistic time-varying stochastic urban logistics task

A fleet of delivery vehicles must serve requests across Manhattan while jointly minimizing total travel time. This task couples vehicle-to-request assignment with time-efficient routing through a stochastic, time-varying network and is precisely the setting that motivated this work. SALT separates the problem into a coordination step (assigning vehicles to requests via optimal transport) and a routing step (navigating each vehicle to its current target via a shared target-conditioned learned

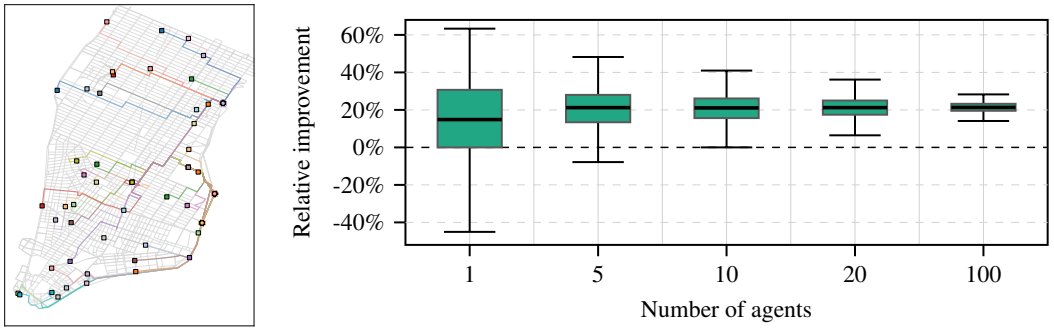


Figure 4: Relative travel-time improvement of SALT-Q over static shortest-path routing on the time-varying Manhattan road network, as a function of fleet size. Boxplots pool 500 evaluation instances from each of 8 training seeds. Boxes span the interquartile range, center lines show medians, and whiskers follow the Tukey 1.5-interquartile-range rule. The dashed line indicates zero improvement. SALT-Q is trained with 20 vehicles and evaluated without retraining on fleets of 1–100 vehicles.

policy). As motivated in Section 4, assignments remain fixed during each training episode to learn the target-conditioned subproblem, and are recomputed at every routing step during evaluation. The environment is a directed southern Manhattan road graph extracted from OpenStreetMap [39], with 2087 intersections and 4255 directed roads (see Figure 4). Travel times depend on road length, road type, traffic-light cycles of 90 seconds (with a road-dependent green-phase duration and a random offset), and stochastic congestion effects. Vehicles start from three depot locations in random proportions and serve delivery requests sampled uniformly across the city. Further environment details are provided in Appendix F.4. We instantiate SALT with tabular Q-learning (SALT-Q) and compare it against a static baseline that assigns vehicles once at $t = 0$ using nominal shortest-path distances and then follows the resulting paths. This standard operational heuristic ignores traffic-light timing and stochastic congestion, since incorporating them exactly requires a much larger spatiotemporal routing problem. We exclude the deep MARL baselines of Section 5.1 because, as demonstrated, they become computationally prohibitive on a graph of this scale and are tied to a fixed training fleet size. SALT-Q policies are trained with $N = 20$ agents and evaluated, without retraining, on fleets of 1, 5, 10, 20, and 100 vehicles.

Figure 4 reports the relative improvement in travel time over the static shortest-path baseline as a function of fleet size. SALT-Q consistently outperforms the shortest-path baseline, with median improvements ranging from 14.8% ($N = 1$) to 21.3% ($N = 100$), reflecting its ability to anticipate traffic-light timing and adapt to stochastic congestion. Performance remains stable across fleet sizes, confirming that SALT generalizes naturally to fleet sizes unseen at training time. The residual variance arises from a natural bias of distributional training: the policy concentrates on high-probability origin–destination pairs, improving sample efficiency at the cost of reduced accuracy on rare assignments. Additional ablations regarding SALT fleet-size transferability and reassignment variants of the shortest-path baseline are reported in Section F.3 and Section F.4, respectively.

6 Conclusion and limitations

We established a separation principle for cooperative MARL and used it to design SALT, a family of algorithms that learn a single-agent policy and use optimal transport to coordinate agents. SALT outperforms existing MARL baselines in grid worlds, both in terms of performance and scalability. In the southern Manhattan road network, where MARL baselines are impractical, SALT improves travel time by 15–21% over a shortest-path heuristic and generalizes to population sizes not seen during training. Our work does not address several important challenges, which we leave to future work. From a theoretical perspective, our separation principle requires decoupled dynamics and an optimal transport terminal cost, ruling out settings with strong interactions, collisions, or non-transport objectives. Our bound on the optimal cost is also not tight in general. From a practical perspective, the target-conditioned single-agent problem must generalize across the target space, which is challenging when targets are high-dimensional or sampled from distributions far from the training one.

Acknowledgments and Disclosure of Funding

This publication was made possible by the ETH AI Center doctoral fellowship to Lucia Pezzetti. This work was also supported as a part of NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40 225155). Nicolas Lanzetti was in part supported by the PIMCO Fellows Program, by the National Science Foundation under Grant #CCF-2326609, and by the Resnick Sustainability Institute.

References

[1] Wei Jing, Di Deng, Yan Wu, and Kenji Shimada. Multi-UAV coverage path planning for the inspection of large and complex structures. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1480–1486, 2020.

[2] T. Lamballais, M. Merschformann, D. Roy, M. B. M. de Koster, K. Azadeh, and L. Suhl. Dynamic policies for resource reallocation in a robotic mobile fulfillment system with time-varying demand. *European Journal of Operational Research*, 300(3):937–952, 2022.

[3] Matthew S. Maxwell, Mateo Restrepo, Shane G. Henderson, and Huseyin Topaloglu. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22(2):266–281, 2010.

[4] Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The World Wide Web Conference*, pages 983–994, 2019.

[5] Zhongjing Ma, Duncan S. Callaway, and Ian A. Hiskens. Decentralized charging control of large populations of plug-in electric vehicles. *IEEE Transactions on Control Systems Technology*, 21 (1):67–78, 2013.

[6] Gioele Zardini, Nicolas Lanzetti, Marco Pavone, and Emilio Frazzoli. Analysis and control of autonomous mobility-on-demand systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):633–658, 2022.

[7] Javier Alonso-Mora, Samitha Samaranayake, Alex Wallar, Emilio Frazzoli, and Daniela Rus. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462–467, 2017.

[8] Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems! In *International Conference on Learning Representations*, 2019.

[9] 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*, pages 321–384. Springer, Cham, 2021.

[10] Aleksandar Krnjaic, Raul D. Steleac, Jonathan D. Thomas, Georgios Papoudakis, Lukas Schäfer, Andrew Wing Keung To, Kuan-Ho Lao, Murat Cubuktepe, Matthew Haley, Peter Börsting, and Stefano V. Albrecht. Scalable multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 677–684, 2024.

[11] Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1774–1783, 2018.

[12] Yi Shen, Benjamin McClosky, Joseph W. Durham, and Michael M. Zavlanos. Multi-agent reinforcement learning for resource allocation in large-scale robotic warehouse sortation centers. In *IEEE Conference on Decision and Control*, pages 7137–7143, 2023.

[13] 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, Datasets and Benchmarks Track*, volume 35, pages 24611–24624, 2022.

[14] 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.

[15] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius 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 *International Conference on Autonomous Agents and Multiagent Systems*, pages 2085–2087, 2018.

[16] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[17] 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.

[18] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

[19] Abubakr O. Al-Abbasi, Arnob Ghosh, and Vaneet Aggarwal. DeepPool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4714–4727, 2019.

[20] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[21] 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 *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5887–5896, 2019.

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

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

[24] Enshu Wang, Rong Ding, Zhaoxing Yang, Haiming Jin, Chenglin Miao, Lu Su, Fan Zhang, Chunming Qiao, and Xinbing Wang. Joint charging and relocation recommendation for e-taxi drivers via multi-agent mean field hierarchical reinforcement learning. *IEEE Transactions on Mobile Computing*, 21(4):1274–1290, 2022.

[25] Batuhan Yardim, Artur Goldman, and Niao He. When is mean-field reinforcement learning tractable and relevant? In *International Conference on Autonomous Agents and Multiagent Systems*, pages 2038–2046, 2024.

[26] Mingyu Yang, Jian Zhao, Xunhan Hu, Wengang Zhou, Jiangcheng Zhu, and Houqiang Li. LDSA: Learning dynamic subtask assignment in cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1698–1710, 2022.

[27] Antonio Terpin, Nicolas Lanzetti, and Florian Dörfler. Dynamic programming in probability spaces via optimal transport. *SIAM Journal on Control and Optimization*, 62(2):1183–1206, 2024.

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[29] Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Conference on Theoretical Aspects of Rationality and Knowledge*, pages 195–210, 1996.

[30] Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. Wiley, New York, 1994.

[31] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic optimal control: The discrete-time case*. Athena Scientific, Belmont, MA, 1996.

[32] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis: A hitchhiker's guide*. Springer, Berlin, 3 edition, 2006.

[33] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

[34] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016.

[35] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, 2016.

[36] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013.

[37] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the StarCraft multi-agent challenge? arXiv preprint arXiv:2011.09533, 2020.

[38] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022.

[39] OpenStreetMap contributors. OpenStreetMap. `https://www.openstreetmap.org`, 2025. Planet dump retrieved from `https://planet.osm.org`.

[40] Paul R. Halmos. *Measure theory*. Graduate Texts in Mathematics. Springer, New York, 1974.

[41] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2 edition, 2002.

[42] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, anniversary edition, 2012.

[43] Kenneth L. Judd. The law of large numbers with a continuum of IID random variables. *Journal of Economic Theory*, 35(1):19–25, 1985.

[44] Yeneng Sun. The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126(1):31–69, 2006.

[45] Alexander S. Kechris. *Classical descriptive set theory*, volume 156 of *Graduate Texts in Mathematics*. Springer, New York, 1995.

[46] Cédric Villani. *Optimal transport: Old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009.

[47] Garrett Birkhoff. Tres observaciones sobre el álgebra lineal. *Universidad Nacional de Tucumán, Revista, Serie A*, 5:147–151, 1946.

[48] Liviu Aolaritei, Nicolas Lanzetti, Hongruyu Chen, and Florian Dörfler. Distributional uncertainty propagation via optimal transport. *IEEE Transactions on Automatic Control*, 70(9): 6080–6095, 2025.

[49] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[51] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with CUDA. *Queue*, 6(2):40–53, 2008. doi: 10.1145/1365490.1365500.

[52] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. `http://github.com/jax-ml/jax`, 2018.

Table 1: Notation for Appendices A– F.

Object	Definition	Intuition
$\hat{\xi}_t^{s,a}$	Random element of $\mathcal{P}(W)$	Distribution of noise experienced by agents in (s,a) at time t
$\eta_t^{s,a}$	Distribution of $\hat{\xi}_t^{s,a}$	Describes the probability distribution associated with the noise realizations
η_t	Probability measure in $\mathcal{P}(\mathcal{P}(W)^{S\times A})$	Collects all $\eta_t^{s,a}$ across bins, encoding population-level randomness
ρ_t	$\lambda_t \otimes \hat{\xi}_t^{s,a}$	Joint state-action-noise distribution

A Noise models and averaging property

This appendix develops the measure-theoretic framework used to handle stochasticity in the multi-agent system. We first introduce the relevant general notation, and illustrate how common noise models fit into this framework through a series of examples. We conclude with Lemma A.10, which will be a key technical tool used in the proof of Theorem 3.1, Part I in Appendix C. Throughout, the space of probability measures on a Polish space X, $\mathcal{P}(X)$, is endowed with the topology of weak convergence and its Borel σ-algebra. Since X is a Polish space, $\mathcal{P}(X)$ is itself a Polish space [32, Theorem 15.15]. The product measure of two independent random variables with marginals $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ is the unique measure $\mu \otimes \nu \in \mathcal{P}(X \times Y)$ [40, Theorem B, Section 35] such that $(\mu \otimes \nu)(A \times B) = \mu(A)\,\nu(B)$ for all Borel sets $A \subseteq X$, $B \subseteq Y$.

A.1 General framework

In multi-agent systems, randomness enters at the *individual* level, through the noise realizations for each individual agent. At the *population* level, these realizations induce randomness in the population distribution and therefore in the next population state. To capture this in a unified way, we work with distributions over noise distributions. Concretely, following the notation in Section 2.2, let $\lambda_t \in \mathcal{P}(S \times A)$ denote the joint state-action distribution of the population at time t, and denote by $\xi_t(\cdot \mid s,a) \in \mathcal{P}(W)$ the *nominal kernel*, i.e., the law of the noise experienced by an agent in state s taking action a. For agents sharing the same state-action pair (s,a), their noise realizations form the *realized kernel* $\hat{\xi}_t^{s,a} \in \mathcal{P}(W)$. Unlike the nominal kernel $\xi_t(\cdot \mid s,a)$, which is a fixed object, $\hat{\xi}_t^{s,a}$ is random, fluctuating across episodes depending on the specific noise draws of the agents in that bin. Therefore, we can talk about the distribution of $\hat{\xi}_t^{s,a}$, which we denote by $\eta_t^{s,a} \in \mathcal{P}(\mathcal{P}(W))$, and collect these across all state-action pairs into the joint object $\eta_t \in \mathcal{P}(\mathcal{P}(W)^{S\times A})$. Finally, to describe the full population transition in this stochastic setting, it is not enough to track where agents are and what actions they take; one also needs to account for the noise they experience. This motivates the introduction of the joint state-action-noise distribution, defined as the gluing

$$\rho_t = \lambda_t(\mu_t) \otimes \hat{\xi}_t^{s,a} \in \mathcal{P}(S \times A \times W).$$

Intuitively, $\rho_t(\{(s,a,w)\})$ is the fraction of agents that are at state s, use action a, and experience the noise realization w. Since the noise realizations are random, $\hat{\xi}_t^{s,a}$ is random and, consequently, so is ρ_t. At this point, the next population is a pushforward of ρ_t under the system dynamics f_t:

$$\mu_{t+1} = f_{t\#}\rho_t.$$

Different noise models below correspond to different distributional assumptions on $\hat{\xi}_t^{s,a}$, and hence on ρ_t. The objects introduced above are summarized in Table 1.

The gluing $\rho_t = \lambda_t \otimes \hat{\xi}_t^{s,a}$ is well-defined per realization only if the realized field is a measurable kernel. We impose this as a standing assumption. It holds directly for the individual and global models, while the only subtle case is discussed in Example A.6.

Assumption A.1 (Well-posedness of the realized noise field)**.** Fix a time step t and an admissible state-action distribution λ_t. The realized noise field, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, admits

a version for which there exists a Borel set $D_t \subseteq S \times A$ with $\lambda_t(D_t) = 1$ such that

$$\big((s,a),\omega\big) \mapsto \hat{\xi}_t^{s,a}(\omega)(B)$$

is jointly measurable on $D_t \times \Omega$ for every Borel set $B \subseteq W$.

In particular, almost surely, $(s,a) \mapsto \hat{\xi}_t^{s,a}$ is a Borel kernel on D_t, so ρ_t is a well-defined random element of $\mathcal{P}(S \times A \times W)$, and expectations over the field can be interchanged with λ_t-integrals by Tonelli's theorem.

We additionally impose the following consistency condition on the realized kernel.

Assumption A.2. Fix a time step t and an admissible state-action distribution λ_t, together with its realized noise field. There exists a Borel set $E_t \subseteq S \times A$ with $\lambda_t(E_t) = 1$ such that for all $(s,a) \in E_t$ and for every bounded Borel measurable $\varphi : W \to \mathbb{R}$

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W \varphi(w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right] = \int_W \varphi(w)\mathrm{d}\xi_t(w \mid s,a). \tag{7}$$

Condition (7) should be read as an unbiased or barycenter condition between the realized kernel $\hat{\xi}_t^{s,a}$ and the nominal kernel $\xi_t(\cdot \mid s,a)$. More precisely, the assumption requires that these two descriptions agree after averaging over the realization of $\hat{\xi}_t^{s,a}$. As we will show below, Assumption A.2 is satisfied by several standard realized-kernel constructions. Specifically, in Proposition A.9 we show that it holds true for the noise models presented in Examples A.4–A.7. More generally, Assumption A.2 is satisfied whenever the realized kernel is chosen from a family of probability kernels whose mixture is the nominal kernel $\xi_t(\cdot \mid s,a)$.

What (7) excludes are biased realized-kernel constructions, as we show in the next example.

Example A.3. *Let $W = \{0,1\}$ and $\xi_t = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Suppose that we artificially bias all noise realizations towards 1, so that*

$$\hat{\xi}_t^{s,a} = \delta_1$$

almost surely. Then

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W w\mathrm{d}\hat{\xi}_t^{s,a}(w)\right] = 1 \neq \frac{1}{2} = \int_W w\mathrm{d}\xi_t(w \mid s,a).$$

Hence, (7) *fails.*

More generally, (7) of Assumption A.2 fails whenever the realized population-level kernel systematically overweights some noise realizations relative to the nominal kernel used in the single-agent Bellman equation. In the remainder of this section, we specialize this general framework to the individual, local, and global noise models considered in this work (Appendix A.2). Finally, in Appendix A.3, we prove a key property of all these noise models: we establish, in Lemma A.10, that the expected cost under the realized kernel $\hat{\xi}_t^{s,a}$ equals the cost under the nominal kernel $\xi_t(\cdot \mid s,a)$. Ultimately, this unbiasedness property, verified for all three noise models in Proposition A.9, is what allows us to eliminate randomness from the backward recursion in Appendix C and recover the single-agent Bellman equation.

A.2 Noise models

We start with the case where the noise is i.i.d. over all agents with the same state-action pair. We start with the case of finitely many agents:

Example A.4 (Individual noise, finitely many agents)**.** *Consider N finite agents, each with its "own" noise sample*

$$w_t^i \sim \xi_t(\cdot \mid s_t^i, a_t^i).$$

We assume noise draws to be independent across agents and across state-action pairs. Given the state-action distribution λ_t, its support $\operatorname{supp}\lambda_t$ *is finite (at most N atoms, since $\lambda_t = \frac{1}{N}\sum_{i=1}^N \delta_{(s_t^i,a_t^i)}$). For each bin $(s,a) \in \operatorname{supp}\lambda_t$, let*

$$N_{s,a} \coloneqq N \cdot \lambda_t(\{(s,a)\}) \geq 1$$

be the number of agents in that bin. This is a strictly positive integer, determined by λ_t. Define the empirical measure map

$$\begin{aligned}\Psi^{s,a}_{N_{s,a}} \colon W^{N_{s,a}} &\to \mathcal{P}(W)\\ w^{s,a} &\mapsto \frac{1}{N_{s,a}} \sum_{i=1}^{N_{s,a}} \delta_{w^{s,a,i}},\end{aligned}$$

where $w^{s,a} = (w^{s,a,i})_{i\in[N_{s,a}]}$ collects the noise draws of the agents in bin (s,a). The realized noise distribution for the state-action pair (s,a) is

$$\hat{\xi}^{s,a}_t = \Psi^{s,a}_{N_{s,a}}(w^{s,a}) \in \mathcal{P}(W),$$

with distribution

$$\eta^{s,a}_t = \Psi^{s,a}_{N_{s,a}\#}\left(\bigotimes_{i\in[N_{s,a}]} \xi_t(\cdot \mid s,a)\right) \in \mathcal{P}(\mathcal{P}(W)).$$

Since $\operatorname{supp}\lambda_t$ is finite and noise realizations are independent across state-action pairs, the joint law is the finite product

$$\eta_t = \bigotimes_{(s,a)\in\operatorname{supp}\lambda_t} \eta^{s,a}_t \in \mathcal{P}(\mathcal{P}(W)^{\operatorname{supp}\lambda_t}).$$

We extend η_t to all state-action pairs by setting, for $(s,a)\notin\operatorname{supp}\lambda_t$, $\hat{\xi}^{s,a}_t = \delta_{w^0}$, for an arbitrarily chosen $w^0\in W$. Clearly, the specific choice of w^0 is inconsequential: since λ_t assigns zero mass to these bins, they do not contribute to λ_t-weighted integrals, which are exactly the type of integrals in our problem setting. The map $(s,a)\mapsto\eta^{s,a}_t$ is Borel measurable: it takes finitely many distinct values (one per element of $\operatorname{supp}\lambda_t$, plus the constant extension outside), and the preimage of each value is a finite union of singletons or the complement of $\operatorname{supp}\lambda_t$, all of which are Borel sets.

Example A.5 (Individual noise, infinitely many agents)**.** *In the limit case of infinitely many agents, the two-level randomization (each agent drawing a private noise sample, and the collection of noise realizations across the population) collapses into a single draw from the deterministic coupling*

$$\rho_t = \lambda_t \otimes \xi_t(\cdot \mid s,a).$$

Intuitively, the population is so numerous that its empirical profile of noise realizations matches the underlying noise kernel itself. Rigorously, in the case of finite state and action spaces, we can assume that the number of agents in each state-action pair (s,a) grows to infinity and Varadarajan's theorem [41, Theorem 11.4.1.] in the space $\mathcal{P}(W)$ guarantees

$$\hat{\xi}^{s,a}_t \xrightarrow{a.s.} \xi_t(\cdot \mid s,a), \quad \text{as } N_{s,a}\to\infty.$$

Consequently, the laws of the realized kernels $\eta^{s,a}_t$ converge weakly to $\delta_{\xi_t(\cdot\mid s,a)}$ [41, Theorem 9.2.1, Proposition 9.3.5]. We therefore set $\hat{\xi}^{s,a}_t = \xi_t(\cdot\mid s,a)$ almost surely and

$$\eta^{s,a}_t = \delta_{\xi_t(\cdot\mid s,a)} \in \mathcal{P}(\mathcal{P}(W)).$$

For general (possibly uncountable) Polish spaces S and A, we construct the joint law via Kolmogorov's extension theorem [42, Theorems 36.1 and 36.2]. Specifically, for each finite subset $J\subset S\times A$, define a probability measure

$$\eta^J_t = \bigotimes_{(s,a)\in J} \delta_{\xi_t(\cdot\mid s,a)} \in \mathcal{P}(\mathcal{P}(W)^J).$$

These finite-dimensional marginals are consistent (the marginal of η^J_t onto any $J'\subset J$ equals $\eta^{J'}_t$, since products of Dirac masses marginalize correctly). Thus, Kolmogorov's extension theorem guarantees that there exists a unique probability measure

$$\eta_t = \bigotimes_{(s,a)\in S\times A} \delta_{\xi_t(\cdot\mid s,a)},$$

and the realized joint state-action-noise distribution collapses to the deterministic coupling $\rho_t = \lambda_t\otimes\xi_t(\cdot\mid s,a)$.

Example A.6 (Local noise)**.** *Suppose the noise realizations are shared among agents with the same state-action pair. For each fixed $(s, a) \in S \times A$ we draw*

$$w_t^{s,a} \sim \xi_t(\cdot \mid s, a),$$

independently across state-action pairs. Then, all agents in bin (s, a) experience the same noise realization $w_t^i = w_t^{s,a}$. In this framework, we define the empirical measure map

$$\begin{aligned} \Psi^{s,a} &\colon W \to \mathcal{P}(W) \\ w_t^{s,a} &\mapsto \delta_{w_t^{s,a}}. \end{aligned}$$

The realized noise distribution for bin (s, a) collapses to

$$\hat{\xi}_t^{s,a} = \delta_{w_t^{s,a}},$$

with distribution

$$\eta_t^{s,a} = {\Psi^{s,a}}_{\#}\xi_t(\cdot \mid s, a) \in \mathcal{P}(\mathcal{P}(W)).$$

Since noise samples are independent across state-action pairs, the joint law is constructed by the same Kolmogorov extension argument as in Example A.5 (applied with $\eta_t^{s,a}$ in place of $\delta_{\xi_t(\cdot \mid s,a)}$). For every finite $J \subset S \times A$, the marginal $\eta_t^J = \bigotimes_{(s,a)\in J} \eta_t^{s,a}$ is consistent, yielding the unique extension[1]

$$\eta_t = \bigotimes_{(s,a)\in S\times A} \eta_t^{s,a} \in \mathcal{P}(\mathcal{P}(W)^{S\times A}).$$

Example A.7 (Global noise)**.** *Consider the case in which all agents are subject to the same noise draw $w_t \sim \xi_t \in \mathcal{P}(W)$, regardless of their state and action. Since every state-action pair (s, a) experiences the same noise realization, we have $\hat{\xi}_t^{s,a} = \delta_{w_t}$ for all (s, a). We then define the diagonal embedding*

$$\begin{aligned} \Phi &\colon W \to \mathcal{P}(W)^{S\times A} \\ w &\mapsto ((s, a) \mapsto \delta_w)\,. \end{aligned}$$

The entire random family $\left(\hat{\xi}_t^{s,a}\right)_{(s,a)\in S\times A} = \Phi(w_t)$ is a $\mathcal{P}(W)^{S\times A}$-valued random variable and its law is

$$\eta_t = \Phi_{\#}\xi_t \in \mathcal{P}(\mathcal{P}(W)^{S\times A}).$$

In words, $\xi_t \in \mathcal{P}(W)$ is pushed forward through Φ to a measure on $\mathcal{P}(W)^{S\times A}$, concentrated on the diagonal $\{\Phi(w) \colon w \in W\}$. We refer the reader to Remark A.8 for a discussion on the well-posedness of Φ.

Remark A.8 (Well-posedness of Example A.7). We verify here that the map $\Phi \colon W \to \mathcal{P}(W)^{S\times A}$ is well-defined and a Borel measurable map between Polish spaces, so that the pushforward $\eta_t = \Phi_{\#}\xi_t$ is well-defined.

Step 1: Each coordinate map is continuous. We start by noting that each coordinate map $w \mapsto \delta_w \colon W \to \mathcal{P}(W)$ is continuous with respect to the weak topology on $\mathcal{P}(W)$ [32, Theorem 15.8].

Step 2: Φ is Borel measurable. To show that the full map Φ is Borel measurable we consider two cases depending on whether $S \times A$ is countable or uncountable. In the countable case, $\mathcal{P}(W)^{S\times A}$ is a countable product of Polish spaces, which is again a Polish space [45, Proposition 3.3] under the product topology. A map into a countable product is measurable if and only if each coordinate projection composed with it is measurable, which holds by Step 1. So Φ is Borel measurable. If,

[1] Assumption A.1 holds whenever λ_t has countable support, including every finite population. When λ_t is atomless, η_t still exists, but a continuum of independent draws admits no jointly measurable version with nondegenerate coordinates on an ordinary product probability space [43, 44] and so $\lambda_t \otimes \hat{\xi}_t^{s,a}$ cannot be formed pathwise. In that case, we do not construct ρ_t but take it as the primitive disturbance, requiring only that its state-action marginal is λ_t almost surely and that $\mathbb{E}[\rho_t] = \lambda_t \otimes \xi_t$. On a rich Fubini extension, an enlargement of the product space that retains the Fubini property and is rich enough to carry the continuum of independent noise draws itself, the exact law of large numbers of Sun [44] then gives $\rho_t = \lambda_t \otimes \xi_t$ almost surely on the atomless part, while atoms keep their shared draws [44]. We record this only to show that the formulation is consistent, indeed, the separation proof uses just the marginal and average properties above, so its conclusions also hold in this formulation.

instead, $S \times A$ is uncountable, then $\mathcal{P}(W)^{S\times A}$ might fail to be Polish. However, this does not pose issues as the image of Φ is contained in the diagonal subset

$$\mathcal{D} = \{\Phi(w) \colon w \in W\} = \{((s,a) \mapsto \delta_w) \colon w \in W\} \subset \mathcal{P}(W)^{S\times A}.$$

The map $\Phi \colon W \to \mathcal{D}$ is a homeomorphism: both Φ and its inverse $\Phi^{-1} \colon \mathcal{D} \to W$ (which evaluates any coordinate and then inverts $w \mapsto \delta_w$) are continuous. Hence, $\mathcal{D}$ is homeomorphic to W, which is Polish. In particular, $\mathcal{D}$ is a Polish space in the subspace topology inherited from $\mathcal{P}(W)^{S\times A}$. Since $\Phi \colon W \to \mathcal{D}$ is continuous between Polish spaces it is Borel measurable, and $\eta_t = \Phi_{\#}\xi_t$ is a well-defined Borel probability measure on $\mathcal{D}$ and so also on $\mathcal{P}(W)^{S\times A}$. We finally observe that η_t is fully concentrated on $\mathcal{D}$, so all subsequent integrals only involve elements of $\mathcal{D}$ and the uncountability of $S \times A$ does not pose an issue.

We now show that these canonical noise models satisfy Assumption A.2.

Proposition A.9 (The noise models satisfy Assumption A.2)**.** *For each of the noise models in Examples A.4–A.7, the realized kernel $\hat{\xi}_t^{s,a}$ and its distribution η_t satisfy Assumption A.2. That is, there exists a Borel set $E_t \subseteq S \times A$ with $\lambda_t(E_t) = 1$ such that for all $(s,a) \in E_t$ and for every bounded Borel measurable $\varphi : W \to \mathbb{R}$ simultaneously,*

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W \varphi(w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right] = \int_W \varphi(w)\mathrm{d}\xi_t(w \mid s,a).$$

Proof. We prove the statement for each noise model separately.

Individual noise (Example A.4). Fix $(s,a) \in \operatorname{supp}\lambda_t$, hence $N_{s,a} \geq 1$. By construction,

$$\hat{\xi}_t^{s,a} = \frac{1}{N_{s,a}} \sum_{i=1}^{N_{s,a}} \delta_{w_t^{s,a,i}}$$

where $w_t^{s,a,i}$ are i.i.d. with law $\xi_t(\cdot \mid s,a)$. Hence, for any bounded measurable $\varphi \colon W \to \mathbb{R}$,

$$\int_W \varphi(w)\mathrm{d}\hat{\xi}_t^{s,a}(w) = \frac{1}{N_{s,a}} \sum_{i=1}^{N_{s,a}} \varphi(w_t^{s,a,i})$$

and therefore, by linearity of expectation and since each $w_t^{s,a,i}$ has law $\xi_t(\cdot \mid s,a)$,

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W \varphi(w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right] = \frac{1}{N_{s,a}} \sum_{i=1}^{N_{s,a}} \mathbb{E}_{\xi_t(\cdot \mid s,a)}\left[\varphi(w)\right] = \int_W \varphi(w)\mathrm{d}\xi_t(w \mid s,a).$$

Thus (7) holds on $E_t = \operatorname{supp}\lambda_t$, which clearly satisfies $\lambda_t(E_t) = 1$.

Infinite-population model (Example A.5). By construction $\hat{\xi}_t^{s,a} = \xi_t(\cdot \mid s,a)$ almost surely for every $(s,a) \in S \times A$. Therefore, for every bounded measurable φ

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W \varphi(w)\mathrm{d}\hat{\xi}_t^{s,a}\right] = \int_W \varphi(w)\mathrm{d}\xi_t(w \mid s,a).$$

Thus, (7) holds with $E_t = S \times A$.

Local noise (Example A.6). For each fixed (s,a), a single draw $w_t^{s,a} \sim \xi_t(\cdot \mid s,a)$ is taken, independently across state-action pairs, and the realized noise distribution is $\hat{\xi}_t^{s,a} = \delta_{w_t^{s,a}}$. So, for every bounded measurable φ, the inner integral equals $\varphi(w_t^{s,a})$ and, since $w_t^{s,a} \sim \xi_t(\cdot \mid s,a)$,

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W \varphi(w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right] = \mathbb{E}_{\xi_t(\cdot \mid s,a)}\left[\varphi(w)\right] = \int_W \varphi(w)\mathrm{d}\xi_t(w \mid s,a).$$

Thus, (7) holds again with $E_t = S \times A$.

Global-noise (Example A.7). This case is identical to the local noise one, with the same draw w_t shared across all (s,a). □

A.3 Averaging property of realized noise kernels

We now show that even if $\hat{\xi}_t^{s,a}$ is random, the expected cost computed under its law η_t equals the cost computed directly under the nominal kernel $\xi_t(\cdot \mid s,a)$. In other words, replacing the realized kernel $\hat{\xi}_t^{s,a}$ by the nominal kernel $\xi_t(\cdot \mid s,a)$ is unbiased in expectation. In particular, this lemma will allow us to remove randomness from the backward recursion in the proof of Theorem 3.1, Part I.

Lemma A.10 (Averaging realized kernels)**.** *Let S, A, W be Polish spaces and fix a time step t. Let λ_t denote the joint state-action distribution and $\xi_t : S \times A \to \mathcal{P}(W)$ be a Borel stochastic kernel. For each (s,a), let $\hat{\xi}_t^{s,a}$ be a $\mathcal{P}(W)$-valued random variable with distribution $\eta_t^{s,a} \in \mathcal{P}(\mathcal{P}(W))$, such that Assumption A.2 holds and $(s,a) \mapsto \eta_t^{s,a}$ is a Borel stochastic kernel. Then, for every nonnegative lower semianalytic $c : S \times A \times W \to \bar{\mathbb{R}}_{\geq 0}$ (in particular, for every nonnegative Borel measurable c),*

$$\int_{S\times A} \mathbb{E}_{\eta_t^{s,a}}\left[\int_W c(s,a,w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right]\mathrm{d}\lambda_t(s,a) = \int_{S\times A}\int_W c(s,a,w)\mathrm{d}\xi_t(w \mid s,a)\mathrm{d}\lambda_t(s,a). \quad (8)$$

Proof. For $n \in \mathbb{N}$, define

$$c_n(s,a,w) \coloneqq \min\{c(s,a,w), n\}.$$

By construction, each c_n is bounded and lower semianalytic, and $c_n \uparrow c$ pointwise.

To start, we prove that the two integrands in (8) are (universally) measurable. For the left-hand side, for each n, the map $(s,a,\zeta) \mapsto \int_W c_n(s,a,w)\mathrm{d}\zeta(w)$ is lower semianalytic on $S \times A \times \mathcal{P}(W)$: the evaluation map $q(\mathrm{d}w \mid s,a,\zeta) \coloneqq \zeta(\mathrm{d}w)$ is a Borel stochastic kernel (since $\zeta \mapsto \zeta(B)$ is Borel measurable for every Borel set $B \subseteq W$ [31, Proposition 7.28]), and integration of a lower semianalytic function against a Borel stochastic kernel preserves lower semianalyticity [31, Proposition 7.48]. Since $(s,a) \mapsto \eta_t^{s,a}$ is a Borel stochastic kernel, a second application of [31, Proposition 7.48] yields lower semianalyticity, hence universal measurability, of

$$(s,a) \mapsto \mathbb{E}_{\eta_t^{s,a}}\left[\int_W c_n(s,a,w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right].$$

Letting $n \uparrow \infty$, and using that a pointwise increasing limit of universally measurable functions is universally measurable, we obtain the measurability of

$$(s,a) \mapsto \mathbb{E}_{\eta_t^{s,a}}\left[\int_W c(s,a,w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right].$$

For the right-hand side, instead, the measurability of

$$(s,a) \mapsto \int_W c(s,a,w)\mathrm{d}\xi_t(w \mid s,a),$$

follows from the same argument, applied with the Borel stochastic kernel ξ_t [31, Proposition 7.48].

We now prove the equality (8). Fix n and $(s,a) \in E_t$, and consider the bounded test function $\varphi_{s,a}(w) \coloneqq c_n(s,a,w)$, viewed as a function of w alone for fixed (s,a). Assumption A.2 is stated for bounded *Borel* test functions, while $\varphi_{s,a}$ is only lower semianalytic, hence universally measurable; we extend (7) to $\varphi_{s,a}$ as follows. Let $m \coloneqq \xi_t(\cdot \mid s,a)$. Since $\varphi_{s,a}$ is universally measurable, it is measurable with respect to the m-completion of the Borel σ-algebra, so there exist a bounded Borel function φ' and a Borel set $\mathcal{N} \subseteq W$ with $m(\mathcal{N}) = 0$ and $\{\varphi_{s,a} \neq \varphi'\} \subseteq \mathcal{N}$. Applying (7) to the bounded Borel function $\mathbf{1}_{\mathcal{N}}$ gives $\mathbb{E}_{\eta_t^{s,a}}\left[\hat{\xi}_t^{s,a}(\mathcal{N})\right] = m(\mathcal{N}) = 0$, and since $\hat{\xi}_t^{s,a}(\mathcal{N}) \geq 0$, we conclude $\hat{\xi}_t^{s,a}(\mathcal{N}) = 0$ almost surely. Consequently, $\int \varphi_{s,a}\mathrm{d}\hat{\xi}_t^{s,a} = \int \varphi'\mathrm{d}\hat{\xi}_t^{s,a}$ almost surely and $\int \varphi_{s,a}\mathrm{d}m = \int \varphi'\mathrm{d}m$; applying (7) to φ' then yields (7) for $\varphi_{s,a}$ itself. This gives

$$\mathbb{E}_{\eta_t^{s,a}}\left[\int_W c_n(s,a,w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right] = \int_W c_n(s,a,w)\mathrm{d}\xi_t(w \mid s,a)$$

for all $(s,a) \in E_t$. Since $\lambda_t(E_t) = 1$, integrating both sides with respect to λ_t gives

$$\int_{S\times A} \mathbb{E}_{\eta_t^{s,a}}\left[\int_W c_n(s,a,w)\mathrm{d}\hat{\xi}_t^{s,a}(w)\right]\mathrm{d}\lambda_t(s,a) = \int_{S\times A}\int_W c_n(s,a,w)\mathrm{d}\xi_t(w \mid s,a)\mathrm{d}\lambda_t(s,a).$$

Finally, let $n \to \infty$. Since $c_n \uparrow c$, the monotone convergence theorem yields (8). □

B Preliminary lemmata

We now collect the auxiliary results required for the main proof in Appendix C.

Measurability convention. Throughout the appendices, primitive data such as the dynamics, stage costs, and terminal cost are assumed to be Borel measurable, as in the main text. Since every nonnegative Borel measurable function is lower semianalytic, these data also satisfy the analytic-measurability assumptions used in the auxiliary results below. We use lower semianalytic functions because the dynamic programming recursion involves pointwise infima over the action space, and such infima need not be Borel measurable on general Polish spaces; see [31] for a detailed discussion. Accordingly, value functions and pointwise Bellman infima are interpreted as lower semianalytic, hence universally measurable, functions. This is sufficient for the integrals and measurable-selection arguments used in the proof of Theorem 3.1. For this reason, the new auxiliary lemmata below (Lemmas B.3, B.5, and B.6) state their regularity requirements explicitly in terms of lower semianalytic costs: this is the regularity actually available when they are invoked (the costs there involve the value functions j_t). The results recalled from [27] are stated and proved there under the standing assumption that all spaces are Polish and all maps are Borel; their proofs (change-of-variables, gluing, and disintegration arguments) apply verbatim when the transportation cost is only lower semianalytic, hence universally measurable (all integrals remain well-defined by nonnegativity), and we use them in this form. We work at this level of generality to simultaneously capture systems with finite and continuous state and action spaces. Nevertheless, when all spaces are finite, measurability is automatic and all measurability arguments can therefore be skipped.

We begin by recalling from [27] a result that lets us "move" the pushforward from the set of couplings into the transportation cost, and vice versa.

Lemma B.1 (Pushforward and optimal transport [27, Lemma 6.1])**.** *Given a transportation cost $j_{t+1} : S \times Z \to \bar{\mathbb{R}}_{\geq 0}$, a map $f_t \colon S \times A \times W \to S$, and a probability measure $\rho_t \in \mathcal{P}(S \times A \times W)$, it holds that*

$$\inf_{(\mathrm{proj}^{S\times A\times W\times Z}_{S\times A\times W})_{\#}\gamma=\rho_t} \int_{S\times A\times W\times Z} j_{t+1} \circ (f_t \times \mathrm{Id}_Z)\mathrm{d}\gamma \;=\; \inf_{(\mathrm{proj}^{S\times Z}_{S})_{\#}\tilde{\gamma}=f_{t\#}\rho_t} \int_{S\times Z} j_{t+1}\mathrm{d}\tilde{\gamma},$$

where Id_Z is the identity map over Z.

In our setting, this lemma tells us that evaluating the cost after applying the system dynamics is equivalent to pushing forward the distribution first: we can "apply the dynamics first" and then evaluate the population-level cost without changing the result.

Next, we recall how to merge additive costs into a single transport objective.

Lemma B.2 (Sum of optimal transport discrepancies [27, Lemma 6.2])**.** *Given transportation costs $c_1 \colon X \to \bar{\mathbb{R}}_{\geq 0}$, $c_2 \colon X \times Z \to \bar{\mathbb{R}}_{\geq 0}$, and probability measures $\rho \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Z)$, it holds that*

$$\mathbb{E}_\rho\,[c_1] + \mathcal{K}[c_2](\rho,\nu) = \mathcal{K}[c](\rho,\nu),$$

with $c : X \times Z \to \bar{\mathbb{R}}_{\geq 0}$ defined as $c(x,z) = c_1(x) + c_2(x,z)$.

This result allows us to collapse stage costs and terminal transport costs into a unified expression, which is crucial for dynamic programming formulations.

We now introduce stochasticity which, in our setting, will appear due to the stochastic dynamics. Through a standard argument, we show that we can safely exchange expectation over random noise with the infimum over couplings, at the price of an inequality.

Lemma B.3 (Infimum and expectation)**.** *Let X, W, Z be Polish spaces, $\lambda \in \mathcal{P}(X)$, and $\nu \in \mathcal{P}(Z)$. Let $\hat{\xi} = (\hat{\xi}^x)_{x\in X}$ be a random family of probability measures $\hat{\xi}^x \in \mathcal{P}(W)$, defined on a common probability space and such that, almost surely, $x \mapsto \hat{\xi}^x$ is a Borel measurable kernel on a set of full λ-measure (cf. Assumption A.1). Denote its distribution by η. Let $c : X \times W \times Z \to \bar{\mathbb{R}}_{\geq 0}$ be measurable. Whenever the integrand of the left-hand side below fails to be measurable in $\hat{\xi}$, the outer expectation is understood; in our applications, it is universally measurable (see the proof of*

Theorem 3.1). Then

$$\mathbb{E}_\eta\left[\inf_{\gamma\in\Gamma(\lambda,\nu)}\int_{X\times Z}\int_W c(x,w,z)\mathrm{d}\hat{\xi}^x(w)\mathrm{d}\gamma(x,z)\right] \leq \inf_{\gamma\in\Gamma(\lambda,\nu)}\mathbb{E}_\eta\left[\int_{X\times Z}\int_W c(x,w,z)\mathrm{d}\hat{\xi}^x(w)\mathrm{d}\gamma(x,z)\right].$$

Proof. Fix any $\bar{\gamma}\in\Gamma(\lambda,\nu)$. Pointwise in $\hat{\xi}$,

$$\inf_{\gamma\in\Gamma(\lambda,\nu)}\int_{X\times Z}\int_W c(x,w,z)\mathrm{d}\hat{\xi}^x(w)\mathrm{d}\gamma(x,z)\leq\int_{X\times Z}\int_W c(x,w,z)\mathrm{d}\hat{\xi}^x(w)\mathrm{d}\bar{\gamma}(x,z).$$

Taking expectations over $\hat{\xi}$ preserves the inequality (also when the left-hand side is interpreted as an outer expectation, by its monotonicity). Since the resulting inequality holds for every coupling $\bar{\gamma}$, taking the infimum over $\bar{\gamma}\in\Gamma(\lambda,\nu)$ yields the claim. □

With this inequality, randomness can be "pushed inside" the transport problem, a step that will later allow us to handle noisy transitions.

The next tool, recalled from [27], refines our handling of couplings: it shows how an optimizer can be disintegrated with respect to one marginal, yielding conditional distributions.

Lemma B.4 (Disintegration of the optimizer [27, Lemma 6.4])**.** *Given a transportation cost $c: S\times A\times Z\to\bar{\mathbb{R}}_{\geq 0}$ and probability measures $\mu\in\mathcal{P}(S)$, $\nu\in\mathcal{P}(Z)$, it holds that*

$$\inf_{(\mathrm{proj}^{S\times A\times Z}_{S\times Z})_\#\gamma\in\Gamma(\mu,\nu)}\int_{S\times A\times Z}c(s,a,z)\mathrm{d}\gamma(s,a,z) = \inf_{\gamma'\in\Gamma(\mu,\nu)}\inf_{\{\alpha^{s,z}\}\in\Lambda(A)}\int_{S\times Z}\int_A c(s,a,z)\mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma'(s,z), \quad (9)$$

where $\Lambda(A):=\{\{\alpha^{s,z}\}_{(s,z)\in S\times Z}\subseteq\mathcal{P}(A)\,|\,S\times Z\ni(s,z)\mapsto\alpha^{s,z}\in\mathcal{P}(A)$ Borel$\}$.

This separation will allow us to express the population-level policy as a distribution over conditional agent-level actions.

Next, we study more in detail the right-hand side in (9), showing that the infimum over the family $\{\alpha^{s,z}\}\in\Lambda(A)$ can be "pushed" into the integral.

Lemma B.5 (Reduction to pointwise infimum over deterministic policies)**.** *Let S, A, Z be Polish spaces and let $c\colon S\times A\times Z\to\bar{\mathbb{R}}_{\geq 0}$ be lower semianalytic. Let $\mu\in\mathcal{P}(S)$ and $\nu\in\mathcal{P}(Z)$. Define*

$$\Lambda(A):=\left\{\{\alpha^{s,z}\}_{(s,z)\in S\times Z}\subseteq\mathcal{P}(A)\;\middle|\;S\times Z\ni(s,z)\mapsto\alpha^{s,z}\in\mathcal{P}(A)\text{ is Borel measurable}\right\}.$$

Then, the function $(s,z)\mapsto\inf_{a\in A}c(s,a,z)$ is universally measurable, and for every coupling $\gamma\in\Gamma(\mu,\nu)$,

$$\inf_{\{\alpha^{s,z}\}\in\Lambda(A)}\int_{S\times Z}\int_A c(s,a,z)\mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma(s,z)=\int_{S\times Z}\inf_{a\in A}c(s,a,z)\mathrm{d}\gamma(s,z). \quad (10)$$

Proof. We start by assessing the measurability of the pointwise infimum. Since c is lower semianalytic and A is a Polish space, the function $(s,z)\mapsto\inf_{a\in A}c(s,a,z)$ is lower semianalytic by [31, Proposition 7.47], hence universally measurable. In particular, the right-hand side of (10) is a well-defined element of $[0,+\infty]$.

We then prove (10) by showing "$\leq$" and "$\geq$" separately.

Lower bound "$\geq$". For any $\{\alpha^{s,z}\}\in\Lambda(A)$ and every $(s,z)\in S\times Z$, since $\alpha^{s,z}\in\mathcal{P}(A)$ is a probability measure

$$\int_A c(s,a,z)\mathrm{d}\alpha^{s,z}(a)\geq\inf_{a\in A}c(s,a,z).$$

Integrating both sides with respect to γ and taking the infimum over all $\{\alpha^{s,z}\} \in \Lambda(A)$ leads to

$$\inf_{\{\alpha^{s,z}\}\in\Lambda(A)} \int_{S\times Z} \int_A c(s,a,z)\mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma(s,z) \geq \int_{S\times Z} \inf_{a\in A} c(s,a,z)\mathrm{d}\gamma(s,z).$$

Upper bound "$\leq$". Let $\varepsilon > 0$. Since $c \geq 0$ is lower semianalytic and A is a Polish space, [31, Proposition 7.50] guarantees the existence of a universally measurable ε-optimal selector $a^\varepsilon \colon S \times Z \to A$ satisfying

$$c(s, a^\varepsilon(s,z), z) \leq \inf_{a\in A} c(s,a,z) + \varepsilon \quad \forall (s,z) \in S \times Z.$$

Now, since γ is a σ-finite probability measure on $S \times Z$, by [31, Lemma 7.27] there exists a Borel measurable map $b^\varepsilon \colon S \times Z \to A$ with $b^\varepsilon = a^\varepsilon$ γ-almost everywhere, so that

$$c(s, b^\varepsilon(s,z), z) \leq \inf_{a\in A} c(s,a,z) + \varepsilon \quad \gamma\text{-almost everywhere.}$$

The Dirac embedding $A \to \mathcal{P}(A)$ defined by $a \mapsto \delta_a$ is continuous in the weak topology [32, Theorem 15.8], hence Borel measurable. Therefore, the composition $(s,z) \mapsto \delta_{b^\varepsilon(s,z)}$ is Borel measurable, and so is the family $\{\delta_{b^\varepsilon(s,z)}\}_{(s,z)} \in \Lambda(A)$. Thus,

$$\begin{aligned}\inf_{\{\alpha^{s,z}\}\in\Lambda(A)} \int_{S\times Z} \int_A c(s,a,z)\mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma(s,z) &\leq \int_{S\times Z} c(s, b^\varepsilon(s,z), z)\mathrm{d}\gamma(s,z) \\ &\leq \int_{S\times Z} \inf_{a\in A} c(s,a,z)\mathrm{d}\gamma(s,z) + \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we conclude

$$\inf_{\{\alpha^{s,z}\}\in\Lambda(A)} \int_{S\times Z} \int_A c(s,a,z)\mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma(s,z) \leq \int_{S\times Z} \inf_{a\in A} c(s,a,z)\mathrm{d}\gamma(s,z).$$

This completes the proof of the upper bound and, with it, the proof of the lemma. □

This lemma will allow us to relate the dynamic programming over probability measures with the standard dynamic programming over the state and action spaces.

Finally, we present a lemma that upper bounds the transport cost when replacing "full couplings" with "marginal plans". This inequality will be used to simplify the optimization by relaxing the coupling constraints, enabling us to separate noise from the population-level allocation.

Lemma B.6 (Suboptimality of marginal plans)**.** *Let X, W, Z be Polish spaces. Given a lower semianalytic transportation cost $c : X \times W \times Z \to \bar{\mathbb{R}}_{\geq 0}$ and probability measures $\rho \in \mathcal{P}(X \times W)$, $\nu \in \mathcal{P}(Z)$, let $\rho = \lambda \otimes q^x$ be the disintegration of ρ with respect to its first marginal $\lambda := (\mathrm{proj}_X^{X\times W})_\# \rho$, where $q^x \in \mathcal{P}(W)$ is a (deterministic) Borel kernel (we use the symbol q rather than $\hat{\xi}$ to stress that no randomness is involved here). Then,*

$$\inf_{\gamma\in\Gamma(\rho,\nu)} \int_{X\times W\times Z} c(x,w,z)\mathrm{d}\gamma(x,w,z) \leq \inf_{\gamma_1\in\Gamma(\lambda,\nu)} \int_{X\times Z} \int_W c(x,w,z)\mathrm{d}q^x(w)\mathrm{d}\gamma_1(x,z). \quad (11)$$

Proof. Consider the disintegration of ρ with respect to its first marginal $\lambda = (\mathrm{proj}_X^{X\times W})_\# \rho \in \mathcal{P}(X)$, $\rho = \lambda \otimes q^x$. Let $\gamma_1 \in \Gamma(\lambda, \nu)$ and consider the gluing $\bar{\gamma} = \gamma_1 \otimes q^x \in \mathcal{P}(X \times Z \times W)$. To "reorder" marginals, consider the permutation

$$\sigma : X \times Z \times W \to X \times W \times Z, \quad \sigma(x,z,w) = (x,w,z),$$

and set $\gamma = \sigma_\#(\gamma_1 \otimes q^x) \in \mathcal{P}(X \times W \times Z)$.

We now show that γ is a coupling between ρ and ν and, thus, a valid upper bound for the infimum in (11). We check each marginal by testing against an arbitrary bounded measurable function. Let

$\varphi\colon X \times W \to \mathbb{R}$ be bounded and measurable, we have

$$\begin{aligned}
&\int_{X\times W} \varphi(x,w)\mathrm{d}\left(\left(\mathrm{proj}_{X\times W}^{X\times W\times Z}\right)_{\#}\gamma\right)(x,w)\\
&\qquad = \int_{X\times W\times Z} \varphi\left(\mathrm{proj}_{X\times W}^{X\times W\times Z}(x,w,z)\right)\mathrm{d}\gamma(x,w,z)\\
&\qquad = \int_{X\times W\times Z} \varphi\big(\mathrm{proj}_{X\times W}^{X\times W\times Z}(x,w,z)\big)\mathrm{d}\left(\sigma_{\#}(\gamma_1\otimes q^x)\right)(x,w,z)\\
&\qquad = \int_{X\times Z\times W} \varphi\left(\mathrm{proj}_{X\times W}^{X\times W\times Z}(\sigma(x,z,w))\right)\mathrm{d}(\gamma_1\otimes q^x)\,(x,z,w)\\
&\qquad = \int_{X\times Z}\int_{W} \varphi(x,w)\mathrm{d}q^x(w)\mathrm{d}\gamma_1(x,z)\\
&\qquad = \int_{X}\int_{W} \varphi(x,w)\mathrm{d}q^x(w)\mathrm{d}\lambda(x).
\end{aligned}$$

For every bounded and measurable φ, since q^x is the disintegration of ρ on W given x, and $\gamma_1 \in \Gamma(\lambda,\nu)$ has first marginal λ, Fubini–Tonelli [42, Theorem 18.3] gives

$$\int_X\int_W \varphi(x,w)\mathrm{d}q^x(w)\mathrm{d}\lambda(x) = \int_{X\times W}\varphi(x,w)\,\mathrm{d}(\lambda\otimes q^x)\,(x,w) = \int_{X\times W}\varphi(x,w)\mathrm{d}\rho(x,w).$$

Hence $\left(\mathrm{proj}_{X\times W}^{X\times W\times Z}\right)_{\#}\gamma = \rho$. Similarly, for the marginal over Z, let $\psi\colon Z\to\mathbb{R}$ be bounded and measurable. Then,

$$\begin{aligned}
&\int_Z \psi(z)\mathrm{d}\left(\left(\mathrm{proj}_{Z}^{X\times W\times Z}\right)_{\#}\gamma\right)(z)\\
&\qquad = \int_{X\times W\times Z} \psi\left(\mathrm{proj}_{Z}^{X\times W\times Z}(x,w,z)\right)\mathrm{d}\gamma(x,w,z)\\
&\qquad = \int_{X\times W\times Z} \psi\left(\mathrm{proj}_{Z}^{X\times W\times Z}(x,w,z)\right)\mathrm{d}\left(\sigma_{\#}(\gamma_1\otimes q^x)\right)(x,w,z)\\
&\qquad = \int_{X\times Z\times W} \psi\left(\mathrm{proj}_{Z}^{X\times W\times Z}(\sigma(x,z,w))\right)\mathrm{d}(\gamma_1\otimes q^x)\,(x,z,w)\\
&\qquad = \int_{X\times Z}\int_{W} \psi(z)\mathrm{d}q^x(w)\mathrm{d}\gamma_1(x,z).
\end{aligned}$$

Since $\int_W \mathrm{d}q^x(w) = 1$ for all x and ν is the marginal over Z of $\gamma_1\in\Gamma(\lambda,\nu)$, this reduces to

$$\int_{X\times Z}\psi(z)\mathrm{d}\gamma_1(x,z) = \int_Z \psi(z)\mathrm{d}\nu(z).$$

Combining the two chains of equalities, $\int_Z \psi\mathrm{d}\big(\left(\mathrm{proj}_{Z}^{X\times W\times Z}\right)_{\#}\gamma\big) = \int_Z\psi\mathrm{d}\nu$ for every bounded measurable ψ, and hence $\left(\mathrm{proj}_{Z}^{X\times W\times Z}\right)_{\#}\gamma = \nu$. Overall, we have proven that, for all $\gamma_1\in\Gamma(\lambda,\nu)$, $\gamma = \sigma_{\#}(\gamma_1\otimes q^x)\in\Gamma(\rho,\nu)$.

We are now ready to prove (11). Fix $\gamma_1\in\Gamma(\lambda,\nu)$. The coupling $\gamma = \sigma_{\#}(\gamma_1\otimes q^x)\in\Gamma(\rho,\nu)$ is a valid candidate for the left-hand side. Thus, we obtain the upper bound

$$\begin{aligned}
\inf_{\tilde\gamma\in\Gamma(\rho,\nu)}\int_{X\times W\times Z} c(x,w,z)\mathrm{d}\tilde\gamma(x,w,z) &\le \int_{X\times W\times Z} c(x,w,z)\mathrm{d}\gamma(x,w,z)\\
&= \int_{X\times W\times Z} c(x,w,z)\,\mathrm{d}\left(\sigma_{\#}(\gamma_1\otimes q^x)\right)(x,w,z)\\
&= \int_{X\times Z\times W} c(x,w,z)\,\mathrm{d}(\gamma_1\otimes q^x)\,(x,z,w)\\
&= \int_{X\times Z}\int_W c(x,w,z)\mathrm{d}q^x(w)\mathrm{d}\gamma_1(x,z),
\end{aligned}$$

where the last equality uses the disintegration of $\gamma_1 \otimes q^x$ and Tonelli's theorem. Since $\gamma_1 \in \Gamma(\lambda, \nu)$ is arbitrary, we may take infima to obtain

$$\inf_{\gamma \in \Gamma(\rho,\nu)} \int c(x,w,z) \mathrm{d}\gamma(x,w,z) \leq \inf_{\gamma_1 \in \Gamma(\lambda,\nu)} \int_{X \times Z} \int_W c(x,w,z) \mathrm{d}q^x(w) \mathrm{d}\gamma_1(x,z),$$

which is the desired statement. □

Through the following example, we show that the inequality (11) is generally strict.

Example B.7 (Strictness of Lemma B.6)**.** *Let $X = W = Z = \mathbb{R}$. Set $\lambda = \mathcal{N}(0,1)$ on X. For each $x \in X$, let $q = q^x = \mathcal{N}(0,1)$ on W, independent of x. Then, the product measure $\rho = \lambda \otimes q$ is the law of a pair (x, w) with $x, w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $\nu = \mathcal{N}(0,1)$ on Z and define the cost*

$$c(x,w,z) = (w-z)^2.$$

To evaluate the left-hand side of (11)*, we aim to solve*

$$\inf_{\gamma \in \Gamma(\rho,\nu)} \int_{\mathbb{R}^3} c(x,w,z) \mathrm{d}\gamma(x,w,z).$$

To do so, we show that there exists a plan $\tilde{\gamma} \in \Gamma(\rho,\nu)$ that attains a cost of 0, which, by non-negativity of the cost, implies optimality. Consider $T : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $T(x,w) = (x,w,w)$ and set $\tilde{\gamma} = T_{\#}\rho$. By construction, its (x,w)-marginal is ρ and its z-marginal is given by $w \sim \mathcal{N}(0,1) = \nu$, hence $\tilde{\gamma} \in \Gamma(\rho,\nu)$. Under this coupling

$$\int_{\mathbb{R}^3} (w-z)^2 \mathrm{d}\tilde{\gamma}(x,w,z) = \int_{\mathbb{R}^2} (w-w)^2 \mathrm{d}\rho(x,w) = 0$$

and so

$$\inf_{\gamma \in \Gamma(\rho,\nu)} \int_{\mathbb{R}^3} c(x,w,z) \mathrm{d}\gamma(x,w,z) = 0.$$

We now consider the right-hand side in (11)*. For any $\gamma_1 \in \Gamma(\lambda,\nu)$, we evaluate*

$$\int_{X \times Z} \int_W c(x,w,z) \mathrm{d}q^x(w) \mathrm{d}\gamma_1(x,z).$$

Since the disintegration of ρ is $q = \mathcal{N}(0,1)$ (in particular, it is independent of x), we have

$$\int_{\mathbb{R}} c(x,w,z) \mathrm{d}q(w) = \int_{\mathbb{R}} (w-z)^2 \mathrm{d}q(w) = \int_{\mathbb{R}} w^2 \mathrm{d}q(w) - 2z \int_{\mathbb{R}} w \mathrm{d}q(w) + z^2 = 1 + z^2,$$

and so

$$\begin{aligned} \inf_{\gamma_1 \in \Gamma(\lambda,\nu)} \int_{\mathbb{R}^2} \int_{\mathbb{R}} c(x,w,z) \mathrm{d}q(w) \mathrm{d}\gamma_1(x,z) &= \inf_{\gamma_1 \in \Gamma(\lambda,\nu)} \int_{\mathbb{R}^2} 1 + z^2 \mathrm{d}\gamma_1(x,z) \\ &= 1 + \int_{\mathbb{R}} z^2 \mathrm{d}\nu(z) \\ &= 2, \end{aligned}$$

where the last equality follows from the fact that the second marginal of γ_1 is $\nu = \mathcal{N}(0,1)$ and thus its second moment is 1. Clearly, the left- and right-hand sides do not agree, and the inequality (11) *is strict in general.*

Example B.7 shows that the inequality of Lemma B.6, and hence the bound of Theorem 3.1, Part I, is generally strict. Informally, the gap corresponds to the value of being able to match noise realizations directly to targets, rather than averaging over them. In control problems where noise is small relative to the nominal cost (as in the low-noise regime of our grid-world experiments), we expect this gap to be small.

C Proofs of Section 3

We now prove the main results presented in Section 3. Throughout the following appendices, we write for the set of admissible state-action distributions at $\mu \in \mathcal{P}(S)$

$$M(\mu) \coloneqq \left\{ \lambda \in \mathcal{P}(S \times A) \colon (\mathrm{proj}_S^{S\times A})_{\#}\lambda = \mu \right\},$$

that is, the state-action distributions whose S-marginal is μ. An admissible population control in the sense of Problem 2 thus satisfies $\lambda_t(\mu) \in M(\mu)$ for every μ. Recall the single-agent problem and the dynamic programming for the cost-to-go function in (2):

$$\begin{aligned} j_T(s_T, z) &= g_T(s_T, z); \\ j_t(s_t, z) &= \inf_{a_t \in A} \mathbb{E}_{w_t}\left[g_t(s_t, a_t, w_t) + j_{t+1}(f_t(s_t, a_t, w_t), z) \right]. \end{aligned}$$

C.1 Proof of Theorem 3.1

We aim to show that the population-level cost-to-go function in Problem 2 is upper bounded by an optimal transport problem whose cost is the single-agent cost-to-go function in (2). Concretely, we show by backward induction that

$$J_t(\mu_t, \nu) \leq \inf_{\gamma \in \Gamma(\mu_t, \nu)} \int j_t(s, z) \mathrm{d}\gamma(s, z).$$

We first prove the theorem under the field-based setting of Assumptions A.1 and A.2. The constructions in Examples A.4–A.7 and Proposition A.9 cover individual noise with finitely or infinitely many agents, global noise, and local noise on countably supported state-action distributions. For atomless local noise, the footnote in Example A.6 gives the corresponding aggregate random-measure construction and explains how the same proof carries over. In the field-based setting, Lemma A.10 allows us to replace the realized noise kernels by the nominal kernel after taking expectations. Throughout, Assumption A.1 makes the gluing $\rho_t = \lambda_t \otimes \hat{\xi}_t^{s,a}$ a well-defined random measure and justifies interchanging expectations over the realized field with λ_t-integrals. All integrands below are lower semianalytic (see Appendix B), so all expectations are well-defined.

We start with the proof of Part I of the theorem.

Proof. The proof proceeds by induction, in three steps. First, we use the definition of the cost-to-go function and the induction hypothesis at time $t+1$. Second, we express the conditional expectations over noise realizations in terms of the averaged single-agent cost. Finally, we reorganize the infima over couplings by means of Lemma B.4, which allows us to recover the desired distributional dynamic program and conclude the induction.

The base case of the induction is straightforward. By definition $J_T(\mu_T, \nu) = \mathcal{K}[g_T](\mu_T, \nu)$. Since $j_T(s, z) = g_T(s, z)$, we have $\mathcal{K}[j_T](\mu_T, \nu) = \mathcal{K}[g_T](\mu_T, \nu) = J_T(\mu_T, \nu)$. Fix $t < T$, using the induction hypothesis at time $t+1$, i.e.,

$$J_{t+1}(f_{t\#}\rho_t, \nu) \leq \inf_{\gamma \in \Gamma(f_{t\#}\rho_t, \nu)} \int_{S\times Z} j_{t+1}(s^+, z) \mathrm{d}\gamma,$$

the backward recursion for the cost-to-go function at time t gives

$$\begin{aligned} J_t(\mu_t, \nu) &= \inf_{(\mathrm{proj}_S^{S\times A})_{\#}\lambda_t = \mu_t} \mathbb{E}_{\eta_t}\left[\mathbb{E}_{\rho_t}\left[g_t(s, a, w) \right] + J_{t+1}(f_{t\#}\rho_t, \nu) \right] \\ &\leq \inf_{(\mathrm{proj}_S^{S\times A})_{\#}\lambda_t = \mu_t} \mathbb{E}_{\eta_t}\left[\underbrace{\int_{S\times A\times W} g_t(s, a, w) \mathrm{d}\rho_t + \inf_{\gamma \in \Gamma(f_{t\#}\rho_t, \nu)} \int_{S\times Z} j_{t+1}(s^+, z) \mathrm{d}\gamma}_{(*)} \right]. \end{aligned} \tag{12}$$

We now focus on the term denoted by $(*)$. We analyze $(*)$ pointwise for each realization of $\hat{\xi}_t$ before taking expectations. By Lemma B.1, we can "push" the dynamics into the transportation cost, so that

$$(*) = \int_{S\times A\times W} g_t(s,a,w)\mathrm{d}\rho_t + \inf_{\gamma'\in\Gamma(\rho_t,\nu)} \int_{S\times A\times W\times Z} j_{t+1}(f_t(s,a,w),z)\mathrm{d}\gamma'.$$

Here we use that the coupling correspondence in Lemma B.1 preserves the target marginal: since $f_t \times \mathrm{Id}_Z$ acts as the identity on the Z-coordinate, it restricts to a bijection between the couplings with Z-marginal ν, so imposing $\gamma' \in \Gamma(\rho_t, \nu)$ (with ν fixed) is equivalent to the corresponding fixed-ν constraint on $f_{t\#}\rho_t$. Since ρ_t is the first marginal of γ', Lemma B.2, with $X = S \times A \times W$, $c_1 = g_t$ and $c_2 = j_{t+1} \circ (f_t \times \mathrm{Id}_Z)$, allows us to merge the two costs into a single transport objective on $S \times A \times W \times Z$. We therefore have

$$(*) = \inf_{\gamma'\in\Gamma(\rho_t,\nu)} \int_{S\times A\times W\times Z} g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\gamma'.$$

We now use Lemma B.6 with $X = S \times A$, $\rho_t = \lambda_t \otimes \hat{\xi}_t^{s,a}$, and $c(s,a,w,z) = g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)$, to replace the couplings of the full state-action-noise distribution ρ_t with couplings of its $(S \times A)$-marginal λ_t. This yields the inequality

$$(*) \le \inf_{\tilde{\gamma}\in\Gamma(\lambda_t,\nu)} \int_{S\times A\times Z} \int_W g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\hat{\xi}_t^{s,a}\mathrm{d}\tilde{\gamma}.$$

This relaxation removes the dependence on the realized noise from the optimal assignment, thus restricting it to depend only on (s,a).

We now plug the upper bound obtained for $(*)$ inside the expectation over $\hat{\xi}_t^{s,a}$, as in (12). Since λ_t is fixed and the infimum is now over couplings $\tilde{\gamma} \in \Gamma(\lambda_t, \nu)$ that do not depend on $\hat{\xi}_t^{s,a}$, Lemma B.3 applies, allowing us to interchange expectation and infimum and yielding the upper bound

$$\begin{aligned}
\mathbb{E}_{\eta_t}[(*)] &\le \mathbb{E}_{\eta_t}\left[\inf_{\tilde{\gamma}\in\Gamma(\lambda_t,\nu)} \int_{S\times A\times Z}\int_W g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\hat{\xi}_t^{s,a}\mathrm{d}\tilde{\gamma}\right] \\
&\le \inf_{\tilde{\gamma}\in\Gamma(\lambda_t,\nu)} \mathbb{E}_{\eta_t}\left[\int_{S\times A\times Z}\int_W g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\hat{\xi}_t^{s,a}\mathrm{d}\tilde{\gamma}\right] \\
&= \inf_{\tilde{\gamma}\in\Gamma(\lambda_t,\nu)} \int_{S\times A\times Z} \mathbb{E}_{\eta_t^{s,a}}\left[\int_W g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\hat{\xi}_t^{s,a}\right]\mathrm{d}\tilde{\gamma},
\end{aligned}$$

where the last equality follows from Tonelli's theorem, which applies by nonnegativity of the costs and joint measurability of the realized field (Assumption A.1). Informally, the second inequality is precisely where we transition from "choose the assignment after seeing the realized noise" to "choose one assignment against average noise". We now focus on the inner expectation. In particular, consider a fixed $\tilde{\gamma} \in \Gamma(\lambda_t, \nu)$. We can disintegrate the coupling with respect to its $(S \times A)$-marginal λ_t as $\tilde{\gamma} = \lambda_t \otimes \kappa_{s,a}$ where $\kappa_{s,a}$ is a (conditional) probability over Z. By non-negativity of all functions, Tonelli's theorem gives

$$\begin{aligned}
&\int_{S\times A\times Z} \mathbb{E}_{\eta_t^{s,a}}\left[\int_W g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\hat{\xi}_t^{s,a}\right]\mathrm{d}\tilde{\gamma} \\
&\qquad = \int_{S\times A} \mathbb{E}_{\eta_t^{s,a}}\left[\int_W\int_Z g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\kappa_{s,a}\mathrm{d}\hat{\xi}_t^{s,a}\right]\mathrm{d}\lambda_t.
\end{aligned}$$

We can apply Lemma A.10 (treating (s,a) as parameters) to the function

$$\tilde{c}(s,a,w) \coloneqq \int_Z g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\kappa_{s,a}(z).$$

Note, in particular, that Lemma A.10 can be applied since $\kappa_{s,a}$ is fixed and not a function of the noise distribution $\hat{\xi}_t^{s,a}$. Thus, we get

$$\begin{aligned}
&\int_{S\times A} \mathbb{E}_{\eta_t^{s,a}}\left[\int_W\int_Z g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\kappa_{s,a}\mathrm{d}\hat{\xi}_t^{s,a}\right]\mathrm{d}\lambda_t \\
&\qquad = \int_{S\times A} \mathbb{E}_w\left[\int_Z g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\mathrm{d}\kappa_{s,a}\right]\mathrm{d}\lambda_t \\
&\qquad = \int_{S\times A\times Z} \mathbb{E}_w\left[g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\right]\mathrm{d}\tilde{\gamma}
\end{aligned}$$

where the last step follows again from Tonelli's theorem and the definition of $\tilde{\gamma}$. Combined with the previous chain of inequalities, we therefore have

$$\mathbb{E}_{\eta_t}\left[(*)\right] \leq \inf_{\tilde{\gamma}\in\Gamma(\lambda_t,\nu)} \int_{S\times A\times Z} \mathbb{E}_w\left[g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\right] \mathrm{d}\tilde{\gamma}.$$

We can now plug back our bound in (12) to get

$$\begin{aligned} J_t(\mu_t,\nu) &\leq \inf_{(\mathrm{proj}_S^{S\times A})_{\#}\lambda_t=\mu_t} \inf_{\tilde{\gamma}\in\Gamma(\lambda_t,\nu)} \int_{S\times A\times Z} \mathbb{E}_w\left[g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\right] \mathrm{d}\tilde{\gamma} \\ &= \inf_{(\mathrm{proj}_{S\times Z}^{S\times A\times Z})_{\#}\tilde{\gamma}\in\Gamma(\mu_t,\nu)} \int_{S\times A\times Z} \mathbb{E}_w\left[g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\right] \mathrm{d}\tilde{\gamma} \\ &=: (\star). \end{aligned}$$

Now, we can use Lemma B.4 to disintegrate the optimizer in $(\star)$, which gives

$$(\star) = \inf_{\substack{\tilde{\gamma}\in\Gamma(\mu_t,\nu)\\ \alpha^{s,z}\in\Lambda(A)}} \int_{S\times Z}\int_A \mathbb{E}_w\left[g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\right] \mathrm{d}\alpha^{s,z}\mathrm{d}\tilde{\gamma}$$

The integrand is lower semianalytic: g_t, g_T, and f_t are Borel, and, by backward induction, j_{t+1} is lower semianalytic, since lower semianalyticity is preserved by composition with Borel maps [31, Lemma 7.30], integration against the Borel kernel ξ_t [31, Proposition 7.48], and pointwise infima over A [31, Proposition 7.47]. Therefore, we can apply Lemma B.5, which gives

$$(\star) = \inf_{\tilde{\gamma}\in\Gamma(\mu_t,\nu)} \int_{S\times Z} \inf_{a\in A} \mathbb{E}_w\left[g_t(s,a,w) + j_{t+1}(f_t(s,a,w),z)\right] \mathrm{d}\tilde{\gamma}.$$

At this point, we recognize that the term in the integral coincides with the Bellman recursion for j_t:

$$j_t(s_t,z) = \inf_{a_t\in A} \mathbb{E}_{w_t}\left[g_t(s_t,a_t,w_t) + j_{t+1}(f_t(s_t,a_t,w_t),z)\right].$$

Hence,

$$J_t(\mu_t,\nu) \leq (\star) = \inf_{\tilde{\gamma}\in\Gamma(\mu_t,\nu)} \int_{S\times Z} j_t(s,z)\mathrm{d}\tilde{\gamma}.$$

This concludes the inductive step and the proof of the theorem. □

Two comments on our proof are in order. First, while the proof follows the inductive strategy of the proof of the deterministic separation in [27], there is a key difference: the next population state is random and depends on the realized empirical noise kernel $\hat{\xi}_t$. This requires various additional steps (and preliminary lemmata) in the proof. All these steps trivialize in the absence of noise and, thus, our proof recovers the deterministic separation in [27] when the dynamics are deterministic. Second, our result is an upper bound on the cost-to-go function. Intuitively, this conservatism originates from the following fact: The population state-action distribution can, in principle, choose future couplings *after* observing this realization, while our upper bound posits that couplings are chosen *before* the noise is realized, averaging over the kernel $\xi_t(\cdot \mid s,a)$. This is precisely where Lemmas B.3 and B.6 yield inequalities rather than equalities. In the deterministic case, all these steps trivialize and we obtain equality as in [27].

We now prove Part II of the theorem.

Proof. For each time step t, consider the function

$$\mathcal{K}[j_t](\mu_t,\nu) = \inf_{\gamma\in\Gamma(\mu_t,\nu)} \int_{S\times Z} j_t(s,z)\mathrm{d}\gamma.$$

Let $\varepsilon > 0$. At each time t, we can construct a "separation-based" state-action distribution λ_t^ε as follows: given a population distribution μ_t, choose an $\frac{\varepsilon}{2T}$-optimal coupling $\gamma_t^\varepsilon \in \Gamma(\mu_t,\nu)$ satisfying

$$\int_{S\times Z} j_t(s,z)\mathrm{d}\gamma_t^\varepsilon \leq \mathcal{K}[j_t](\mu_t,\nu) + \frac{\varepsilon}{2T}, \tag{13}$$

and a universally measurable $\frac{\varepsilon}{2T}$-optimal action selector $a_t^\varepsilon \colon S \times Z \to A$ satisfying, for all (s, z),

$$\mathbb{E}_w \left[g_t(s, a_t^\varepsilon(s,z), w) + j_{t+1}(f_t(s, a_t^\varepsilon(s,z), w), z) \right] \leq j_t(s,z) + \frac{\varepsilon}{2T}. \tag{14}$$

Formally, we have

$$\lambda_t^\varepsilon \coloneqq \left(\mathrm{proj}_S^{S\times Z}, a_t^\varepsilon\right)_\# \gamma_t^\varepsilon.$$

In words, this state-action distribution assigns agents to targets through γ_t^ε and controls the individual agents through a_t^ε. Since the first marginal of γ_t^ε is μ_t, we have $\lambda_t^\varepsilon \in M(\mu_t)$. In the rest of the proof, we will show that the resulting control sequence satisfies the bound,

$$J_0^\varepsilon(\mu_0, \nu) \coloneqq J_0^{\lambda_0^\varepsilon, \ldots, \lambda_{T-1}^\varepsilon}(\mu_0, \nu) \leq \mathcal{K}[j_0](\mu_0, \nu) + \varepsilon. \tag{15}$$

Before doing so, we verify well-posedness of all the objects. The $\frac{\varepsilon}{2T}$-optimal universally measurable selector a_t^ε exists by [31, Proposition 7.50], since the Bellman integrand in (14) is lower semianalytic and nonnegative (see the induction argument in the proof of Part I). For the existence of the $\frac{\varepsilon}{2T}$-optimal coupling we work as follows. Since j_t is lower semianalytic (by backward induction on (2)), the map $\gamma \mapsto \int_{S\times Z} j_t \mathrm{d}\gamma$ is lower semianalytic by [31, Corollary 7.48.1] (jointly in (μ, γ), as it does not depend on μ). Further, the set $\{(\mu, \gamma) : \gamma \in \Gamma(\mu, \nu)\}$ is closed. Indeed, fixing $\nu \in \mathcal{P}(Z)$, define

$$\mathcal{G}_Z \coloneqq \{(\mu, \gamma) \in \mathcal{P}(S) \times \mathcal{P}(S \times Z) : \gamma \in \Gamma(\mu, \nu)\}.$$

The maps $F_S(\gamma) \coloneqq \left(\mathrm{proj}_S^{S\times Z}\right)_\# \gamma$ and $F_Z(\gamma) \coloneqq \left(\mathrm{proj}_Z^{S\times Z}\right)_\# \gamma$ are continuous for the weak topologies, being the pushforward under a continuous map. Hence,

$$\mathcal{G}_Z = \{(\mu, \gamma) : F_S(\gamma) = \mu\} \cap \{(\mu, \gamma) : F_Z(\gamma) = \nu\}.$$

The first set is the preimage of the diagonal in $\mathcal{P}(S) \times \mathcal{P}(S)$ under the continuous map $(\mu, \gamma) \mapsto (\mu, F_S(\gamma))$, and the second is the preimage of the closed singleton $\{\nu\}$ under the continuous map $(\mu, \gamma) \mapsto F_Z(\gamma)$. Since these spaces are Polish, the diagonal and singleton are closed. By continuity of F_S and F_Z, the set $\mathcal{G}_Z$ is closed, hence Borel and in particular analytic. Thus, we can apply [31, Proposition 7.50] to obtain the existence of a universally measurable map

$$\mu \mapsto \gamma_t^\varepsilon(\mu) \in \Gamma(\mu, \nu) \quad \text{with} \int_{S\times Z} j_t \mathrm{d}\gamma_t^\varepsilon(\mu) \leq \mathcal{K}[j_t](\mu_t, \nu) + \frac{\varepsilon}{2T}.$$

This shows the existence of a universally measurable selector $\gamma_t^\varepsilon(\mu)$. In particular, for fixed μ, $\gamma_t^\varepsilon(\mu)$ is a Borel probability measure, and λ_t^ε is a well-defined Borel probability measure: it is the pushforward of $\gamma_t^\varepsilon(\mu)$ under the universally measurable map $(\mathrm{proj}_S^{S\times Z}, a_t^\varepsilon)$, which coincides $\gamma_t^\varepsilon(\mu)$-almost everywhere with a Borel map [31, Lemma 7.27].

We now turn our attention to the desired bound (15). We now show by backward induction the more general bound

$$J_t^\varepsilon(\mu_t, \nu) \coloneqq J_t^{\lambda_t^\varepsilon, \ldots, \lambda_{T-1}^\varepsilon}(\mu_t, \nu) \leq \mathcal{K}[j_t](\mu_t, \nu) + \frac{T-t}{T}\varepsilon, \tag{16}$$

which, evaluated at time $t = 0$, readily implies (15). The base case $t = T$ holds exactly by definition of terminal cost and $j_T = g_T$:

$$J_T^\varepsilon(\mu_T, \nu) = \mathcal{K}[g_T](\mu_T, \nu) = \inf_{\gamma \in \Gamma(\mu_T, \nu)} \int_{S\times Z} g_T(s,z) \mathrm{d}\gamma = \mathcal{K}[j_T](\mu_T, \nu).$$

Now fix $t < T$ and assume (16) holds at $t + 1$. By the policy-cost recursion (which uses the independence of the noise across time) and the induction hypothesis at $t + 1$ applied inside the expectation, with $\rho_t^\varepsilon = \lambda_t^\varepsilon \otimes \hat{\xi}_t^{s,a}$,

$$J_t^\varepsilon(\mu_t, \nu) \leq \mathbb{E}_{\eta_t} \left[\int_{S\times A\times W} g_t(s, a, w) \mathrm{d}\rho_t^\varepsilon + \mathcal{K}[j_{t+1}](\mu_{t+1}, \nu) \right] + \frac{T-t-1}{T}\varepsilon. \tag{17}$$

Next, we define

$$\bar{\gamma}_t^\varepsilon \coloneqq (f_t, \mathrm{Id}_Z)_\# (\sigma_\# \tilde{\gamma}_t^\varepsilon) \in \mathcal{P}(S \times Z)$$

where

$$\tilde{\gamma}_t^\varepsilon \coloneqq (\mathrm{Id}_{S\times Z}, a_t^\varepsilon)_\# \gamma_t^\varepsilon \otimes \hat{\xi}_t^{s, a_t^\varepsilon(s,z)} \in \mathcal{P}(S \times Z \times A \times W).$$

Where we use the permutation

$$\sigma : S \times Z \times A \times W \to S \times A \times W \times Z, \quad \sigma(s,z,a,w) = (s,a,w,z),$$

to "re-order" the marginals. (The gluing is well-posed: $(s,z) \mapsto \hat{\xi}_t^{s,a_t^\varepsilon(s,z)}$ is a universally measurable kernel, as the composition of the field of Assumption A.1 with the universally measurable map $(\mathrm{Id}_S, a_t^\varepsilon)$.) Informally, $\bar{\gamma}_t^\varepsilon$ is the law of $(f_t(s, a_t^\varepsilon(s,z), w), z)$ where $(s,z) \sim \gamma_t^\varepsilon$ and $w \sim \hat{\xi}_t^{s,a_t^\varepsilon(s,z)}$. We now show $\bar{\gamma}_t^\varepsilon \in \Gamma(\mu_{t+1}, \nu)$. Specifically, the second marginal is ν, since the z-coordinate is unchanged by σ and by (f_t, Id_Z), and $(\mathrm{proj}_Z^{S\times Z})_\# \gamma_t^\varepsilon = \nu$. For the first marginal, the $(S \times A \times W)$-marginal of $\sigma_\# \tilde{\gamma}_t^\varepsilon$ equals ρ_t^ε, since the noise kernel depends on (s,z) only through the state-action pair $(s, a_t^\varepsilon(s,z))$, whose law under $(\mathrm{Id}_{S\times Z}, a_t^\varepsilon)_\# \gamma_t^\varepsilon$ is λ_t^ε by definition. Hence the first marginal of $\bar{\gamma}_t^\varepsilon$ is $f_{t\#}\rho_t^\varepsilon = \mu_{t+1}$, completing the proof of $\bar{\gamma}_t^\varepsilon \in \Gamma(\mu_{t+1}, \nu)$. By definition of the infimum, we then obtain

$$\mathcal{K}[j_{t+1}](\mu_{t+1}, \nu) \le \int_{S\times Z} j_{t+1}(s,z)\mathrm{d}\bar{\gamma}_t^\varepsilon = \int_{S\times Z}\int_W j_{t+1}(f_t(s, a_t^\varepsilon(s,z), w), z)\mathrm{d}\hat{\xi}_t^{s,a_t^\varepsilon(s,z)}\mathrm{d}\gamma_t^\varepsilon.$$

Taking expectations over the realized noise field and applying Tonelli's theorem and Lemma A.10 as in the proof of Part I (with λ_t^ε in place of λ_t), we get

$$\begin{aligned}
&\mathbb{E}_{\eta_t}\left[\int_{S\times A\times W} g_t(s,a,w)\mathrm{d}\rho_t^\varepsilon + \mathcal{K}[j_{t+1}](\mu_{t+1},\nu)\right] \\
&\quad\le \mathbb{E}_{\eta_t}\left[\int_{S\times Z}\int_W g_t(s, a_t^\varepsilon(s,z), w) + j_{t+1}(f_t(s, a_t^\varepsilon(s,z), w), z)\mathrm{d}\hat{\xi}_t^{s,a_t^\varepsilon(s,z)}\mathrm{d}\gamma_t^\varepsilon\right] \\
&\quad= \int_{S\times Z} \mathbb{E}_w\left[g_t(s, a_t^\varepsilon(s,z), w) + j_{t+1}(f_t(s, a_t^\varepsilon(s,z), w), z)\right]\mathrm{d}\gamma_t^\varepsilon.
\end{aligned}$$

Then, we use the $\frac{\varepsilon}{2T}$-optimality of a_t^ε in (14) and of γ_t^ε in (13) to get

$$\begin{aligned}
&\int_{S\times Z} \underbrace{\mathbb{E}_w\left[g_t(s, a_t^\varepsilon(s,z), w) + j_{t+1}(f_t(s, a_t^\varepsilon(s,z), w), z)\right]}_{\le j_t(s,z) + \varepsilon/(2T)}\mathrm{d}\gamma_t^\varepsilon \\
&\qquad\qquad \le \int_{S\times Z} j_t(s,z)\mathrm{d}\gamma_t^\varepsilon + \frac{\varepsilon}{2T} \le \mathcal{K}[j_t](\mu_t,\nu) + \frac{\varepsilon}{2T} + \frac{\varepsilon}{2T}.
\end{aligned}$$

We can now combine this bound with (17) and get

$$J_t^\varepsilon(\mu_t,\nu) \le \mathcal{K}[j_t](\mu_t,\nu) + \frac{\varepsilon}{T} + \frac{T-t-1}{T}\varepsilon = \mathcal{K}[j_t](\mu_t,\nu) + \frac{T-t}{T}\varepsilon.$$

This concludes the inductive step. The bound at $t=0$ gives $J_0^\varepsilon(\mu_0,\nu) \le \mathcal{K}[j_0](\mu_0,\nu) + \varepsilon$. □

Remark C.1 (Choosing $\varepsilon = 0$ under stronger regularity). The proof of Part II only requires universally measurable ε-optimal coupling selectors, which exist under the lower semianalytic assumptions used throughout Appendix B. Under the stronger assumption that the cost-to-go functions $j_t : S \times Z \to \bar{\mathbb{R}}_{\ge 0}$ are lower semicontinuous and the spaces are compact, the infimum in $\mathcal{K}[j_t](\mu,\nu)$ is attained for each fixed (μ,ν) by the standard existence theorem for optimal couplings [46, Theorem 4.1]. This pointwise existence is, however, not sufficient to choose $\varepsilon = 0$ in Theorem 3.1, Part II: the construction requires *measurable selectors* of exact optimizers. These exist when all spaces are compact, g_t, g_T, f_t are continuous, and ξ_t is weakly continuous: then j_t and the Bellman integrand are lower semicontinuous by backward induction, and exact Borel minimizing selectors exist [31, Proposition 7.33] (see also [32, Theorem 18.19]); the same argument applies to the coupling problem, where $\mu \mapsto \Gamma(\mu,\nu)$ is a compact-valued continuous correspondence.

C.2 Proof of Corollary 3.3

Proof. We construct, for every $\varepsilon > 0$, an admissible separation policy whose cost is within $\frac{(T-t)\varepsilon}{T}$ of $\mathcal{K}[j_t](\mu_t^N, \nu^N)$. Since $\varepsilon > 0$ is arbitrary, and J_t^N is the infimum of the expected cost over all admissible policies, the claim follows. We start by giving an ε-optimal separation policy $\pi_{\mathrm{sep},t}^\varepsilon$ which

attains the upper bound. We then show that this policy is admissible, from which we directly obtain the desired inequality.

To define $\pi^{\varepsilon}_{\mathrm{sep},t}$ we proceed as follows. First, since μ^N_t and ν^N are both empirical measures over N atoms, the optimal transport problem $\mathcal{K}[j_t](\mu^N_t, \nu^N)$ is a finite-dimensional transportation linear program. We work at the level of agent and target indices, i.e., we replace μ^N_t and ν^N by the uniform measures on $\{1, \dots, N\}$ and transport the cost $c_{ij} = j_t(s^i_t, z^j)$. With this identification, every coupling between μ^N_t and ν^N can be represented by a matrix $\Gamma = (\gamma_{ij})_{i,j=1}^N$ satisfying $\gamma_{ij} \geq 0$, $\sum_{i=1}^N \gamma_{ij} = \sum_{j=1}^N \gamma_{ij} = \frac{1}{N}$; equivalently, $N\Gamma$ is doubly stochastic. By the Birkhoff–von Neumann theorem [47] applied to $N\Gamma$, every such matrix is a convex combination of $\frac{1}{N}$-scaled permutation matrices; since the objective is linear, the minimum is attained by a $\frac{1}{N}$-scaled permutation matrix. Thus,

$$\mathcal{K}[j_t](\mu^N_t, \nu^N) = \frac{1}{N} \min_{\sigma_t \in S_N} \sum_{i=1}^N j_t(s^i_t, z^{\sigma_t(i)}),$$

where S_N denotes the group of permutations of $\{1, \dots, N\}$. Since S_N is finite and each map $\boldsymbol{s} \mapsto \sum_{i=1}^N j_t(s^i, z^{\sigma(i)})$ is universally measurable, a minimizing permutation $\sigma^*_t \coloneqq \sigma^*_t(\boldsymbol{s}_t)$ can be selected measurably (e.g., the first minimizer in a fixed ordering of S_N); write $z^{*,i}_t \coloneqq z^{\sigma^*_t(i)}$ for the target assigned to agent i. Given that agents are now assigned to targets, we can now consider the single-agent optimal control problem from s^i_t to the corresponding target $z^{*,i}$. Specifically, for $\varepsilon > 0$, there exists a universally measurable $\frac{\varepsilon}{T}$-optimal action selector $a^{\varepsilon}_t \colon S \times Z \to A$, satisfying, for all $(s, z) \in S \times Z$:

$$\mathbb{E}_w \left[g_t(s, a^{\varepsilon}_t(s, z), w) + j_{t+1}(f_t(s, a^{\varepsilon}_t(s, z), w), z) \right] \leq j_t(s, z) + \frac{\varepsilon}{T}.$$

At this point, we define the separation policy $\pi^{\varepsilon}_{\mathrm{sep},t}$ at time t. An agent at s^i_t executes the action $a^{\varepsilon}_t(s^i_t, z^{*,i}_t)$. Informally, the optimal transport problem indicates the target, whereas the single-agent optimal control problem determines the optimal action to reach that target. This policy is deterministic, Markovian (it depends only on $\boldsymbol{s}_t$), and universally measurable, being the composition of a^{ε}_t with the assignment map $\boldsymbol{s}_t \mapsto (s^i_t, z^{*,i}_t)$; hence $\pi^{\varepsilon}_{\mathrm{sep}} \in \Pi$.

At this point, it suffices to verify that this policy yields the bound on the cost. In particular, we show that the expected cost-to-go function under this policy, denoted by $J^{N,\varepsilon}_t(\boldsymbol{s}_t, \nu^N)$, is upper bounded as follows:

$$J^{N,\varepsilon}_t(\boldsymbol{s}_t, \nu^N) \leq \mathcal{K}[j_t](\mu^N_t, \nu^N) + \frac{T-t}{T}\varepsilon,$$

which gives the desired upper bound. We proceed by induction. The equality at time $t = T$ is trivial, as the terminal cost does not depend on the policy but only on the terminal state. Suppose then the statement holds at time $t+1$. By the policy-cost recursion (using the independence of the noise across time) and the induction hypothesis inside the expectation,

$$\begin{aligned}
J^{N,\varepsilon}_t(\boldsymbol{s}_t, \nu^N) &\leq \mathbb{E}_{\boldsymbol{w}_t} \left[\frac{1}{N} \sum_{i=1}^N g_t(s^i_t, a^i_t, w^i_t) + \mathcal{K}[j_{t+1}](\mu^N_{t+1}, \nu^N) \right] + \frac{T-t-1}{T}\varepsilon \\
&= \frac{1}{N} \mathbb{E}_{\boldsymbol{w}_t} \left[\sum_{i=1}^N g_t(s^i_t, a^i_t, w^i_t) + \min_{\sigma_{t+1} \in S_N} \sum_{i=1}^N j_{t+1}(f_t(s^i_t, a^i_t, w^i_t), z^{\sigma_{t+1}(i)}) \right] \\
&\qquad + \frac{T-t-1}{T}\varepsilon \\
&\leq \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{w^i_t} \left[g_t(s^i_t, a^i_t, w^i_t) + j_{t+1}(f_t(s^i_t, a^i_t, w^i_t), z^{\sigma^*_t(i)}) \right] + \frac{T-t-1}{T}\varepsilon \\
&\leq \frac{1}{N} \sum_{i=1}^N j_t(s^i_t, z^{\sigma^*_t(i)}) + \frac{\varepsilon}{T} + \frac{T-t-1}{T}\varepsilon \\
&= \frac{1}{N} \min_{\sigma \in S_N} \sum_{i=1}^N j_t(s^i_t, z^{\sigma(i)}) + \frac{T-t}{T}\varepsilon,
\end{aligned}$$

where the first inequality combines the policy-cost recursion with the inductive assumption, the second inequality follows since σ_t^* is a potentially suboptimal permutation for the assignment at time $t+1$, the last inequality follows from the $\frac{\varepsilon}{T}$-optimality of the action selector a_t^ε used by the policy $\pi_{\mathrm{sep},t}^\varepsilon$, and the last equality follows from the definition of σ_t^*. Note that the expectation step above only uses the marginal law $w_t^i \sim \xi_t(\cdot \mid s_t^i, a_t^i)$ of each agent's noise together with the linearity of expectation; it is therefore valid under all noise models considered in this work (individual, local, and global), regardless of the correlation of the noise across agents. This gives

$$J_t^{N,\varepsilon}(\boldsymbol{s}_t, \nu^N) \le \mathcal{K}[j_t](\mu_t^N, \nu^N) + \frac{T-t}{T}\varepsilon,$$

which completes the induction.

Since $\varepsilon > 0$ is arbitrary, and J_t^N is the infimum of the expected cost over all admissible policies, letting $\varepsilon \downarrow 0$ gives

$$J_t^N(\boldsymbol{s}_t, \nu^N) \le \mathcal{K}[j_t](\mu_t^N, \nu^N).$$

□

D Quantifying the tightness of the optimal transport upper bound

The inequality in the proof of Theorem 3.1, Part I emerges from the fact that we remove the dependence of the target assignment on the realized noise. More precisely, we first apply Lemma B.6 to remove the transport plan dependence on the noise in the term $(*)$ of (12), and then apply Lemma B.3 to allow the transport plan to be chosen before the noise is realized. In this appendix, we quantify the loss introduced by these two steps.

Throughout the appendix, we retain Assumptions A.1 and A.2 and write $\hat{\xi}_\tau = \big(\hat{\xi}_\tau^{s,a}\big)_{(s,a)}$ for the realized noise field. Recall that Assumption A.2 reads

$$\mathbb{E}_{\eta_\tau}\left[\int_W \varphi(w)\mathrm{d}\hat{\xi}_\tau^{s,a}(w)\right] = \int_W \varphi(w)\mathrm{d}\xi_\tau(w \mid s,a)$$

for every bounded Borel function φ and almost every (s,a). By truncation and monotone convergence, the same identity holds for every nonnegative Borel φ, which is the form we use below.

Assumption D.1 (Lipschitz single-agent cost-to-go)**.** Equip S with a metric d_S. We assume that for every $\tau \in \{1,\dots,T\}$ and for some $L_\tau < \infty$,

$$|j_\tau(s,z) - j_\tau(s',z)| \le L_\tau d_S(s,s') \tag{18}$$

for all $s, s' \in S$ and $z \in Z$.

We need the constant to be uniform in z because the target assigned to an agent may itself be reassigned once the noise has been observed. The assumption holds, for instance, whenever all the underlying spaces are finite and the single-agent cost-to-go functions are finite-valued.

For each $\tau < T$, we measure the spread of the transition by

$$D_\tau := \sup_{(s,a)\in S\times A}\int_{W\times W} d_S(f_\tau(s,a,w), f_\tau(s,a,\tilde{w}))\,\mathrm{d}\xi_\tau(w \mid s,a)\mathrm{d}\xi_\tau(\tilde{w} \mid s,a), \tag{19}$$

that is, by the expected distance between two conditionally independent next states generated from the same state-action pair. We assume throughout that D_τ, together with the expectations and transport costs appearing in the comparison below, is finite. This is the case, for example, when $w \mapsto f_\tau(s,a,w)$ is Lipschitz uniformly in (s,a), the nominal kernels have uniformly bounded first moments, and $\mathcal{K}[j_\tau](\mu,\nu) < \infty$, or more simply when all the underlying spaces are compact and data are continuous.

For the following analysis, it is convenient to write the recursion (4) in operator form. For a next-stage function F and $\rho_\tau^\lambda := \lambda \otimes \hat{\xi}_\tau^{s,a}$, we hence set

$$(\mathcal{T}_\tau F)(\mu,\nu) := \inf_{\lambda\in M(\mu)} \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau(s,a,w)\mathrm{d}\rho_\tau^\lambda(s,a,w) + F\big(f_{\tau\#}\rho_\tau^\lambda, \nu\big)\right],$$

so that $J_\tau = \mathcal{T}_\tau J_{\tau+1}$ by (4). The population Bellman operator $\mathcal{T}_\tau$ satisfies the monotonicity and constant-shift properties that are standard for dynamic programming operators [31, Ch. 8]. More precisely, the following lemma holds.

Lemma D.2 (Monotonicity and constant shift of the population Bellman operator)**.** *Let $F, \widetilde{F}\colon \mathcal{P}(S)\times \mathcal{P}(Z) \to \bar{\mathbb{R}}$ be universally measurable and bounded below, and let $c \in \mathbb{R}$. Then*

(i) $F \le \widetilde{F}$ pointwise implies $\mathcal{T}_\tau F \le \mathcal{T}_\tau \widetilde{F}$ pointwise;

(ii) $\mathcal{T}_\tau[F + c] = \mathcal{T}_\tau[F] + c$.

Proof. Both properties are immediate from the definition of $\mathcal{T}_\tau$: the next-stage function enters the right-hand side additively and monotonically, while a constant depends neither on the realized next population nor on λ, so it factors out of the expectation and out of the infimum. The argument uses only the structure of the definition. □

Finally, we also recall that

$$\mathcal{K}[j_\tau](\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{S\times Z} j_\tau(s,z)\mathrm{d}\gamma(s,z).$$

Here, γ assigns the mass in the current state distribution μ to the target distribution ν, and $j_\tau(s,z)$ is the single-agent cost associated with each state–target pair. This assignment is made before the noise at time τ is observed. In contrast, $\mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]]$ assigns targets after the realized noise has determined the next population. The difference in (20) is the cost of not being able to use that realized noise when assigning targets.

Lemma D.3 (One-step Bellman defect)**.** *Under Assumptions A.1, A.2, and D.1, for every $\tau < T$, $\mu \in \mathcal{P}(S)$, and $\nu \in \mathcal{P}(Z)$, it holds that*

$$0 \le \mathcal{K}[j_\tau](\mu,\nu) - \mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]]\,(\mu,\nu) \le L_{\tau+1} D_\tau. \tag{20}$$

Proof. Throughout this proof we abbreviate the state-action pair as $x = (s,a)$. We start by showing that the two quantities in (20) have the same expected stage cost. As a consequence, we only need to compare how the next-stage single-agent cost-to-go $j_{\tau+1}(f_\tau(s,a,w),z)$ is minimized in the two transport problems. In $\mathcal{K}[j_\tau](\mu,\nu)$, the target assignment depends only on x, whereas in $\mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]]\,(\mu,\nu)$ it may also depend on the realized noise. The second minimization is therefore over a larger set of assignments and cannot have a larger value. In the following proof, we show that the difference between the two values is at most $L_{\tau+1}D_\tau$.

Step 1: Define the two next-stage transport problems. Fix $\lambda \in M(\mu)$. If the target is assigned using x alone, the expected next-stage cost-to-go is

$$A_\tau(\lambda) \coloneqq \inf_{\gamma\in\Gamma(\lambda,\nu)} \int_{(S\times A)\times Z} \int_W j_{\tau+1}\big(f_\tau(s,a,\tilde{w}),z\big)\mathrm{d}\xi_\tau(\tilde{w}\,|\,x)\mathrm{d}\gamma(x,z).$$

Here, the plan γ assigns a target z to each $x=(s,a)$, before the noise $\tilde{w}$ is drawn from the nominal kernel. Hence, this target assignment does not depend on $\tilde{w}$.

On the other hand, if the noise field is observed before the target is assigned, the transport plan may couple the target with the full pair (x,w). For a fixed realization of the noise field, the smallest next-stage cost-to-go is then

$$A_\tau(\lambda;\hat{\xi}_\tau) \coloneqq \inf_{\gamma'\in\Gamma(\lambda\otimes\hat{\xi}_\tau^x,\nu)} \int_{(S\times A)\times W\times Z} j_{\tau+1}\big(f_\tau(s,a,w),z\big)\mathrm{d}\gamma'(x,w,z).$$

Indeed, since the first marginal of γ' is $\lambda\otimes\hat{\xi}_\tau^x$, the transport plan γ' assigns a target after both x and its realized noise w are known.

Step 2: Decompose the transport cost. We now verify that these two quantities are exactly the next-stage terms that appear in (20). We begin with $\mathcal{K}[j_\tau](\mu,\nu)$. Define

$$c_\tau(s,a,z) \coloneqq \int_W g_\tau(s,a,w) + j_{\tau+1}\big(f_\tau(s,a,w),z\big)\mathrm{d}\xi_\tau(w\,|\,s,a).$$

By the single-agent Bellman recursion we have

$$j_\tau(s,z) = \inf_{a\in A} c_\tau(s,a,z),$$

and plugging it into the definition of the transport cost gives

$$\mathcal{K}[j_\tau](\mu,\nu) = \inf_{\gamma_1\in\Gamma(\mu,\nu)} \int_{S\times Z} \inf_{a\in A} c_\tau(s,a,z)\mathrm{d}\gamma_1(s,z).$$

The function c_τ is lower semianalytic under the conventions of Appendix B, therefore we can apply Lemma B.5. In particular, for every fixed $\gamma_1 \in \Gamma(\mu,\nu)$,

$$\int_{S\times Z} \inf_{a\in A} c_\tau(s,a,z)\mathrm{d}\gamma_1(s,z) = \inf_{\{\alpha^{s,z}\}\in\Lambda(A)} \int_{S\times Z}\int_A c_\tau(s,a,z)\mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma_1(s,z).$$

Here, $\alpha^{s,z}$ is a Borel kernel that chooses an action for each state–target pair. Now, the pair $(\gamma_1,\alpha^{s,z})$ defines a measure γ on $(S\times A)\times Z$ by

$$\mathrm{d}\gamma(s,a,z) = \mathrm{d}\alpha^{s,z}(a)\mathrm{d}\gamma_1(s,z),$$

and we can rewrite its marginals in the form used by the population Bellman problem. To do so, let λ be the (s,a)-marginal of γ. Since the state marginal of γ_1 is μ, the state marginal of λ is also μ, and therefore $\lambda \in M(\mu)$. Combining this with the fact that the target marginal of γ is ν, we obtain $\gamma \in \Gamma(\lambda,\nu)$. Conversely, from every pair consisting of $\lambda \in M(\mu)$ and $\gamma \in \Gamma(\lambda,\nu)$, we can obtain a coupling $\gamma_1 \in \Gamma(\mu,\nu)$ by taking the (s,z)-marginal of γ, and a Borel action kernel $\alpha^{s,z}$ by disintegrating γ with respect to (s,z). Therefore, optimizing over $(\gamma_1,\alpha^{s,z})$ is equivalent to optimizing over $\lambda \in M(\mu)$ and $\gamma \in \Gamma(\lambda,\nu)$.

The transport problem can therefore be written as

$$\mathcal{K}[j_\tau](\mu,\nu) = \inf_{\lambda\in M(\mu)} \inf_{\gamma\in\Gamma(\lambda,\nu)} \int_{(S\times A)\times Z} c_\tau(s,a,z)\mathrm{d}\gamma(s,a,z).$$

The outer infimum chooses an admissible state-action distribution λ, while the inner infimum assigns its mass to the target distribution ν.

We can now expand the integral of c_τ. Since the stage cost does not depend on the target z, its integral depends only on the (s,a)-marginal λ. We denote it as

$$G_\tau(\lambda) \coloneqq \int_{S\times A}\int_W g_\tau(s,a,w)\mathrm{d}\xi_\tau(w\,|\,x)\mathrm{d}\lambda(x).$$

Therefore

$$\int_{(S\times A)\times Z} c_\tau(s,a,z)\mathrm{d}\gamma(x,z) = G_\tau(\lambda) + \int_{(S\times A)\times Z}\int_W j_{\tau+1}\big(f_\tau(s,a,\tilde{w}),z\big)\mathrm{d}\xi_\tau(\tilde{w}\,|\,x)\mathrm{d}\gamma(x,z).$$

For a given λ, the first term is fixed. Minimizing the second term over $\gamma \in \Gamma(\lambda,\nu)$ gives $A_\tau(\lambda)$, and minimizing again over $\lambda \in M(\mu)$ gives

$$\mathcal{K}[j_\tau](\mu,\nu) = \inf_{\lambda\in M(\mu)} \left\{G_\tau(\lambda) + A_\tau(\lambda)\right\}. \tag{21}$$

Step 3: Decompose the population Bellman term. We now turn to the second quantity in (20). From the definition of $\mathcal{T}_\tau$,

$$\begin{aligned}&\mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]](\mu,\nu)\\ &\qquad= \inf_{\lambda\in M(\mu)} \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau(s,a,w)\,\mathrm{d}\left(\lambda\otimes\hat{\xi}_\tau^x\right)(x,w) + \mathcal{K}[j_{\tau+1}](f_{\tau\#}\lambda\otimes\hat{\xi}_\tau^x,\nu)\right].\end{aligned}$$

We look at the two terms in the expectation separately. The expectation of the stage cost can be rewritten using Assumption A.2 as

$$\mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau(s,a,w)\,\mathrm{d}\left(\lambda\otimes\hat{\xi}_\tau^x\right)(x,w)\right] = G_\tau(\lambda).$$

For the transport term, we fix $\lambda \in M(\mu)$ and a realization of the noise field. Since the term couples the next-state distribution $f_{\tau\#}\lambda\otimes\hat{\xi}_\tau^x$ with the target distribution ν, we can apply Lemma B.1. The

lemma allows us to describe the same transport problem before applying the dynamics by coupling (x, w) directly with a target z and evaluating the cost at the resulting next state $f_\tau(s, a, w)$. By doing so we directly obtain $A_\tau(\lambda; \hat{\xi}_\tau)$. In particular, we have

$$\begin{aligned}\mathcal{K}[j_{\tau+1}](f_{\tau\#}\lambda \otimes \hat{\xi}_\tau^x, \nu) &= \inf_{\gamma' \in \Gamma(\lambda \otimes \hat{\xi}_\tau^x, \nu)} \int_{(S\times A)\times W\times Z} j_{\tau+1}\big(f_\tau(s, a, w), z\big)\mathrm{d}\gamma'(x, w, z) \\ &= A_\tau(\lambda; \hat{\xi}_\tau).\end{aligned}$$

Substituting these two equalities into the definition of $\mathcal{T}_\tau$ gives

$$\mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]]\,(\mu, \nu) = \inf_{\lambda \in M(\mu)} \Big\{ G_\tau(\lambda) + \mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] \Big\}. \tag{22}$$

Step 4: Compare the next-stage costs for a fixed λ. Now, since (21) and (22) contain the same stage-cost integrals $G_\tau(\lambda)$, it is enough to show that, for every $\lambda \in M(\mu)$,

$$0 \le A_\tau(\lambda) - \mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] \le L_{\tau+1} D_\tau. \tag{23}$$

Indeed, after adding the common stage-cost term, the inequalities remain valid for every λ, and taking the infimum over $\lambda \in M(\mu)$ gives (20).

Upper bound. We first prove the upper bound. Fix a realization of the noise field and take any γ' admissible for $A_\tau(\lambda; \hat{\xi}_\tau)$. Since $x = (s, a)$ and the (x, z)-marginal of γ' has marginals λ and ν, respectively, it is admissible for $A_\tau(\lambda)$. We can therefore evaluate $A_\tau(\lambda)$ at this (x, z)-marginal and draw $\tilde{w}$ independently from $\xi_\tau(\cdot \,|\, x)$, which gives

$$\begin{aligned}A_\tau(\lambda) &\le \int_{(S\times A)\times W\times Z} \int_W j_{\tau+1}\big(f_\tau(s, a, \tilde{w}), z\big)\mathrm{d}\xi_\tau(\tilde{w} \,|\, x)\mathrm{d}\gamma'(x, w, z) \\ &\le \int_{(S\times A)\times W\times Z} j_{\tau+1}\big(f_\tau(s, a, w), z\big)\mathrm{d}\gamma'(x, w, z) \\ &\quad + L_{\tau+1} \int_{(S\times A)\times W} \int_W d_S(f_\tau(s, a, w), f_\tau(s, a, \tilde{w}))\,\mathrm{d}\xi_\tau(\tilde{w} \,|\, x)\,\mathrm{d}\Big(\lambda \otimes \hat{\xi}_\tau^x\Big)\,(x, w),\end{aligned}$$

where the second inequality follows from Assumption D.1.

Looking at the obtained upper bound, we can observe that the first integral is the value of a fixed γ' in the definition of $A_\tau(\lambda; \hat{\xi}_\tau)$. The last integral, instead, does not contain z, and so does not depend on how γ' assigns targets. We can therefore minimize over γ' and obtain

$$\begin{aligned}A_\tau(\lambda) &\le A_\tau(\lambda; \hat{\xi}_\tau) \\ &\quad + L_{\tau+1} \int_{(S\times A)\times W} \int_W d_S(f_\tau(s, a, w), f_\tau(s, a, \tilde{w}))\,\mathrm{d}\xi_\tau(\tilde{w} \,|\, x)\,\mathrm{d}\Big(\lambda \otimes \hat{\xi}_\tau^x\Big)\,(x, w).\end{aligned}$$

Averaging over the noise field and applying Assumption A.2 to replace $\hat{\xi}_\tau^x$ by $\xi_\tau(\cdot \,|\, x)$ in the last integral gives

$$\begin{aligned}&\mathbb{E}_{\eta_\tau}\left[\int_{(S\times A)\times W} \int_W d_S(f_\tau(s, a, w), f_\tau(s, a, \tilde{w}))\,\mathrm{d}\xi_\tau(\tilde{w} \,|\, x)\,\mathrm{d}\Big(\lambda \otimes \hat{\xi}_\tau^x\Big)\,(x, w)\right] \\ &\quad = \int_{S\times A} \int_{W\times W} d_S(f_\tau(s, a, w), f_\tau(s, a, \tilde{w}))\,\mathrm{d}\xi_\tau(w \,|\, x)\mathrm{d}\xi_\tau(\tilde{w} \,|\, x)\mathrm{d}\lambda(x) \\ &\quad \le D_\tau,\end{aligned}$$

where the last inequality follows from the supremum in (19). We have therefore proved the upper bound

$$A_\tau(\lambda) \le \mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] + L_{\tau+1} D_\tau.$$

Lower bound. For the lower bound in (23), fix any $\gamma \in \Gamma(\lambda, \nu)$. After the noise field is realized, we can couple this plan with the realized noise kernel by setting

$$\mathrm{d}\gamma'_{\hat{\xi}_\tau}(x, w, z) \coloneqq \mathrm{d}\gamma(x, z)\mathrm{d}\hat{\xi}^x_\tau(w).$$

The resulting plan is admissible for $A_\tau(\lambda; \hat{\xi}_\tau)$ since its (x, w)-marginal is $\lambda \otimes \hat{\xi}^x_\tau$, because the x-marginal of γ is λ, and its target marginal is ν. Evaluating the infimum at this plan, averaging over the noise field, and then applying Lemma A.10 with (x, z) as the base variable and $j_{\tau+1}(f_\tau(s, a, w), z)$ as the cost, gives

$$\begin{aligned}\mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] &\leq \mathbb{E}_{\eta_\tau}\left[\int_{(S\times A)\times W\times Z} j_{\tau+1}\big(f_\tau(s, a, w), z\big)\mathrm{d}\gamma'_{\hat{\xi}_\tau}(x, w, z)\right] \\ &= \int_{(S\times A)\times Z}\int_W j_{\tau+1}\big(f_\tau(s, a, \tilde{w}), z\big)\mathrm{d}\xi_\tau(\tilde{w} \,|\, x)\mathrm{d}\gamma(x, z).\end{aligned}$$

The right-hand side corresponds exactly to the value of the fixed γ in the definition of $A_\tau(\lambda)$, and, since the inequality holds for every $\gamma \in \Gamma(\lambda, \nu)$, minimizing over γ yields

$$\mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] \leq A_\tau(\lambda).$$

Step 5: Return to the two Bellman quantities. This proves (23). To conclude, for every $\lambda \in M(\mu)$, adding the common stage-cost expectation to (23) gives

$$\begin{aligned}G_\tau(\lambda) + \mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] &\leq G_\tau(\lambda) + A_\tau(\lambda) \\ &\leq G_\tau(\lambda) + \mathbb{E}_{\eta_\tau}\left[A_\tau(\lambda; \hat{\xi}_\tau)\right] + L_{\tau+1}D_\tau.\end{aligned}$$

Taking the infimum over λ in all three terms and using (21) and (22) gives exactly (20). □

Lemma D.3 bounds, for one time step, the difference between assigning targets before and after the noise is observed. Applying this bound successively from time t to time $T-1$ gives the total difference stated in the following theorem.

Theorem D.4 (Quantification of the separation gap)**.** *Under Assumptions A.1, A.2, and D.1, for every $t \in \{0, \dots, T\}$, every $\mu_t \in \mathcal{P}(S)$, and every $\nu \in \mathcal{P}(Z)$,*

$$0 \leq \mathcal{K}[j_t](\mu_t, \nu) - J_t(\mu_t, \nu) \leq \sum_{\tau=t}^{T-1} L_{\tau+1}D_\tau. \tag{24}$$

Proof. We proceed by backward induction. At the horizon the two sides agree, since $j_T = g_T$ and therefore $\mathcal{K}[j_T](\mu, \nu) = J_T(\mu, \nu)$. Assume next that (24) holds at time $\tau + 1$, pointwise in the population and target distributions, so that

$$0 \leq \mathcal{K}[j_{\tau+1}] - J_{\tau+1} \leq \sum_{r=\tau+1}^{T-1} L_{r+1}D_r.$$

Since $\mathcal{T}_\tau$ is monotone and commutes with the addition of constants by Lemma D.2, applying it to both sides gives

$$0 \leq \mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]] - \mathcal{T}_\tau J_{\tau+1} \leq \sum_{r=\tau+1}^{T-1} L_{r+1}D_r. \tag{25}$$

We can now compare $\mathcal{K}[j_\tau] - J_\tau$. In particular, using $J_\tau = \mathcal{T}_\tau J_{\tau+1}$ and (20) and (25), we obtain

$$\begin{aligned}0 &\leq \mathcal{K}[j_\tau] - J_\tau \\ &\leq (\mathcal{K}[j_\tau] - \mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]]) + (\mathcal{T}_\tau[\mathcal{K}[j_{\tau+1}]] - \mathcal{T}_\tau J_{\tau+1}) \\ &\leq L_{\tau+1}D_\tau + \sum_{r=\tau+1}^{T-1} L_{r+1}D_r = \sum_{r=\tau}^{T-1} L_{r+1}D_r,\end{aligned}$$

which is (24) at time τ and completes the induction. □

Corollary D.5 (Finite-agent consequence)**.** *Let $\boldsymbol{s}_t = (s_t^1, \dots, s_t^N) \in S^N$ be the agent states at time t, with empirical distribution $\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{s_t^i}$, and let ν^N be an N-atom target distribution as in Corollary 3.3. Under the assumptions of Theorem D.4, it holds that*

$$0 \le \mathcal{K}[j_t](\mu_t^N, \nu^N) - J_t^N(\boldsymbol{s}_t, \nu^N) \le \sum_{\tau=t}^{T-1} L_{\tau+1} D_\tau. \tag{26}$$

Proof. For a joint noise law and costs that are invariant under permutations of agent labels, the finite-agent value depends only on the empirical state. We may therefore choose an ε-optimal policy whose empirical state–action measure is a feedback function of that state. This measure is admissible for the population problem and has the same cost, so the population problem is a relaxation of the agent-based one. Together with Corollary 3.3, this gives

$$J_t(\mu_t^N, \nu^N) \le J_t^N(\boldsymbol{s}_t, \nu^N) \le \mathcal{K}[j_t](\mu_t^N, \nu^N).$$

Subtracting the middle term from the transport cost and applying (24) at (μ_t^N, ν^N) yields (26). □

Interpretation. The two sides of (20) differ because the population Bellman recursion may reassign targets after seeing the realized next population, whereas the transport bound commits to an assignment before the nominal noise is integrated. The bound in Theorem D.4 gives a clear interpretation of this loss. In particular, $L_{\tau+1}$ measures how sensitive the remaining cost is to a displacement of an agent, while D_τ measures how far the noise can push apart two agents that start from the same state-action pair. The gap therefore closes for deterministic dynamics, and more generally whenever the optimal assignment does not depend on the realized noise.

Bounding the dispersion term. The definition of D_τ involves the distance between next states. This quantity can be bounded directly in terms of the noise when W carries a metric d_W and the dynamics are K_τ-Lipschitz in their noise argument, uniformly in (s, a):

$$d_S(f_\tau(s, a, w), f_\tau(s, a, \tilde{w})) \le K_\tau d_W(w, \tilde{w}).$$

Indeed, applying this inequality to the two independent noise draws in the definition of D_τ gives

$$D_\tau \le K_\tau \sup_{(s,a)} \mathbb{E}_{\xi_\tau(\cdot \mid s,a)^{\otimes 2}} \left[d_W(w_\tau, \tilde{w}_\tau) \right].$$

If W is also a normed space and the conditional mean exists, we can apply the triangle inequality through that mean and, since the two noise draws have the same conditional distribution, obtain

$$D_\tau \le 2K_\tau \sup_{(s,a)} \mathbb{E}_{\xi_\tau(\cdot \mid s,a)} \left[\left\| w_\tau - \mathbb{E}_{\xi_\tau(\cdot \mid s,a)} [w_\tau] \right\| \right].$$

Finally, for additive dynamics $f_\tau(s, a, w) = \widetilde{f}_\tau(s, a) + w$, with S and W carrying the same norm metric, the distance between the two next states is exactly $\|w - \tilde{w}\|$, and the first estimate holds with equality and $K_\tau = 1$.

Tightness. We here show through an example that the linear dependence on D_τ in Lemma D.3 and Theorem D.4 cannot be improved without further assumptions.

Example D.6 (The linear coefficient is tight)**.** *Let $\tau = T - 1$, $A = \{0\}$, and $S = Z = \{0, 1\}$, equipped with the discrete metric. Set zero stage cost $g_t(s, a, z) = 0$, and the following terminal cost and dynamics:*

$$g_T(s, z) = |s - z|, \qquad f_\tau(s, a, w) = w,$$

where $w \sim \mathrm{Bernoulli}(\frac{1}{2})$. Consider an infinite population and use the individual noise model in Example A.5. Suppose that initially the population is concentrated at $s = 0$, i.e. $\mu_0 = \delta_0$, and the target distribution is selected as $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Before the noise is observed, the only possible coupling assigns half the mass to each target. Moreover, since the next state equals w regardless of the current state, the single-agent Bellman recursion gives

$$j_\tau(0, z) = \mathbb{E}_w \left[|w - z| \right] = \tfrac{1}{2}|0 - z| + \tfrac{1}{2}|1 - z| = \tfrac{1}{2},$$

for every $z \in Z$. Hence $j_\tau(0, 0) = j_\tau(0, 1) = \frac{1}{2}$, and

$$\mathcal{K}[j_\tau](\mu, \nu) = \frac{1}{2}.$$

After the transition, on the other hand, the realized population is exactly ν, and the population Bellman step matches equal states and targets at zero cost:

$$\mathcal{T}_\tau[\mathcal{K}[j_T]]\,(\mu,\nu) = 0.$$

Since $L_T = 1$ and two independent Bernoulli($\frac{1}{2}$) *draws differ with probability $1/2$, we have $D_\tau = 1/2$, so equality holds in* (20)*.*

E Robustness to weakly coupled dynamics

In this appendix, we study what happens when the separation-based policy is applied to a model in which the agents are weakly coupled, that is, a system whose dynamics and stage cost may depend on the current population distribution. Since the separation principle of Section 3 requires decoupled agents, we construct the policy from a decoupled surrogate model, deploy it in the coupled system, and bound the resulting loss of optimality. Because the admissible controls remain $M(\mu)$ in both models, the following result covers soft population effects, such as congestion or queueing, but not hard capacity or collision constraints that change the set of feasible controls.

Models and assumptions. Throughout this appendix, we denote the decoupled surrogate with a superscript D, so its dynamics, stage cost, single-agent value, and population cost-to-go are written as

$$f_\tau^D(s,a,w), \qquad g_\tau^D(s,a,w), \qquad j_\tau^D(s,z), \qquad J_\tau^D(\mu,\nu).$$

The corresponding quantities for the coupled model are denoted by a superscript C:

$$f_\tau^C(s,a,\mu,w), \qquad g_\tau^C(s,a,\mu,w), \qquad J_\tau^C(\mu,\nu).$$

Here, the dynamics and stage costs also depend on the population, so there is no direct counterpart of j_τ^D. Indeed, because f_τ^C depends on μ, the cost-to-go of a single agent cannot be defined without fixing the trajectory of the population. At the terminal time T, we denote by $J_T^C(\mu,\nu)$ the terminal objective, evaluated at the terminal fleet state μ and target distribution ν. Apart from the population dependence displayed above, the two models share the same basic setting. The models have the same state, action, noise, and target spaces, the same nominal noise kernel, and the same admissible controls $M(\mu)$. We also drive them with the same realized noise field, which satisfies Assumptions A.1 and A.2, so their next populations are

$$\mu_{\tau+1}^{C,\lambda} \coloneqq f_\tau^C(\cdot,\cdot,\mu,\cdot)_{\#}\rho_\tau^\lambda, \qquad \mu_{\tau+1}^{D,\lambda} \coloneqq f_\tau^D{}_{\#}\rho_\tau^\lambda, \tag{27}$$

where $\rho_\tau^\lambda \coloneqq \lambda \otimes \hat{\xi}_\tau^{s,a}$ denotes the joint state-action-noise distribution for a fixed population μ and a control $\lambda \in M(\mu)$. Finally, the terminal objective of the coupled and decoupled models is common and denoted by

$$J_T^C(\mu,\nu) = J_T^D(\mu,\nu) = \mathcal{K}[g_T](\mu,\nu). \tag{28}$$

We next specify how far the coupled model may be from the decoupled one at each time step.

Assumption E.1 (Bounded one-step model mismatch)**.** Fix a reference point $s_0 \in S$ and assume that every population distribution $\mu \in \mathcal{P}(S)$ at which the coupled and decoupled models are compared below has finite first moment,

$$\int_S d_S(s,s_0)\mathrm{d}\mu(s) < \infty.$$

For every $\tau < T$, let $\varepsilon_\tau^f, \delta_\tau \geq 0$ be constants such that the following hold:

(i) *Dynamics mismatch.* When the coupled and decoupled dynamics are evaluated at the same state, action, and noise, their outputs satisfy

$$d_S\big(f_\tau^C(s,a,\mu,w), f_\tau^D(s,a,w)\big) \leq \varepsilon_\tau^f \tag{29}$$

uniformly in $\mu \in \mathcal{P}(S)$ and $(s,a,w) \in S \times A \times W$.

(ii) *Stage-cost mismatch.* The expected stage costs, evaluated at the same population and the same admissible state-action distribution, satisfy

$$\left|\mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W}\big(g_\tau^C(s,a,\mu,w) - g_\tau^D(s,a,w)\big)\,\mathrm{d}\rho_\tau^\lambda(s,a,w)\right]\right| \leq \delta_\tau \tag{30}$$

uniformly in $\mu \in \mathcal{P}(S)$ and $\lambda \in M(\mu)$.

These two constants control the two parts of one Bellman step. Indeed, δ_τ bounds the difference in the current stage cost, while ε_τ^f bounds the difference between the next populations.

Note also that condition (30) is satisfied if there exists a pointwise bound $|g_\tau^C - g_\tau^D| \le \delta_\tau$. Additionally, for a fixed initial population, then it is sufficient that (29) and (30) hold on the population states reachable under any admissible control.

These one-step bounds enter the comparison through the Bellman recursions. Recall the Bellman operator $\mathcal{T}_\tau$ introduced in Appendix D. We now rewrite the two population recursions, for the coupled and decoupled models, in a common form. For $m \in \{C, D\}$, define

$$(\mathcal{T}_\tau^m F)(\mu,\nu) \coloneqq \inf_{\lambda \in M(\mu)} \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau^m(s,a,\mu,w)\mathrm{d}\rho_\tau^\lambda(s,a,w) + F(\mu_{\tau+1}^{m,\lambda},\nu)\right].$$

Here and below, we use the convention $g_\tau^D(s,a,\mu,w) \coloneqq g_\tau^D(s,a,w)$, thus neither f_τ^D nor g_τ^D depends on the population argument. With this notation, we have the Bellman recursions $J_\tau^m = \mathcal{T}_\tau^m J_{\tau+1}^m$. In particular, $\mathcal{T}_\tau^D = \mathcal{T}_\tau$, since the decoupled stage cost and dynamics are exactly the data g_τ and f_τ used to define $\mathcal{T}_\tau$ in Appendix D, so every result proved there for $\mathcal{T}_\tau$ applies verbatim to $\mathcal{T}_\tau^D$. Additionally, if π is a feedback policy with $\pi_\tau(\mu) \in M(\mu)$, we drop the infimum and denote the same expression evaluated at $\lambda = \pi_\tau(\mu)$ as

$$(\mathcal{T}_\tau^{m,\pi} F)(\mu,\nu) \coloneqq \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau^m(s,a,\mu,w)\mathrm{d}\rho_\tau^{\pi_\tau(\mu)}(s,a,w) + F(\mu_{\tau+1}^{m,\pi_\tau(\mu)},\nu)\right].$$

The cost of this policy satisfies

$$J_T^{m,\pi} = \mathcal{K}[g_T], \qquad J_\tau^{m,\pi} = \mathcal{T}_\tau^{m,\pi} J_{\tau+1}^{m,\pi}.$$

The next-stage terms in these recursions may be evaluated at different population distributions, so to compare those distributions, we use the 1-Wasserstein distance induced by d_S. Explicitly, for any $\mu, \nu \in \mathcal{P}(S)$ with finite first moments, the 1-Wasserstein distance induced by d_S is denoted by $W_1(\mu,\nu)$ and defined as

$$W_1(\mu,\nu) \coloneqq \inf_{q\in\Gamma(\mu,\nu)} \int_{S\times S} d_S(s,\widetilde{s})\mathrm{d}q(s,\widetilde{s}).$$

To compare the next populations in (27), we need an estimate for two maps applied to the same measure. We formalize this in the following lemma. Note that this differs from [48, Proposition 3], which considers one map applied to two measures.

Lemma E.2 (Wasserstein bound for pushforwards of a common measure)**.** *Let U be a measurable space, let $\mu \in \mathcal{P}(U)$, and let $F, G\colon U \to S$ be measurable. Then*

$$W_1(F_\#\mu, G_\#\mu) \le \int_U d_S\big(F(u), G(u)\big)\mathrm{d}\mu(u). \tag{31}$$

Proof. Consider the coupling $q_{F,G} \coloneqq (F,G)_\#\mu$, where (F,G) maps u to $(F(u),G(u))$. Its first marginal is $F_\#\mu$, since for every Borel set $A \subseteq S$,

$$q_{F,G}(A\times S) = \mu\big(F^{-1}(A)\big) = (F_\#\mu)(A).$$

The same argument shows that its second marginal is $G_\#\mu$, so $q_{F,G} \in \Gamma(F_\#\mu, G_\#\mu)$. Evaluating the definition of W_1 at this plan and applying the change-of-variables formula for the pushforward $q_{F,G} = (F,G)_\#\mu$ gives

$$\begin{aligned} W_1(F_\#\mu, G_\#\mu) &\le \int_{S\times S} d_S(s,\widetilde{s})\mathrm{d}q_{F,G}(s,\widetilde{s}) \\ &= \int_U d_S\big(F(u),G(u)\big)\mathrm{d}\mu(u). \end{aligned}$$

This concludes the proof. □

In other words, when two distributions are obtained from the same measure through two different maps, their Wasserstein distance can be bounded by comparing the outputs of the maps pointwise.

It remains to relate this distance between populations to the next-stage cost. To do so, we assume that the decoupled single-agent cost-to-go j_τ^D satisfies Assumption D.1, with constants L_τ, and use D_τ (19) for the dispersion of the decoupled dynamics f_τ^D. As in Appendix D, we assume that all transport costs appearing below are finite.

We can now proceed with the study of the robustness of the separation-based policy. The first step is to show that the transport cost is Lipschitz in the population distribution. This will allow us, in Lemma E.4 below, to turn the state mismatch in (29) into a bound on the next-stage cost.

Lemma E.3 (Population sensitivity of the transport cost)**.** *Suppose Assumption D.1 holds for j_τ^D with constant L_τ. Then, for every $\tau \in \{1, \dots, T\}$, every $\mu, \widetilde{\mu} \in \mathcal{P}(S)$ with finite first moments, and every $\nu \in \mathcal{P}(Z)$,*

$$\left|\mathcal{K}[j_\tau^D](\mu, \nu) - \mathcal{K}[j_\tau^D](\widetilde{\mu}, \nu)\right| \leq L_\tau W_1(\mu, \widetilde{\mu}). \tag{32}$$

Proof. The proof consists of transferring a nearly optimal target assignment for μ to $\widetilde{\mu}$. Fix $\varepsilon > 0$. By the definition of the transport cost and the Wasserstein distance, we can choose $\gamma_\varepsilon \in \Gamma(\mu, \nu)$ and $q_\varepsilon \in \Gamma(\mu, \widetilde{\mu})$ such that

$$\int_{S\times Z} j_\tau^D(s, z) \mathrm{d}\gamma_\varepsilon(s, z) \leq \mathcal{K}[j_\tau^D](\mu, \nu) + \varepsilon,$$
$$\int_{S\times S} d_S(s, \widetilde{s}) \mathrm{d}q_\varepsilon(s, \widetilde{s}) \leq W_1(\mu, \widetilde{\mu}) + \varepsilon.$$

Since the two plans have the same s-marginal μ, the gluing lemma gives a measure r_ε on $S \times S \times Z$ with (s, z)-marginal γ_ε and $(s, \widetilde{s})$-marginal q_ε. Note, in particular, that its $(\widetilde{s}, z)$-marginal couples $\widetilde{\mu}$ with ν, so we can use it in the transport problem for $\mathcal{K}[j_\tau^D](\widetilde{\mu}, \nu)$. The first inequality below evaluates the transport cost at the $(\widetilde{s}, z)$-marginal of r_ε, the second applies (18) pointwise, and then we use the (s, z) and $(s, \widetilde{s})$-marginals of r_ε to separate the two integrals:

$$\begin{aligned}
\mathcal{K}[j_\tau^D](\widetilde{\mu}, \nu) &\leq \int_{S\times S\times Z} j_\tau^D(\widetilde{s}, z) \mathrm{d}r_\varepsilon(s, \widetilde{s}, z) \\
&\leq \int_{S\times S\times Z} \left[j_\tau^D(s, z) + L_\tau d_S(s, \widetilde{s})\right] \mathrm{d}r_\varepsilon(s, \widetilde{s}, z) \\
&= \int_{S\times Z} j_\tau^D(s, z) \mathrm{d}\gamma_\varepsilon(s, z) + L_\tau \int_{S\times S} d_S(s, \widetilde{s}) \mathrm{d}q_\varepsilon(s, \widetilde{s}) \\
&\leq \mathcal{K}[j_\tau^D](\mu, \nu) + \varepsilon + L_\tau\big(W_1(\mu, \widetilde{\mu}) + \varepsilon\big) \\
&= \mathcal{K}[j_\tau^D](\mu, \nu) + L_\tau W_1(\mu, \widetilde{\mu}) + (1 + L_\tau)\varepsilon,
\end{aligned} \tag{33}$$

where the last inequality uses the two ε-optimality bounds in the choice of γ_ε and q_ε. Since the construction holds for all $\varepsilon > 0$, letting $\varepsilon \downarrow 0$ gives

$$\mathcal{K}[j_\tau^D](\widetilde{\mu}, \nu) - \mathcal{K}[j_\tau^D](\mu, \nu) \leq L_\tau W_1(\mu, \widetilde{\mu}).$$

Exchanging the roles of μ and $\widetilde{\mu}$ throughout the construction above, that is, choosing $\gamma_\varepsilon \in \Gamma(\widetilde{\mu}, \nu)$ that is ε-optimal for $\mathcal{K}[j_\tau^D](\widetilde{\mu}, \nu)$ and $q_\varepsilon \in \Gamma(\widetilde{\mu}, \mu)$ that is ε-optimal for $W_1(\widetilde{\mu}, \mu)$, and then repeating the gluing step and (33) with s and $\widetilde{s}$ swapped, gives

$$\mathcal{K}[j_\tau^D](\mu, \nu) - \mathcal{K}[j_\tau^D](\widetilde{\mu}, \nu) \leq L_\tau W_1(\widetilde{\mu}, \mu).$$

Since $W_1(\widetilde{\mu}, \mu) = W_1(\mu, \widetilde{\mu})$, the two bounds give (32). □

Lemma E.3 allows us to compare one Bellman step in the two models. To do so, we first keep the control λ fixed, so that the only differences come from the stage cost and the next population, and then minimize over λ.

Lemma E.4 (One-step model perturbation)**.** *Under Assumptions A.1, A.2, D.1, and E.1, for every $\tau < T$, every $\mu \in \mathcal{P}(S)$ with finite first moment, every $\nu \in \mathcal{P}(Z)$, and every admissible feedback policy π,*

$$\left|\mathcal{T}_\tau^C\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu, \nu) - \mathcal{T}_\tau^D\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu, \nu)\right| \leq \delta_\tau + L_{\tau+1}\varepsilon_\tau^f, \tag{34}$$
$$\left|\mathcal{T}_\tau^{C,\pi}\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu, \nu) - \mathcal{T}_\tau^{D,\pi}\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu, \nu)\right| \leq \delta_\tau + L_{\tau+1}\varepsilon_\tau^f. \tag{35}$$

Proof. **Step 1: Compare the next populations for a fixed control.** Fix $\lambda \in M(\mu)$ and a realization of the noise field. By (27), we can obtain the two next populations by pushing forward the same state-action-noise distribution ρ_τ^λ, once through f_τ^C and once through f_τ^D. We can then apply (31) with $\mu = \rho_\tau^\lambda$, $F = f_\tau^C(\cdot, \cdot, \mu, \cdot)$, and $G = f_\tau^D$ and obtain

$$W_1(\mu_{\tau+1}^{C,\lambda}, \mu_{\tau+1}^{D,\lambda}) \leq \int_{S\times A\times W} d_S\big(f_\tau^C(s,a,\mu,w), f_\tau^D(s,a,w)\big)\,\mathrm{d}\rho_\tau^\lambda(s,a,w) \leq \varepsilon_\tau^f. \tag{36}$$

The last inequality follows directly from the uniform dynamics-mismatch bound (29) in Assumption E.1.

Step 2: Compare the next-stage transport costs. For the fixed realization of the noise field selected in the previous step, Lemma E.3 applied at time $\tau+1$, together with (36), gives

$$\Big|\mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\lambda}, \nu) - \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\lambda}, \nu)\Big| \leq L_{\tau+1} W_1\Big(\mu_{\tau+1}^{C,\lambda}, \mu_{\tau+1}^{D,\lambda}\Big) \leq L_{\tau+1}\varepsilon_\tau^f.$$

Since this bound holds for every realization of the field, we can average it out and obtain

$$\mathbb{E}_{\eta_\tau}\Big[\Big|\mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\lambda}, \nu) - \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\lambda}, \nu)\Big|\Big] \leq L_{\tau+1}\varepsilon_\tau^f.$$

Step 3: Add the stage-cost difference. By (30) in Assumption E.1, the absolute difference between the two expected stage costs is at most δ_τ. We now add this difference to the difference between the next-stage transport costs. The triangle inequality and the estimate from Step 2 give

$$\begin{aligned}
&\Bigg|\mathbb{E}_{\eta_\tau}\Big[\int_{S\times A\times W} \big(g_\tau^C(s,a,\mu,w) - g_\tau^D(s,a,w)\big)\mathrm{d}\rho_\tau^\lambda(s,a,w) \\
&\qquad + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\lambda}, \nu) - \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\lambda}, \nu)\Big]\Bigg| \\
&\leq \Bigg|\mathbb{E}_{\eta_\tau}\Big[\int_{S\times A\times W} \big(g_\tau^C(s,a,\mu,w) - g_\tau^D(s,a,w)\big)\mathrm{d}\rho_\tau^\lambda(s,a,w)\Big]\Bigg| \\
&\qquad + \mathbb{E}_{\eta_\tau}\Big[\Big|\mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\lambda}, \nu) - \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\lambda}, \nu)\Big|\Big] \\
&\leq \delta_\tau + L_{\tau+1}\varepsilon_\tau^f.
\end{aligned} \tag{37}$$

Step 4: Pass to the Bellman operators. By linearity of expectation, one of the two one-sided inequalities implied by (37) is

$$\begin{aligned}
&\mathbb{E}_{\eta_\tau}\Big[\int_{S\times A\times W} g_\tau^C(s,a,\mu,w)\mathrm{d}\rho_\tau^\lambda(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\lambda}, \nu)\Big] \\
&\quad \leq \mathbb{E}_{\eta_\tau}\Big[\int_{S\times A\times W} g_\tau^D(s,a,w)\mathrm{d}\rho_\tau^\lambda(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\lambda}, \nu)\Big] + \delta_\tau + L_{\tau+1}\varepsilon_\tau^f.
\end{aligned} \tag{38}$$

The left- and right-hand sides are precisely the fixed-control objectives minimized in the definitions of $\mathcal{T}_\tau^C$ and $\mathcal{T}_\tau^D$. Since (38) holds for every admissible λ, we can take the infimum over the controls in $M(\mu)$ on both sides, and obtain

$$\begin{aligned}
&\mathcal{T}_\tau^C\big[\mathcal{K}[j_{\tau+1}^D]\big](\mu,\nu) \\
&\quad = \inf_{\lambda\in M(\mu)} \mathbb{E}_{\eta_\tau}\Big[\int_{S\times A\times W} g_\tau^C(s,a,\mu,w)\mathrm{d}\rho_\tau^\lambda(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\lambda}, \nu)\Big] \\
&\quad \leq \inf_{\lambda\in M(\mu)} \mathbb{E}_{\eta_\tau}\Big[\int_{S\times A\times W} g_\tau^D(s,a,w)\mathrm{d}\rho_\tau^\lambda(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\lambda}, \nu)\Big] + \delta_\tau + L_{\tau+1}\varepsilon_\tau^f \\
&\quad = \mathcal{T}_\tau^D\big[\mathcal{K}[j_{\tau+1}^D]\big](\mu,\nu) + \delta_\tau + L_{\tau+1}\varepsilon_\tau^f.
\end{aligned}$$

Similarly, taking the infimum over $M(\mu)$ in the reverse fixed-control inequality of (37) gives the other one-sided bound

$$\mathcal{T}_\tau^D\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) \le \mathcal{T}_\tau^C\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) + \delta_\tau + L_{\tau+1}\varepsilon_\tau^f.$$

Combining these two inequalities proves (34). Finally, for the fixed-policy operators, it is enough to evaluate the same objectives at the same control $\lambda = \pi_\tau(\mu)$. Indeed, substituting this value into (37) gives directly (35). □

The separation-based policy. We now construct the separation-based policy from the decoupled model, following the same steps as in Section 4 and Theorem 3.1, Part II. We will then evaluate this policy in the coupled model, so to bound the performance loss due to the model mismatch. Briefly, the policy construction proceeds as follows: at an observed population μ, the policy first chooses a $\theta_\tau/2$-optimal target assignment $\gamma_\tau^\theta \in \Gamma(\mu,\nu)$ for the decoupled transport problem, so that

$$\int_{S\times Z} j_\tau^D(s,z)\mathrm{d}\gamma_\tau^\theta(s,z) \le \mathcal{K}[j_\tau^D](\mu,\nu) + \frac{\theta_\tau}{2}. \tag{39}$$

It then chooses a measurable action $a_\tau^\theta(s,z)$ whose single-agent cost is within $\theta_\tau/2$ of the decoupled optimum. Writing $a = a_\tau^\theta(s,z)$, this means that

$$\int_W \left[g_\tau^D(s,a,w) + j_{\tau+1}^D\big(f_\tau^D(s,a,w),z\big)\right]\mathrm{d}\xi_\tau(w\,|\,s,a) \le j_\tau^D(s,z) + \frac{\theta_\tau}{2}. \tag{40}$$

Since this inequality holds for every $(s,z) \in S\times Z$, the target plan and the selected actions together induce the following population control

$$\pi_{\mathrm{sep},\tau}^\theta(\mu) \coloneqq \big(\mathrm{proj}_S^{S\times Z}, a_\tau^\theta\big)_{\#}\gamma_\tau^\theta \in M(\mu). \tag{41}$$

Note that the control $\pi_{\mathrm{sep},\tau}^\theta(\mu)$ is a measurable map from the population space to the set of admissible controls, as required by the definition of a feedback policy. The construction ensures that the policy is well-defined and can be implemented in practice, as it relies on measurable selections of near-optimal target assignments and actions. The two errors of size $\theta_\tau/2$ add up to the per-stage tolerance θ_τ. If, additionally, the target plan and the actions can both be chosen as measurable exact minimizers, then the construction holds when we take $\theta_\tau = 0$.

Finally, recall that ρ_τ^λ was defined immediately after (27), for every $\lambda \in M(\mu)$, as the corresponding state-action-noise distribution. Since $\pi_{\mathrm{sep},\tau}^\theta(\mu) \in M(\mu)$, we can apply that definition with $\lambda = \pi_{\mathrm{sep},\tau}^\theta(\mu)$, obtaining $\rho_\tau^{\pi_{\mathrm{sep},\tau}^\theta(\mu)}$.

Before evaluating this separation-based policy in the coupled model, we show that when working in the decoupled model, one Bellman step under this policy costs at most the decoupled transport surrogate plus θ_τ. This is the estimate that will enter the backward argument of the main theorem.

Lemma E.5 (One-step bound for the separation-based policy)**.** *Under Assumptions A.1 and A.2, the separation-based policy defined by* (39)–(41) *satisfies, for every $\tau < T$, every $\mu \in \mathcal{P}(S)$, and every $\nu \in \mathcal{P}(Z)$,*

$$\mathcal{T}_\tau^{D,\pi_{\mathrm{sep}}^\theta}\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) \le \mathcal{K}[j_\tau^D](\mu,\nu) + \theta_\tau. \tag{42}$$

Proof. **Step 1: Lift the policy to a measure on (s,z,w).** We start by fixing μ and a realization of the noise field and, to obtain an admissible target assignment at the next time, we keep the target attached to each state while the state moves according to its selected action and noise. More precisely, consider the measure

$$\mathrm{d}\gamma_\tau^\theta(s,z)\mathrm{d}\hat{\xi}_\tau^{s,a_\tau^\theta(s,z)}(w)$$

on (s,z,w). It generates the same state, action, and noise distribution $\rho_\tau^{\pi_{\mathrm{sep},\tau}^\theta(\mu)}$ as the separation-based policy. More precisely, for every nonnegative measurable function h on $S\times A\times W$,

$$\begin{aligned}&\int_{S\times A\times W} h(s,a,w)\mathrm{d}\rho_\tau^{\pi_{\mathrm{sep},\tau}^\theta(\mu)}(s,a,w)\\ &\quad= \int_{S\times Z}\int_W h\big(s,a_\tau^\theta(s,z),w\big)\mathrm{d}\hat{\xi}_\tau^{s,a_\tau^\theta(s,z)}(w)\mathrm{d}\gamma_\tau^\theta(s,z).\end{aligned} \tag{43}$$

Indeed, (41) says that applying $(s,z) \mapsto (s, a_\tau^\theta(s,z))$ to γ_τ^θ gives the state-action distribution chosen by the policy. Adding the realized noise kernel then gives $\rho_\tau^{\pi^\theta_{\mathrm{sep},\tau}(\mu)}$, which proves the identity.

Step 2: Build a feasible next-stage transport plan. We now push the measure on (s,z,w) forward under the following map:

$$(s,z,w) \longmapsto \big(f_\tau^D(s, a_\tau^\theta(s,z), w), z\big).$$

Its first marginal is $\mu_{\tau+1}^{D,\pi^\theta_{\mathrm{sep},\tau}(\mu)}$ by (27) and the identity (43), while its second marginal is ν, because the map leaves the target coordinate unchanged. The pushforward is therefore an admissible coupling in the transport problem

$$\mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\pi^\theta_{\mathrm{sep},\tau}(\mu)}, \nu).$$

Hence, this coupling handles the next-stage term of the Bellman objective, it remains to show that the coupling also accounts for the current stage cost. To do so, take $h = g_\tau^D$ in (43) and write the stage cost using the same measure on (s,z,w):

$$\begin{aligned}
&\int_{S\times A\times W} g_\tau^D(s,a,w)\mathrm{d}\rho_\tau^{\pi^\theta_{\mathrm{sep},\tau}(\mu)}(s,a,w) \\
&\quad = \int_{S\times Z}\int_W g_\tau^D\big(s, a_\tau^\theta(s,z), w\big)\mathrm{d}\hat{\xi}_\tau^{s,a_\tau^\theta(s,z)}(w)\mathrm{d}\gamma_\tau^\theta(s,z).
\end{aligned}$$

Thus, both parts of the Bellman objective can be bounded using one integral with respect to the measure constructed above.

Step 3: Bound the cost of this coupling. We now combine the two parts of the Bellman objective and use the near-optimality of the target assignment and the actions to bound the total cost. Start by observing that the next-stage transport cost $\mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\pi^\theta_{\mathrm{sep},\tau}(\mu)}, \nu)$ is the infimum over all admissible couplings, so evaluating it at the coupling just constructed gives an upper bound. Consequently, for the fixed realization of the noise field in the previous step, and writing $a = a_\tau^\theta(s,z)$, the Bellman objective of the separation-based policy is bounded by

$$\begin{aligned}
&\int_{S\times A\times W} g_\tau^D(s,a,w)\mathrm{d}\rho_\tau^{\pi^\theta_{\mathrm{sep},\tau}(\mu)}(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\pi^\theta_{\mathrm{sep},\tau}(\mu)}, \nu) \\
&\quad \le \int_{S\times Z}\int_W \big[g_\tau^D(s,a,w) + j_{\tau+1}^D\big(f_\tau^D(s,a,w),z\big)\big]\,\mathrm{d}\hat{\xi}_\tau^{s,a}(w)\mathrm{d}\gamma_\tau^\theta(s,z).
\end{aligned}$$

To average this bound over the noise field, fix μ and γ_τ^θ. The universally measurable selector a_τ^θ has a Borel version $b : S\times Z \to A$ that agrees with it γ_τ^θ-almost everywhere. Such a version follows by applying [31, Lemma 7.28(c)] to the selector's Dirac kernel and inverting the Dirac embedding on the resulting full-measure set. Replacing the selector by b changes neither the control nor the integrals. With $a = b(s,z)$, the cost in square brackets is nonnegative and lower semianalytic. The nominal kernel and the law of the realized kernel are Borel after this substitution on a γ_τ^θ-full set, where Assumptions A.1 and A.2 hold. Extending the kernels off that set, we may apply Lemma A.10 with base measure γ_τ^θ on $S\times Z$ and a singleton action space. Since the coupling and selector were fixed before the noise realization, taking expectations and recognizing the Bellman operator gives

$$\begin{aligned}
&\mathcal{T}_\tau^{D,\pi^\theta_{\mathrm{sep}}}\big[\mathcal{K}[j_{\tau+1}^D]\big]\,(\mu,\nu) \\
&\quad = \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau^D(s,a,w)\mathrm{d}\rho_\tau^{\pi^\theta_{\mathrm{sep},\tau}(\mu)}(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{D,\pi^\theta_{\mathrm{sep},\tau}(\mu)}, \nu)\right] \\
&\quad \le \mathbb{E}_{\eta_\tau}\left[\int_{S\times Z}\int_W \big[g_\tau^D(s,a,w) + j_{\tau+1}^D\big(f_\tau^D(s,a,w),z\big)\big]\,\mathrm{d}\hat{\xi}_\tau^{s,a}(w)\mathrm{d}\gamma_\tau^\theta(s,z)\right] \\
&\quad = \int_{S\times Z}\int_W \big[g_\tau^D(s,a,w) + j_{\tau+1}^D\big(f_\tau^D(s,a,w),z\big)\big]\,\mathrm{d}\xi_\tau(w\,|\,s,a)\mathrm{d}\gamma_\tau^\theta(s,z),
\end{aligned}$$

where the last equality is exactly the application of Lemma A.10 with the expression in square brackets as the cost.

By (40), the inner integral is at most $j_\tau^D(s,z) + \theta_\tau/2$. Therefore,

$$\mathcal{T}_\tau^{D,\pi_{\mathrm{sep}}^\theta}\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) \le \int_{S\times Z}\left[j_\tau^D(s,z) + \frac{\theta_\tau}{2}\right]\mathrm{d}\gamma_\tau^\theta(s,z) \le \mathcal{K}[j_\tau^D](\mu,\nu) + \theta_\tau,$$

where the last inequality uses the near-optimality of the target assignment in (39) to contribute another $\theta_\tau/2$. This is exactly (42). □

The previous lemma controls the cost of the chosen policy relative to the decoupled transport surrogate. To compare that surrogate with the decoupled optimal Bellman step, we also use the result proved in Appendix D. In particular, applying Lemma D.3 to the decoupled model gives

$$0 \le \mathcal{K}[j_\tau^D](\mu,\nu) - \mathcal{T}_\tau^D\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) \le L_{\tau+1}D_\tau. \tag{44}$$

We now have the three one-step estimates needed to derive the full performance bound. (i) Lemma E.5 controls the separation-based policy in the decoupled model, (ii) Lemma E.4 transfers a Bellman step between the coupled and decoupled models, and (iii) (44) compares the decoupled transport surrogate with the decoupled optimal Bellman step. The theorem below accumulates these errors over the remaining time steps.

Theorem E.6 (Performance of separation under weak coupling)**.** *Consider the coupled and decoupled models defined above. Assume that the field satisfies Assumptions A.1 and A.2, that the decoupled single-agent cost-to-go satisfies Assumption D.1 with constants L_τ, and that Assumption E.1 holds with constants ε_τ^f and δ_τ. Let $\pi_{\mathrm{sep}}^\theta$ be the policy defined in* (39)–(41)*. Then, for every $\tau \in \{0,\dots,T\}$, $\mu \in \mathcal{P}(S)$ with finite first moment, and $\nu \in \mathcal{P}(Z)$,*

$$0 \le J_\tau^{C,\pi_{\mathrm{sep}}^\theta}(\mu,\nu) - J_\tau^C(\mu,\nu) \le \sum_{r=\tau}^{T-1}\left[\theta_r + L_{r+1}D_r + 2\delta_r + 2L_{r+1}\varepsilon_r^f\right]. \tag{45}$$

When the transport and action minima are attained measurably, we can take $\theta_r = 0$ for every r. In this case, at the initial population $\bar{\mu}_0$, the bound becomes

$$J_0^{C,\pi_{\mathrm{sep}}^0}(\bar{\mu}_0,\nu) - J_0^C(\bar{\mu}_0,\nu) \le \sum_{r=0}^{T-1} L_{r+1}D_r + 2\sum_{r=0}^{T-1}\left(\delta_r + L_{r+1}\varepsilon_r^f\right). \tag{46}$$

Proof. We use the decoupled transport cost as a common reference. First, we bound from above the cost incurred by the separation-based policy in the coupled model. Second, we bound the coupled optimal value from below using the same reference. Finally, we subtract the second estimate from the first to obtain the claimed performance bound. Both estimates are proved backward in time. At $\tau = T$, the policy value, the coupled optimum, and $\mathcal{K}[j_T^D]$ all equal the common terminal cost in (28), so both inductions have the same base case. Explicitly, for every (μ,ν),

$$J_T^{C,\pi_{\mathrm{sep}}^\theta}(\mu,\nu) = J_T^C(\mu,\nu) = \mathcal{K}[j_T^D](\mu,\nu) = \mathcal{K}[g_T](\mu,\nu).$$

Since the sums in (47) and (48) are empty at $\tau = T$, these equalities directly prove both base cases.

In both inductions that follow, we use the monotonicity and constant-shift properties of the Bellman operators (Lemma D.2). Indeed, the lemma is stated for $\mathcal{T}_\tau^D = \mathcal{T}_\tau$, and since its proof uses only the structure of the definition, it also applies verbatim to $\mathcal{T}_\tau^C$ and to the fixed-policy operators $\mathcal{T}_\tau^{m,\pi}$.

Step 1: Upper bound for the separation-based policy. We first show that the policy cost in the coupled model is at most the decoupled transport cost plus the near-optimality and model-mismatch errors. More precisely, for every (μ,ν),

$$J_\tau^{C,\pi_{\mathrm{sep}}^\theta}(\mu,\nu) \le \mathcal{K}[j_\tau^D](\mu,\nu) + \sum_{r=\tau}^{T-1}\left(\theta_r + \delta_r + L_{r+1}\varepsilon_r^f\right). \tag{47}$$

Assume, by backward induction, that (47) holds at time $\tau+1$ for every population. The induction hypothesis says that, from any such population, the cost of following the separation-based policy is at

most the corresponding decoupled transport cost plus the errors accumulated from times $\tau+1$ to $T-1$.

Now fix (μ,ν) and apply the induction hypothesis at the realized next population $\mu_{\tau+1}^{C,\pi^\theta_{\mathrm{sep},\tau}(\mu)}$. We obtain

$$\begin{aligned}
J_\tau^{C,\pi^\theta_{\mathrm{sep}}}(\mu,\nu) &= \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau^C(s,a,\mu,w)\mathrm{d}\rho_\tau^{\pi^\theta_{\mathrm{sep},\tau}(\mu)}(s,a,w) + J_{\tau+1}^{C,\pi^\theta_{\mathrm{sep}}}\left(\mu_{\tau+1}^{C,\pi^\theta_{\mathrm{sep},\tau}(\mu)},\nu\right)\right] \\
&\le \mathbb{E}_{\eta_\tau}\left[\int_{S\times A\times W} g_\tau^C(s,a,\mu,w)\mathrm{d}\rho_\tau^{\pi^\theta_{\mathrm{sep},\tau}(\mu)}(s,a,w) + \mathcal{K}[j_{\tau+1}^D](\mu_{\tau+1}^{C,\pi^\theta_{\mathrm{sep},\tau}(\mu)},\nu)\right] \\
&\quad + \sum_{r=\tau+1}^{T-1}\left(\theta_r+\delta_r+L_{r+1}\varepsilon_r^f\right) \\
&= \mathcal{T}_\tau^{C,\pi^\theta_{\mathrm{sep}}}\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) + \sum_{r=\tau+1}^{T-1}\left(\theta_r+\delta_r+L_{r+1}\varepsilon_r^f\right) \\
&\le \mathcal{T}_\tau^{D,\pi^\theta_{\mathrm{sep}}}\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) + \delta_\tau + L_{\tau+1}\varepsilon_\tau^f + \sum_{r=\tau+1}^{T-1}\left(\theta_r+\delta_r+L_{r+1}\varepsilon_r^f\right) \\
&\le \mathcal{K}[j_\tau^D](\mu,\nu) + \sum_{r=\tau}^{T-1}\left(\theta_r+\delta_r+L_{r+1}\varepsilon_r^f\right).
\end{aligned}$$

The second inequality is (35) in Lemma E.4, which costs $\delta_\tau + L_{\tau+1}\varepsilon_\tau^f$, while the last inequality is (42) in Lemma E.5, which contributes θ_τ. After adding these current errors to the errors from later times, we obtain (47) at time τ and prove the upper bound.

Step 2: Lower bound for the coupled optimum. We now compare the coupled optimal value with the same decoupled transport cost. Here the transport cost provides a lower bound after we subtract the separation and model-mismatch errors:

$$J_\tau^C(\mu,\nu) \ge \mathcal{K}[j_\tau^D](\mu,\nu) - \sum_{r=\tau}^{T-1}\left[L_{r+1}D_r+\delta_r+L_{r+1}\varepsilon_r^f\right]. \tag{48}$$

Now fix (μ,ν). Since the induction hypothesis is a pointwise lower bound on $J_{\tau+1}^C$, applying it inside $\mathcal{T}_\tau^C$ and using Lemma D.2 gives

$$\begin{aligned}
J_\tau^C(\mu,\nu) &= \mathcal{T}_\tau^C\left[J_{\tau+1}^C\right](\mu,\nu) \\
&\ge \mathcal{T}_\tau^C\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) - \sum_{r=\tau+1}^{T-1}\left[L_{r+1}D_r+\delta_r+L_{r+1}\varepsilon_r^f\right] \\
&\ge \mathcal{T}_\tau^D\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) - \delta_\tau - L_{\tau+1}\varepsilon_\tau^f - \sum_{r=\tau+1}^{T-1}\left[L_{r+1}D_r+\delta_r+L_{r+1}\varepsilon_r^f\right] \\
&\ge \mathcal{K}[j_\tau^D](\mu,\nu) - \sum_{r=\tau}^{T-1}\left[L_{r+1}D_r+\delta_r+L_{r+1}\varepsilon_r^f\right].
\end{aligned}$$

Here, the second inequality is (34) in Lemma E.4, which contributes $\delta_\tau + L_{\tau+1}\varepsilon_\tau^f$. Rearranging (44) gives

$$\mathcal{T}_\tau^D\left[\mathcal{K}[j_{\tau+1}^D]\right](\mu,\nu) \ge \mathcal{K}[j_\tau^D](\mu,\nu) - L_{\tau+1}D_\tau,$$

which is the last inequality in the display. This proves (48) at time τ.

Step 3: Combine the two bounds. Subtracting the obtained lower bound (48) from the upper bound (47) cancels the common decoupled transport cost and gives

$$J_\tau^{C,\pi_{\mathrm{sep}}^\theta}(\mu,\nu) - J_\tau^C(\mu,\nu) \le \sum_{r=\tau}^{T-1} \left(\theta_r + \delta_r + L_{r+1}\varepsilon_r^f\right) + \sum_{r=\tau}^{T-1} \left(L_{r+1}D_r + \delta_r + L_{r+1}\varepsilon_r^f\right)$$
$$= \sum_{r=\tau}^{T-1} \left[\theta_r + L_{r+1}D_r + 2\delta_r + 2L_{r+1}\varepsilon_r^f\right].$$

That corresponds exactly to the upper bound in (45). To conclude the proof, it is enough to note that the zero lower bound is immediate because $\pi_{\mathrm{sep}}^\theta$ is admissible in the coupled model and J_τ^C is the infimum of the costs over all admissible policies there. Finally, if exact minimizers can be selected measurably, setting $\tau = 0$, $\mu = \bar{\mu}_0$, and $\theta_r = 0$ gives (46). □

Interpretation. The three terms in the bound correspond to the three comparisons used in the proof. The tolerance θ_r comes from the near-optimal target plan and action used to define the separation-based policy in (39)–(41). The term $L_{r+1}D_r$ is already present in Lemma D.3 for the decoupled problem and bounds the difference between assigning targets before the noise is observed and optimizing the population Bellman step after the resulting population is known. Finally, $\delta_r + L_{r+1}\varepsilon_r^f$ is the cost of replacing one coupled Bellman step by its decoupled counterpart, as bounded in Lemma E.4. This last term appears twice because the proof makes that replacement once for the deployed policy and once for the coupled optimal value.

Bounding the model mismatch. Note that Assumption E.1 only quantifies the model mismatch in terms of the abstract mismatch ε_τ^f. To use the theorem in a concrete model, it is often more convenient to relate the quantity to the strength of the population interaction. This is the case, for example, when the coupled dynamics are Lipschitz in the population distribution:

$$d_S\big(f_\tau^C(s,a,\mu,w), f_\tau^C(s,a,\widetilde{\mu},w)\big) \le \iota_\tau W_1(\mu,\widetilde{\mu}), \tag{49}$$

uniformly in (s,a,w). In this case, indeed, the constant ι_τ measures how strongly the population affects the dynamics. If, in addition, the decoupled surrogate is obtained by freezing that population argument at a reference profile μ_τ^{ref}:

$$f_\tau^D(s,a,w) = f_\tau^C(s,a,\mu_\tau^{\mathrm{ref}},w),$$

and it holds that $W_1(\mu,\mu_\tau^{\mathrm{ref}}) \le R_\tau$ at every population where the two models are compared, then we can apply (49) with $\widetilde{\mu} = \mu_\tau^{\mathrm{ref}}$. By the definition of f_τ^D, this allows us to bound the general model mismatch in Assumption E.1 by the more informative $\varepsilon_\tau^f \le \iota_\tau R_\tau$. Under this class of models, (46) becomes

$$J_0^{C,\pi_{\mathrm{sep}}^0}(\bar{\mu}_0,\nu) - J_0^C(\bar{\mu}_0,\nu) \le \sum_{\tau=0}^{T-1} L_{\tau+1}D_\tau + 2\sum_{\tau=0}^{T-1} \left(\delta_\tau + L_{\tau+1}\iota_\tau R_\tau\right).$$

If, in addition, the stage costs coincide, so that $\delta_\tau = 0$, then

$$J_0^{C,\pi_{\mathrm{sep}}^0}(\bar{\mu}_0,\nu) - J_0^C(\bar{\mu}_0,\nu) \le \sum_{\tau=0}^{T-1} L_{\tau+1}D_\tau + 2\sum_{\tau=0}^{T-1} L_{\tau+1}\iota_\tau R_\tau. \tag{50}$$

The second sum in (50) corresponds to the additional error due to weak coupling: it grows linearly in ι_τ and vanishes when the dynamics stop depending on the population, in which case (50) reduces to the bound of Theorem D.4.

F Additional numerical results

This appendix provides the implementation details and complementary results for the experiments in Section 5. We first describe the grid-world experiments and report the additional performance metrics used in our evaluation (Appendix F.1). We then give the details of the scalability study (Appendix F.2) and examine how the origin–destination pairs covered during training affect transfer across population sizes (Appendix F.3). Finally, we describe the Manhattan road network (Appendix F.4) and report the computational resources used in the experiments.

F.1 Additional details for the grid-world experiments in Section 5.1

Environment. We use an 11×30 grid with horizon $T = 100$ and $N = 7$ agents. At the beginning of each episode, the agents are placed uniformly at random in the first column, while the N targets are sampled uniformly from the last ten columns. The action space is $A = \{\texttt{up}, \texttt{down}, \texttt{left}, \texttt{right}, \texttt{stay}\}$. At each time step, the selected action is executed with probability $1 - p$ and is replaced uniformly by one of the other four actions with probability p. We consider $p \in \{0, 0.1, 0.25\}$ under the individual, local, and global noise models defined in Appendix A. Specifically, under individual noise, each agent receives an independent noise draw, under local noise, agents with the same state-action pair share the same draw, while under global noise, the entire population shares one draw. We finally clip the resulting positions of the agents to make them remain in the grid after each transition.

An agent becomes inactive after reaching a relevant target and enters an absorbing state. In the case of a MARL agent, a relevant target for an agent is any previously unoccupied target, in the case of SALT, instead, each agent is assigned a target z before the rollout, and the agent becomes inactive when it reaches that specific target.[2] For SALT with a fixed assignment, both the routing policy and the stopping rule depend only on each agent's own state and target. Conditional on the initial states and assignments, the trajectories are independent under individual noise, but may be correlated under local or global noise.

Reward structure. An active agent incurs a stage cost of 1 on every transition until it does not reach a target. At the horizon, the MARL training objective also includes the empirical optimal transport cost $\mathcal{K}[c](\mu_T, \nu)$, where c is the Manhattan distance. In the target-conditioned single-agent problems used by SALT, the corresponding terminal penalty is the Manhattan distance from the agent to its assigned target. At evaluation, we report travel time, reach rate, and empirical terminal optimal transport cost separately.

Baselines. We compare SALT with MAPPO, IPPO, HAPPO, VDN, QMIX, and MFQ. All baselines use the same episode budget and training seeds. All neural policies have two hidden layers with 64 units and $\tanh$ activations. The separation-based neural methods are trained as goal-conditioned single-agent controllers. For SALT-PPO, SALT-A2C, and SALT-DDQN, the policy receives the input

$$\left(\frac{r}{H_g}, \frac{c}{W_g}, \frac{r_z}{H_g}, \frac{c_z}{W_g}, \frac{t}{T-1}, \frac{\|s - z\|_1}{H_g + W_g} \right),$$

where $s = (r, c)$ is the current agent position, $z = (r_z, c_z)$ is its assigned target, t is the time step, and $H_g = 10$ and $W_g = 29$ are the maximum row and column indices. The separation policy therefore selects an action without observing the full population. Coordination between agents occurs only through solving the optimal transport problem, whose solution selects z.

MAPPO, IPPO, HAPPO, QMIX, VDN, and MFQ instead receive target-relative local observations. For agent i, this observation has dimension $5 + 3N$:

$$\left(\frac{r_i}{H_g}, \frac{c_i}{W_g}, \frac{i}{N-1}, \frac{t}{T-1}, a_i, \frac{\Delta r_{i1}}{H_g}, \frac{\Delta c_{i1}}{W_g}, b_1, \ldots, \frac{\Delta r_{iN}}{H_g}, \frac{\Delta c_{iN}}{W_g}, b_N \right).$$

Here a_i indicates whether agent i is active, $\Delta r_{ij} = r_{z_j} - r_i$ and $\Delta c_{ij} = c_{z_j} - c_i$ are the coordinates of target j relative to agent i, and b_j indicates whether target j remains unoccupied. Thus, each agent observes the target set as an ordered list of relative offsets and availability flags. MAPPO also uses a centralized critic state of dimension $6N + 1$. This state contains all normalized agent positions, all active-agent flags, the time step, all normalized target positions, and all target-availability flags. HAPPO uses the same centralized critic state, but has one non-shared actor per agent and updates the actors sequentially. QMIX passes the global state to its monotone mixing network, whose embedding dimension is 32. IPPO uses a local critic, VDN uses additive value decomposition, and MFQ augments each agent's action value with the global mean action distribution of the currently active agents. All neural methods are optimized with Adam [49].

[2] This target-dependent stopping rule is an experimental convention and is not part of the target-independent motion map $f_t(s, a, w)$ in Section 2. The simulator also holds an inactive agent at its arrival position and masks subsequent actions, perturbations, and stage costs.

Training and evaluation. For each algorithm, noise model, and noise probability, we train five policies using seeds 0–4 and evaluate each policy on 50 rollout instances. Within a given condition, evaluation seed $s \in \{0, \dots, 49\}$ fixes the starts, targets, and initial state of the transition-noise generator for every algorithm.

For each metric, we pool the 50 outcomes from each training seed, and report the arithmetic mean of the resulting 250 outcomes in the heatmaps in Figures 2, 5, and 6. We then summarize all the results and include the corresponding confidence intervals in Figure 7. In particular, the reach-rate bars in Figure 7 show the pooled means together with two-sided 95% confidence intervals, while the terminal-cost and travel-time boxplots show the median and interquartile range (from the boxes), and observed minimum and maximum of the pooled outcomes (from the whiskers).

Hyperparameters. Tables 2, 3, and 4 report all hyperparameters used in the grid-world experiments.

Table 2: Shared grid-world training budget and evaluation settings for Section 5.1.

Quantity	Value
Grid / horizon	11×30, $T = 100$
Agents / targets	$N = 7$
Training seeds	0–4 (5 independent runs)
Evaluation instances	50 per training seed, 250 pooled
Noise probabilities	$0, 0.1, 0.25$
Noise modes	individual, local, global
Training episodes	320,000
Environment interactions	$320{,}000 \times 7 \times 100 = 2.24 \times 10^8$ per run
Batch episodes	64
Update batches for neural methods	5,000
Step cost	1
Goal bonus	0
Evaluation target region	columns 20–29
Evaluation rematching for SALT	every time step

Table 3: Hyperparameters used for SALT in Section 5.1.

Hyperparameter	SALT-Q	SALT-DDQN	SALT-PPO	SALT-A2C
Discount γ	0.98	0.98	0.98	0.98
Learning rate	$0.25 \to 0.02$	3×10^{-4}	actor 3×10^{-4} critic 10^{-3}	actor 3×10^{-4} critic 10^{-3}
Learning-rate decay	20,000 episodes	–	–	–
Hidden size	–	64	64	64
GAE λ	–	–	0.95	0.95
PPO clip / epochs	–	–	0.2 / 4	–
Entropy coefficient	–	–	0.01	0.01
Maximum gradient norm	–	0.5	0.5	0.5
ε-exploration	$0.5 \to 0.05$	$0.5 \to 0.05$	$0.5 \to 0.05$	$0.5 \to 0.05$
ε-decay	160,000 episodes	1.12×10^8 interactions	1.12×10^8 interactions	1.12×10^8 interactions
Replay capacity	–	500,000	–	–
Replay warmup / minibatch	–	5,000 / 256	–	–
Learner steps per batch	–	64	–	–
Target update period	–	500 learner steps	–	–

For SALT-PPO and SALT-A2C, the ε schedule in Table 3 controls random exploration of the target assignment, while action exploration is induced by the stochastic actor. SALT-Q and SALT-DDQN use the same schedule for random assignments and ε-greedy action exploration.

Additional performance metrics. In addition to average travel time, we report two further metrics. Figure 5 reports the empirical optimal transport cost between the final agent positions and the targets, that is, the minimum average Manhattan distance obtained by matching agents to targets. Figure 6

Table 4: Hyperparameters used for MARL baselines in Section 5.1.

Hyperparameter	MAPPO	IPPO	HAPPO	QMIX	VDN	MFQ
Discount γ	0.98	0.98	0.98	0.98	0.98	0.98
Learning rate	actor 3×10^{-4} critic 10^{-3}	actor 3×10^{-4} critic 10^{-3}	actor 3×10^{-4} critic 10^{-3}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Hidden size	64	64	64	64	64	64
GAE λ	0.95	0.95	0.95	–	–	–
PPO clip / epochs	0.2 / 4	0.2 / 4	0.2 / 4	–	–	–
Entropy coefficient	$0.01 \rightarrow 0.001$	$0.01 \rightarrow 0.001$	$0.01 \rightarrow 0.001$	–	–	–
Maximum gradient norm	0.5	0.5	0.5	0.5	0.5	0.5
Actor minibatch	4,096	4,096	4,096	–	–	–
Critic minibatch	1,024	4,096	1,024	–	–	–
Replay capacity	–	–	–	100,000	100,000	100,000
Minimum replay size	–	–	–	2,048	2,048	2,048
Replay batch size	–	–	–	512	512	512
Updates per batch	–	–	–	8	8	8
Target update period	–	–	–	50 batches	50 batches	50 batches
Exploration	–	–	–	ε: $0.5 \rightarrow 0.05$	ε: $0.5 \rightarrow 0.05$	τ: $5 \rightarrow 0.1$ ε: $0.3 \rightarrow 0.02$
Exploration decay	–	–	–	160,000 episodes	160,000 episodes	320,000 episodes

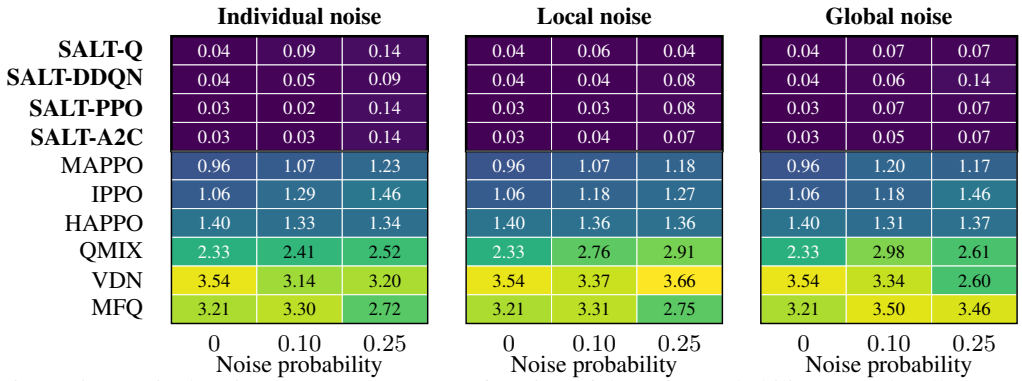


Figure 5: Terminal optimal transport cost as a function of the noise probability p. Each cell shows the pooled arithmetic mean under the training and evaluation protocol described above. Figure 7 shows the corresponding distributions. Brighter cells indicate higher cost.

reports the fraction of agents that reach a target before the horizon. More precisely, let

$$\mu_T^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{s_T^i}, \qquad \nu^N = \frac{1}{N}\sum_{j=1}^{N} \delta_{z^j}.$$

With this notation, the empirical terminal optimal transport cost is

$$\mathcal{K}[d](\mu_T^N, \nu^N) = \min_{\sigma \in S_N} \frac{1}{N}\sum_{i=1}^{N} d\Big(s_T^i, z^{\sigma(i)}\Big),$$

where S_N is the set of permutations of $\{1, \ldots, N\}$, s_T^i is the final position of agent i, z^j is target j, $d(\cdot, \cdot)$ is the Manhattan distance, and σ assigns agents to targets. This cost is zero exactly when every target is occupied and, unlike the reach rate, it also distinguishes near misses from terminal configurations that remain far from the target set.

The three metrics complement each other in describing performance and should therefore be read together. Average travel time tells us how long the agents travel, but an unsuccessful trajectory may have the same travel time as a slow but successful one. The reach rate tells us how many agents actually reach a target, but treats all unsuccessful agents in the same way. The terminal optimal transport cost adds the remaining information by distinguishing final configurations that are close to full target coverage from those that remain far from it. Figure 7 reports the three metrics together. To interpret it recall that higher values indicate better performance for the reach rate, while lower values indicate better performance for terminal optimal transport cost and average travel time. We can observe that across all three metrics, the SALT variants achieve higher reach rates, lower terminal costs, and shorter travel times than the MARL baselines.

	Individual noise			Local noise			Global noise		
SALT-Q	100	99.9	99.9	100	99.9	99.9	100	99.9	99.9
SALT-DDQN	99.8	100	100	99.8	100	100	99.8	100	100
SALT-PPO	100	100	100	100	100	100	100	100	100
SALT-A2C	100	100	99.8	100	100	100	100	100	100
MAPPO	84.9	83.9	81.1	84.9	84.4	80.2	84.9	82.2	81.4
IPPO	87.7	85.9	80.7	87.7	85	81.4	87.7	85.5	80.2
HAPPO	79	79.4	77.7	79	78.9	76.6	79	79.4	77.9
QMIX	61.9	61.3	63.7	61.9	58.7	59.1	61.9	57	58.3
VDN	46.2	49.9	52.7	46.2	46.5	46.7	46.2	44	50.6
MFQ	59.3	56.5	58.4	59.3	54.9	59.3	59.3	60.1	57.7
	0 0.10 0.25			0 0.10 0.25			0 0.10 0.25		
	Noise probability			Noise probability			Noise probability		

Figure 6: Reach rate as a function of the noise probability p. Each cell shows the pooled arithmetic mean under the training and evaluation protocol described above. Figure 7 reports the corresponding 95% confidence intervals. Darker cells indicate higher reach rate.

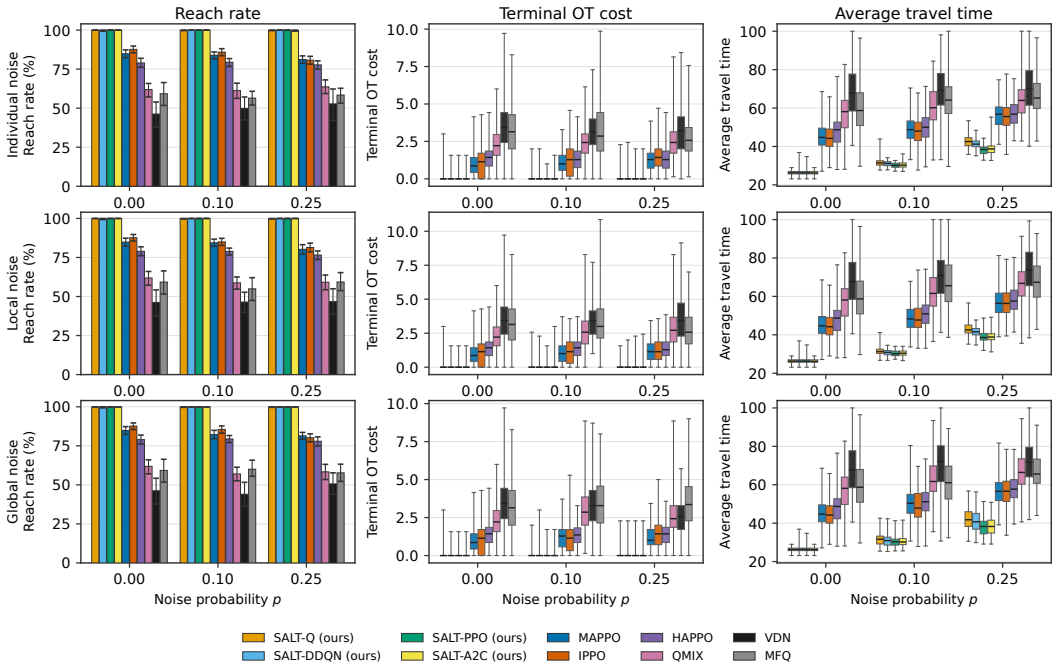


Figure 7: Reach rate, terminal optimal transport cost, and average travel time under the evaluation protocol described above. Reach-rate bars show pooled means and two-sided 95% crossed two-way bootstrap confidence intervals. The remaining panels show the pooled outcome distributions. Boxes span the interquartile range, center lines are medians, and whiskers show the observed minimum and maximum.

F.2 Additional details for the scalability experiment

We next study how the number of environment interactions required for training changes with the population size. For this experiment, we use square grids and individual noise with probability $p = 0.1$. Each agent starts from an independently sampled row of the first column, while the targets are sampled independently and uniformly over the full grid.

To avoid confounding population size with target density, we scale the grid area approximately in proportion to N, the resulting grid sizes are reported in Table 5. We select PPO as the single-agent learner in SALT and compare the resulting method with MAPPO. The hyperparameters are the same as in Tables 3 and 4. For each population size in Table 5, we choose the maximum number of batches

Table 5: Population sizes and grid sizes used in the scalability experiment.

Agents N	2	4	6	8	16	32	64	128
Grid dimensions	10×10	14×14	17×17	20×20	28×28	40×40	57×57	80×80
Horizon	20	28	34	40	56	80	114	160
Maximum planned batches	781,250	279,018	153,186	97,656	34,877	12,207	4,283	1,526

so that the per-agent-transition budget is approximately 2×10^9. We record a reach-rate milestone at the first of 10 consecutive batches whose 64-episode reach rate meets the stated threshold in Figure 3.

F.3 Origin–destination coverage under population-size transfer

The transportation cost used by SALT is learned from target-conditioned trajectories. Consequently, an origin–destination (OD) pair selected only a few times during training may therefore have a less accurate cost-to-go estimate than a pair selected more often, and this can affect performance at evaluation especially when the population size differs from the one used during training. Indeed, as training progresses, a larger population gives the assignment layer more opportunities to select low-cost matchings, whereas some OD pairs may then be selected less often and become underrepresented in the training data. With fewer agents, however, we have a limited number of available matchings, and some unfavorable assignments cannot be avoided. As a result, the OD pairs encountered at evaluation may differ from those for which the transportation cost was learned accurately.

Here, we examine this effect. To do so, we use a grid world in which the start and target distributions are known and remain the same during training and evaluation, and keep track of the visitation frequency of each OD pair during training.

Setup. We use a 30×30 grid with horizon $T = 100$, unit stage cost, individual action noise with probability $p = 0.1$, and uniform start and target distributions over the full grid. We compare training with $N = 20$ and $N = 5$ and evaluate the performance across fleets of $N \in \{1, 5, 10, 20, 50\}$ agents. Table 6 reports the training settings for the two population sizes.

Table 6: Training settings for the origin–destination coverage experiment.

Quantity	$N = 20$	$N = 5$
Training seeds	0–4	0–4
Nominal interaction budget	2×10^9 agent steps	2×10^9 agent steps
Assignment observations	20,000,000	20,000,000
Population episodes	1,000,000	4,000,000
Learning rate	$0.25 \to 0.05$	$0.25 \to 0.05$
Learning-rate decay	800,000 episodes	3,200,000 episodes
Action exploration	$0.5 \to 0.05$	$0.5 \to 0.05$
Action-exploration decay	500,000 episodes	2,000,000 episodes
Assignment exploration $\varepsilon_{\mathrm{match}}$	$0.5 \to 0$	$0.5 \to 0$
Assignment-exploration decay	800,000 episodes	3,200,000 episodes

Since an episode produces four times as many assignment observations when $N = 20$ as when $N = 5$, the decay periods for $N = 5$ are four times longer. The schedules are therefore the same when measured in assignment observations.

At the beginning of each training episode, follow the assignment exploration rule of Algorithm 1, with $\varepsilon_{\mathrm{match}}$ corresponding to α_k. With probability $\varepsilon_{\mathrm{match}}$, we choose a uniformly random one-to-one agent–target matching, we solve the empirical optimal transport problem using the current Q-based cost-to-go estimate. To measure the coverage of the training data, we record how often each OD pair is selected:

$$n(s, z) = \#\{\text{training episodes in which an agent starting at } s \text{ is assigned to } z\}.$$

Although the learned matching concentrates on lower-cost pairs, assignment exploration mitigates this and gives nonzero support to at least 809,801 of the 810,000 possible OD pairs. The grid-world Q-table is initialized at zero.

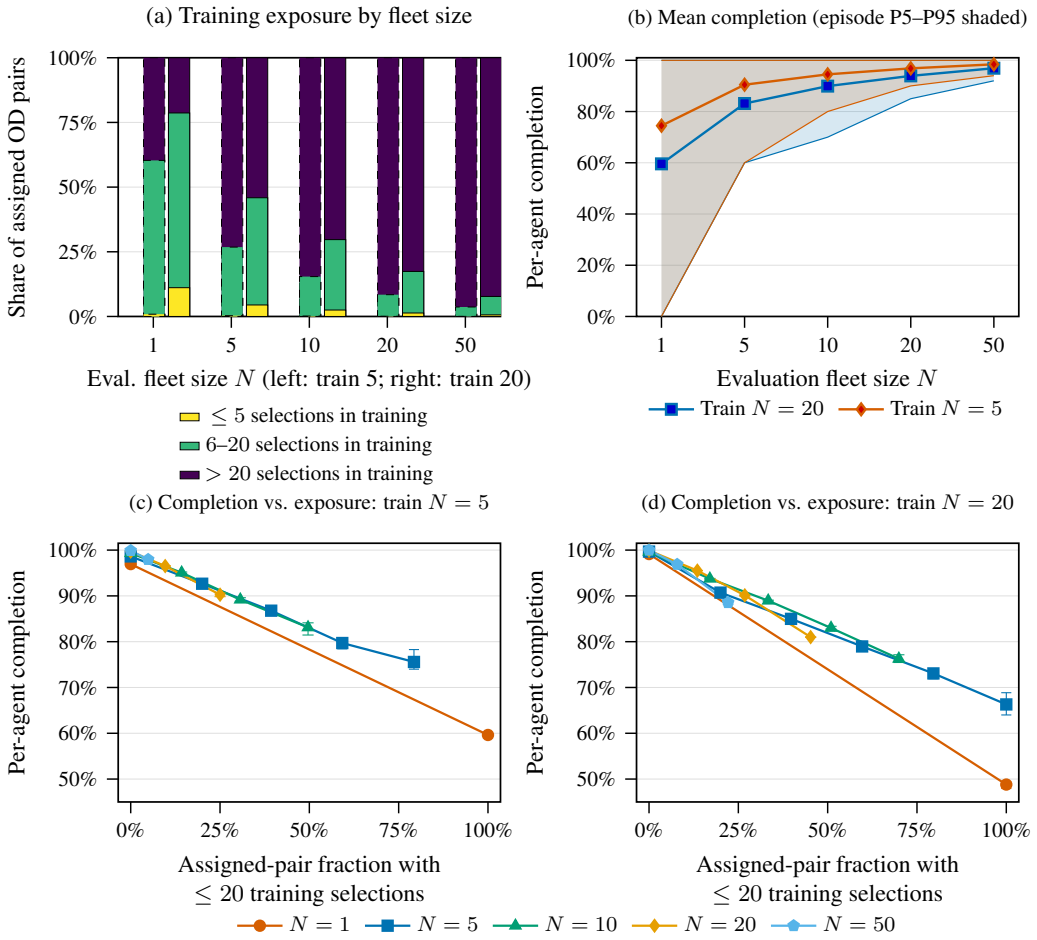


Figure 8: OD-pair coverage and population-size transfer after training SALT-Q with $N = 20$ or $N = 5$ agents at the same nominal interaction budget. (a) Share of nontrivial evaluation assignments grouped by training count; within each evaluation fleet, the left and right bars correspond to training with $N = 5$ and $N = 20$. (b) Overall per-agent completion; lines show pooled episode means and shaded regions the episode-level fifth–ninety-fifth percentiles. (c–d) Per-agent completion conditioned on the fraction of assignments satisfying $n(s, z) \leq 20$, shown separately for the two training fleets. Points are means and error bars span the minimum and maximum over five training seeds. The x coordinate is the mean realized fraction within each bin.

For each trained Q-table, we evaluate fleets of $N \in \{1, 5, 10, 20, 50\}$ agents for 5,000 episodes. During evaluation, we compute the initial optimal transport assignment from the learned Q-table and keep it fixed for the entire rollout, so to better relate training coverage to evaluation performance.

When classifying coverage, we exclude the trivial pairs for which $s = z$ and group every remaining pair (s, z) according to the number of times it was selected during training: $n(s, z) \leq 5$, $6 \leq n(s, z) \leq 20$, or $n(s, z) > 20$. In Figures 8c–d, the horizontal coordinate is the fraction of nontrivial assignments in an episode that satisfy $n(s, z) \leq 20$. We divide this fraction into the bins 0%, $(0, 20\%]$, $(20, 40\%]$, $(40, 60\%]$, $(60, 80\%]$, and $(80, 100\%]$ and a bin is shown only if each of the five training seeds contributes at least 20 evaluation episodes. In this case completion corresponds to the fraction of agents that reach their assigned target by the horizon.

Results. Figure 8a reports the percentage of evaluation OD pairs in each of the three training-count ranges described above. For both training fleets, the shares satisfying $n(s, z) \leq 5$ and $6 \leq n(s, z) \leq 20$ decrease as the evaluation fleet grows. We can see that, with the same interaction budget, training with $N = 5$ agents gives broader coverage with respect to training with $N = 20$

agents. For example, for an evaluation fleet of size $N = 1$, the share satisfying $n(s, z) > 20$ increases from 21.3% after training with $N = 20$ to 39.7% after training with $N = 5$. At the same time, the share satisfying $n(s, z) \leq 5$ decreases from 11.1% to 0.9%.

Figure 8b shows how this difference in coverage impacts the completion rate. Training with $N = 20$ leads to completion rates that increase from 59.6% for an evaluation fleet of size $N = 1$ to 96.9% for $N = 50$. Training with $N = 5$ raises these values to 74.5% and 98.4%, respectively, and also improves completion at every intermediate fleet size. The shaded areas show the episode-level fifth–ninety-fifth percentiles after pooling the evaluation episodes from all five training seeds.

Figures 8c–d examine the relation more directly by plotting completion against the fraction of evaluation assignments satisfying $n(s, z) \leq 20$. For each fleet size and under both training regimes, completion decreases as this fraction increases. For an evaluation fleet of size $N = 1$, the assignment layer has no alternative target. In this case, in episodes involving an under-covered OD pair, the agent reaches its target in 48.8% of cases after training with $N = 20$, compared with 59.6% after training with $N = 5$. A similar difference is present for $N = 5$: when 20% of the assignments are under-covered, completion increases from 90.7% to 92.7%, while when their mean fraction is approximately 60%, it increases from 79.0% to 79.7%. The minimum–maximum bars show that these trends are consistent across the five training seeds.

These results show that the OD-pair coverage obtained during training has an important impact on the transfer of SALT across fleet sizes. When the evaluation fleet is smaller than the training fleet, the assignment layer may select pairs that were underrepresented in the training data, leading to worse performance. At a fixed interaction budget, training with fewer agents improves coverage and reduces this mismatch. However, it requires proportionally more population episodes and therefore a longer training time. We believe that adapting the fleet size during training may help balance these two effects, and point it out as a promising direction for future work.

F.4 Towards a realistic, time-varying urban logistics task

Manhattan environment and dynamics. We model southern Manhattan as a directed road graph extracted from OpenStreetMap [39], whose data are available under the Open Data Commons Open Database License (ODbL). We restrict the graph to 38 taxi zones in southern Manhattan, retain its largest strongly connected component, and remove self-loops. The resulting graph contains $|V| = 2087$ nodes and 4255 directed edges. Each node in the graph represents an intersection, while each directed edge represents an admissible road segment. The nominal edge travel times are computed from the road length and a road-type-dependent speed limit.

To model variable traffic conditions, we include traffic signals and stochastic congestion noise. We model edge-level traffic signals using a 90-second cycle whose green-phase duration depends on the road type. Primary roads are green for 50% of the cycle, whereas secondary and residential roads are green for 30%. The remaining road types (motorway, trunk, tertiary, and unclassified roads) are treated as always green. Finally, the signal offset of each intersection is randomly initialized. The impact of traffic signals is that an agent may have to wait at an intersection if it arrives during a red phase. More precisely, when an agent traverses edge e at time t, its realized transition time is the sum of the travel time and the waiting time determined by the signal phase.

To represent time-varying traffic conditions, we also add stochastic congestion noise. When edge e is traversed, its travel time is multiplied by

$$1 + u, \qquad u \sim \mathrm{Unif}(0, \eta_e),$$

where $\eta_e = 0.1$ for primary roads, $\eta_e = 0.15$ for secondary roads, and $\eta_e = 0$ otherwise. A new value of u is drawn for every agent-edge traversal. We exclude operationally unrealistic immediate U-turns from the urban routing task, in both training and evaluation. After a vehicle traverses $u \to v$, the reverse action $v \to u$ is masked whenever v has another admissible successor.

Time discretization and Q-table. The Manhattan simulator evolves in event time, thus, one decision step corresponds to traversing one outgoing road segment, rather than advancing time by a fixed one-second increment. The tabular Q-learner instead uses a discrete time coordinate with resolution $\Delta t = 1$ second. Since all traffic lights follow a 90-second cycle, we record in the time coordinate of the Q-table the traffic phase instead of unbounded physical time. Thus, the Q-table

contains $T_{\text{disc}} = 90/\Delta t = 90$ time slices. At physical time t, the corresponding table index is

$$k(t) = \text{clip}(\text{floor}(t/\Delta t), 0, 89).$$

More precisely, we consider a target-conditioned Q-table, whose entries are indexed by the current node v, the request node z, the traffic-phase index k, and the outgoing action a: $Q(v, z, k, a)$. For the Manhattan graph, this gives a table of shape

$$|V| \times |V| \times T_{\text{disc}} \times \max_v |\text{out}(v)| = 2087 \times 2087 \times 90 \times 4,$$

where $|\text{out}(v)|$ is the out-degree of node v. Invalid padded actions are masked during both action selection and Bellman backups.

The physical edge travel times remain continuous, while the one-second grid is used only to index the Q-table. If $T_{v,a}$ is the nominal travel time of edge (v, a) and u is its congestion perturbation, the realized edge travel time is $\bar{T}_{v,a} = T_{v,a}(1 + u)$. Once the edge has been traversed, the simulator computes the traffic-signal waiting time from the arrival phase. For an edge with period $P_{v,a}$, offset $o_{v,a}$, and green duration $g_{v,a}$, this phase is

$$\phi(t) = (t + \bar{T}_{v,a} + o_{v,a}) \bmod P_{v,a},$$

and the corresponding waiting time is

$$W_{v,a}(t) = \begin{cases} 0, & \phi < g_{v,a}, \\ P_{v,a} - \phi, & \phi \geq g_{v,a}. \end{cases}$$

The next physical time is therefore

$$t' = t + \bar{T}_{v,a} + W_{v,a}(t),$$

and to obtain the next Q-table index, we round t' to the discrete grid and reduce it modulo the common traffic-signal period:

$$k' = \text{round}(t'/\Delta t) \bmod T_{\text{disc}}.$$

The next lookup therefore uses the resulting phase index k'. The reward is based on the same realized delay $\bar{T}_{v,a} + W_{v,a}(t)$, expressed in minutes.

Task setup. We sample the starting positions from the three fixed depot nodes $\{587, 1226, 1972\}$, which have the highest betweenness centrality, while each request is represented by a destination node z sampled uniformly from the graph. During training, we assign the requests at the beginning of each population episode, by solving an epsilon-greedy optimal transport problem, and keep the assignment fixed for the entire training episode, as required to learn the target-conditioned subproblem. At evaluation, instead, the online-assignment SALT policy recomputes the optimal transport problem from the agents' current nodes and traffic phases at every routing step, and then takes one greedy Q-learning action. Figure 4 compares this policy with a static nominal shortest-path baseline. The baseline computes a single distance-based optimal transport assignment at $t = 0$ and follows nominal shortest paths without accounting for the signal phase or congestion.

Learning algorithm. For the Manhattan experiment, we use tabular Q-learning with learning rate 0.1 and no discount factor. Action exploration is ε-greedy, with ε decreasing linearly from 1.0 to 0.01 over 90% of the total training steps. The training budget is $5{,}000{,}000$ population episodes with 20 agents. The maximum shortest-path length in the processed graph is 66 edges, thus during both training and the reported final evaluation, we cap each trajectory at twice this value, or 132 road traversals. To provide a warm start to the learning algorithm, we initialize the Q-table using nominal shortest-path travel-time estimates, hence ignoring congestion noise or traffic-signal phases.

Reward. At each step, the stage reward of an active agent is the negative realized delay in minutes:

$$r_t = -\frac{\bar{T}_{v,a} + W_{v,a}(t)}{60}.$$

Inactive agents enter an absorbing state and receive zero reward thereafter. If the 132-traversal cap is reached before the target, the terminal Bellman target also accounts for the remaining distance: it includes $-d_{\text{SP}}(v, z)/60$, the negative nominal shortest-path travel time from the current node v to the target z, rather than an arbitrary fixed timeout penalty.

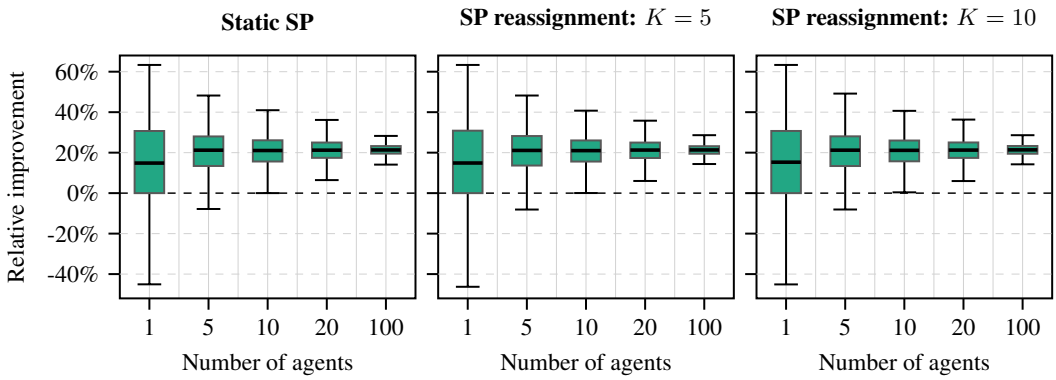


Figure 9: Relative travel-time improvement of SALT-Q against nominal shortest-path routing with a static assignment (left), reassignment every five routing steps (center), or reassignment every ten routing steps (right). The evaluation samples and boxplot conventions match Figure 4.

Evaluation and logging. We report results for 8 training seeds (0–7). For each seed and each fleet size $N \in \{1, 5, 10, 20, 100\}$, we draw 500 evaluation instances, giving 4000 outcomes per fleet size. Within each evaluation run, SALT-Q and all shortest-path baselines share starts, requests, and the fixed traffic-signal configuration, while the separate rollouts use independent congestion-noise streams. We then report the continuous travel-time improvements over the baseline as $100(T_{\text{baseline}} - T_{\text{SALT-Q}})/T_{\text{baseline}}$ per instance. Figures 4 and 9 show medians and interquartile ranges after pooling outcomes across training seeds. Here, whiskers follow the Tukey 1.5-interquartile-range rule and outliers remain in the underlying samples but are not represented in the plots.

We evaluate the performance of SALT over three shortest-path baselines differing in the reassignment frequency they use. Static shortest path computes the optimal transport assignment once at the beginning of the episode and keep the assignment fixed until the end of the episode itself. K-reassignment shortest path instead recomputes the optimal transport assignment every K routing steps from the vehicles' current nodes to the still-unserved requests, served requests instead remain locked to the vehicles that already reached them. Between reassignments, routing follows the same nominal shortest paths as the static baseline. In Figure 9 we show the performance of 5-reassignment and 10-reassignment shortest path. As visible, the reassignment improvement is minimal. To measure it directly, define $G_K(N) = 100(\bar{T}_{\text{static}} - \bar{T}_{\text{reassign},K})/\bar{T}_{\text{static}}$, where bars denote means over the available continuous-time outcomes. For $N = (5, 10, 20, 100)$, we obtain $G_5 = (-0.0014,\ 0.0008,\ 0.0031,\ 0.0086)\%$ and $G_{10} = (-0.0480,\ -0.0243,\ -0.0406,\ -0.0045)\%$. The marginal differences are extremely small and can be traced back to the stochasticity of the environment. Therefore, we can conclude that for $N = 1$ reassignment cannot change the matching, and the small aggregate differences can only be due to different stochastic congestion realizations. Therefore, we can conclude that periodic target reassignment has essentially no practical improvement to nominal shortest-path routing in this experiment, and that the substantial advantage in Figure 4 is not an artifact of comparing online SALT assignment against a static assignment heuristic.

F.5 Compute resources

We ran all reported experiments on the Clariden supercomputer at the Swiss National Supercomputing Centre (CSCS). Each Clariden compute node contains four NVIDIA GH200 modules, each with a 72-core Grace CPU and 96 GB of HBM3 memory. An independent rank requested one GH200 and 72 CPU cores.

For the grid-world experiments, we used PyTorch [50] with CUDA [51] and assigned one independent rank to each module. Within a grid-world condition, the recorded training times summed across methods range from 3.9 to 4.9 hours, with a mean of 4.2 hours, and the 50 final rollouts per method were generated in the same jobs and required little additional time relative to training. The eight-condition scalability sweep used the same allocation and at most 96 additional active GPU-hours under the 12-hour job limit.

For the Manhattan experiments, we used JAX [52] with CUDA and the same one-GH200, 72-core allocation for each training seed. We ran four training seeds on each node. Training one policy for $5{,}000{,}000$ episodes required approximately 77.5 active GPU-hours.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction summarize the paper's scope and contributions: a separation principle for cooperative MARL with decoupled stochastic dynamics and terminal optimal transport costs, the SALT algorithmic framework, and empirical validation in grid worlds and a Manhattan road-network task. The claims are tied to the stated theoretical assumptions and to the experimental settings reported in Sections 3–5 and Appendices C–F.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 explicitly discusses limitations, including the need for decoupled dynamics and optimal transport terminal costs, the exclusion of interactions/collisions and non-transport objectives, the non-tightness of the stochastic upper bound, and the challenge of generalizing target-conditioned policies to high-dimensional or out-of-distribution targets.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The theoretical results are stated with the relevant problem assumptions in Sections 2–3. Formal proofs of the main theorem and corollaries are provided in Appendix C, with supporting measure-theoretic and noise-model assumptions developed in Appendices A and B.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper describes the SALT training/evaluation procedure, the baselines, the grid-world and Manhattan environments, training budgets, evaluation protocols, and hyperparameters in Section 5 and Appendix F. Code is available at `https://github.com/lucia-pezzetti/SALT`, as provided in Section 5.

   Guidelines:

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

4.4 We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

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

   [Yes]

   Justification: The code is released at the public repository linked in Section 5.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies the main training and evaluation details in Appendix F.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The paper reports results over repeated evaluation seeds and evaluation instances. Error bars for the grid-world experiments are provided in Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Compute time and type of workers are specified in Appendix F.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research uses simulated control environments and a public road-network source, does not involve human subjects, and does not release high-risk models or sensitive personal data.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

Justification: This work develops a generic algorithmic method for cooperative multi-agent reinforcement learning rather than a domain-specific deployed system. In the guidelines it is reported that algorithms to optimize neural networks do not need any societal impact justification.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not introduce or release a high-risk model. The experiments concern simulated routing/control policies and standard road-network data.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The Manhattan road-network experiment uses map data extracted from OpenStreetMap. We credit OpenStreetMap contributors in the main text, Appendix F, and bibliography.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We documented the code with instructions in README files and dedicated comments in the code itself.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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