ROBUST ACTOR-CRITIC LEARNING UNDER DISTRIBUTIONAL UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study finite-horizon Markov decision processes in which the transition dynamics are only known to lie within a Wasserstein ball around a reference model. Our goal is to learn policies that remain effective under such distributional uncertainty. A central difficulty is that the worst-case transition kernel depends on the policy itself, so the usual policy-gradient expression does not fully capture the sensitivity of the robust objective. Our main contribution is a characterization of one-sided directional derivatives of the robust value function with respect to the policy parameters, obtained via the Lagrangian dual of the inner robust problem. These derivatives replace the classical policy gradient and determine both the policy update and the auxiliary quantities that must be estimated. Building on this result, we develop a robust dynamic programming principle that applies jointly to the value function and its sensitivity, yielding backward recursions that a critic can learn to approximate. The resulting actor–critic algorithm uses three parametrized components, with separate networks for the policy, the robust value function, and its sensitivity. Under suitable regularity conditions, the latter is vector-valued and provides the policy gradient. On finite-state benchmarks the method recovers the exact robust dynamic programming solution, and on a linear–quadratic control task in continuous state space, it produces policies that remain reliable under shifted dynamics.

1 INTRODUCTION

In many reinforcement learning and control problems, system dynamics are uncertain or misspecified, so policies performing well under one model may fail under another. This motivates robust reinforcement learning, which seeks policies stable across model variations. Such challenges arise in inventory management and finance, where uncertainty and regime shifts are common. Robust approaches range from classical methods Bertsimas and Thiele (2006) to distributionally robust formulations using Wasserstein ambiguity sets Kim and Chung (2024). These ideas also lead to regularized portfolio optimization Blanchet et al. (2022) and adaptive control under uncertainty Bhudisaksang et al. (2026). In this framework, uncertainty is captured via sets of transition laws, yielding policies that are more robust to model misspecification.

Our approach builds on two main strands of literature: robust MDPs and distributionally robust optimization (DRO). In robust MDPs, Iyengar (2005); Nilim and El Ghaoui (2005) established tractable dynamic programming under rectangular ambiguity sets. Extensions to broader uncertainty models appear in Lim et al. (2013); Neufeld et al. (2023), while recent works develop policy gradient methods Wang et al. (2023) and robust Q-learning under Wasserstein ambiguity Neufeld and Sester (2024). Related actor–critic methods for risk-aware settings are proposed in Coache and Jaimungal (2026).

On the DRO side, convex formulations and tractable reformulations are introduced in Wiesemann et al. (2014), with Wasserstein-based guarantees in Mohajerin Esfahani and Kuhn (2018) and structural insights in Gao and Kleywegt (2023). The key duality result we use is from Blanchet and Murthy (2019), with links to regularization in Blanchet et al. (2019). Sensitivity analyses are developed in Bartl et al. (2021); Sauldubois and Touzi (2024); Compoint et al. (2025).

The closest related work is Wang and Zou (2022), which develops a policy gradient framework for robust MDPs but restricts attention to R-contamination ambiguity sets. In that setting, the robust

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

Bellman operator associated with contamination sets admits a closed-form expression. This makes the policy gradient derivation relatively direct. However, under more general ambiguity sets such as Wasserstein or Kullback–Leibler ambiguity sets, the worst-case transition kernel depends on the value function through an optimization problem, and differentiating through it would require regularity conditions on the optimal kernel that do not hold in general.

Table 1: Comparison with related robust reinforcement learning and Wasserstein control methods.

Work	Ambiguity set	Finite horizon	Actor–critic	Wasserstein dual	Sensitivity term
Wang and Zou (2022)	R-contamination	yes	policy gradient / actor–critic	no	closed form
Wang et al. (2023)	rectangular transition uncertainty	yes	policy gradient	no	robust PG inner loop
Kumar et al. (2023)	rectangular transition uncertainty	yes	policy gradient	no	robust gradient
Sun et al. (2024)	transition-kernel uncertainty	average-cost	policy mirror descent	no	robust sub-gradient
Yang et al. (2023)	distributional shift	contextual bandit	policy gradient	problem-specific	no Bellman sensitivity
Neufeld and Sester (2024)	Wasserstein transition ambiguity	discounted tabular	Q-learning	yes	no actor gradient
Yang (2021); Kim and Yang (2023)	Wasserstein LQ	finite-horizon control	no	Riccati / dual	special LQ structure
This paper	Wasserstein transition kernels	yes	yes	yes	one-sided envelope over worst-case laws

Our contributions. We develop a policy gradient framework for robust RL under distributional ambiguity around the reference transition kernels. We work with the dual formulation of Blanchet et al. (2019), which avoids differentiating the optimal kernel directly. Our contributions are as follows.

- **Robust dynamic programming.** We establish a robust DPP for finite-horizon MDPs under general set-valued transition kernels and randomized policies, and show that worst-case transition selectors exist (Theorem 2.5).
- **Policy gradient representation.** We prove that the robust value functions are continuous and admit directional derivatives in the policy parameter (Propositions 3.3 and 3.4), characterized by a primal envelope formula over the worst-case transition laws (Theorem 3.5 and Corollary 3.7). The transport-based dual expression is in general only a bound, exact when the worst-case transport is unique, in which case the derivative is a vector-valued gradient (Proposition B.3). In finite spaces the recursion is exact without further assumptions (Theorem 3.6).
- **Robust actor–critic algorithm.** We propose a tractable directional actor–critic algorithm that estimates the robust value function and its one-sided directional sensitivity in a given direction through parametric families, updated via a backward recursion based on the dual Bellman operator and an exact transport program over the worst-case laws, with the vector-gradient actor–critic as its specialization when the worst-case transport is unique. We validate it against exact robust dynamic programming on coin-toss and inventory-control problems and on sample-supported linear-quadratic and inverted-pendulum tasks, with gradient checks showing that optimal transport is needed to recover the robust gradient.
- **Exact finite-model ascent.** A near-active branch linear program selects the direction and a safeguarded step gives monotone ascent. Every accumulation point of the refinement iterates is directionally stationary, with an outer-iteration bound for a tolerance-stationarity criterion (Algorithm 1, Theorem 3.8, and Corollary C.4). These guarantees concern exact finite-model evaluation, not the sampled actor–critic.

2 Problem formulation

This section introduces the key definitions and assumptions. Standard notation ($\mathcal{P}(E)$, $\mathcal{C}(E;\mathbb{R})$, $\|\cdot\|_\infty$, the q-Wasserstein distance W_q, $\mathcal{T} := \{0,\dots,T-1\}$, $\bar{\mathcal{T}} := \mathcal{T}\cup\{T\}$, $\mathcal{T}_t := \{t,\dots,T-1\}$) is collected in Appendix A.1. Let $\mathcal{X}\subset\mathbb{R}^d$ and $A\subset\mathbb{R}^m$ be compact state and action spaces, endowed with their Borel σ-algebras. Let $T\in\mathbb{N}^\star$ be a finite horizon. We work on the canonical path space

$$\Omega := (\mathcal{X}\times A)^T\times\mathcal{X}, \quad \omega := (x_0,a_0,\dots,x_{T-1},a_{T-1},x_T)\in\Omega, \tag{1}$$

with the coordinate maps $X_t(\omega) := x_t$, $a_{t'}(\omega) = a_{t'}$ for $(t,t')\in\bar{\mathcal{T}}\times\mathcal{T}$ and the trajectory map $S(\omega) := \big(X_0(\omega),a_0(\omega),\dots,X_{T-1}(\omega),a_{T-1}(\omega),X_T(\omega)\big)$ for any $\omega\in\Omega$. At each time $t\in\mathcal{T}$, the set-valued or correspondence transition kernel

$$\mathcal{P}_t : \mathcal{X}\times A \rightrightarrows \mathcal{P}(\mathcal{X}), \quad (x,a)\mapsto\mathcal{P}_t(x,a)\subset\mathcal{P}(\mathcal{X}), \tag{2}$$

captures distributional uncertainty: the conditional law $X_{t+1}|(X_t=x,a_t=a)$ is only known to belong to $\mathcal{P}_t(x,a)$. We refer to Appendix A.3 for an introduction on the general theory of

correspondences and to Aliprantis and Border (2006) (Chapter 17) for an extensive introduction. We now make the following assumptions for the set-valued maps $(\mathcal{P}_t)_{t\in\mathcal{T}}$ which are essential in the sequel for measurable selection argument in the derivation of the dynamic programming principle.

Assumption 2.1. (Assumptions on the set of measures). Fix $p \in \{0, 1\}$.

(1) For any $t \in \mathcal{T}$, the correspondence $\mathcal{X} \times A \ni (x, a) \mapsto \mathcal{P}_t(x, a) \subset \mathcal{P}(\mathcal{X})$ is weakly measurable, non-empty, compact valued with closed graph and continuous in the sense of Appendix A.3;

(2) There exists $C \geq 0$ such that $\int_{\mathcal{X}}(1+\|y\|^p)\mathbb{P}(\mathrm{d}y) \leq C(1+\|x\|^p+\|a\|^p)$ for any $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$ and $\mathbb{P} \in \mathcal{P}_t(x, a)$.

In this work, we use randomized policies $\pi = (\pi_t)_{t\in\mathcal{T}}$, i.e., we consider for any $t \in \mathcal{T}$, $\pi_t \in L^0(\mathcal{X}; \mathcal{P}(A))$, which represents the conditional law of the random variable $a_t|X_t$. This is the standard setting of reinforcement learning control problems in comparison with deterministic policies where π_t would belong to $L^0(\mathcal{X}; A)$. We now make the following assumptions on the policy $\pi = (\pi_t)_{t\in\mathcal{T}}$ which will be used in the sequel for deriving some regularity results.

Assumption 2.2. (Admissible set of randomized policies). Fix the same p as in Assumption 2.1. A policy $\pi = (\pi_t)_{t\in\mathcal{T}}$ is admissible if for any $t \in \mathcal{T}$:

(1) $\mathcal{X} \ni x \mapsto \pi_t(x, \mathrm{d}a) \in \mathcal{P}(A)$ is Borel measurable;

(2) $x \mapsto \pi_t(x, \mathrm{d}a)$ is weakly continuous, i.e. $\int_A g(a)\pi_t(x_n, \mathrm{d}a) \to \int_A g(a)\pi_t(x, \mathrm{d}a)$ for all $g \in \mathcal{C}_b(A; \mathbb{R})$ whenever $x_n \to x$;

(3) $\int_A \|a\|^p \pi_t(x, \mathrm{d}a) \leq C\big(1 + \|x\|^p\big)$ for some $C \geq 0$ and all $x \in \mathcal{X}$. We denote by Π the set of admissible randomized policies.

We now consider the set of admissible measures on the path space Ω constructed from the measurable correspondences $(\mathcal{P}_t)_{t\in\mathcal{T}}$ and the randomized policies $(\pi_t)_{t\in\mathcal{T}}$.

Definition 2.3. (Admissible set of measures on the path space). Let Assumptions 2.1 and 2.2 hold. For $t \in \mathcal{T}$, the set of admissible transition selectors is $\mathfrak{S}_t := \{\mathbb{P}_t : \mathcal{X} \times A \to \mathcal{P}(\mathcal{X})$ Borel measurable with $\mathbb{P}_t(x, a) \in \mathcal{P}_t(x, a),\ \forall (x, a) \in \mathcal{X} \times A\}$. For $x \in \mathcal{X}$, $\pi \in \Pi$ and $t \in \bar{\mathcal{T}}$, the set of admissible path laws from (t, x) is

$$\mathcal{B}_{x,\pi|t} := \Big\{\mathbb{P} \in \mathcal{P}((\mathcal{X} \times A)^{T-t} \times \mathcal{X}) : \mathbb{P}(\mathrm{d}\omega) = \delta_x(\mathrm{d}x_t) \textstyle\prod_{l=t}^{T-1} \pi_l(x_l, \mathrm{d}a_l)\, \mathbb{P}_l(x_l, a_l, \mathrm{d}x_{l+1}),\ \mathbb{P}_l \in \mathfrak{S}_l\Big\},$$

with the convention that an empty product equals 1, and we write $\mathcal{B}_{x,\pi} := \mathcal{B}_{x,\pi|0}$. This type of ambiguity set is called rectangular in the literature and is convenient for deriving a robust dynamic programming principle (see Remark A.1).

We now introduce the maps $f : \mathcal{T} \times \mathcal{X} \times A \times \mathcal{X} \ni (t, x, a, x') \mapsto f(t, x, a, x') \in \mathbb{R}$ and $g : \mathcal{X} \ni x \mapsto g(x) \in \mathbb{R}$ as the cumulative and terminal rewards on which we make the following assumptions.

Assumption 2.4. (Assumptions on the reward functions). Fix the same p as in Assumption 2.1.

(1) For each $t \in \mathcal{T}$, the map $\mathcal{X} \times A \times \mathcal{X} \ni (x, a, x') \mapsto f(t, x, a, x') \in \mathbb{R}$ is continuous and the map $\mathcal{X} \ni x \mapsto g(x) \in \mathbb{R}$ is also continuous;

(2) There exists a constant $C_f > 0$ such that for every $(t, x, a, x') \in \mathcal{T} \times \mathcal{X} \times A \times \mathcal{X}$, we have $|f(t, x, a, x')| \leq C_f\big(1 + \|x\|^p + \|a\|^p + \|x'\|^p\big)$;

(3) There exists a constant $C_g > 0$ such that for every $x \in \mathcal{X}$, we have $|g(x)| \leq C_g\big(1 + \|x\|^p\big)$.

We define the total reward by $R : \Omega \ni \omega = (x_0, a_0, \ldots, x_{T-1}, a_{T-1}, x_T) \mapsto \sum_{t=0}^{T-1} f(t, x_t, a_t, x_{t+1}) + g(x_T)$. The objective is to maximize $\mathbb{E}^{\mathbb{P}}[R(S)]$ under the worst-case measure of $\mathbb{P} \in \mathcal{B}_{x,\pi}$ and over the admissible randomized policies set Π. For any $t \in \bar{\mathcal{T}}$ and $\pi \in \Pi$, we hence define the value function V_t^π as

$$\mathcal{X} \ni x \mapsto V_t^\pi(x) := \inf_{\mathbb{P}\in\mathcal{B}_{x,\pi|t}} \mathbb{E}^{\mathbb{P}}\big[\sum_{l=t}^{T-1} f(l, X_l, a_l, X_{l+1}) + g(X_T)\big]. \tag{3}$$

The corresponding robust value function is then defined for any $t \in \bar{\mathcal{T}}$ by $V_t(x) := \sup_{\pi \in \Pi} V_t^\pi(x)$ and note that for $t = 0$, we have $V_0(x) = \sup_{\pi \in \Pi} \inf_{\mathbb{P} \in \mathcal{B}_{x,\pi}} \mathbb{E}^{\mathbb{P}}\big[R(S)\big]$.

We end this subsection by stating the robust dynamic programming principle for Markov decision processes with randomized policies.

Theorem 2.5. (Robust dynamic programming principle I).

Let Assumptions 2.1, 2.2 and 2.4 hold. Then, for any $\pi \in \Pi$ and for any $(t, x) \in \mathcal{T} \times \mathcal{X}$, we have

$$\begin{cases} V_T^\pi(x) & = g(x), \\ V_t^\pi(x) & = \mathbb{E}_{a_t \sim \pi_t(x)}\Big[\inf_{\mathbb{P}_t \in \mathcal{P}_t(x,a_t)} \mathbb{E}^{\mathbb{P}_t}[f(t,x,a_t,x_{t+1}) + V_{t+1}^\pi(x_{t+1})]\Big], \end{cases} \tag{4}$$

the inner infimum is attained by a Borel measurable selector $\mathbb{P}_t^{\star,\pi} \in \mathfrak{S}_t$, and for any $t \in \bar{\mathcal{T}}$ and any $\pi \in \Pi$, we have $V_t^\pi \in \mathcal{C}_p(\mathcal{X}; \mathbb{R})$.

The proof follows standard arguments (see Iyengar (2005) and Neufeld et al. (2023)) and is sketched in Appendix B.1. Each local problem in (4) is an optimization over an infinite set of probability measures, which motivates ambiguity sets with a tractable dual representation (Remark A.2).

3 Actor-critic algorithms

3.1 Wasserstein ambiguity sets

In this section, we specify the form of the ambiguity set $\mathcal{P}_t(x, a)$ for every $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$ to make the solvability of the optimization problem (4) more tractable. We work under a distributional ambiguity given by the Wasserstein distance W_q for a given $q \in \mathbb{N}^\star$. In this way, we will define for any $t \in \mathcal{T}$ a reference probability kernel $\mathbb{P}_t^0$ such that the mapping $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}_t^0(x, a, \cdot) \in \mathcal{P}(\mathcal{X})$ is measurable and satisfies $\int_{\mathcal{X}} \|x'\|^p \mathbb{P}_t^0(x, a, \mathrm{d}x') < \infty$, for all $(x, a) \in \mathcal{X} \times A$ and we specify $p = 0$ in our previous Assumption 2.1. In this case, we can define for any $\varepsilon > 0$ and for any $t \in \mathcal{T}$ the following set-valued map

$$\mathcal{X} \times A \ni (x, a) \mapsto \mathcal{B}^{\varepsilon,q}\big(\mathbb{P}_t^0(x, a)\big) := \big\{\mathbb{P} \in \mathcal{P}(\mathcal{X}) : W_q(\mathbb{P}, \mathbb{P}_t^0(x, a)) \leq \varepsilon\big\}. \tag{5}$$

Now, as it is standard, we will parametrize our admissible set of randomized policies Π. Namely, we will consider that our policy $\pi = (\pi_t)_{t \in \mathcal{T}}$ admit densities with respect to some reference measure ν on $\mathcal{P}(A)$ and are parametrized by some $\theta \in \Theta$ where Θ denotes a given compact set (e.g. the parameter space of a neural network). With a slight abuse of notation, we shall identify the density of the measure $(\pi_t^\theta(x, \mathrm{d}a))_{t \in \mathcal{T}, \theta \in \Theta}$ with its density $(\pi_t^\theta(x, a))_{t \in \mathcal{T}, \theta \in \Theta}$. We now make some assumptions on the reference probability kernels $(\mathbb{P}_t^0)_{t \in \mathcal{T}}$ and the family $(\pi_t^\theta)_{t \in \mathcal{T}}$.

Assumption 3.1. (Assumptions on the reference measure). We assume that the following holds:

(1) For each $t \in \mathcal{T}$, the map $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}_t^0(x, a) \in \mathcal{P}(\mathcal{X})$ is continuous for the weak topology, i.e, $\int_{\mathcal{X}} \varphi(x') \mathbb{P}_t^0(x_n, a_n, \mathrm{d}x') \to \int_{\mathcal{X}} \varphi(x')\, \mathbb{P}_t^0(x, a, \mathrm{d}x'), \quad \forall \varphi \in \mathcal{C}_b(\mathcal{X}; \mathbb{R})$ and for any sequence $(x_n, a_n) \underset{n \to \infty}{\to} (x, a)$.

(2) For each $t \in \mathcal{T}$, the map $\Theta \times \mathcal{X} \ni (\theta, x) \mapsto \pi_t^\theta(x, \mathrm{d}a) \in \mathcal{P}(A)$ is weakly continuous. Moreover, for each $\varphi \in \mathcal{C}(\mathcal{X} \times A; \mathbb{R})$, the map $(\theta, x) \mapsto \int_A \varphi(x, a) \pi_t^\theta(x, \mathrm{d}a)$ is continuously differentiable in θ on a neighborhood of Θ, with gradient $\nabla_\theta \int_A \varphi(x, a) \pi_t^\theta(x, \mathrm{d}a) = \int_A \varphi(x, a) \nabla_\theta \log \pi_t^\theta(x, a) \pi_t^\theta(x, \mathrm{d}a)$ jointly continuous in (θ, x), and for a constant $C_\pi \geq 0$,

$$\sup_{\theta \in \Theta} \sup_{(t,x,a) \in \mathcal{T} \times \mathcal{X} \times A} \|\nabla_\theta \log \pi_t^\theta(x, a)\| \leq C_\pi < \infty.$$

We recall that the ambiguity set (5) satisfies the requirements of Assumption 2.1 (see Proposition 3.1 in Neufeld et al. (2023)). In particular, since $\mathcal{X}$ is compact, W_q metrizes weak convergence on $\mathcal{P}(\mathcal{X})$ and each ball $\mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x, a))$ is a nonempty weakly compact subset of $\mathcal{P}(\mathcal{X})$. To tackle the ambiguity sets defined by (5), we will rely on some duality results borrowed from Theorem 1 in Blanchet and Murthy (2019), which we summarize in the Proposition below.

Proposition 3.2. *Let $c : \mathcal{X} \times \mathcal{X} \ni (x, y) \mapsto c(x, y) = \|x - y\|^q$ and $\lambda \geq 0$. We define the operator $\mathcal{F}^\lambda$ as $\mathcal{F}^\lambda : L^0(\mathcal{X}; \mathbb{R}) \ni V \mapsto \big(x \mapsto \mathcal{F}^\lambda(V)(x) := \sup_{y \in \mathcal{X}} \{V(y) - \lambda c(x, y)\}\big)$. Then, for every $V \in \mathcal{C}(\mathcal{X}; \mathbb{R})$ and $\mathbb{P}^0 \in \mathcal{P}(\mathcal{X})$,*

$$\min_{\mathbb{P} \in \mathcal{B}^{\varepsilon,q}(\mathbb{P}^0)} \mathbb{E}^{\mathbb{P}}\big[V(X)\big] = \sup_{\lambda \geq 0} \mathbb{E}^{\mathbb{P}^0}\big[- \mathcal{F}^\lambda(-V)(X) - \varepsilon^q \lambda\big],$$

where the minimum on the left is attained. Moreover, if $\mathbb{P}^\star$ attains the minimum, $\pi^\star \in \Pi(\mathbb{P}^0, \mathbb{P}^\star)$ is an optimal coupling for W_q, and $\lambda^\star$ attains the supremum, then $\lambda^\star\big(\varepsilon^q - \int c, \mathrm{d}\pi^\star\big) = 0$ and $y \in \arg\min_{z \in \mathcal{X}} \{V(z) + \lambda^\star c(x', z)\}$ for $\pi^\star$-a.e. (x', y).

The attainment follows from the weak compactness of the ball, and the complementary slackness from $\mathbb{E}^{\mathbb{P}^\star}[V] \geq \int (V(y) + \lambda^\star c(x', y))\pi^\star(\mathrm{d}x', \mathrm{d}y) - \lambda^\star \varepsilon^q \geq \mathbb{E}^{\mathbb{P}^0}[-\mathcal{F}^{\lambda^\star}(-V)] - \lambda^\star \varepsilon^q$, whose ends coincide by strong duality. Proposition 3.2 reduces an infinite-dimensional optimization over probability measures to a scalar one, and the dynamic programming principle can be rewritten as

$$\begin{cases} V_T^{\pi^\theta}(x) & = g(x), \\ V_t^{\pi^\theta}(x) & = \mathbb{E}_{a_t \sim \pi_t^\theta(x)}\Big[\sup_{\lambda \geq 0}\big(\mathbb{E}^{\mathbb{P}_t^0(x,a_t)}\big[- \mathcal{F}^\lambda(-f(t, x, a_t, \cdot) - V_{t+1}^{\pi^\theta})(x_{t+1})\big] - \varepsilon^q \lambda\big)\Big] \end{cases} \tag{6}$$

3.2 Policy gradient representation

We now establish the regularity needed to differentiate the robust criterion with respect to $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$. In the sequel, we will write indifferently $V_\cdot^{\pi^\theta}$ or $V_\cdot^\theta$. We write, for $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$,

$$\begin{aligned} G_t^\theta(x, a) :&= \min_{\mathbb{P} \in \mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x,a))} \int_{\mathcal{X}} (f(t, x, a, y) + V_{t+1}^\theta(y))\mathbb{P}(\mathrm{d}y) \\ &= \sup_{\lambda \geq 0} \big(\mathbb{E}_{X \sim \mathbb{P}_t^0(x,a,\cdot)}[-\mathcal{F}^\lambda(-f(t, x, a, \cdot) - V_{t+1}^\theta)(X)] - \varepsilon^q \lambda\big), \end{aligned}$$

the two expressions being equal by Proposition 3.2, so that $V_t^\theta(x) = \int_A G_t^\theta(x, a)\pi_t^\theta(x, \mathrm{d}a)$ by (6). We denote by

$$\mathcal{P}_t^\star(x, a, \theta) := \operatorname*{argmin}_{\mathbb{P} \in \mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x,a))} \int_{\mathcal{X}} (f(t, x, a, y) + V_{t+1}^\theta(y))\mathbb{P}(\mathrm{d}y)$$

the set of worst-case transition laws at $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$.

Proposition 3.3. *Under Assumption 3.1, for every $\theta \in \Theta$ and every $t \in \bar{\mathcal{T}}$, $V_t^\theta \in \mathcal{C}(\mathcal{X}; \mathbb{R})$ and $G_t^\theta \in \mathcal{C}(\mathcal{X} \times A; \mathbb{R})$. Moreover, for every (t, x, a), $\mathcal{P}_t^\star(x, a, \theta)$ is nonempty and weakly compact, the correspondence $(x, a) \rightrightarrows \mathcal{P}_t^\star(x, a, \theta)$ is upper hemicontinuous, $\Lambda_t^{\star,\theta}(x, a)$ is a nonempty compact subset of $[0, \Lambda]$ with $\Lambda := 2\sup_{t,\theta,x,a}\|f(t, x, a, \cdot) + V_{t+1}^\theta\|_\infty/\varepsilon^q$, and $Y_{t,\theta,\lambda}^\star(x, a; x')$ is nonempty and compact.*

The proof of Proposition 3.3 is presented in Appendix B.2. We now quantify how the robust value function varies with the policy parameter. Because the worst-case law is selected by a minimization, we cannot expect classical differentiability and we work with one-sided directional derivatives. For a feasible direction $r \in \mathbb{R}^{d_\theta}$ at θ (i.e. $\theta + hr \in \Theta$ for all small $h > 0$), we write

$$D_\theta^+ V_t^\theta(x)[r] := \lim_{h \downarrow 0} \frac{V_t^{\theta+hr}(x) - V_t^\theta(x)}{h}, \quad D_\theta^- V_t^\theta(x)[r] := \lim_{h \uparrow 0} \frac{V_t^{\theta+hr}(x) - V_t^\theta(x)}{h},$$

whenever the limits exist, and similarly for $G_t^\theta(x, a)$. The results below are stated for the right derivative; the left derivative follows by replacing r with $-r$, which exchanges minima and maxima.

Proposition 3.4. *Let Assumption 3.1 hold and suppose that, for every $t \in \mathcal{T}$, the correspondence $\mathcal{X} \times A \ni (x, a) \rightrightarrows \mathcal{P}_t^\star(x, a, \theta)$ is lower hemicontinuous at $\theta \in \Theta$. Then for every $t \in \bar{\mathcal{T}}$ and every feasible direction $r \in \mathbb{R}^{d_\theta}$ at θ, the right directional derivatives $D_\theta^+ V_t^\theta(\cdot)[r] \in \mathcal{C}(\mathcal{X}; \mathbb{R})$ and $D_\theta^+ G_t^\theta(\cdot, \cdot)[r] \in \mathcal{C}(\mathcal{X} \times A; \mathbb{R})$ exist, the convergence of the difference quotients is uniform,*

$$\left\| \frac{V_t^{\theta+hr} - V_t^\theta}{h} - D_\theta^+ V_t^\theta(\cdot)[r] \right\|_\infty + \left\| \frac{G_t^{\theta+hr} - G_t^\theta}{h} - D_\theta^+ G_t^\theta(\cdot, \cdot)[r] \right\|_\infty \xrightarrow[h \downarrow 0]{} 0, \tag{7}$$

and the difference quotients are uniformly bounded for h small enough. The same holds for the left derivatives.

The proof of Proposition 3.4 is presented in Appendix B.3. Next, we characterize the one-sided derivatives of the one-step worst-case value G_t^θ as an envelope over the worst-case laws.

Theorem 3.5. (Policy gradient). *Under the assumptions of Proposition 3.4, for every $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$ and feasible direction $r \in \mathbb{R}^{d_\theta}$ at θ,*

$$D_\theta^\pm G_t^\theta(x,a)[r] = \min_{\mathbb{P} \in \mathcal{P}_t^\star(x,a,\theta)} / \max \int_{\mathcal{X}} D_\theta^\pm V_{t+1}^\theta(y)[r]\, \mathbb{P}(\mathrm{d}y), \tag{8}$$

with the minimum for D_θ^+ and the maximum for D_θ^-. Moreover, the transport-based dual expressions satisfy the one-sided bounds

$$D_\theta^+ G_t^\theta(x,a)[r] \geq \mathbb{E}_{X \sim \mathbb{P}_t^0(x,a,\cdot)}\Big[\inf_{y \in Y_{t,\theta,\lambda^\star}^\star(x,a;X)} D_\theta^+ V_{t+1}^\theta(y)[r] \Big], \tag{9}$$

$$D_\theta^- G_t^\theta(x,a)[r] \leq \mathbb{E}_{X \sim \mathbb{P}_t^0(x,a,\cdot)}\Big[\sup_{y \in Y_{t,\theta,\lambda^\star}^\star(x,a;X)} D_\theta^- V_{t+1}^\theta(y)[r] \Big], \tag{10}$$

for every dual optimizer $\lambda^\star \in \Lambda_t^{\star,\theta}(x,a)$, where

$$\begin{cases} \Lambda_t^{\star,\theta}(x,a) & := \operatorname{argmax}_{\lambda \geq 0} \Big\{ \mathbb{E}_{X \sim \mathbb{P}_t^0(x,a,\cdot)} \big[-\mathcal{F}^\lambda(-f(t,x,a,\cdot) - V_{t+1}^\theta(\cdot))(X) \big] - \varepsilon^q \lambda \Big\}, \\ Y_{t,\theta,\lambda}^\star(x,a;x') & := \operatorname{argmin}_{y \in \mathcal{X}} \big\{ f(t,x,a,y) + V_{t+1}^\theta(y) + \lambda\, c(x',y) \big\}, \qquad x' \in \mathcal{X}. \end{cases} \tag{11}$$

The bounds in (9) *may be strict; see Remark B.1.*

Formula (8) is a one-sided Danskin theorem applied to the primal form of G_t^θ, differentiated at the frozen worst-case laws.

Theorem 3.6. (Finite state–action spaces). *Suppose that $\mathcal{X}$ and A are finite, $\varepsilon > 0$, and that Assumption 3.1(2) holds. Fix $\theta \in \Theta$ and a feasible direction r. Then for every $t \in \bar{\mathcal{T}}$ the one-sided derivatives $D_\theta^\pm V_t^\theta(x)[r]$ exist for every $x \in \mathcal{X}$, and the recursion of Corollary 3.7 holds with*

$$D_\theta^+ G_t^\theta(x,a)[r] = \min_{\mathbb{P}^\star \in \mathcal{P}_t^\star(x,a,\theta)} \sum_{y \in \mathcal{X}} \mathbb{P}^\star(\{y\}) D_\theta^+ V_{t+1}^\theta(y)[r], \tag{12}$$

$$D_\theta^- G_t^\theta(x,a)[r] = \max_{\mathbb{P}^\star \in \mathcal{P}_t^\star(x,a,\theta)} \sum_{y \in \mathcal{X}} \mathbb{P}^\star(\{y\}) D_\theta^- V_{t+1}^\theta(y)[r]. \tag{13}$$

Moreover, with $H_\theta(y) := f(t,x,a,y) + V_{t+1}^\theta(y)$, $\mathcal{P}_t^\star(x,a,\theta)$ is the polytope of second marginals of the optimal face of the transport linear program

$$\min_{\gamma \geq 0} \sum_{x',y \in \mathcal{X}} \gamma(x',y) H_\theta(y) \text{ s.t. } \sum_{y \in \mathcal{X}} \gamma(x',y) = \mathbb{P}_t^0(x,a,x'), \ \ \forall x' \in \mathcal{X}, \ \sum_{x',y \in \mathcal{X}} \gamma(x',y) c(x',y) \leq \varepsilon^q, \tag{14}$$

so that the minima and maxima in (12) *are linear programs over that face.*

The proof is given in Appendix B.5. In finite state–action spaces, the one-sided derivatives exist without additional topological assumptions, and the worst-case laws form the optimal face of a finite transport linear program (Theorem 3.6, Appendix B.5), yielding the exact directional recursion used in Section 3.3.

Theorems 2.5 and 3.5 give the following recursion, which is the key argument behind Algorithm 2.

Corollary 3.7. (Robust dynamic programming principle II). *Under the assumptions of Theorem 3.5, for all $(t,x) \in \mathcal{T} \times \mathcal{X}$, every feasible direction $r \in \mathbb{R}^{d_\theta}$ and either one-sided derivative,*

$$\begin{cases} D_\theta V_T^\theta(x)[r] = 0, \\ D_\theta V_t^\theta(x)[r] = \mathbb{E}_{a \sim \pi_t^\theta(x,\cdot)} \left[G_t^\theta(x,a) \left\langle \nabla_\theta \log \pi_t^\theta(x,a), r \right\rangle + D_\theta G_t^\theta(x,a)[r] \right]. \end{cases} \tag{15}$$

Why the classical policy gradient is incomplete. For every direction r, Corollary 3.7 gives

$$D_\theta V_t^\theta(x)[r] = \underbrace{\mathbb{E}_{a \sim \pi_t^\theta} \left[G_t^\theta(x,a) \left\langle \nabla_\theta \log \pi_t^\theta(x,a), r \right\rangle \right]}_{\text{score-function term}} + \underbrace{\mathbb{E}_{a \sim \pi_t^\theta} \left[D_\theta G_t^\theta(x,a)[r] \right]}_{\text{robust sensitivity term}}. \tag{16}$$

The second term is the usual continuation sensitivity, but in the robust setting it is evaluated under the worst-case transition law, so the classical policy-gradient recursion does not directly apply. When the transport argmin is unique for some dual optimizer and $\mathbb{P}_t^0(x,a,\cdot)$-a.e. x', the worst-case law is unique and the one-sided derivatives coincide. If this holds for all $(s,x,a) \in \mathcal{T} \times \mathcal{X} \times A$, then $\nabla_\theta V_s^\theta$ satisfies (52), with the continuation gradient evaluated at the worst-case transported state rather than the nominal next state. Appendix E evaluates the role of this sensitivity term.

3.3 Robust actor-critic algorithm

Corollary 3.7 evaluates the sensitivity in a given direction. We now specify how to choose the direction and the step. In this subsection only, $\mathcal{X}$ and A are finite, $\varepsilon > 0$, $\Theta = [-B, B]^{d_\theta}$ with $B > 0$, and the policies are strictly positive and C^2 on a neighborhood of Θ. Policy evaluation, branch gradients and inner optimization are exact. Write $J(\theta) = \mathbb{E}_{X_0 \sim \mu_0}[V_0^\theta(X_0)]$.

Let Σ be the finite set of row-wise vertex selections of the Wasserstein balls, and let J_σ be the return under the fixed transition kernels $\sigma \in \Sigma$. For $\zeta \geq 0$, retain the locally near-optimal selections

$$S_\zeta(\theta) = \left\{ \sigma \in \Sigma : \ \left\langle \sigma(t,x,a), f(t,x,a,\cdot) + V_{t+1}^\theta \right\rangle - G_t^\theta(x,a) \leq \zeta, \ \forall (t,x,a) \right\}. \tag{17}$$

The pairing denotes a finite-state expectation. Define the feasible direction set

$$T_\zeta(\theta) = \left\{ r : \ \|r\|_\infty \leq 1, \quad \begin{array}{ll} r_i \leq 0 & \text{if } \theta_i \geq B - \zeta, \\ r_i \geq 0 & \text{if } \theta_i \leq -B + \zeta \end{array} \right\}. \tag{18}$$

The direction is selected by the finite linear program

$$\Delta_\zeta(\theta) = \max_{r \in T_\zeta(\theta), \, \delta \in \mathbb{R}} \delta \quad \text{s.t.} \quad \delta \leq \langle \nabla_\theta J_\sigma(\theta), r \rangle, \qquad \sigma \in S_\zeta(\theta). \tag{19}$$

Here each branch gradient uses that branch's own value function, not the robust value when the branch is only near-optimal. Appendix C gives the fixed-kernel gradient formula and the complete analysis.

Algorithm 1 Exact robust policy ascent

Require: $\theta_0 \in \Theta$, $\zeta_0 > 0$, $\nu \in (0,1)$; positive global bounds C_1, C_2 from Theorem C.1(iv)

for $k = 0, 1, 2, \ldots$ **do**
  Evaluate $V_t^{\theta_k}, G_t^{\theta_k}$ by exact robust policy evaluation; construct $S_{\zeta_k}(\theta_k)$ using the near-optimal vertices of (14).
  Evaluate $\nabla_\theta J_\sigma(\theta_k)$ for every $\sigma \in S_{\zeta_k}(\theta_k)$; solve (19) to obtain (Δ_k, r_k).
  if $\Delta_k \leq \zeta_k$ **then**
    $\theta_{k+1} \leftarrow \theta_k$, $\zeta_{k+1} \leftarrow \nu \zeta_k$. ▷ refine the tolerance
  else
    $h_k \leftarrow \min\{\zeta_k, \, \zeta_k / C_1, \, \Delta_k / (2C_2)\}$.
    $\theta_{k+1} \leftarrow \theta_k + h_k r_k$, $\zeta_{k+1} \leftarrow \zeta_k$. ▷ feasible ascent
  end if
end for

Theorem 3.8. (Exact finite-model ascent). *Let $(\theta_k, \zeta_k)_{k \geq 0}$ be generated by Algorithm 1 and let $\{k \geq 0 : \ \Delta_k \leq \zeta_k\}$ be the set of refinement iterations. $(J(\theta_k))_{k \geq 0}$ is nondecreasing and convergent, refinements are infinitely many, $\zeta_k \to 0$ as $k \to \infty$, and every accumulation point $\bar{\theta}$ of the refinement iterates is directionally stationary, i.e, $D_\theta^+ J(\bar{\theta})[r] \leq 0$ for every $r \in T_0(\bar{\theta})$.*

Convergence and scope. Corollary C.4 gives $O(\zeta^{-2})$ outer iterations to the tolerance-stationarity criterion of Definition C.3. It is not a global-optimality or sampling-complexity guarantee. The branch set can be exponentially large, so this count is not a polynomial running-time bound. The C_1, C_2 used in the step must be valid global bounds, not local estimates.

Algorithm 2 in Appendix D retains the sampled value/sensitivity-critic implementation and the vector specialization under Proposition B.3. Its approximation errors are not covered by the exact ascent theorem. Finite disturbance support makes its local transport problem finite, but does not by itself extend that theorem to continuous-state learning.

4 NUMERICAL EXPERIMENTS

In this section, we illustrate the validity of our theoretical results and the performance of Algorithm 2. A supply-chain example, exact-DP recovery statistics, ablations, gradient checks, a robust Q-learning baseline, uniqueness and sample-sensitivity studies, shifted-dynamics evaluations and an inventory problem are in Appendix E; overhead, scaling and training details in Appendix F. Our code is available at the project code repository.

4.1 A COIN TOSS EXAMPLE

We use the coin-toss benchmark. The state is $\mathcal{X} = \{0, \dots, n\}$, where X_t is the number of heads in a block of n Bernoulli trials, the action set is $A = \{-1, 0, 1\}$, and $T = 10$. Actions $a = -1, 0, 1$ correspond respectively to betting that the next block is lower, abstaining, and betting that it is higher. Under the reference model $X_{t+1} \sim \mathbb{P}_t^0(x, a, \cdot) = \mathrm{Binomial}(n, p_0)$ for all (t, x, a), where $p_0 \in (0, 1)$ is the reference success probability. The reward is $f(t, x, a, x') = a\mathbb{1}_{\{x'>x\}} - a\mathbb{1}_{\{x'<x\}} - |a|\mathbb{1}_{\{x'=x\}}$ and $g = 0$, so correct bets earn $+1$, wrong bets earn -1, and ties penalize nonzero bets. We use cost $c(x, y) = |x - y|$, Wasserstein ambiguity around $\mathbb{P}_t^0$, and a tabular softmax policy $\pi_t^\theta(x, a) = \exp(\theta_{t,x,a}) / \sum_{a' \in A} \exp(\theta_{t,x,a'})$ with parameters $\theta = (\theta_{t,x,a})_{t,x,a} \in \mathbb{R}^{|\mathcal{T}| \times |\mathcal{X}| \times |A|}$. As ε increases, the betting region shrinks and the policy abstains on more states. The learned policies coincide with exact robust dynamic programming (DP) on every state–time cell and with the robust Q-iteration of Neufeld and Sester (2024) (Appendix E.9). Exact agreement with robust dynamic programming depends on the inner Wasserstein solver. Appendix E.3 compares the grid-based and exact versions and reports any remaining mismatch, while Appendix E.9 compares against robust Q-iteration.

To assess robustness to misspecification, Figure 1 compares the cumulative profits obtained by the different policies under perturbed coin biases. For each policy, we simulate 10^5 rounds while varying the true success probability p_{true} around the reference value p_0.

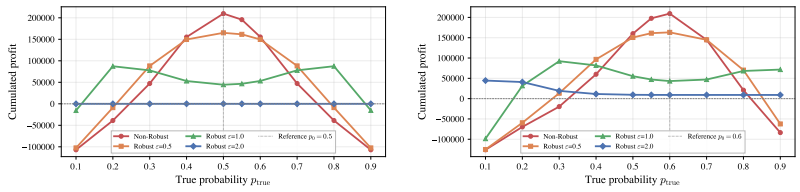


Figure 1: Cumulated profit under model misspecification for $p_0 = 0.5$ (left) and $p_0 = 0.6$ (right).

As ε increases, the learned policy becomes more conservative: profits near the reference model p_0 decrease, but performance varies less across misspecified values of p_{true}.

4.2 A LINEAR-QUADRATIC CONTROL PROBLEM

We consider a finite-horizon linear-quadratic control problem under distributional uncertainty in the noise driving the state dynamics. This continuous-state benchmark lies beyond the compact state-action and bounded-score assumptions of the theory, and is used to assess whether the proposed method can still perform well in a standard control setting. The reference state dynamics are $X_{t+1} = AX_t + Ba_t + \Xi w_t$, $t \in \mathcal{T}$, $X_0 = x$, where $A \in \mathbb{R}^{d \times d}$, $B \in \mathbb{R}^{d \times m}$, $\Xi \in \mathbb{R}^{d \times k}$. This falls in our setting by considering the reference kernel $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}_t^0(x, a, \cdot) = m(x, a, \cdot)\sharp(\frac{1}{N}\sum_{i=1}^N \delta_{\hat{w}^{(i)}})$ constructed from samples of the noise $(\hat{w}^{(i)})_{1 \le i \le N}$ where we set $\mathcal{X} \times A \times \mathbb{R}^k \ni (x, a, w) \mapsto m(x, a, w) = Ax + Ba + \Xi w$ and where we take the Wasserstein ambiguity set around the reference transition kernel $(\mathbb{P}_t^0)_{t \in \mathcal{T}}$ where we restrict attention to measures constructed from the samples $(\hat{w}^{(i)})_{1 \le i \le n}$, i.e., of the forms $(m(x, a, \cdot)\sharp\gamma)$ where $\gamma \in \mathcal{P}(\{(\hat{w}^{(1)}, \dots, \hat{w}^{(n)})\})$. The running and terminal costs are $f(t, x, a, x') := x^\top Q x + a^\top R a$ and $g(x) := x^\top P x$ with $Q \in \mathbb{S}_+^d(\mathbb{R})$, $R \in \mathbb{S}_{>+}^m(\mathbb{R})$ and $P \in \mathbb{S}_+^d(\mathbb{R})$ where $\mathbb{S}_+^d(\mathbb{R})$ and $\mathbb{S}_{>+}^m(\mathbb{R})$ denote respectively the set of $\mathbb{R}$-valued positive semidefinite and positive definite symmetric matrices. This subsection uses the minimax cost convention $\inf_\pi \sup_{\mathbb{P}} \mathbb{E}[\sum_t f + g]$. To express the same problem under the reward-maximization convention used earlier, replace the stage and terminal costs by their negatives.

For reference, we compare against the Wasserstein penalty framework of Yang (2021) and Kim and Yang (2023), in which the adversary pays $\lambda W_q(\gamma_s, \nu)^q$ per period instead of facing a budget and which admits a closed-form solution, quadratic in the state with a linear optimal control (Proposition 4.1 in Kim and Yang (2023)). Motivated by this structure, we use the scalar case $d = m = 1$ as a visualization benchmark. Since the state and action spaces are continuous, we use a Gaussian policy

$$\pi_t^\theta(x, \cdot) \sim \mathcal{N}\big(\mu_\theta(t, x), \sigma_\theta^2 I_m\big), \quad (t, x) \in \mathcal{T} \times \mathbb{R}^d.$$

The mean μ_θ is a two-layer feedforward network with ReLU activation, and $\log \sigma_\theta$ is learned jointly with the network weights. We report the learned mean feedback $x \mapsto \mu_{\theta^\star}(0, x)$ alongside the corresponding value function returned by Algorithm 2.

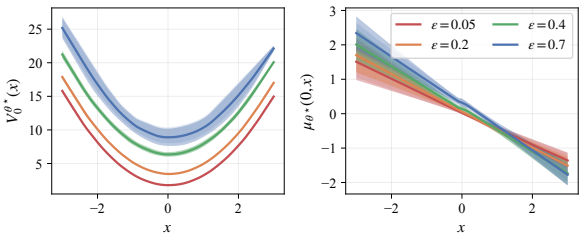


Figure 2: Value function (left) and mean of the optimal policy (right) for different ε.

The learned value function is approximately quadratic and the learned mean policy is nearly linear, as expected in the LQ setting. The learned variance also collapses during training, indicating convergence toward an almost deterministic feedback policy, $\sigma_{\theta^\star} \approx 0$.

Shifted dynamics and further experiments. Under distribution shifts in the noise law, robust policies improve performance when the shift is represented by the ambiguity set, while gains can disappear under shifts outside its support (Appendix E.7). On an inverted-pendulum task with sampled torque disturbances, robust training also outperforms the non-robust actor–critic and domain randomisation across the tested shifts. Additional sensitivity and inventory experiments are reported in Appendices E.5 and E.8.

5 Conclusion and Future Work

We propose a general framework for robust actor–critic learning under distributional uncertainty, based on a robust dynamic programming principle for both the value function and its sensitivity. Our main contribution is a characterization of the one-sided policy sensitivities as an envelope over the worst-case transition laws. When the worst-case transport is unique, the one-sided sensitivities coincide and yield a vector-valued policy gradient. At ties, the exact directional sensitivity instead requires optimization over the optimal coupling face. For finite models, Algorithm 1 supplements this sensitivity representation with a complete direction-selection rule and the exact ascent guarantee of Theorem 3.8. The experiments show that ties occur frequently at the exact dual optimum, that the nominal-continuation target can reach the same tabular fixed points while failing to represent the gradient of the robust objective, that DP recovery depends on the solver and the residual cases reported in Appendix E.3, and that a sample-supported adversary protects against reweightings of the sampled disturbance atoms, but not against shifts to a wider support.

Several limitations remain. The compact-space theory requires lower hemicontinuity of the worst-case-law correspondence, which is automatic in finite spaces but remains unverified for general continuous parametrizations. The convergence guarantee concerns the exact finite-model scheme; sampling and function approximation are not covered, including with sample-supported ambiguity. The sampled actor update is heuristic and incurs higher per-iteration computational cost than the non-robust baseline. Extending the approach beyond ambiguity classes with tractable dual formulations remains an open problem.

AI USE STATEMENT

Generative AI tools were used for language editing and for code assistance and debugging. They were not used to formulate mathematical claims, develop proofs, design the methodology or experiments, propose hypotheses, or interpret the reported results. All AI-assisted code and text were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper and its accompanying artifacts.

ETHICS STATEMENT

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

REPRODUCIBILITY STATEMENT

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

REFERENCES

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 3 edition, 2006.

Daniel Bartl, Samuel Drapeau, Jan Obłój, and Johannes Wiesel. Sensitivity analysis of wasserstein distributionally robust optimization problems. *Proceedings of the Royal Society A*, 477(2256): 20210176, 2021.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2 edition, 1999.

Dimitris Bertsimas and Aurélie Thiele. A robust optimization approach to inventory theory. *Operations Research*, 54(1):150–168, 2006.

Theerawat Bhudisaksang, Álvaro Cartea, and Leandro Sánchez-Betancourt. Adaptive-robust portfolio optimisation. *Mathematics and Financial Economics*, 20:171–202, 2026.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with wasserstein distances. *Management Science*, 68(9):6382–6410, 2022.

Jean-François Bonnans and Alexander Shapiro. Optimization problems with perturbations: A guided tour. *SIAM Review*, 40(2):228–264, 1998.

Anthony Coache and Sebastian Jaimungal. Robust reinforcement learning with dynamic distortion risk measures. *SIAM Journal on Mathematics of Data Science*, 8(1):1–22, 2026.

Arthur Compoint, Nathan Sauldubois, and Nizar Touzi. Sensitivity analysis of distributionally robust bsdes and rbsdes, 2025.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Kihyun Kim and Insoon Yang. Distributional robustness in minimax linear quadratic control with wasserstein distance. *SIAM Journal on Control and Optimization*, 61(2):458–483, 2023.

Yun Geon Kim and Byung Do Chung. Data-driven wasserstein distributionally robust dual-sourcing inventory model under uncertain demand. *Omega*, 127:103112, 2024.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y. Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Advances in Neural Information Processing Systems*, volume 36, pages 59477–59501, 2023.

Kazimierz Kuratowski and Czeslaw Ryll-Nardzewski. A general theorem on selectors. *Bulletin de l'Académie Polonaise des Sciences, Série des Sciences Mathématiques, Astronomiques et Physiques*, 13:397–403, 1965.

Shie H. Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust markov decision processes. In *Advances in Neural Information Processing Systems*, volume 26, pages 701–709, 2013.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Ariel Neufeld and Julian Sester. Robust q-learning algorithm for markov decision processes under wasserstein uncertainty. *Automatica*, 168:111825, 2024.

Ariel Neufeld, Julian Sester, and Mario Šikić. Markov decision processes under model uncertainty. *Mathematical Finance*, 33(3):618–665, 2023.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Elijah Polak. *Optimization: Algorithms and Consistent Approximations*, volume 124 of *Applied Mathematical Sciences*. Springer, New York, 1997.

Nathan Sauldubois and Nizar Touzi. First-order martingale model risk and semi-static hedging, 2024.

Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy optimization for robust average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pages 17348–17372, 2024. doi: 10.52202/079017-0551.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 35763–35797, 2023.

Yining Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 23484–23526, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870, 2021.

Zhouhao Yang, Yihong Guo, Pan Xu, Anqi Liu, and Animashree Anandkumar. Distributionally robust policy gradient for offline contextual bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 6443–6462, 2023.

A Appendix

A.1 Notation

For a real Hilbert space H, we denote by $\langle \cdot, \cdot \rangle_H$ its scalar product and by $\|h\|_H := \sqrt{\langle h, h \rangle_H}$ its associated norm. When there is no ambiguity on H, we simply write $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$. We denote by $\mathcal{P}(E)$ the space of Borel probability measures on a Borel space E. We recall that $\mathcal{P}(E)$ is Borel when E is itself a Borel space and we will denote by $\mathcal{B}(E)$ the Borel σ-algebra on E. Given two measurable spaces $(E, \mathcal{E})$, $(F, \mathcal{F})$ and $\mu \in \mathcal{P}(E)$, we denote by $L^0(E; F)$ the space of measurable maps from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ and by $L^1(\mu) := \big\{ f : E \to \mathbb{R} \text{ Borel-measurable} : \int_E |f(x)| \mu(\mathrm{d}x) < \infty \big\}$. We denote by $\mathcal{C}(E; \mathbb{R})$ ($\mathcal{C}_b(E; \mathbb{R})$) the space of (bounded) continuous real-valued maps on E and by $\mathcal{C}_p(E; \mathbb{R})$ the set of continuous maps on E with polynomial growth at most $p \in \mathbb{R}_+$. Whenever E is compact, we endow $\mathcal{C}(E; \mathbb{R})$ with its uniform metric $\|\varphi\|_\infty := \sup_{x \in E} \big|\varphi(x)\big|$ for any $\varphi \in \mathcal{C}(E; \mathbb{R})$. Consider (E, d) a metric space, $\mu, \nu \in \mathcal{P}(E)$ and $q \in \mathbb{N}^\star$, the q-Wasserstein distance W_q is defined as $W_q^q(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{E \times E} d(x, y)^q \pi(\mathrm{d}x, \mathrm{d}y)$ where $\Pi(\mu, \nu)$ is the set of probability measures on $E \times E$ such that $\mathrm{pr}_1 \sharp \pi = \mu$ and $\mathrm{pr}_2 \sharp \pi = \nu$ where pr_1 and pr_2 are the canonical projection maps. Given $T \in \mathbb{N}^\star$, we will denote by $\mathcal{T} := \big\{0, 1, \dots, T-1\big\}$, $\bar{\mathcal{T}} := \mathcal{T} \cup \big\{T\big\}$ and for any $t \in \mathcal{T}$, $\mathcal{T}_t := \big\{t, t+1, \dots, T-1\big\}$.

A.2 Some remarks on the choice of ambiguity sets

Remark A.1. (Rectangular ambiguity sets and separability of the infimum).

It is standard in the literature to consider *rectangular ambiguity sets*, as they encode an independence structure in the choice of admissible transition kernels across time. In particular, rectangularity ensures that admissible selections can be made independently at each stage.

An ambiguity set $\mathcal{B}$ is said to be rectangular if it factorizes as

$$\mathcal{B} = \prod_{l \in \mathcal{T}} \mathfrak{S}_l,$$

where, in our setting, for any $l \in \mathcal{T}$,

$$\mathfrak{S}_l := \big\{ \mathbb{P}_l : \mathcal{X} \times A \to \mathcal{P}(\mathcal{X}) \ : \ \mathbb{P}_l \text{ Borel measurable and } \mathbb{P}_l(x, a) \in \mathcal{P}_l(x, a), \ \forall (x, a) \in \mathcal{X} \times A \big\}.$$

As a consequence of this product structure, for any function $\Phi : \mathcal{B} \to \mathbb{R}$, one has

$$\inf_{(\mathbb{P}_l)_{l \in \mathcal{T}} \in \mathcal{B}} \Phi\big((\mathbb{P}_l)_{l \in \mathcal{T}}\big) = \inf_{\mathbb{P}_{T-1} \in \mathfrak{S}_{T-1}} \inf_{\mathbb{P}_{T-2} \in \mathfrak{S}_{T-2}} \cdots \inf_{\mathbb{P}_0 \in \mathfrak{S}_0} \Phi\big((\mathbb{P}_l)_{l \in \mathcal{T}}\big).$$

This identity is a direct consequence of the fact that $\mathcal{B}$ is a Cartesian product. The role of rectangularity is therefore to ensure that the global ambiguity set admits such a product structure, which is a key property underlying the validity of dynamic programming in robust control problems.

Remark A.2. Following Blanchet and Murthy (2019), Wasserstein duality extends to ambiguity balls defined by general transport costs $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The dual representation replaces the infinite-dimensional optimization over probability measures by a finite-dimensional optimization over the dual variable, which is the form used in our actor–critic recursion.

A.3 Background on theory of set-valued maps

Definition A.3. (Weak measurability and graph).

Let X be a Polish space, $\mathcal{B}(X)$ its Borel σ-algebra and $(\Omega, \mathcal{F})$ be a measurable space. Let $\psi : \Omega \rightrightarrows X$, i.e., for any $\omega \in \Omega$, $\psi(\omega) \subset X$. We say that the set-valued map or correspondence ψ is weakly-measurable if for every open subset $U \subset X$, we have

$$\big\{ \omega \in \Omega : \psi(\omega) \cap U \neq \emptyset \big\} \in \mathcal{F}. \tag{20}$$

Given a correspondence $\varphi : X \rightrightarrows Y$, we define its graph as

$$\mathrm{Gr}(\varphi) := \big\{ (x, y) \in X \times Y : y \in \varphi(x) \big\} \tag{21}$$

Definition A.4. (Upper and lower hemicontinuity).

Let $\varphi : X \rightrightarrows Y$ be a correspondence between two topological spaces.

(1) φ is called upper hemicontinuous if $\{x \in X : \varphi(x) \subset A\}$ is an open set for all open sets $A \subset Y$.

(2) φ is called lower hemicontinuous if $\{x \in X : \varphi(x) \cap A \neq \emptyset\}$ is an open set for all open sets $A \subset Y$.

(3) We say that φ is continuous if φ is upper and lower hemicontinuous.

A.4 Background on directional differentiability and envelope theorems

Definition A.5. (One-sided directional and Gâteaux derivatives). Let $\Theta \subset \mathbb{R}^{d_\theta}$, $\theta \in \Theta$ and let $r \in \mathbb{R}^{d_\theta}$ be a feasible direction at θ, i.e. $\theta + hr \in \Theta$ for all $h > 0$ small enough. For $\phi : \Theta \to \mathbb{R}$, the right directional derivative of ϕ at θ in the direction r is

$$D^+\phi(\theta)[r] := \lim_{h \downarrow 0} \frac{\phi(\theta + hr) - \phi(\theta)}{h}, \tag{22}$$

whenever the limit exists, and the left directional derivative is $D^-\phi(\theta)[r] := \lim_{h\uparrow 0}(\phi(\theta + hr) - \phi(\theta))/h = -D^+\phi(\theta)[-r]$ when $-r$ is also feasible. If θ is an interior point, $D^+\phi(\theta)[r]$ exists for every r and $r \mapsto D^+\phi(\theta)[r]$ is linear, then ϕ is Gâteaux differentiable at θ and we write $D^+\phi(\theta)[r] = \langle \nabla_\theta \phi(\theta), r \rangle$.

Theorem A.6. (Danskin's theorem)

Let $\Theta \subset \mathbb{R}^{d_\theta}$ be open, let K be a nonempty compact metric space, and let

$$\Phi : \Theta \times K \to \mathbb{R}$$

be continuous. For $\theta \in \Theta$, define the value function

$$\phi(\theta) := \sup_{\lambda \in K} \Phi(\theta, \lambda), \text{ and } \Lambda^\star(\theta) := \operatorname*{argmax}_{\lambda \in K} \Phi(\theta, \lambda).$$

Assume that, for every $\lambda \in K$, the map $\theta \mapsto \Phi(\theta, \lambda)$ is directionally differentiable at θ in every direction $r \in \mathbb{R}^{d_\theta}$, and that the directional derivative

$$D_\theta \Phi(\theta, \lambda)[r] := \lim_{h \downarrow 0} \frac{\Phi(\theta + hr, \lambda) - \Phi(\theta, \lambda)}{h}$$

exists and is continuous in λ, locally uniformly in r. Then ϕ is directionally differentiable at θ and

$$D\phi(\theta)[r] = \sup_{\lambda \in \Lambda^\star(\theta)} D_\theta \Phi(\theta, \lambda)[r].$$

In particular, if $\Lambda^\star(\theta) = \{\lambda^\star(\theta)\}$ is a singleton and $\theta \mapsto \Phi(\theta, \lambda)$ is continuously differentiable in a neighborhood of θ, then ϕ is Gâteaux differentiable at θ and

$$\nabla_\theta \phi(\theta) = \nabla_\theta \Phi(\theta, \lambda^\star(\theta)).$$

The analogous statement for a minimum follows by applying the theorem to $-\Phi$.

Proof. See Proposition 4.12 in Bonnans and Shapiro (1998) and Proposition B.22 in Bertsekas (1999). □

The continuity of $\lambda \mapsto D_\theta \Phi(\theta, \lambda)[r]$ cannot be dropped in Theorem A.6, and it fails for the dual objective of Proposition 3.2 when the transport argmin is not a singleton (Remark B.1). This is why the proof of Theorem 3.5 works directly with the primal problem and uniform convergence of difference quotients.

A.5 ADDITIONAL RESULTS

Lemma A.7. (Continuity of the integral).

Let H be a continuous map over the product of compact sets $\mathcal{X} \times A$ and π be a kernel probability from $\mathcal{X}$ into $\mathcal{P}(A)$ such that for any sequence $(x_n)_{n\in\mathbb{N}^\star} \subset \mathcal{X}$ converging to $x \in \mathcal{X}$, the weak convergence of $(\pi(x_n, \cdot))_{n\in\mathbb{N}^\star}$ towards $\pi(x, \cdot)$ holds. Denote now φ as

$$\varphi : \mathcal{X} \ni x \mapsto \int_A H(x, a)\pi(x, \mathrm{d}a) \in \mathbb{R}. \tag{23}$$

Then, the map φ belongs to $\mathcal{C}(\mathcal{X}; \mathbb{R})$.

Proof. Let $(x_n)_{n\in\mathbb{N}^\star} \subset \mathcal{X}$ be a sequence converging to some $x \in \mathcal{X}$. Then,

$$\Big|\int_A H(x_n, a)\pi(x_n, \mathrm{d}a) - \int_A H(x, a)\pi(x, \mathrm{d}a)\Big| \leq \int_A |H(x_n, a) - H(x, a)|\pi(x_n, \mathrm{d}a) + \Big|\int_A H(x, a)\pi(x_n, \mathrm{d}a) - \int_A H(x, a)\pi(x, \mathrm{d}a)\Big|$$

The map $A \ni a \mapsto H(x, a)$ is bounded continuous since A is compact, which implies that the second term vanishes as $n \to \infty$ since the weak convergence of $\pi(x_n, \cdot) \underset{n\to\infty}{\to} \pi(x, \cdot)$ holds. For the first term, since H is continuous on the compact set $\mathcal{X} \times A$, it is uniformly continuous, so that

$$\big|H(x_n, a) - H(x, a)\big| \leq \sup_{a\in A}\big|H(x_n, a) - H(x, a)\big| \underset{n\to\infty}{\to} 0, \tag{24}$$

which concludes the proof. □

Lemma A.8. (Berge's Maximum Principle).

Let $\varphi : \mathcal{X} \rightrightarrows \mathcal{Y}$ be a upper and lower hemicontinuous correspondence between topological spaces with nonempty compact values, and suppose that $f : \{(x, y) \in \mathcal{X} \times \mathcal{Y} : y \in \varphi(x)\} \to \mathbb{R}$ is continuous. Then the map $m : \mathcal{X} \ni x \mapsto \max_{y\in\varphi(x)} f(x, y)$ is continuous.

Proof. The proof of this result can be found in Theorem 17.31 in Aliprantis and Border (2006). □

Lemma A.9. (Measurable maximum theorem).

Let X be a Polish space, (S, Σ) be a measurable space and let $\varphi : S \rightrightarrows X$ be a weakly measurable correspondence with non empty compact values and suppose $f : S \times X \to \mathbb{R}$ is a Carathéodory map, i.e, a map satisfying the following assumptions

(1) For any $x \in X$, the map $f(\cdot, x)$ is Σ-measurable

(2) For any $s \in S$, the map $f(s, \cdot)$ is continuous.

Define the map $m : S \to \mathbb{R}$

$$S \ni s \mapsto m(s) := \inf_{x\in\varphi(s)} f(s, x), \tag{25}$$

and the correspondence of minimizers by the map μ :

$$S \ni s \mapsto \mu(s) := \big\{x \in \varphi(s) : f(s, x) = m(s)\big\}. \tag{26}$$

Then, the following holds

(1) The value function m is measurable

(2) The argmin correspondence μ has nonempty and compact values

(3) The argmin correspondence μ is weakly measurable and admits a measurable selector

Proof. The proof of this result can be found in Theorem 18.19 in Aliprantis and Border (2006). □

Theorem A.10. (Kuratowski and Ryll-Nardzewski measurable selection theorem).

Let X be a Polish space, $\mathcal{B}(X)$ its Borel $\sigma-$algebra and $(\Omega, \mathcal{F})$ be a measurable space. Let ψ be a weakly-measurable correspondence in Ω taking values in the set of nonempty closed subsets of X. Then ψ admits a measurable selection, i.e a $(\mathcal{B}(X), \mathcal{F})$ measurable map.

Proof. The proof of this known result can be found in Theorem 1 (page 398) in Kuratowski and Ryll-Nardzewski (1965) □

B PROOFS OF THE MAIN RESULTS

B.1 PROOF OF THEOREM 2.5

Proof. As we are working with rectangular sets (see Remark A.1), the proof of the dynamic programming principle and the growth property for the robust value function follows from standard arguments (see Iyengar (2005) and Neufeld et al. (2023)) based on the measurable selection of Theorem A.10, our assumptions and since our randomized setting doesn't affect the type of ambiguity sets. Moreover, the existence of the measurable selector is guaranteed by Lemma A.9 so we just give a sketch of proof. Indeed, from our assumptions, we are able to show the continuity of the map F_{t-1} for any $t \in \{1, \dots, T\}$ defined from $\mathrm{Gr}(\mathcal{P}_{t-1}) := \big\{(x, a, \mathbb{P}) : x \in \mathcal{X}, a \in A, \mathbb{P} \in \mathcal{P}_{t-1}(x, a)\big\}$ into $\mathbb{R}$ as

$$F_{t-1} : \mathrm{Gr}(\mathcal{P}_{t-1}) \ni (x, a, \mathbb{P}) \mapsto \int_{\mathcal{X}} \big(f(t-1, x, a, x') + V_t^\pi(x')\big)\mathbb{P}(\mathrm{d}x'). \tag{27}$$

by backward induction. Indeed, we rely on the proof done in Neufeld et al. (2023). First, $V_T^\pi = g$ is continuous by assumption. Suppose now that the result holds for some $t \in \{1, \dots, T\}$, i.e. $V_t^\pi \in \mathcal{C}(\mathcal{X}; \mathbb{R})$. Let $(x, a, \mathbb{P}) \in \mathrm{Gr}(\mathcal{P}_{t-1})$ and $(x_n, a_n, \mathbb{P}_n)_{n\in\mathbb{N}^\star}$ a sequence in $\mathrm{Gr}(\mathcal{P}_{t-1})$ such that $(x_n, a_n, \mathbb{P}_n) \underset{n\to\infty}{\to} (x, a, \mathbb{P})$ where the convergence is understood with respect to the topology of the product space $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X})$. Therefore, we have

$$\begin{aligned} &|F_{t-1}(x_n, a_n, \mathbb{P}_n) - F_{t-1}(x, a, \mathbb{P})| && (28)\\ \leq &|F_{t-1}(x_n, a_n, \mathbb{P}_n) - F_{t-1}(x, a, \mathbb{P}_n)| + |F_{t-1}(x, a, \mathbb{P}_n) - F_{t-1}(x, a, \mathbb{P})| && (29)\end{aligned}$$

The second term vanishes as $n \to \infty$ since the topology of the W_q distance is equivalent to the topology of the weak convergence since $\mathcal{X}$ is compact and since $f(t-1, x, a, \cdot)$ and V_t^π are continuous and bounded on $\mathcal{X}$, while the first term goes to 0 by the uniform continuity of $f(t-1, \cdot)$ on the compact set $\mathcal{X} \times A \times \mathcal{X}$. Hence, F_{t-1} is continuous. Therefore, we now apply Berge's maximum theorem (Lemma A.8) to obtain that the map

$$\mathcal{X} \times A \ni (x, a) \mapsto \inf_{\mathbb{P}\in\mathcal{P}_{t-1}(x,a)} F_{t-1}(x, a, \mathbb{P}) \in \mathbb{R}, \tag{30}$$

is continuous. Moreover, since the extension of F_{t-1} to $\mathcal{X} \times A \times \mathfrak{S}_{t-1}$ is a Caratheodory map, we can apply the measurable maximum theorem (see Lemma A.9) which implies the existence of a measurable selector $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}_{t-1}^\star(x, a) \in \mathcal{P}_{t-1}(x, a)$ such that $F_{t-1}(x, a, \mathbb{P}_{t-1}^\star(x, a)) = \inf_{\mathbb{P}\in\mathcal{P}_{t-1}(x,a)} F_{t-1}(x, a, \mathbb{P})$ for any $(x, a) \in \mathcal{X} \times A$. It implies also that $\mathbb{P}_{t-1}^\star \in \mathfrak{S}_{t-1}$. Moreover, we have $h_{t-1}(x, a) = \inf_{\mathbb{P}\in\mathcal{P}_{t-1}(x,a)} F_{t-1}(x, a, \mathbb{P})$ where we set

$$h_{t-1} : \mathcal{X} \times A \ni (x, a) \mapsto \inf_{\mathbb{P}_{t-1}\in\mathfrak{S}_{t-1}} \int_{\mathcal{X}} \big(f(t-1, x, a, x') + V_t^\pi(x')\big)\mathbb{P}_{t-1}(x, a, \mathrm{d}x') \in \mathbb{R}. \tag{31}$$

Indeed, if $\mathbb{P}_{t-1} \in \mathfrak{S}_{t-1}$, then $\mathbb{P}_{t-1}(x, a) \in \mathcal{P}_{t-1}(x, a)$, we hence have

$$\inf_{\mathbb{P}\in\mathcal{P}_{t-1}(x,a)} F_{t-1}(x, a, \mathbb{P}) \leq h_{t-1}(x, a), \tag{32}$$

and we get the reverse inequality by the existence of the measurable selector since we have

$$\begin{aligned} \inf_{\mathbb{P}\in\mathcal{P}_{t-1}(x,a)} F_{t-1}(x, a, \mathbb{P}) &= F_{t-1}(x, a, \mathbb{P}_{t-1}^\star(x, a)) && (33)\\ &= \int_{\mathcal{X}} \big(f(t-1, x, a, x') + V_t^\pi(x')\big)\mathbb{P}_{t-1}^\star(x, a, \mathrm{d}x') && (34)\\ &\geq h_{t-1}(x, a). && (35)\end{aligned}$$

Hence, by the existence of the measurable selector $\mathbb{P}_t^\star \in \mathfrak{S}_t$ for any $t \in \mathcal{T}$ and by the backward induction, we get the result defining the measure

$$\mathbb{P}^{\star,\pi}(\mathrm{d}\omega) = \delta_{x_0}(\mathrm{d}x_0) \prod_{t=0}^{T-1} \pi_t(x_t, \mathrm{d}a_t) \prod_{t=0}^{T-1} \mathbb{P}_t^\star(x_t, a_t, \mathrm{d}x_{t+1}), \in \mathcal{P}(\Omega). \tag{36}$$

such that $\mathbb{E}^{\mathbb{P}^{\star,\pi}}[R(S)] = V_0^\pi(x_0)$; the continuity of V_{t-1}^π follows from that of h_{t-1} and Lemma A.7, which closes the induction. □

B.2 Proof of Proposition 3.3

Proof. We prove it by backward induction. It is clear for $t = T$ since for any $\theta \in \Theta$, we have $V_T^\theta = g \in \mathcal{C}(\mathcal{X}; \mathbb{R})$ by assumption. Suppose that $V_{t+1}^\theta \in \mathcal{C}(\mathcal{X}; \mathbb{R})$ for any $\theta \in \Theta$ for some $t \in \mathcal{T}$, and fix $\theta \in \Theta$. Throughout, $\mathcal{P}(\mathcal{X})$ is endowed with the topology of weak convergence, which is compact and metrized by W_q since $\mathcal{X}$ is compact.

Set $H_\theta(x, a, y) := f(t, x, a, y) + V_{t+1}^\theta(y)$, a continuous bounded map on $\mathcal{X} \times A \times \mathcal{X}$, and $\Phi_t^\theta(x, a, \mathbb{P}) := \int_{\mathcal{X}} H_\theta(x, a, y)\mathbb{P}(\mathrm{d}y)$ on $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X})$. The map Φ_t^θ is continuous: if $(x_n, a_n, \mathbb{P}_n) \to (x, a, \mathbb{P})$, then $|\Phi_t^\theta(x_n, a_n, \mathbb{P}_n) - \Phi_t^\theta(x, a, \mathbb{P})| \le \sup_y |H_\theta(x_n, a_n, y) - H_\theta(x, a, y)| + |\int H_\theta(x, a, \cdot)\mathrm{d}(\mathbb{P}_n - \mathbb{P})|$, where the first term vanishes by uniform continuity of H_θ on a compact set and the second by weak convergence. The correspondence $(x, a) \rightrightarrows \mathcal{B}(x, a) := \mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x, a))$ is continuous with nonempty compact values (Proposition 3.1 in Neufeld et al. (2023), which only uses the weak continuity of $(x, a) \mapsto \mathbb{P}_t^0(x, a)$ from Assumption 3.1(1) and the compactness of $\mathcal{X}$). By Proposition 3.2, $G_t^\theta(x, a) = \min_{\mathbb{P} \in \mathcal{B}(x,a)} \Phi_t^\theta(x, a, \mathbb{P})$, and Berge's maximum theorem (Lemma A.8, applied to $-\Phi_t^\theta$) yields that G_t^θ is continuous on $\mathcal{X} \times A$ and that the argmin correspondence $(x, a) \rightrightarrows \mathcal{P}_t^\star(x, a, \theta)$ has nonempty compact values and is upper hemicontinuous (Theorem 17.31 in Aliprantis and Border (2006)).

Let $\Lambda > 0$ and define the map F_t^θ by

$$F_t^\theta(x, a, \lambda) := \int_{\mathcal{X}} -\mathcal{F}^\lambda(-f(t, x, a, \cdot) - V_{t+1}^\theta(\cdot))(x')\mathbb{P}_t^0(x, a, \mathrm{d}x') - \varepsilon^q\lambda, \quad \forall (x, a, \lambda) \in \mathcal{X} \times A \times [0, \Lambda]. \tag{37}$$

We show that F_t^θ is continuous on $\mathcal{X} \times A \times [0, \Lambda]$ for any $\Lambda > 0$. First, the map

$$\mathcal{X} \times A \times [0, \Lambda] \times \mathcal{X} \ni (x, a, \lambda, x') \mapsto -\mathcal{F}^\lambda(-f(t, x, a, \cdot) - V_{t+1}^\theta(\cdot))(x') = \inf_{y \in \mathcal{X}} \Psi(x, a, \lambda, x', y)$$

is continuous, by Berge's maximum principle (Lemma A.8) applied to the continuous map

$$\mathcal{X} \times A \times [0, \Lambda] \times \mathcal{X} \times \mathcal{X} \ni (x, a, \lambda, x', y) \mapsto \Psi(x, a, \lambda, x', y) := f(t, x, a, y) + V_{t+1}^\theta(y) + \lambda c(x', y) \tag{38}$$

with the constant correspondence $\varphi(x, a, \lambda, x') = \mathcal{X}$, which is continuous and compact-valued. The same theorem gives that $Y_{t,\theta,\lambda}^\star(x, a; x')$ is nonempty and compact. Since $\mathcal{X} \times A \times [0, \Lambda] \times \mathcal{X}$ is compact, this continuity is uniform, so that $\mathcal{X} \times A \times [0, \Lambda] \ni (x, a, \lambda) \mapsto -\mathcal{F}^\lambda(-f(t, x, a, \cdot) - V_{t+1}^\theta(\cdot)) \in \mathcal{C}(\mathcal{X}; \mathbb{R})$ is continuous for $\|\cdot\|_\infty$. Moreover, the map $\mathcal{X} \times A \times \mathcal{C}(\mathcal{X}; \mathbb{R}) \ni (x, a, h) \mapsto T(x, a, h) := \int_{\mathcal{X}} h(x')\mathbb{P}_t^0(x, a, \mathrm{d}x') \in \mathbb{R}$ is continuous: if $(x_n, a_n, h_n) \to (x, a, h)$ with $\mathcal{C}(\mathcal{X}; \mathbb{R})$ endowed with its uniform metric, then

$$\begin{aligned} |T(x_n, a_n, h_n) - T(x, a, h)| &\le \int_{\mathcal{X}} |h_n(x') - h(x')|\, \mathbb{P}_t^0(x_n, a_n, \mathrm{d}x') + \Big|\int_{\mathcal{X}} h(x')\mathbb{P}_t^0(x_n, a_n, \mathrm{d}x') - \int_{\mathcal{X}} h(x')\mathbb{P}_t^0(x, a, \mathrm{d}x')\Big|, \\ &\le \|h_n - h\|_\infty + \Big|\int_{\mathcal{X}} h(x')\, \mathbb{P}_t^0(x_n, a_n, \mathrm{d}x') - \int_{\mathcal{X}} h(x')\mathbb{P}_t^0(x, a, \mathrm{d}x')\Big| \end{aligned}$$

which goes to 0 as $n \to \infty$ by Assumption 3.1(1). Hence F_t^θ is continuous on $\mathcal{X} \times A \times [0, \Lambda]$. Moreover, since $\inf_{y \in \mathcal{X}} \Psi(x, a, \lambda, x', y) \le f(t, x, a, x') + V_{t+1}^\theta(x')$,

$$F_t^\theta(x, a, \lambda) \le \int_{\mathcal{X}} \big(f(t, x, a, x') + V_{t+1}^\theta(x')\big)\mathbb{P}_t^0(x, a, \mathrm{d}x') - \varepsilon^q\lambda \le \|H_\theta\|_\infty - \varepsilon^q\lambda. \tag{39}$$

Therefore, F_t^θ is coercive in λ uniformly in (x, a, θ) while $F_t^\theta(x, a, 0) \ge -\|H_\theta\|_\infty$. Consequently every maximizer of $\lambda \mapsto F_t^\theta(x, a, \lambda)$ over $\mathbb{R}_+$ lies in $[0, 2\|H_\theta\|_\infty/\varepsilon^q] \subset [0, \Lambda]$ with the constant Λ

of Proposition 3.3, which is finite because $|V_s^\theta| \le (T-s)\|f\|_\infty + \|g\|_\infty$ uniformly in (s,θ). Hence

$$G_t^\theta(x,a) = \sup_{\lambda \ge 0} F_t^\theta(x,a,\lambda) = \max_{\lambda \in [0,\Lambda]} F_t^\theta(x,a,\lambda), \quad \forall (x,a) \in \mathcal{X} \times A, \tag{40}$$

and $\Lambda_t^{\star,\theta}(x,a)$ is a nonempty compact subset of $[0,\Lambda]$. Since F_t^θ is continuous on the compact set $\mathcal{X} \times A \times [0,\Lambda]$, Berge's maximum theorem with the constant correspondence $(x,a) \mapsto [0,\Lambda]$ gives again that G_t^θ is continuous, consistently with the primal argument.

By Assumption 3.1(2) and Lemma A.7, the map $\mathcal{X} \ni x \mapsto \int_A G_t^\theta(x,a)\pi_t^\theta(x,\mathrm{d}a)$ belongs to $\mathcal{C}(\mathcal{X};\mathbb{R})$, which concludes the proof by the dynamic programming principle (6). □

B.3 Proof of Proposition 3.4 and Theorem 3.5

Proof. We prove the right-sided statements; the left-sided ones follow by replacing r with $-r$ (when $-r$ is feasible), which turns minima into maxima. Fix $\theta \in \Theta$ satisfying the lower hemicontinuity assumption of Proposition 3.4 and a feasible direction r at θ. All limits are taken as $h \downarrow 0$ with $\theta + hr \in \Theta$. We use the notation H_θ, Φ_t^θ, $\mathcal{B}(x,a)$ of the proof of Proposition 3.3, so that $G_t^\theta(x,a) = \min_{\mathbb{P} \in \mathcal{B}(x,a)} \Phi_t^\theta(x,a,\mathbb{P})$ and $\mathcal{P}_t^\star(x,a,\theta) = \operatorname{argmin}_{\mathbb{P} \in \mathcal{B}(x,a)} \Phi_t^\theta(x,a,\mathbb{P})$.

For $t = T$ and any $h > 0$, $V_T^{\theta+hr} - V_T^\theta = g - g = 0$, so $D_\theta^+ V_T^\theta(\cdot)[r] = 0 \in \mathcal{C}(\mathcal{X};\mathbb{R})$ and (7) holds trivially at $t = T$.

Fix $t \in \mathcal{T}$ and assume that $D_\theta^+ V_{t+1}^\theta(\cdot)[r] \in \mathcal{C}(\mathcal{X};\mathbb{R})$ exists with

$$\delta_h := \left\| \frac{V_{t+1}^{\theta+hr} - V_{t+1}^\theta}{h} \; - \; D_\theta^+ V_{t+1}^\theta(\cdot)[r] \right\|_\infty \xrightarrow[h \downarrow 0]{} 0. \tag{41}$$

In particular $\sup_{0<h\le h_0} \|(V_{t+1}^{\theta+hr} - V_{t+1}^\theta)/h\|_\infty \le \|D_\theta^+ V_{t+1}^\theta(\cdot)[r]\|_\infty + 1 < +\infty$ for h_0 small enough.

Fix $(x,a) \in \mathcal{X} \times A$ and set

$$\Phi'(\theta,\mathbb{P})[r] := \int_{\mathcal{X}} D_\theta^+ V_{t+1}^\theta(y)[r]\, \mathbb{P}(\mathrm{d}y).$$

Since f does not depend on θ,

$$\frac{\Phi_t^{\theta+hr}(x,a,\mathbb{P}) - \Phi_t^\theta(x,a,\mathbb{P})}{h} - \Phi'(\theta,\mathbb{P})[r] = \int_{\mathcal{X}} \Big(\frac{V_{t+1}^{\theta+hr} - V_{t+1}^\theta}{h} - D_\theta^+ V_{t+1}^\theta(\cdot)[r] \Big)(y)\, \mathbb{P}(\mathrm{d}y), \tag{42}$$

whose absolute value is bounded by δ_h uniformly in $\mathbb{P} \in \mathcal{B}(x,a)$. Let

$$m_t(x,a;\theta') := \min_{\mathbb{P} \in \mathcal{P}_t^\star(x,a,\theta')} \int_{\mathcal{X}} D_\theta^+ V_{t+1}^\theta(y)[r]\, \mathbb{P}(\mathrm{d}y).$$

For $\mathbb{P}^\star \in \mathcal{P}_t^\star(x,a,\theta)$,

$$\frac{G_t^{\theta+hr}(x,a) - G_t^\theta(x,a)}{h} \le \int_{\mathcal{X}} D_\theta^+ V_{t+1}^\theta(y)[r]\, \mathbb{P}^\star(\mathrm{d}y) + \delta_h,$$

so

$$\limsup_{h \downarrow 0} \frac{G_t^{\theta+hr}(x,a) - G_t^\theta(x,a)}{h} \le m_t(x,a;\theta).$$

Conversely, let $\mathbb{P}_h \in \mathcal{P}_t^\star(x,a,\theta+hr)$ attain $m_t(x,a;\theta+hr)$. Then

$$\frac{G_t^{\theta+hr}(x,a) - G_t^\theta(x,a)}{h} \ge m_t(x,a;\theta+hr) - \delta_h.$$

By weak compactness, along any sequence $h_n \downarrow 0$ we may extract a subsequence such that $\mathbb{P}_{h_n} \to \bar{\mathbb{P}}$. Since $\|V_{t+1}^{\theta+h_n r} - V_{t+1}^\theta\|_\infty \to 0$ and Φ_t^θ is continuous, $\bar{\mathbb{P}} \in \mathcal{P}_t^\star(x,a,\theta)$. Hence, by weak continuity of $\mathbb{P} \mapsto \Phi'(\theta,\mathbb{P})[r]$,

$$\liminf_{n \to \infty} m_t(x,a;\theta+h_n r) \ge m_t(x,a;\theta).$$

Therefore $D_\theta^+ G_t^\theta(x,a)[r]$ exists and $D_\theta^+ G_t^\theta(x,a)[r] = m_t(x,a;\theta)$, which is (8). Moreover,

$$\left|\frac{G_t^{\theta+hr}(x,a) - G_t^\theta(x,a)}{h} - D_\theta^+ G_t^\theta(x,a)[r]\right| \le \delta_h + \big(m_t(x,a;\theta) - m_t(x,a;\theta+hr)\big)_+. \quad (43)$$

Also $|G_t^{\theta+hr} - G_t^\theta| \le \|V_{t+1}^{\theta+hr} - V_{t+1}^\theta\|_\infty$, so the difference quotients of G_t^θ are uniformly bounded by the bound of Step 1.

By Proposition 3.3 and the lower hemicontinuity assumption of Proposition 3.4, the correspondence $(x,a) \rightrightarrows \mathcal{P}_t^\star(x,a,\theta)$ is continuous with nonempty compact values, and $(x,a,\mathbb{P}) \mapsto \Phi'(\theta,\mathbb{P})[r]$ is continuous (it does not depend on (x,a)). Berge's maximum theorem (Lemma A.8) therefore gives that $(x,a) \mapsto D_\theta^+ G_t^\theta(x,a)[r] = m_t(x,a;\theta)$ is continuous on $\mathcal{X}\times A$. We now show that $\sup_{(x,a)}(m_t(x,a;\theta) - m_t(x,a;\theta+hr))_+ \to 0$, which together with (43) gives the uniform convergence of the G-quotients in (7). Suppose not: there exist $\eta>0$, $h_n \downarrow 0$ and $(x_n,a_n) \in \mathcal{X}\times A$ with $m_t(x_n,a_n;\theta) - m_t(x_n,a_n;\theta+h_n r) \ge \eta$. By compactness we may assume $(x_n,a_n)\to(\bar{x},\bar{a})$ and, choosing $\mathbb{P}_n \in \mathcal{P}_t^\star(x_n,a_n,\theta+h_n r)$ attaining $m_t(x_n,a_n;\theta+h_n r)$, that $\mathbb{P}_n \to \bar{\mathbb{P}}$ weakly, with $\bar{\mathbb{P}} \in \mathcal{B}(\bar{x},\bar{a})$ by the closed graph of $\mathcal{B}$. Then

$$\Phi_t^\theta(x_n,a_n,\mathbb{P}_n) \le \Phi_t^{\theta+h_n r}(x_n,a_n,\mathbb{P}_n) + \|V_{t+1}^{\theta+h_n r} - V_{t+1}^\theta\|_\infty \quad (44)$$

$$= G_t^{\theta+h_n r}(x_n,a_n) + o(1) \le G_t^\theta(x_n,a_n) + 2\|V_{t+1}^{\theta+h_n r} - V_{t+1}^\theta\|_\infty, \quad (45)$$

and passing to the limit using the continuity of Φ_t^θ and of G_t^θ (Proposition 3.3) yields $\Phi_t^\theta(\bar{x},\bar{a},\bar{\mathbb{P}}) \le G_t^\theta(\bar{x},\bar{a})$, i.e. $\bar{\mathbb{P}} \in \mathcal{P}_t^\star(\bar{x},\bar{a},\theta)$. Hence $m_t(x_n,a_n;\theta+h_n r) = \Phi'(\theta,\mathbb{P}_n)[r] \to \Phi'(\theta,\bar{\mathbb{P}})[r] \ge m_t(\bar{x},\bar{a};\theta)$, while $m_t(x_n,a_n;\theta) \to m_t(\bar{x},\bar{a};\theta)$ by the continuity established above. Thus $\liminf_n \big(m_t(x_n,a_n;\theta) - m_t(x_n,a_n;\theta+h_n r)\big) \le 0 < \eta$, a contradiction. We have shown

$$\sup_{(x,a)\in\mathcal{X}\times A}\left|\frac{G_t^{\theta+hr}(x,a) - G_t^\theta(x,a)}{h} - D_\theta^+ G_t^\theta(x,a)[r]\right| \xrightarrow[h\downarrow 0]{} 0. \quad (46)$$

By the dynamic programming principle (6),

$$\frac{V_t^{\theta+hr}(x) - V_t^\theta(x)}{h} = \int_A \frac{G_t^{\theta+hr}(x,a) - G_t^\theta(x,a)}{h}\pi_t^{\theta+hr}(x,\mathrm{d}a) + \frac{1}{h}\Big(\int_A G_t^\theta(x,a)\pi_t^{\theta+hr}(x,\mathrm{d}a) - \int_A G_t^\theta(x,a)\pi_t^\theta(x,\mathrm{d}a)\Big) =: I_h(x) + J_h(x).$$

For the first term, since $\pi_t^{\theta+hr}(x,\cdot)$ is a probability measure, (46) gives

$$\sup_{x\in\mathcal{X}}\Big|I_h(x) - \int_A D_\theta^+ G_t^\theta(x,a)[r]\,\pi_t^{\theta+hr}(x,\mathrm{d}a)\Big| \xrightarrow[h\downarrow 0]{} 0,$$

and since $(x,a) \mapsto D_\theta^+ G_t^\theta(x,a)[r]$ is continuous on the compact set $\mathcal{X}\times A$, the map $(\theta',x) \mapsto \int_A D_\theta^+ G_t^\theta(x,a)[r]\,\pi_t^{\theta'}(x,\mathrm{d}a)$ is continuous on the compact set $\Theta\times\mathcal{X}$ (by Assumption 3.1(2) and the argument of Lemma A.7), hence uniformly continuous, so that

$$\sup_{x\in\mathcal{X}}\Big|I_h(x) - \int_A D_\theta^+ G_t^\theta(x,a)[r]\,\pi_t^{\theta}(x,\mathrm{d}a)\Big| \xrightarrow[h\downarrow 0]{} 0.$$

For the second term, Assumption 3.1(2) applied to $\varphi := G_t^\theta \in \mathcal{C}(\mathcal{X}\times A;\mathbb{R})$ gives that $(\theta',x) \mapsto \int_A G_t^\theta(x,a)\pi_t^{\theta'}(x,\mathrm{d}a)$ is continuously differentiable in θ' with gradient $\int_A G_t^\theta(x,a)\nabla_\theta \log \pi_t^{\theta'}(x,a)\pi_t^{\theta'}(x,\mathrm{d}a)$ jointly continuous in (θ',x). By the mean value theorem,

$$J_h(x) = \int_0^1 \Big\langle \int_A G_t^\theta(x,a)\nabla_\theta \log \pi_t^{\theta+shr}(x,a)\,\pi_t^{\theta+shr}(x,\mathrm{d}a), r\Big\rangle \mathrm{d}s,$$

and the uniform continuity of the gradient on the compact set $\Theta\times\mathcal{X}$ yields

$$\sup_{x\in\mathcal{X}}\Big|J_h(x) - \int_A G_t^\theta(x,a)\big\langle\nabla_\theta \log \pi_t^\theta(x,a), r\big\rangle\,\pi_t^\theta(x,\mathrm{d}a)\Big| \xrightarrow[h\downarrow 0]{} 0,$$

and $\sup_{0<h\le h_0}\|J_h\|_\infty \le C_\pi \|G_t^\theta\|_\infty \|r\|$ for h_0 small enough. Collecting terms yields the uniform convergence in x of $(V_t^{\theta+hr} - V_t^\theta)/h$ to the continuous limit

$$D_\theta^+ V_t^\theta(x)[r] := \int_A D_\theta^+ G_t^\theta(x,a)[r]\pi_t^\theta(x,\mathrm{d}a) + \int_A G_t^\theta(x,a)\langle\nabla_\theta \log \pi_t^\theta(x,a), r\rangle\pi_t^\theta(x,\mathrm{d}a), \quad \forall x\in\mathcal{X},$$

together with the uniform bound on the quotients, which is the induction hypothesis (41) at index t and the recursion of Corollary 3.7. Applying this backward induction down to $t = 0$ proves Proposition 3.4, the envelope formula (8) and Corollary 3.7.

Fix (x, a), $\lambda^\star \in \Lambda_t^{\star,\theta}(x, a)$ and $\mathbb{P}^\star \in \mathcal{P}_t^\star(x, a, \theta)$, and let $\pi^\star \in \Pi(\mathbb{P}_t^0(x, a, \cdot), \mathbb{P}^\star)$ be an optimal coupling for W_q, which exists since c is continuous on the compact set $\mathcal{X} \times \mathcal{X}$. By the complementary slackness statement of Proposition 3.2 applied to $V := H_\theta(x, a, \cdot)$, we have $y \in Y^\star_{t,\theta,\lambda^\star}(x, a; x')$ for $\pi^\star$-a.e. (x', y). Therefore

$$\int_{\mathcal{X}} D_\theta^+ V_{t+1}^\theta(y)[r]\, \mathbb{P}^\star(\mathrm{d}y) = \int_{\mathcal{X}\times\mathcal{X}} D_\theta^+ V_{t+1}^\theta(y)[r]\, \pi^\star(\mathrm{d}x', \mathrm{d}y) \tag{47}$$

$$\geq \int_{\mathcal{X}} \inf_{y \in Y^\star_{t,\theta,\lambda^\star}(x,a;x')} D_\theta^+ V_{t+1}^\theta(y)[r]\, \mathbb{P}_t^0(x, a, \mathrm{d}x'), \tag{48}$$

where the integrand on the right is Borel measurable and bounded by Lemma A.9 (the map $(x', y) \mapsto H_\theta(x, a, y) + \lambda^\star c(x', y)$ is continuous, so the argmin correspondence $x' \rightrightarrows Y^\star_{t,\theta,\lambda^\star}(x, a; x')$ is weakly measurable with compact values, and the infimum of the continuous map $D_\theta^+ V_{t+1}^\theta(\cdot)[r]$ over it is measurable by the same lemma). Taking the minimum over $\mathbb{P}^\star \in \mathcal{P}_t^\star(x, a, \theta)$ and using (8) gives the first inequality in (9); the second follows by symmetry. This completes the proof of Theorem 3.5. □

Remark B.1. The inequality (9) can be strict even when the dual optimizer is unique. $\mathcal{X} = \{0, 1, 2\}$, $c(x, y) = |x - y|$, $\varepsilon = 1/2$, $\mathbb{P}_t^0(x, a, \cdot) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2$, and suppose that $H_\theta(x, a, \cdot) = (1, 0, 1)$ and $D_\theta^+ V_{t+1}^\theta(\cdot)[r] = (0, -1, 0)$ on $(0, 1, 2)$. The worst-case laws move as much mass as the budget allows onto the state 1: $\mathcal{P}_t^\star(x, a, \theta) = \{(\frac{1}{2} - \alpha, \frac{1}{2}, \alpha) : \alpha \in [0, \frac{1}{2}]\}$, all with value $G_t^\theta(x, a) = \frac{1}{2}$, and the envelope formula gives $D_\theta^+ G_t^\theta(x, a)[r] = \min_\alpha\big(-\frac{1}{2}\big) = -\frac{1}{2}$; indeed, perturbing $H_\theta(x, a, 1)$ to $-h$ leaves the optimal transport plans unchanged and gives $G_t^{\theta+hr}(x, a) = \frac{1}{2} - \frac{h}{2}$. On the dual side, $F_t^\theta(x, a, \lambda) = \min(1, \lambda) - \lambda/2$ has the unique maximizer $\lambda^\star = 1$, for which $Y^\star_{t,\theta,1}(x, a; 0) = \{0, 1\}$ and $Y^\star_{t,\theta,1}(x, a; 2) = \{1, 2\}$, so the right-hand side of (9) equals $\frac{1}{2}\min(0, -1) + \frac{1}{2}\min(-1, 0) = -1 < -\frac{1}{2}$. The pointwise worst tie-break sends the whole mass to 1, which the transport budget forbids. Equivalently, $\lambda \mapsto D_\theta^+ F_t^\theta(x, a, \lambda)[r]$ jumps from -1 to 0 at $\lambda^\star = 1$ and is only lower semicontinuous, so the continuity hypothesis of Danskin's theorem (Theorem A.6) fails for the scalar dual problem. This is why the analysis is carried out on the primal problem.

Remark B.2. The lower hemicontinuity assumption of Proposition 3.4 is not needed for the pointwise envelope formula (8). It is used in Step 3 to make the argmin correspondence

$$(x, a) \rightrightarrows \mathcal{P}_t^\star(x, a, \theta)$$

continuous, which by Berge's theorem yields continuity of $(x, a) \mapsto D_\theta^+ G_t^\theta(x, a)[r]$ and allows the convergence of the difference quotients to be upgraded from pointwise to uniform in (x, a). This uniformity is needed in Step 4 because the quotients are integrated against the moving policy $\pi_t^{\theta+hr}(x, \mathrm{d}a)$.

When $\mathcal{X}$ and A are finite, pointwise convergence is automatically uniform and the assumption is unnecessary. When the worst-case law is unique for every (x, a), upper hemicontinuity of the single-valued argmin map implies continuity, so the assumption also holds automatically.

B.4 The vector-gradient case

In Corollary 3.7, $\mathbb{E}_{a\sim\pi_t^\theta}[G_t^\theta(x, a)\langle\nabla_\theta \log \pi_t^\theta(x, a), r\rangle]$ is the classical score-function term with the robust one-step value, and $\mathbb{E}_{a\sim\pi_t^\theta}[D_\theta^\pm G_t^\theta(x, a)[r]]$ replaces the classical continuation term $\mathbb{E}_a\mathbb{E}_{X\sim\mathbb{P}_t^0(x,a,\cdot)}[\nabla_\theta V_{t+1}^\theta(X)]$ by its worst-case envelope (Theorem 3.5), which under Proposition B.3 averages $\nabla_\theta V_{t+1}^\theta$ at the transported points. Keeping the continuation under the nominal kernel is not the robust gradient (relative error 0.69 against 3×10^{-9}, Appendix E.5), although with tabular softmax policies it reaches the same fixed points, the transport changing only the occupancy weights of the block gradient (Appendix E.4). Dropping it altogether fails.

When for some dual optimizer $\lambda^\star \in \Lambda_t^{\star,\theta}(x, a)$ the transport argmin $Y^\star_{t,\theta,\lambda^\star}(x, a; x')$ is a singleton $\{y_t^\star(x, a, x'; \theta)\}$ for $\mathbb{P}_t^0(x, a, \cdot)$-a.e. x', the worst-case law is unique,

$\mathcal{P}_t^\star(x,a,\theta) = \{y_t^\star(x,a,\cdot;\theta)\sharp\mathbb{P}_t^0(x,a,\cdot)\}$, and both one-sided derivatives in (8) equal $\mathbb{E}_X[D_\theta^\pm V_{t+1}^\theta(y_t^\star(x,a,X;\theta))[r]]$. If this holds at every (s,x,a), $\theta \mapsto V_s^\theta(x)$ is Gâteaux differentiable with the backward recursion

$$\nabla_\theta G_s^\theta(x,a) = \mathbb{E}_{X\sim\mathbb{P}_s^0(x,a,\cdot)}\big[\nabla_\theta V_{s+1}^\theta\big(y_s^\star(x,a,X;\theta)\big)\big], \tag{49}$$

$$\nabla_\theta V_s^\theta(x) = \mathbb{E}_{a\sim\pi_s^\theta(x,\cdot)}\big[G_s^\theta(x,a)\nabla_\theta \log \pi_s^\theta(x,a) + \nabla_\theta G_s^\theta(x,a)\big], \tag{50}$$

$\nabla_\theta V_T^\theta = 0$. Compared with the classical finite-horizon policy gradient theorem, for which $\nabla_\theta G_s^\theta(x,a) = \mathbb{E}_{X\sim\mathbb{P}_s^0(x,a,\cdot)}[\nabla_\theta V_{s+1}^\theta(X)]$, the only change is that the downstream sensitivity is evaluated at the worst-case transport destination instead of the nominal next state. In finite state–action spaces the one-sided derivatives exist without any topological condition, the recursion of Corollary 3.7 holds with the envelope taken over the polytope of worst-case laws, which is the set of second marginals of the optimal face of a finite transport linear program (Theorem 3.6 in Appendix B.5). This is the form used in our tabular experiments.

Proposition B.3. (Vector-gradient). *Let the assumptions of Proposition 3.4 hold and fix (t,x,a). Suppose that for some $\lambda^\star \in \Lambda_t^{\star,\theta}(x,a)$, the set $Y_{t,\theta,\lambda^\star}^\star(x,a;x')$ is a singleton $\{y_t^\star(x,a,x';\theta)\}$ for $\mathbb{P}_t^0(x,a,\cdot)$-a.e. x'. Then $\mathcal{P}_t^\star(x,a,\theta) = \{y_t^\star(x,a,\cdot;\theta)\sharp\mathbb{P}_t^0(x,a,\cdot)\}$ is a singleton, and*

$$D_\theta^\pm G_t^\theta(x,a)[r] = \mathbb{E}_{X\sim\mathbb{P}_t^0(x,a,\cdot)}\big[D_\theta^\pm V_{t+1}^\theta\big(y_t^\star(x,a,X;\theta)\big)[r]\big]. \tag{51}$$

If this uniqueness holds for every $(s,x,a) \in \mathcal{T}_t \times \mathcal{X} \times A$, then $\theta' \mapsto V_s^{\theta'}(x)$ is Gâteaux differentiable at θ for all $s \in \mathcal{T}_t$ and $x \in \mathcal{X}$, with gradients given by the backward recursion

$$\nabla_\theta G_s^\theta(x,a) = \mathbb{E}_{X\sim\mathbb{P}_s^0(x,a,\cdot)}\big[\nabla_\theta V_{s+1}^\theta\big(y_s^\star(x,a,X;\theta)\big)\big], \tag{52}$$

$$\nabla_\theta V_s^\theta(x) = \mathbb{E}_{a\sim\pi_s^\theta(x,\cdot)}\big[G_s^\theta(x,a)\nabla_\theta \log \pi_s^\theta(x,a) + \nabla_\theta G_s^\theta(x,a)\big], \tag{53}$$

with $\nabla_\theta V_T^\theta = 0$.

Proof. Fix (t,x,a) and $\lambda^\star \in \Lambda_t^{\star,\theta}(x,a)$ such that $Y_{t,\theta,\lambda^\star}^\star(x,a;x') = \{y^\star(x')\}$ for $\mathbb{P}_t^0(x,a,\cdot)$-a.e. x', where we write $y^\star(x') := y_t^\star(x,a,x';\theta)$. The map $y^\star$ is Borel measurable on the full-measure set where the argmin is a singleton, being the unique measurable selector of a weakly measurable compact-valued correspondence (Lemma A.9), and we extend it measurably outside this set. Let $\mathbb{P}^\star \in \mathcal{P}_t^\star(x,a,\theta)$ and let $\pi^\star \in \Pi(\mathbb{P}_t^0(x,a,\cdot),\mathbb{P}^\star)$ be an optimal coupling for W_q. By the complementary slackness in Proposition 3.2, $y \in Y_{t,\theta,\lambda^\star}^\star(x,a;x') = \{y^\star(x')\}$ for $\pi^\star$-a.e. (x',y), so that $\pi^\star = (\mathrm{id},y^\star)\sharp\mathbb{P}_t^0(x,a,\cdot)$ and $\mathbb{P}^\star = y^\star\sharp\mathbb{P}_t^0(x,a,\cdot)$. Since $\mathcal{P}_t^\star(x,a,\theta)$ is nonempty (Proposition 3.3), it is the singleton $\{y^\star\sharp\mathbb{P}_t^0(x,a,\cdot)\}$, and (8) reduces to (51) for both one-sided derivatives.

Suppose now that the uniqueness holds for every $(s,x,a) \in \mathcal{T}_t \times \mathcal{X} \times A$ and let θ be an interior point of Θ, so that every r is feasible. We show by backward induction on s that $D_\theta^+ V_s^\theta(x)[r]$ is linear in r and equals $D_\theta^- V_s^\theta(x)[r]$, with the recursion (52). This is clear for $s = T$. If it holds at $s+1$, then by (51) at every (s,x,a), $D_\theta^\pm G_s^\theta(x,a)[r] = \mathbb{E}_X\big[\langle \nabla_\theta V_{s+1}^\theta(y_s^\star(x,a,X;\theta)), r\rangle\big] = \langle \nabla_\theta G_s^\theta(x,a), r\rangle$ with $\nabla_\theta G_s^\theta(x,a)$ given by (52) (the transport map $y_s^\star$ does not depend on r), and Corollary 3.7 gives $D_\theta^\pm V_s^\theta(x)[r] = \langle \nabla_\theta V_s^\theta(x), r\rangle$ with $\nabla_\theta V_s^\theta(x)$ as in (52). By Definition A.5, $\theta' \mapsto V_s^{\theta'}(x)$ is Gâteaux differentiable at θ. □

B.5 Proof of Theorem 3.6

Proof. When $\mathcal{X}$ and A are finite, $\mathcal{P}(\mathcal{X})$ is the probability simplex, weak convergence is convergence of the finitely many coordinates, every real function on $\mathcal{X}$ or $\mathcal{X} \times A$ is continuous and bounded, and the lower hemicontinuity assumption of Proposition 3.4 holds trivially because every correspondence on a finite discrete set is continuous. Assumption 3.1(1) holds trivially, and Assumption 3.1(2) is assumed. The proof of Proposition 3.4 and Theorem 3.5 therefore applies verbatim, with Steps 3 and 4 simplified since pointwise convergence on a finite set is uniform: the one-sided derivatives $D_\theta^\pm V_t^\theta(x)[r]$ exist for every t and x, and the recursion of Corollary 3.7 holds with (12), which is (8) written with sums.

For the description of $\mathcal{P}_t^\star(x,a,\theta)$, consider the transport linear program (14). By definition of W_q, a law $\mathbb{P}$ belongs to $\mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x,a))$ if and only if it is the second marginal of a nonnegative

coupling γ on $\mathcal{X} \times \mathcal{X}$ with first marginal $\mathbb{P}_t^0(x, a, \cdot)$ and transport cost $\sum_{x',y} \gamma(x', y)c(x', y) \leq \varepsilon^q$ (the infimum over couplings is attained since the set of couplings is a polytope). Hence $\min_{\mathbb{P} \in \mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x,a))} \sum_y \mathbb{P}(\{y\}) H_\theta(y)$ is the value of the linear program (14), and $\mathcal{P}_t^\star(x, a, \theta)$ is the image of its optimal face under the linear map $\gamma \mapsto \sum_{x'} \gamma(x', \cdot)$. The feasible set of (14) is a bounded polyhedron, so its optimal face is a polytope and so is its image. Finally, since (12) minimizes or maximizes a linear function of $\mathbb{P}^\star$ over this polytope, it can be computed as a linear program over the optimal face of (14), for instance by adding the constraint $\sum_{x',y} \gamma(x', y) H_\theta(y) = G_t^\theta(x, a)$ to (14) and optimizing $\sum_{x',y} \gamma(x', y) D_\theta^\pm V_{t+1}^\theta(y)[r]$. □

C EXACT FINITE-MODEL ASCENT: CONVERGENCE AND PROOFS

This subsection builds an ascent scheme for J and proves its convergence. The scheme follows the classical pattern of active-set methods with a shrinking tolerance (Polak, 1997, Sections 1.2 and 2.4).

Throughout, $\mathcal{X}$ and A are finite, $\varepsilon > 0$, $\zeta \geq 0$, and $\Theta = [-B, B]^{d_\theta}$. Each $\pi_t^\theta(x, a)$ is strictly positive and C^2 near Θ. Critics, transport programs and dual problems are exact. The ζ-active feasible directions at θ are

$$T_\zeta(\theta) := \big\{ r \in \mathbb{R}^{d_\theta} : \ \|r\|_\infty \leq 1,\ r_i \leq 0 \text{ if } \theta_i \geq B - \zeta,\ r_i \geq 0 \text{ if } \theta_i \leq -B + \zeta \big\},$$

so that $\theta + hr \in \Theta$ for $r \in T_\zeta(\theta)$ and $h \in [0, \zeta]$. By Theorem 3.6, for each $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$ the ball $\mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x, a, \cdot)) \subseteq \mathcal{P}(\mathcal{X})$ is a polytope. Let $\mathcal{V}_t(x, a) \subseteq \mathcal{B}^{\varepsilon,q}(\mathbb{P}_t^0(x, a, \cdot))$ denote its vertex set, which is finite and does not depend on θ. A selection is a map $\sigma : \mathcal{T} \times \mathcal{X} \times A \to \mathcal{P}(\mathcal{X})$, such that $\sigma(t, x, a) \in \mathcal{V}_t(x, a)$, for all $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$, and the set of selections,

$$\Sigma := \prod_{(t,x,a) \in \mathcal{T} \times \mathcal{X} \times A} \mathcal{V}_t(x, a),$$

is finite. Each $\sigma \in \Sigma$ defines a non-robust MDP with the fixed transition kernels $\mathbb{P}_t(x, a, \cdot) := \sigma(t, x, a)$, and we write $V_t^{\sigma,\theta} : \mathcal{X} \to \mathbb{R}$ and $Q_t^{\sigma,\theta} : \mathcal{X} \times A \to \mathbb{R}$ for the value functions of the policy π^θ in this MDP, for $t \in \mathcal{T}$, and $J_\sigma : \Theta \to \mathbb{R}$, $J_\sigma(\theta) := \mathbb{E}_{\mu_0}[V_0^{\sigma,\theta}(X_0)]$, for its performance. With $\langle \mathbb{P}, \phi \rangle := \sum_{y \in \mathcal{X}} \mathbb{P}(\{y\})\phi(y)$ for $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and $\phi : \mathcal{X} \to \mathbb{R}$, define the gaps $g_t^{\sigma,\theta} : \mathcal{X} \times A \to [0, \infty)$ and, for $\zeta \geq 0$, the ζ-active selections $S_\zeta(\theta) \subseteq \Sigma$ by

$$g_t^{\sigma,\theta}(x, a) := \big\langle \sigma(t, x, a),\ f(t, x, a, \cdot) + V_{t+1}^\theta \big\rangle - G_t^\theta(x, a),$$
$$S_\zeta(\theta) := \big\{ \sigma \in \Sigma :\ g_t^{\sigma,\theta}(x, a) \leq \zeta,\ \forall (t, x, a) \in \mathcal{T} \times \mathcal{X} \times A \big\}.$$

Writing $\mathcal{V}_t^{\zeta,\theta}(x, a) := \{\mathbb{P} \in \mathcal{V}_t(x, a) : \langle \mathbb{P}, f(t, x, a, \cdot) + V_{t+1}^\theta \rangle \leq G_t^\theta(x, a) + \zeta\}$ for the ζ-optimal vertices of the transport program at (t, x, a), one has $S_\zeta(\theta) = \prod_{(t,x,a) \in \mathcal{T} \times \mathcal{X} \times A} \mathcal{V}_t^{\zeta,\theta}(x, a)$.

Theorem C.1. *The following results hold:*

(i) Each J_σ is C^2 near Θ, $J = \min_{\sigma \in \Sigma} J_\sigma$ on Θ, $S_0(\theta) \neq \emptyset$, and $S_0(\theta) \subseteq \Sigma^*(\theta) := \{\sigma \in \Sigma : J_\sigma(\theta) = J(\theta)\}$.

(ii) For every direction $r \in \mathbb{R}^{d_\theta}$ with $\theta + hr \in \Theta$ for all small $h > 0$,

$$D_\theta^+ J(\theta)[r] = \min_{\sigma \in \Sigma^*(\theta)} \big\langle \nabla_\theta J_\sigma(\theta), r \big\rangle, \quad D_\theta^- J(\theta)[r] = \max_{\sigma \in \Sigma^*(\theta)} \big\langle \nabla_\theta J_\sigma(\theta), r \big\rangle,$$

the second formula holding for directions feasible on both sides, consistently with Theorem 3.6. Here $\nabla_\theta J_\sigma(\theta) = \mathbb{E}^{\pi^\theta,\sigma}\big[\sum_{t=0}^{T-1} Q_t^{\sigma,\theta}(X_t, a_t) \nabla_\theta \log \pi_t^\theta(X_t, a_t)\big]$ is the classical policy gradient of the σ-MDP.

(iii) $J_\sigma(\theta) - J(\theta) = \mathbb{E}^{\pi^\theta,\sigma}\big[\sum_{t=0}^{T-1} g_t^{\sigma,\theta}(X_t, a_t)\big]$ for every $\sigma \in \Sigma$ and $\theta \in \Theta$, under the path law of $(\mu_0, \pi^\theta, \sigma)$. In particular $J_\sigma(\theta) \leq J(\theta) + T\zeta$ for every $\sigma \in S_\zeta(\theta)$.

(iv) There are C_0, C_1, C_2, depending only on the model data and B, such that all J_σ and J are C_0-Lipschitz on Θ in the norm $\|\cdot\|_\infty$, $|g_t^{\sigma,\theta}(x, a) - g_t^{\sigma,\theta'}(x, a)| \leq C_1 \|\theta - \theta'\|_\infty$ for all $\sigma \in \Sigma$, $(t, x, a) \in \mathcal{T} \times \mathcal{X} \times A$ and $\theta, \theta' \in \Theta$, and $|J_\sigma(\theta + hr) - J_\sigma(\theta) - h\langle \nabla_\theta J_\sigma(\theta), r \rangle| \leq C_2 h^2$ for all $\sigma \in \Sigma$, $\|r\|_\infty \leq 1$ and $\theta, \theta + hr \in \Theta$.

Proof. (i) The representation

$$J_\sigma(\theta) = \sum_{(x_0,a_0,\dots,x_T)} \mu_0(x_0) \prod_{t=0}^{T-1} \pi_t^\theta(x_t,a_t)\,\sigma(t,x_t,a_t)(\{x_{t+1}\}) \times \Big[\textstyle\sum_{t=0}^{T-1} f(t,x_t,a_t,x_{t+1}) + g(x_T)\Big]$$

is a finite sum and $\pi_t^\theta(x_t,a_t)$ is C^2. Hence $J_\sigma \in C^2$. Backward induction gives $V_t^\theta \le V_t^{\sigma,\theta}$. Indeed $G_t^\theta(x,a) \le \langle \sigma(t,x,a), f + V_{t+1}^\theta\rangle \le \langle \sigma(t,x,a), f + V_{t+1}^{\sigma,\theta}\rangle$, and averaging over π^θ preserves the inequality, so $J \le \min_{\sigma\in\Sigma} J_\sigma$. Conversely, (14) reaches its minimum at a vertex, so some $\sigma^* \in \Sigma$ has $g^{\sigma^*,\theta} \equiv 0$, and backward induction gives $V^{\sigma^*,\theta} = V^\theta$. Hence $J_{\sigma^*}(\theta) = J(\theta)$, $S_0(\theta) \neq \emptyset$, and $S_0(\theta) \subseteq \Sigma^*(\theta)$.

(ii) For $\sigma \in \Sigma^*(\theta)$, $J(\theta+hr) \le J_\sigma(\theta+hr) = J(\theta) + h\langle \nabla J_\sigma(\theta), r\rangle + o(h)$, so the upper limit of the quotient as $h \downarrow 0$ is at most the stated minimum. For the lower bound, set $\ell := \liminf_{h\downarrow 0} \big(J(\theta + hr) - J(\theta)\big)/h$ and choose a sequence $(h_n)_{n\ge 0}$ with $h_n \downarrow 0$ along which the quotient converges to ℓ. For each $n \ge 0$, (i) applied at $\theta + h_n r$ provides $\sigma_n \in \Sigma$ with $J_{\sigma_n}(\theta + h_n r) = J(\theta + h_n r)$. Since Σ is finite, some $\sigma \in \Sigma$ satisfies $\sigma_n = \sigma$ for infinitely many $n \ge 0$, and we pass to the corresponding subsequence, still denoted $(h_n)_{n\ge 0}$, so that $J_\sigma(\theta + h_n r) = J(\theta + h_n r)$ for all $n \ge 0$. Both J_σ and J are continuous on Θ, the first being C^2 and the second a finite minimum of continuous functions, so letting $n \to \infty$ in this identity gives $J_\sigma(\theta) = J(\theta)$, that is $\sigma \in \Sigma^*(\theta)$. Along the subsequence the two difference quotients therefore coincide, and differentiability of J_σ gives

$$\frac{J(\theta + h_n r) - J(\theta)}{h_n} = \frac{J_\sigma(\theta + h_n r) - J_\sigma(\theta)}{h_n} \xrightarrow[n\to\infty]{} \big\langle \nabla_\theta J_\sigma(\theta), r\big\rangle.$$

Hence $\ell = \langle \nabla_\theta J_\sigma(\theta), r\rangle \ge \min_{\sigma'\in\Sigma^*(\theta)} \langle \nabla_\theta J_{\sigma'}(\theta), r\rangle$. Together with the upper bound on the upper limit, the quotient converges and $D_\theta^+ J(\theta)[r]$ equals the stated minimum. The left formula follows from the same argument applied to $h \uparrow 0$, with minima replaced by maxima.

(iii) With $\Delta_t := V_t^{\sigma,\theta} - V_t^\theta$ for $t \in \mathcal{T}$ and $\Delta_T \equiv 0$,

$$\Delta_t(x) = \mathbb{E}_{a\sim\pi_t^\theta(x,\cdot)}\big[g_t^{\sigma,\theta}(x,a) + \langle \sigma(t,x,a), \Delta_{t+1}\rangle\big],$$

which telescopes along the (π^θ, σ)-flow.

(iv) Finitely many C^2 functions on the compact Θ, over the finite Σ, give C_0 and C_2. The gap $g_t^{\sigma,\theta}$ depends on θ through V_{t+1}^θ and G_t^θ, which are finite minima of C_0-Lipschitz functions, and this gives C_1. □

Algorithm 1 in the main text is the exact ascent scheme analyzed below.

Lemma C.2. *On an ascent step, $\theta_{k+1} \in \Theta$ and $J(\theta_{k+1}) \ge J(\theta_k) + \frac{1}{2}h_k\Delta_k > J(\theta_k) + \frac{1}{2}h_k\zeta_k$.*

Proof. Feasibility holds since $r_k \in T_{\zeta_k}(\theta_k)$ and $h_k \le \zeta_k$. Write θ, h, r for θ_k, h_k, r_k. Choose $\sigma_h \in \Sigma$ with $J_{\sigma_h}(\theta + hr) = J(\theta + hr)$ and $\sigma^\star \in S_0(\theta)$, both of which exist by Theorem C.1(i), write $\mathbb{P}^h$ for the path law of $(\mu_0, \pi^{\theta+hr}, \sigma_h)$, and call $(t,x) \in \mathcal{T}\times\mathcal{X}$ reached if $\mathbb{P}^h(X_t = x) > 0$. Since $\pi_t^{\theta+hr}(x,a) > 0$ for all $(t,x,a) \in \mathcal{T}\times\mathcal{X}\times A$, one has $\mathbb{P}^h(X_t = x, a_t = a) > 0$ for every reached (t,x) and every $a \in A$. Theorem C.1(iii) applied at $\theta + hr$ gives $\mathbb{E}^{\mathbb{P}^h}\big[\sum_{t=0}^{T-1} g_t^{\sigma_h,\theta+hr}(X_t,a_t)\big] = J_{\sigma_h}(\theta+hr) - J(\theta+hr) = 0$, and the summands being nonnegative, $g_t^{\sigma_h,\theta+hr}(x,a) = 0$ whenever $\mathbb{P}^h(X_t = x, a_t = a) > 0$, that is at every reached triple. Define

$$\sigma'(t,x,a) := \begin{cases} \sigma_h(t,x,a) & \text{if } (t,x) \text{ is reached,} \\ \sigma^\star(t,x,a) & \text{otherwise.} \end{cases} \in \Sigma.$$

The marginals $\mathbb{P}^h(X_t = \cdot)$ are determined recursively by the kernels at reached triples only, so the path law of $(\mu_0, \pi^{\theta+hr}, \sigma')$ equals $\mathbb{P}^h$ and $J_{\sigma'}(\theta+hr) = J(\theta+hr)$. Moreover $\sigma' \in S_{\zeta_k}(\theta)$, since $g_t^{\sigma',\theta}(x,a) = g_t^{\sigma^\star,\theta}(x,a) = 0$ at unreached triples, while at reached triples (iv) gives $g_t^{\sigma',\theta}(x,a) \le g_t^{\sigma',\theta+hr}(x,a) + C_1 h = C_1 h \le \zeta_k$. Hence, by (iv) and $J_{\sigma'}(\theta) \ge J(\theta)$,

$$J(\theta+hr) = J_{\sigma'}(\theta+hr) \;\ge\; J_{\sigma'}(\theta) + h\big\langle \nabla_\theta J_{\sigma'}(\theta), r\big\rangle - C_2h^2 \;\ge\; J(\theta) + h\Delta_k - C_2h^2,$$

where the last step uses $\sigma' \in S_{\zeta_k}(\theta)$ and the definition of Δ_k as a max–min. Finally $C_2 h \le \Delta_k/2$ by the choice of h_k, and $\Delta_k > \zeta_k$ on ascent steps.

□

Definition C.3. A point $\bar\theta \in \Theta$ is directionally stationary if $D_\theta^+ J(\bar\theta)[r] \le 0$ for all $r \in T_0(\bar\theta)$. For $\zeta \ge 0$, a point $\theta \in \Theta$ is ζ-stationary if for every $r \in T_\zeta(\theta)$ there exists $\sigma \in S_\zeta(\theta)$ with $\langle \nabla_\theta J_\sigma(\theta), r\rangle \le \zeta$.

Proof of Theorem 3.8.

Proof. J is bounded on Θ, and each ascent step at tolerance ζ gains at least

$$\tfrac{1}{2}\min\{\zeta, \zeta/C_1, \zeta/(2C_2)\}\zeta > 0$$

by Lemma C.2. Each level therefore holds finitely many ascent steps, refinements are infinite, and $\zeta_k \to 0$.

Let $(\theta_{k_j})_{j\ge 0}$ be refinement iterates with $\theta_{k_j} \to \bar\theta$ as $j \to \infty$, so that $\Delta_{k_j} \le \zeta_{k_j}$ for all $j \ge 0$, and fix $r \in T_0(\bar\theta)$. Two facts hold for all large j. First, $r \in T_{\zeta_{k_j}}(\theta_{k_j})$: if $\theta_{k_j,i} \ge B - \zeta_{k_j}$ along a subsequence then $\bar\theta_i = B$ and $r_i \le 0$ as required, while coordinates with $\bar\theta_i < B$ are eventually unconstrained, the lower bound being symmetric. Second, $S_{\zeta_{k_j}}(\theta_{k_j}) \subseteq \Sigma^*(\bar\theta)$: any $\sigma \notin \Sigma^*(\bar\theta)$ has some gap $g_t^{\sigma,\bar\theta}(x,a) \ge 2\delta > 0$, for otherwise $\sigma \in S_0(\bar\theta) \subseteq \Sigma^*(\bar\theta)$ by (i), so $g_t^{\sigma,\theta_{k_j}}(x,a) \ge 2\delta - C_1\|\theta_{k_j} - \bar\theta\|_\infty > \zeta_{k_j}$ eventually, and the claim follows from finiteness of Σ. Combining,

$$\zeta_{k_j} \;\ge\; \Delta_{k_j} \;\ge\; \min_{\sigma\in S_{\zeta_{k_j}}(\theta_{k_j})} \big\langle \nabla_\theta J_\sigma(\theta_{k_j}), r\big\rangle \;\ge\; \min_{\sigma\in\Sigma^*(\bar\theta)} \big\langle \nabla_\theta J_\sigma(\theta_{k_j}), r\big\rangle \;\xrightarrow[j\to\infty]{}\; D_\theta^+ J(\bar\theta)[r].$$

The second inequality holds because r is feasible for the direction problem by the first fact, the third because a minimum over a subset dominates the minimum over the superset by the second fact, and the limit follows from continuity of the finitely many gradients and (ii). □

Corollary C.4. *Let $\kappa := \min\{1, 1/C_1, 1/(2C_2)\}$ and $B_J := \sup_\Theta J - J(\theta_0)$.*

- *(i) Every accumulation point of $(\theta_k)_{k\ge0}$ has objective value $J_\infty := \lim_{k\to\infty} J(\theta_k)$.*
- *(ii) Each ascent step at tolerance ζ gains at least $(\kappa/2)\zeta^2$. Hence, for every $i \ge 0$, at most $2B_J/(\kappa\nu^{2i}\zeta_0^2)$ ascent steps occur at level $\nu^i\zeta_0$, and the refinement ending this level occurs by iteration $(i+1) + \frac{2B_J}{\kappa(1-\nu^2)}(\nu^i\zeta_0)^{-2}$.*
- *(iii) At every refinement iteration k, the iterate θ_k is ζ_k-stationary. For every $\zeta \in (0, \zeta_0]$, a ζ-stationary iterate is produced within $\lceil \log(\zeta_0/\zeta)/\log(1/\nu)\rceil + 1 + \frac{2B_J}{\kappa(1-\nu^2)\nu^2\zeta^2}$ iterations.*

Proof. (i) $J(\theta_k) \uparrow J_\infty$ as $k \to \infty$ and J is continuous. (ii) On such a step $\Delta_k > \zeta$ and $h_k \ge \kappa\zeta$, so Lemma C.2 gives the gain. With A_l the number of ascent steps at level $l \ge 0$, the bound $\sum_{l\ge0} A_l(\kappa/2)(\nu^l\zeta_0)^2 \le B_J$ gives $A_i \le 2B_J/(\kappa\nu^{2i}\zeta_0^2)$ and $\sum_{l=0}^i A_l \le \frac{2B_J}{\kappa(1-\nu^2)(\nu^i\zeta_0)^2}$, and the refinement ending level i follows these ascent steps and the $i+1$ earlier refinements. (iii) At a refinement, $\Delta_k \le \zeta_k$, and since Δ_k is the maximum over $r \in T_{\zeta_k}(\theta_k)$ of $\min_{\sigma\in S_{\zeta_k}(\theta_k)}\langle \nabla_\theta J_\sigma(\theta_k), r\rangle$, every such r admits $\sigma \in S_{\zeta_k}(\theta_k)$ with $\langle \nabla_\theta J_\sigma(\theta_k), r\rangle \le \zeta_k$, which is ζ_k-stationarity. For the count, take $i := \lceil \log(\zeta_0/\zeta)/\log(1/\nu)\rceil$, so that $\nu\zeta < \nu^i\zeta_0 \le \zeta$. By (ii) the refinement ending level i occurs within the stated bound, and its iterate is $\nu^i\zeta_0$-stationary, hence ζ-stationary. □

Enumerating the finite vertex sets and all selections in $S_\zeta(\theta)$ is an exact, potentially expensive implementation. A reduced working set is equivalent only if an exact separation procedure verifies every omitted branch constraint. Keeping a heuristic subset does not inherit Theorem 3.8. The constants in the step can be any strictly positive global upper bounds satisfying Theorem C.1(iv).

D SAMPLED ACTOR–CRITIC IMPLEMENTATION

The following implementation estimates the sensitivity recursion used in the reported experiments. It takes a candidate direction as input, unlike Algorithm 1, which chooses a direction through a complete finite linear program. Its Monte Carlo and function-approximation errors are not covered by the exact convergence result.

Corollary 3.7 is one-sided and direction-wise: for any feasible direction r, the scalar sensitivities $D_\theta^+ V_t^\theta(\cdot)[r]$ satisfy a backward recursion whose only nonclassical step is optimization over the worst-case laws. Algorithm 2 implements this recursion with a value critic $V_{\psi,t} \approx V_t^\theta$ and a scalar sensitivity critic $U_{\xi,t} \approx D_\theta^+ V_t^\theta(\cdot)[r]$. In finite state spaces and the sample-supported setting of Section 4.2, $D_\theta^+ G_t^\theta[r]$ is computed exactly for the supplied downstream critic and empirical model by optimizing over the optimal face $\widehat{\mathcal{F}}_t(x_t, a_t)$ of the transport program (14). Without ties, this reduces to averaging over the optimal transport map. For neural policies, the score term is computed by a Jacobian–vector product, so no d_θ-dimensional sensitivity is materialized. Line 10 estimates the directional derivative of $J(\theta) = \mathbb{E}_{X_0 \sim \mu_0}[V_0^\theta(X_0)]$ along r, and checks the estimated sign; this is a certificate only with exact evaluation. At $\varepsilon = 0$, the method reduces to a standard finite-horizon actor–critic.

Algorithm 2 Sampled Robust Actor–Critic Algorithm

Require: Policy π^θ, feasible direction r with $\|r\| = 1$, value critics $V_{\psi,t}$, scalar sensitivity critics $U_{\xi,t} : \mathcal{X} \to \mathbb{R}$, samples $a_t \sim \pi_t^\theta(x_t, \cdot)$ and $X_{t+1}^{(i)} \sim \mathbb{P}_t^0(x_t, a_t, \cdot)$, $i = 1, \dots, N$, dual bound Λ, step size η_θ, initial law μ_0

1: Initialize $V_{\psi,T} = g$ and $U_{\xi,T} = 0$

2: **for** $t = T-1, T-2, \dots, 0$ **do**

3: Compute

$$\widehat{G}_t^\theta(x_t, a_t) = \sup_{\lambda \in [0,\Lambda]} \left\{ \frac{1}{N} \sum_{i=1}^N \min_{y \in \mathcal{X}} \left[f(t, x_t, a_t, y) + V_{\psi,t+1}(y) + \lambda c(X_{t+1}^{(i)}, y) \right] - \varepsilon^q \lambda \right\},$$

and let $\widehat{\lambda}_t^\star(x_t, a_t)$ be a maximizer.

4: Update $V_{\psi,t}$ by minimizing $\mathcal{L}_V(\psi) = \mathbb{E}\left[\left(V_{\psi,t}(x_t) - \widehat{G}_t^\theta(x_t, a_t) \right)^2 \right]$.

5: Compute the optimal face $\widehat{\mathcal{F}}_t(x_t, a_t)$ of (14) with source $\frac{1}{N} \sum_i \delta_{X_{t+1}^{(i)}}$, cost

$$H(y) = f(t, x_t, a_t, y) + V_{\psi,t+1}(y),$$

and budget ε^q. If each $y_i^\star \in \operatorname{argmin}_y \left\{ H(y) + \widehat{\lambda}_t^\star c(X_{t+1}^{(i)}, y) \right\}$ is unique, the face reduces to the coupling $i \mapsto y_i^\star$.

6: Compute

$$\widehat{D^+ G_t}(x_t, a_t)[r] = \min_{\gamma \in \widehat{\mathcal{F}}_t(x_t, a_t)} \sum_{i,y} \gamma(i, y)\, U_{\xi,t+1}(y),$$

which reduces to $\frac{1}{N} \sum_i U_{\xi,t+1}(y_i^\star)$ without ties.

7: Form

$$\widehat{z}_t = \widehat{G}_t^\theta(x_t, a_t) \left\langle \nabla_\theta \log \pi_t^\theta(x_t, a_t), r \right\rangle + \widehat{D^+ G_t}(x_t, a_t)[r].$$

8: Update $U_{\xi,t}$ by minimizing $\mathcal{L}_U(\xi) = \mathbb{E}\left[\left(U_{\xi,t}(x_t) - \widehat{z}_t \right)^2 \right]$.

9: **end for**

10: Draw $X_0^{(j)} \sim \mu_0$, $j = 1, \dots, M$, and set

$$\widehat{D^+ J}[r] = \frac{1}{M} \sum_j U_{\xi,0}(X_0^{(j)}).$$

If $\widehat{D^+ J}[r] > 0$, update $\theta \leftarrow \Pi_\Theta(\theta + \eta_\theta r)$.

Remark D.1. Under the uniqueness hypothesis of Proposition B.3, $D_\theta^+ V_s^\theta(x)[r] = \langle \nabla_\theta V_s^\theta(x), r\rangle$, so Algorithm 2 can be run with a vector-valued sensitivity critic $U_{\xi,t} : \mathcal{X} \to \mathbb{R}^{d_\theta}$, $U_{\xi,t} \approx \nabla_\theta V_t^\theta$, using target $\widehat{z}_t = \widehat{G}_t^\theta \nabla_\theta \log \pi_t^\theta + \frac{1}{N}\sum_i U_{\xi,t+1}(y_i^\star)$. This vector specialization uses one backward sweep for all directions but materializes d_θ outputs per state. At a tie point, an arbitrary active coupling need not yield the gradient, so its normalized output is only a candidate direction. Algorithm 2 instead costs one backward sweep per candidate direction and never forms a d_θ-vector. Ties are generic at the exact dual maximizer (13–35% of transport cells, Appendix E.10). There the face recursion in line 6 returns the exact one-sided derivatives, whereas an arbitrary transport selection can be inaccurate or wrong-signed (Appendix E.5). The certificate in line 10 guarantees the sign of the directional step, not a speed-up (Appendix E.6). For unrestricted continuous ambiguity, line 5 requires nonconvex global minimization and the actor update is therefore heuristic.

E Additional Experiments

E.1 Exact finite-model evaluation

These experiments test Algorithm 1 directly and separate it from the candidate-direction implementation of Algorithm 2. We report the safeguarded near-active scheme separately from a near-active variant using line search (LS). The latter uses the same line search as the exact-inner robust policy-gradient (PG) baseline and is not covered by the safeguarded-step convergence theorem. Methods share model instances, initializations, policy classes, and exact evaluation routines.

E.1.1 Transport ties and directional derivatives

We consider both unique and non-unique worst-case laws, including cases with different directional sensitivities. A single optimal coupling fails only when active laws have different sensitivities, while a pointwise dual-argmin selection can fail even with a unique worst-case law because a tied dual destination need not define a budget-feasible transport plan. Thus, nonuniqueness alone does not imply nondifferentiability. Different active sensitivities require the full directional envelope.

Table 2: Median absolute one-step derivative error across the four tie regimes, with wrong-sign rates in parentheses.

Regime	n	Face	One coupling	Dual argmin	Pointwise min.	Nominal
R1	151	2.29×10^{-11} (0%)	2.29×10^{-11} (0%)	1.02×10^{-1} (8%)	1.02×10^{-1} (8%)	2.65×10^{-1} (18%)
R2	22	1.72×10^{-11} (0%)	1.72×10^{-11} (0%)	1.72×10^{-11} (0%)	1.72×10^{-11} (0%)	2.07×10^{-1} (14%)
R3	5	1.44×10^{-11} (0%)	1.44×10^{-11} (0%)	1.44×10^{-11} (0%)	5.00×10^{-1} (0%)	5.00×10^{-1} (100%)
R4	32	1.49×10^{-11} (0%)	9.60×10^{-2} (3%)	1.41×10^{-1} (9%)	8.04×10^{-2} (12%)	4.11×10^{-1} (28%)

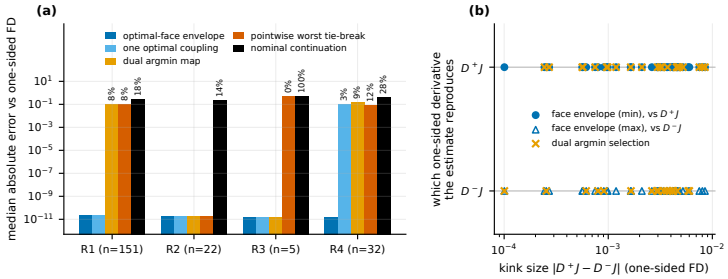


Figure 3: Correctness at transport ties. (a) Median absolute error against one-sided finite differences by regime. (b) Policy-level checks at 40 kinks.

E.1.2 SAFEGUARDED ASCENT AND REFINEMENT

We evaluate the exact outer loop on two-state, two-action models with $T \in \{2, 3, 4\}$, using tabular and time-shared policies. The safeguarded near-active scheme has no objective decreases and no violations of the gain inequality in Lemma C.2. For time-shared policies, the active-only variant without line search has about 1,800 objective decreases per run at $T = 3$ and $T = 4$, averaged over 20 seeds, while line search removes these decreases (Table 3).

The refinement traces follow the decreasing tolerance schedule (Figure 4). The conservative global bounds lead to small steps, so these experiments validate the ascent mechanism rather than practical speed or global optimality.

Table 3: Random two-state, two-action models with time-shared policies and zero initialization (20 seeds). Counts, iterations, and DP gaps are per-run means.

T	Method	Gain-inequality violations	Objective decreases	Iterations	DP gap
2	Near-active, certified step	0	0	11832	0.1715
2	Active-only, no line search	0	0	11809	0.1715
2	Active-only + line search	–	0	17	0.1712
3	Near-active, certified step	0	0	20001	0.3140
3	Active-only, no line search	1316	1789	20001	0.3138
3	Active-only + line search	–	0	26	0.3128
4	Near-active, certified step	0	0	19203	0.5802
4	Active-only, no line search	1906	1803	19188	0.5792
4	Active-only + line search	–	0	107	0.5765

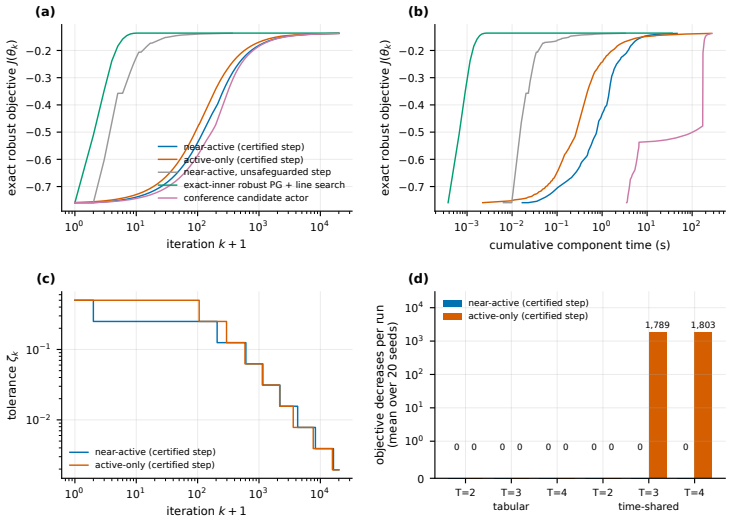


Figure 4: Exact near-active ascent. (a,b) Objective over iterations and cumulative time. (c) Tolerance schedule. (d) Objective decreases across 20 seeds.

E.1.3 NUMERICAL, SAMPLING, AND CRITIC ERROR

On frozen models, refining the scalar-dual grid reduces dual-value error but does not eliminate the derivative error from a single argmin selection, while the optimal-face reference has zero reported derivative error. Increasing the per-row sample size improves the empirical-model approximation to the population objective. Under critic-label noise, false-positive ascent signs remain rare at the tested noise levels (Figure 5). These experiments separate solver, sampling, and critic errors without treating an approximate sensitivity as an exact ascent certificate.

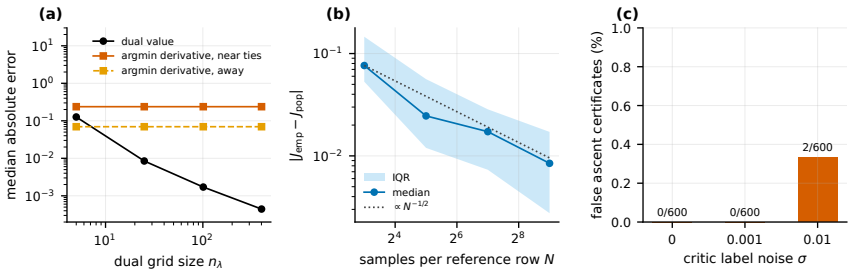


Figure 5: Error-source diagnostics. (a) Dual-grid approximation. (b) Empirical-model approximation with varying per-row sample size. (c) Critic-label noise.

E.2 A SUPPLY CHAIN EXAMPLE

We also consider a finite-horizon inventory control problem with state space $\mathcal{X} := \{0, 1 \ldots, n\}$ where $X_t \in \mathcal{X}$ denotes the on-hand inventory at time $t \in \mathcal{T}$. The action space is $A := \{0, 1, \ldots, n\}$ where the A-valued random variable a_t denotes the order quantity placed at time t. Given a state-action pair $(x, a) \in \mathcal{X} \times A$, the post-inventory is defined as $\bar{x} := \min(n, x + a) \in \mathcal{X}$ and the demand at time $t \in \mathcal{T}$ is represented through a random variable $D_t \sim \text{Unif}(\{0, 1, \ldots, n\})$ such that the next inventory evolves as

$$X_{t+1} = \max(0, \bar{x} - D_t).$$

It falls into our framework by considering the reference kernel $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}_t^0(x, a, \cdot) = m(x, a, \cdot) \sharp \mathcal{U}(\{0, 1, \ldots, n\})$ where $\sharp$ denotes the pushforward map and where we set $\mathcal{X} \times A \times \{0, 1, \ldots, n\} \ni (x, a, d) \mapsto m(x, a, d) := \max(0, \bar{x} - d)$. The reward function f is defined by

$$f(t, x, a, d) := -\Big(h(\bar{x} - d)_+ + p(d - \bar{x})_+ + k\mathbb{1}_{a>0}\Big), \quad \forall (t, x, a, d) \in \mathcal{T} \times \mathcal{X} \times A \times \{0, 1, \ldots, n\}$$

while $g(x) = 0$ for any $x \in \mathcal{X}$. Here, h represents an holding-cost coefficient, $p > 0$ is a shortage cost coefficient and k is a fixed ordering cost. Since the reward function is expressed in terms of (x, a, x'), we equivalently work under the conditional expectation of the reward given the realized next-state inventory level $x' \in \mathcal{X}$, i.e by defining $\tilde{f}(t, x, a, x') := \mathbb{E}\Big[f(t, x, a, D_t) | X_{t+1} = x'\Big]$. The robustness is modeled through the Wasserstein ambiguity set around the reference transition kernels $(\mathbb{P}_t^0)_{t \in \mathcal{T}}$ with the cost $c(x, y) := |x - y|$ on the state space $\mathcal{X}$. In Figure 6, for the learnt policy $\pi_{\theta^\star}$, we plot the greedy policy for each state $x \in \mathcal{X}$ at $t = 0$ (left), and the value V_0 of the learned policy under the reference model, under Wasserstein ambiguity, and under a Kullback–Leibler ambiguity ball of the same radius around $\mathbb{P}_t^0$ computed by exact robust dynamic programming for comparison (right); the KL variant is reported only as a reference value and is not covered by our policy-gradient theory.

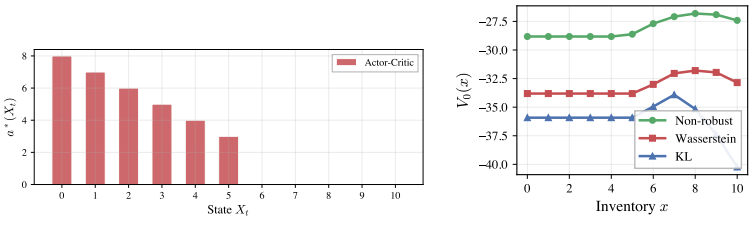


Figure 6: Learned greedy ordering policy $a^*(x) = \underset{a \in A}{\text{argmax}}\, \pi_{\theta^\star}(x, a)$ at $t = 0$ (left); value function $V_0(x)$ for the learned policy under Wasserstein and KL ambiguity, and the non-robust setting (right).

E.3 EXACT DYNAMIC-PROGRAMMING RECOVERY AND RESIDUAL MISMATCH

The experimental labels "grid dual/argmin", "exact dual/optimal coupling", and "direction-wise optimal-face recursion" denote distinct implementations. A single feasible optimal coupling need

not give the exact directional derivative at a heterogeneous tie. The optimal-face envelope is needed there. None of these solver labels implies use of the near-active outer loop of Algorithm 1.

Table 4 compares Algorithm 2 with exact robust DP over 10 seeds, both for the published grid-based implementation and for a variant using the exact dual and the optimal coupling of (14). The 101-point dual grid under-estimates the exact inner problem by up to 2×10^{-2} per stage ($\|V^{\mathrm{grid}} - V^{\mathrm{exact}}\|_\infty = 0.13$ on the supply chain) and can change the recovered policy.

Most discrepancies with exact DP disappear when the exact dual and optimal coupling are used. The remaining mismatches occur either in a near-tie regime or at cells with zero worst-case occupancy. In particular, for the supply chain at $\varepsilon = 2$, the four mismatched cells have zero worst-case occupancy, so $\Delta_\pi^{>0} = 0$, although the greedy value error is 0.0048. At $\varepsilon = 0.5$, the remaining supply-chain discrepancies have very small exact action gaps, below 1.1×10^{-3}.

We report the greedy value error in addition to the policy mismatch because it measures the performance effect of the learned greedy policy, whereas the softmax value error also reflects the actor's remaining temperature after the finite iteration budget of Table 18.

Table 4: Exact robust-DP recovery over 10 seeds, with exact DP as the reference. $\Delta_\pi^{>0}$ is the policy mismatch restricted to cells with positive worst-case occupancy; "zero-occ." is the mean number of cells with zero worst-case occupancy.

Environment	method	ε	Δ_π	$\Delta_\pi^{>0}$	zero-occ. cells	Δ_V (softmax)	Δ_V (greedy)
Coin toss	Algorithm 1 (grid, argmin)	0.5	0.1273 ± 0.0000	0.1273 ± 0.0000	0.0/110	0.0247 ± 0.0001	0.0208 ± 0.0000
Coin toss	Algorithm 1 (grid, argmin)	1.0	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0/110	0.0144 ± 0.0001	0.0000 ± 0.0000
Coin toss	Algorithm 1 (grid, argmin)	2.0	0.0000 ± 0.0000	0.0000 ± 0.0000	32.3/110	0.0173 ± 0.0003	0.0000 ± 0.0000
Coin toss	exact dual + optimal coupling	0.5	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0/110	0.0080 ± 0.0001	0.0000 ± 0.0000
Coin toss	exact dual + optimal coupling	1.0	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0/110	0.0121 ± 0.0000	0.0000 ± 0.0000
Coin toss	exact dual + optimal coupling	2.0	0.0000 ± 0.0000	0.0000 ± 0.0000	26.5/110	0.0173 ± 0.0002	0.0000 ± 0.0000
Supply chain	Algorithm 1 (grid, argmin)	0.5	0.0173 ± 0.0016	0.0173 ± 0.0016	0.0/110	0.0364 ± 0.0001	0.0002 ± 0.0000
Supply chain	Algorithm 1 (grid, argmin)	1.0	0.0000 ± 0.0000	0.0000 ± 0.0000	32.0/110	0.0809 ± 0.0003	0.0000 ± 0.0000
Supply chain	Algorithm 1 (grid, argmin)	2.0	0.0345 ± 0.0012	0.0000 ± 0.0000	9.6/110	0.0526 ± 0.0001	0.0048 ± 0.0000
Supply chain	exact dual + optimal coupling	0.5	0.2845 ± 0.0024	0.2845 ± 0.0024	0.0/110	0.0167 ± 0.0000	0.0006 ± 0.0000
Supply chain	exact dual + optimal coupling	1.0	0.0000 ± 0.0000	0.0000 ± 0.0000	14.3/110	0.0199 ± 0.0000	0.0000 ± 0.0000
Supply chain	exact dual + optimal coupling	2.0	0.0000 ± 0.0000	0.0000 ± 0.0000	14.9/110	0.0306 ± 0.0000	0.0000 ± 0.0000

The residual mismatch is not caused by a transport tie. For the supply chain at $\varepsilon = 2$, transport ties and positive-dimensional optimal faces do occur, as predicted by Theorem 3.6 and Remark B.1. However, the four policy mismatches occur at $(t, x) = (7, 1), \ldots, (7, 4)$ and are not tie-induced. Their worst-case occupancy is zero, so their policy-gradient contribution vanishes and their actions do not affect the robust objective. The learned policy agrees with its exact continuation on every reachable cell.

E.4 Role of the transported continuation

We compare four actor targets that differ in the continuation term:

(a) the full recursion

$$\widehat{z}_t = \widehat{G}_t \nabla_\theta \log \pi_t + \frac{1}{N} \sum_i U_{\xi,t+1}(y_i^\star);$$

(b) the same recursion with nominal continuation

$$\frac{1}{N} \sum_i U_{\xi,t+1}(X_{t+1}^{(i)});$$

(c) the score-only target;

(a′) the full recursion with the exact dual and optimal coupling of (14).

We evaluate all variants on the coin-toss and supply-chain problems with the hyper-parameters of Table 18, $\varepsilon \in \{0, 0.5, 1, 2\}$, and 10 seeds, against exact robust DP (Table 5).

At $\varepsilon = 0$, variants (a), (b), and (a′) coincide and recover the non-robust DP solution exactly. For $\varepsilon > 0$, the score-only target fails because it does not propagate continuation sensitivity to later policy parameters.

The full recursion recovers the grid-DP policy and therefore inherits the discretisation effects discussed in Appendix E.3. With tabular softmax policies, the nominal continuation can reach the same fixed points. For the parameter block corresponding to action b at (t, x), the gradient is

$$d_t(x)\,\pi_t(b|x)\big(\widehat{G}_t(x,b) - V_t(x)\big),$$

so the transported continuation changes the occupancy weight but not the sign pattern determining the greedy fixed point.

This does not imply that the nominal continuation is the robust policy gradient. Appendix E.5 shows that it produces a different gradient.

Variant (a′) recovers the exact-DP policy on the coin toss at $\varepsilon = 0.5$ and on the supply chain at $\varepsilon = 2$. Its only remaining discrepancy is the supply-chain near-tie regime at $\varepsilon = 0.5$, discussed in Appendix E.3.

Table 5: Ablation of the actor target against exact robust DP (10 seeds): (a) full recursion, (b) nominal continuation, (c) score only, and (a′) exact dual with optimal coupling.

Environment	method	ε	Δ_π	Δ_V (softmax)	Δ_V (greedy)
Coin toss	(a) full	0.5	0.1273 ± 0.0000	0.0247 ± 0.0001	0.0208 ± 0.0000
Coin toss	(a) full	1.0	0.0000 ± 0.0000	0.0144 ± 0.0001	0.0000 ± 0.0000
Coin toss	(a) full	2.0	0.0000 ± 0.0000	0.0173 ± 0.0003	0.0000 ± 0.0000
Coin toss	(b) nominal cont.	0.5	0.1273 ± 0.0000	0.0218 ± 0.0000	0.0208 ± 0.0000
Coin toss	(b) nominal cont.	1.0	0.0000 ± 0.0000	0.0067 ± 0.0000	0.0000 ± 0.0000
Coin toss	(b) nominal cont.	2.0	0.0000 ± 0.0000	0.0068 ± 0.0000	0.0000 ± 0.0000
Coin toss	(c) score only	0.5	0.6082 ± 0.0127	4.6996 ± 0.0014	6.4648 ± 0.1782
Coin toss	(c) score only	1.0	0.6045 ± 0.0129	5.8560 ± 0.0022	8.1720 ± 0.1184
Coin toss	(c) score only	2.0	0.6082 ± 0.0130	6.0178 ± 0.0017	8.9569 ± 0.0186
Coin toss	(a′) exact dual + coupling	0.5	0.0000 ± 0.0000	0.0080 ± 0.0001	0.0000 ± 0.0000
Coin toss	(a′) exact dual + coupling	1.0	0.0000 ± 0.0000	0.0121 ± 0.0000	0.0000 ± 0.0000
Coin toss	(a′) exact dual + coupling	2.0	0.0000 ± 0.0000	0.0173 ± 0.0002	0.0000 ± 0.0000
Supply chain	(a) full	0.5	0.0173 ± 0.0016	0.0364 ± 0.0001	0.0002 ± 0.0000
Supply chain	(a) full	1.0	0.0000 ± 0.0000	0.0809 ± 0.0003	0.0000 ± 0.0000
Supply chain	(a) full	2.0	0.0345 ± 0.0012	0.0526 ± 0.0001	0.0048 ± 0.0000
Supply chain	(b) nominal cont.	0.5	0.0000 ± 0.0000	0.0106 ± 0.0000	0.0000 ± 0.0000
Supply chain	(b) nominal cont.	1.0	0.0000 ± 0.0000	0.0113 ± 0.0000	0.0000 ± 0.0000
Supply chain	(b) nominal cont.	2.0	0.0000 ± 0.0000	0.0220 ± 0.0000	0.0000 ± 0.0000
Supply chain	(c) score only	0.5	0.8864 ± 0.0071	23.4938 ± 0.0070	30.7401 ± 2.3119
Supply chain	(c) score only	1.0	0.8864 ± 0.0071	24.4900 ± 0.0075	32.2519 ± 2.4022
Supply chain	(c) score only	2.0	0.8818 ± 0.0051	26.5472 ± 0.0070	33.8977 ± 2.1274
Supply chain	(a′) exact dual + coupling	0.5	0.2845 ± 0.0024	0.0167 ± 0.0000	0.0006 ± 0.0000
Supply chain	(a′) exact dual + coupling	1.0	0.0000 ± 0.0000	0.0199 ± 0.0000	0.0000 ± 0.0000
Supply chain	(a′) exact dual + coupling	2.0	0.0000 ± 0.0000	0.0306 ± 0.0000	0.0000 ± 0.0000

E.5 Gradient check against an exact reference

We compare the analytic recursion with finite differences of the exact robust objective

$$J(\theta) = \mu_0^\top V_0^\theta,$$

where μ_0 denotes the initial-state distribution. On the coin-toss problem at $\varepsilon = 1$, using 20 random parameters $\theta \sim 0.5\,\mathcal{N}(0, I)$, the λ-grid recursion is consistent with the corresponding grid-discretised objective but not with the exact robust objective (Table 6, Figure 7). Solving the dual exactly while selecting a single transport argmin still gives a relative error of about 0.25. Using the optimal coupling reduces the error to approximately 3×10^{-9}.

The discrepancy is due to transport ties. At a breakpoint of the dual, several transport destinations may be optimal, and the derivative is computed under an optimal coupling over the full optimal face. Selecting a single tied destination need not recover the correct robust derivative. The nominal-continuation and score-only targets are also inaccurate, with relative errors 0.69 and 0.98, respectively.

A one-sided finite-difference check gives the same conclusion. At random points, forward and backward derivatives coincide numerically, and the optimal-coupling recursion matches the central difference to numerical precision. Thus, a non-singleton optimal transport face does not necessarily imply nondifferentiability.

Table 6: Relative ℓ_2 error of the analytic gradient against central finite differences of the exact robust objective (coin toss, $\varepsilon = 1$, 20 random θ).

n_λ	rel. ℓ_2 vs exact ($h = 10^{-5}$)	rel. ℓ_2 vs exact ($h = 10^{-6}$)	max abs ($h = 10^{-5}$)	consistency (same grid)
5	4.21e-01 ± 2.2e-02	4.21e-01 ± 2.2e-02	7.93e-02 ± 6.4e-03	1.27e-03 ± 1.3e-03
10	3.74e-01 ± 2.8e-02	3.74e-01 ± 2.8e-02	7.24e-02 ± 5.6e-03	2.44e-09 ± 5.9e-11
25	2.77e-01 ± 1.8e-02	2.77e-01 ± 1.8e-02	5.37e-02 ± 5.3e-03	2.55e-09 ± 6.9e-11
50	2.49e-01 ± 1.5e-02	2.49e-01 ± 1.5e-02	4.63e-02 ± 4.4e-03	1.01e-03 ± 9.8e-04
100	2.36e-01 ± 1.2e-02	2.36e-01 ± 1.2e-02	4.24e-02 ± 3.7e-03	3.06e-03 ± 3.1e-03
200	2.47e-01 ± 1.1e-02	2.47e-01 ± 1.1e-02	4.43e-02 ± 2.8e-03	1.31e-03 ± 1.2e-03
golden-section (argmin transport)	2.46e-01 ± 1.5e-02	2.46e-01 ± 1.5e-02	4.51e-02 ± 3.1e-03	2.46e-01 ± 1.5e-02
golden-section + optimal coupling	2.95e-09 ± 6.2e-11	3.01e-08 ± 6.4e-10	2.55e-10 ± 6.8e-12	2.95e-09 ± 6.2e-11
(b) nominal continuation, exact dual	6.93e-01 ± 7.5e-03	6.93e-01 ± 7.5e-03	1.09e-01 ± 4.3e-03	6.93e-01 ± 7.5e-03
(c) score only, exact dual	9.81e-01 ± 1.2e-03	9.81e-01 ± 1.2e-03	1.58e-01 ± 5.9e-03	9.81e-01 ± 1.2e-03

Table 7: One-sided finite-difference check at 5 random θ per environment.

env	ε	$\|D^+ - D^-\|/\|\nabla J\|$	kink coords	coupling vs central FD	argmin vs central FD	coupling $\in [D^-, D^+]$
coin	1.0	3.90e-06 ± 1.6e-07	0.0000 ± 0.0000	2.86e-09 ± 1.4e-10	2.28e-01 ± 1.1e-02	1.0000 ± 0.0000
supply	2.0	7.41e-06 ± 8.7e-08	0.0000 ± 0.0000	1.75e-08 ± 1.1e-09	2.40e-02 ± 4.4e-03	1.0000 ± 0.0000

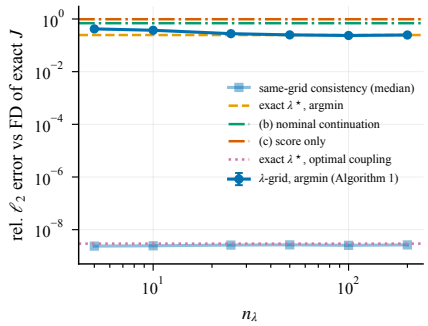


Figure 7: Relative gradient error against the exact robust objective.

We also test random, trained, and engineered tie points using forward and backward finite differences. At a genuine tie,

$$D^+J(\theta)[r] = \min_{\gamma \in \mathcal{F}^\star} \langle \gamma, U_r \rangle, \qquad D^-J(\theta)[r] = \max_{\gamma \in \mathcal{F}^\star} \langle \gamma, U_r \rangle,$$

where $\mathcal{F}^\star$ is the optimal transport face.

We compare (A) the optimal-face recursion, (C_v) a single vertex of the optimal face, a single transport argmin, (D) the pointwise tie-break of the dual formula.

The optimal-face recursion matches finite differences at random and engineered tie points. At engineered ties, $D^+J(\theta)[r] < D^-J(\theta)[r]$ for some directions, confirming nondifferentiability. In contrast, a single vertex or argmin can give incorrect derivative values, signs, and ascent certificates.

Thus, at tie points, the correct one-sided derivative requires optimization over the full optimal transport face rather than an arbitrary tie-breaking rule.

Table 8: One-sided derivatives at random, trained, and engineered tie points.

env	θ-set	(A) err., random r	(A) err., structural r	(C) err., structural r	(C) wrong sign	(C) false cert.	(C_v) err., random r	(D) err., structural r
coin	random θ	3.06e-07 ± 7.8e-08	3.50e-09 ± 3.2e-10	6.32e-02 ± 5.8e-03	0.087 ± 0.022	0.050 ± 0.017	3.06e-07 ± 7.8e-08	1.49e-01 ± 3.8e-03
coin	engineered ties	3.98e-07 ± 1.7e-07	3.75e-09 ± 3.1e-10	6.90e-02 ± 6.7e-03	0.081 ± 0.022	0.037 ± 0.015	4.15e-04 ± 2.0e-04	1.44e-01 ± 4.0e-03
coin	trained policies	4.58e-03 ± 3.6e-03	2.86e-04 ± 2.0e-04	9.81e-02 ± 9.4e-03	0.106 ± 0.024	0.062 ± 0.019	4.81e-02 ± 2.2e-02	1.56e-01 ± 3.3e-03
coin	trained + engineered ties	6.39e-08 ± 1.5e-08	6.15e-10 ± 6.3e-11	7.10e-02 ± 3.9e-03	0.006 ± 0.006	0.000 ± 0.000	1.40e-04 ± 3.5e-05	6.65e-02 ± 3.3e-03
supply	random θ	1.88e-06 ± 4.8e-07	1.63e-08 ± 1.2e-09	2.95e-02 ± 7.1e-04	0.006 ± 0.006	0.000 ± 0.000	1.88e-06 ± 4.8e-07	3.87e-02 ± 1.2e-03
supply	engineered ties	1.58e-06 ± 3.0e-07	1.40e-08 ± 1.1e-09	3.33e-02 ± 7.1e-04	0.006 ± 0.006	0.006 ± 0.006	1.59e-02 ± 3.4e-03	3.86e-02 ± 1.2e-03
supply	trained policies	2.98e-03 ± 1.1e-03	1.13e-05 ± 9.9e-07	5.86e-01 ± 1.6e-02	0.069 ± 0.020	0.037 ± 0.015	2.98e-03 ± 1.1e-03	4.41e-01 ± 8.2e-03
supply	trained + engineered ties	1.30e-03 ± 1.8e-04	1.42e-05 ± 9.8e-07	5.83e-01 ± 1.6e-02	0.081 ± 0.022	0.056 ± 0.018	1.01e-01 ± 4.4e-02	4.38e-01 ± 8.4e-03

Table 9: Structural ascent candidates and their exact one-sided derivatives.

env	θ-set	r_k	n	$D^+J[r_k]$	$D^-J[r_k]$	$\widehat{D}^+$ (A)	$\widehat{D}$ (C)	$D^+ < 0$	certified (A)	(C)> 0	false cert. (C)	sign(A)=sign(D^+)	rel. err. (A)
coin	random	r1	20	+4.399e-01 ± 9.7e-03	+4.399e-01 ± 9.7e-03	+4.399e-01 ± 9.7e-03	+4.774e-01 ± 8.7e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	4.23e-09 ± 8.2e-10
coin	random	r2	20	+4.429e-01 ± 9.3e-03	+4.429e-01 ± 9.3e-03	+4.429e-01 ± 9.3e-03	+4.006e-01 ± 1.2e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	3.06e-09 ± 4.2e-10
coin	random	r3	20	+4.603e-01 ± 8.5e-03	+4.603e-01 ± 8.5e-03	+4.603e-01 ± 8.5e-03	+4.578e-01 ± 9.5e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	2.79e-09 ± 6.1e-10
coin	random	r4	20	+4.562e-01 ± 8.4e-03	+4.562e-01 ± 8.4e-03	+4.562e-01 ± 8.4e-03	+4.568e-01 ± 8.7e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	3.92e-09 ± 6.0e-10
coin	engineered	r1	20	+4.279e-01 ± 1.0e-02	+4.279e-01 ± 1.0e-02	+4.279e-01 ± 1.0e-02	+4.741e-01 ± 9.4e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	3.76e-09 ± 7.1e-10
coin	engineered	r2	20	+4.319e-01 ± 9.6e-03	+4.319e-01 ± 9.6e-03	+4.319e-01 ± 9.6e-03	+3.891e-01 ± 1.3e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	3.98e-09 ± 4.6e-10
coin	engineered	r3	20	+4.505e-01 ± 8.8e-03	+4.506e-01 ± 8.8e-03	+4.505e-01 ± 8.8e-03	+4.523e-01 ± 1.0e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	2.53e-09 ± 5.8e-10
coin	engineered	r4	20	+4.474e-01 ± 8.8e-03	+4.474e-01 ± 8.8e-03	+4.474e-01 ± 8.8e-03	+4.506e-01 ± 9.5e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	4.72e-09 ± 6.8e-10
coin	trained	r1	20	+3.902e-03 ± 5.7e-05	+3.902e-03 ± 5.7e-05	+3.902e-03 ± 5.7e-05	+4.616e-03 ± 3.9e-05	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	2.15e-07 ± 2.0e-07
coin	trained	r2	20	+3.955e-03 ± 4.3e-05	+3.955e-03 ± 4.3e-05	+3.953e-03 ± 4.4e-05	+3.467e-03 ± 6.4e-05	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	7.23e-04 ± 7.2e-04
coin	trained	r3	20	+4.197e-03 ± 3.8e-05	+4.197e-03 ± 3.8e-05	+4.196e-03 ± 3.8e-05	+4.319e-03 ± 3.8e-05	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	3.44e-05 ± 3.4e-05
coin	trained	r4	20	+4.183e-03 ± 3.5e-05	+4.183e-03 ± 3.5e-05	+4.182e-03 ± 3.5e-05	+4.308e-03 ± 2.9e-05	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	3.86e-04 ± 3.9e-04
coin	engineered_trained	r1	20	+2.724e-01 ± 4.6e-03	+2.725e-01 ± 4.6e-03	+2.724e-01 ± 4.6e-03	+2.741e-01 ± 6.8e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	6.35e-10 ± 1.4e-10
coin	engineered_trained	r2	20	+2.724e-01 ± 4.6e-03	+2.725e-01 ± 4.6e-03	+2.724e-01 ± 4.6e-03	+2.741e-01 ± 6.8e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	5.93e-10 ± 1.2e-10
coin	engineered_trained	r3	20	+2.724e-01 ± 4.6e-03	+2.725e-01 ± 4.6e-03	+2.724e-01 ± 4.6e-03	+2.741e-01 ± 6.8e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	6.66e-10 ± 1.3e-10
coin	engineered_trained	r4	20	+2.724e-01 ± 4.6e-03	+2.725e-01 ± 4.6e-03	+2.724e-01 ± 4.6e-03	+2.741e-01 ± 6.8e-03	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	5.66e-10 ± 1.3e-10
supply	random	r1	20	+1.709e+00 ± 4.2e-02	+1.709e+00 ± 4.2e-02	+1.709e+00 ± 4.2e-02	+1.761e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.66e-08 ± 3.5e-09
supply	random	r2	20	+1.707e+00 ± 4.2e-02	+1.707e+00 ± 4.2e-02	+1.707e+00 ± 4.2e-02	+1.757e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.57e-08 ± 1.8e-09
supply	random	r3	20	+1.709e+00 ± 4.2e-02	+1.709e+00 ± 4.2e-02	+1.709e+00 ± 4.2e-02	+1.760e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.74e-08 ± 2.2e-09
supply	random	r4	20	+1.707e+00 ± 4.2e-02	+1.707e+00 ± 4.2e-02	+1.707e+00 ± 4.2e-02	+1.757e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.57e-08 ± 1.8e-09
supply	engineered	r1	20	+1.701e+00 ± 4.2e-02	+1.707e+00 ± 4.2e-02	+1.701e+00 ± 4.2e-02	+1.759e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.66e-08 ± 2.4e-09
supply	engineered	r2	20	+1.700e+00 ± 4.2e-02	+1.705e+00 ± 4.2e-02	+1.700e+00 ± 4.2e-02	+1.755e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.30e-08 ± 2.3e-09
supply	engineered	r3	20	+1.701e+00 ± 4.2e-02	+1.707e+00 ± 4.2e-02	+1.701e+00 ± 4.2e-02	+1.758e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.41e-08 ± 2.1e-09
supply	engineered	r4	20	+1.700e+00 ± 4.2e-02	+1.705e+00 ± 4.2e-02	+1.700e+00 ± 4.2e-02	+1.755e+00 ± 4.4e-02	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.30e-08 ± 2.3e-09
supply	trained	r1	20	+2.422e-03 ± 1.1e-06	+2.422e-03 ± 1.1e-06	+2.422e-03 ± 1.1e-06	+4.346e-03 ± 2.3e-06	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.16e-05 ± 1.9e-06
supply	trained	r2	20	+2.010e-03 ± 2.5e-06	+2.010e-03 ± 2.5e-06	+2.010e-03 ± 2.5e-06	+2.920e-03 ± 4.7e-06	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.01e-05 ± 2.2e-06
supply	trained	r3	20	+2.424e-03 ± 1.3e-06	+2.424e-03 ± 1.3e-06	+2.424e-03 ± 1.3e-06	+3.973e-03 ± 2.5e-06	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.25e-05 ± 1.7e-06
supply	trained	r4	20	+2.030e-03 ± 1.3e-05	+2.030e-03 ± 1.3e-05	+2.030e-03 ± 1.3e-05	+2.961e-03 ± 2.6e-05	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.08e-05 ± 2.1e-06
supply	engineered_trained	r1	20	+2.421e-03 ± 1.1e-06	+2.423e-03 ± 1.1e-06	+2.421e-03 ± 1.1e-06	+4.347e-03 ± 2.3e-06	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.29e-05 ± 1.8e-06
supply	engineered_trained	r2	20	+1.995e-03 ± 2.4e-06	+2.008e-03 ± 2.4e-06	+1.995e-03 ± 2.4e-06	+2.884e-03 ± 4.5e-06	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.43e-05 ± 1.8e-06
supply	engineered_trained	r3	20	+2.422e-03 ± 1.3e-06	+2.428e-03 ± 1.3e-06	+2.422e-03 ± 1.3e-06	+3.964e-03 ± 2.5e-06	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.57e-05 ± 2.5e-06
supply	engineered_trained	r4	20	+2.019e-03 ± 1.5e-05	+2.032e-03 ± 1.4e-05	+2.019e-03 ± 1.5e-05	+2.935e-03 ± 3.0e-05	0.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	1.39e-05 ± 1.7e-06

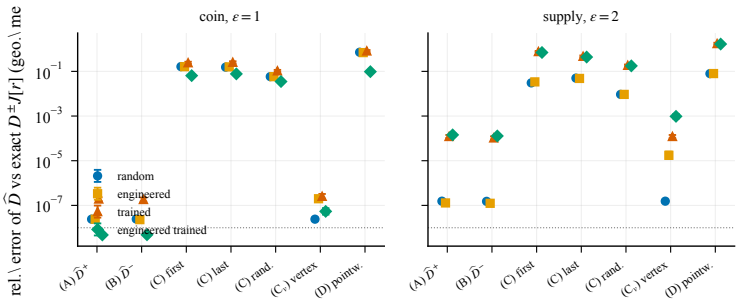


Figure 8: Relative error of the one-sided derivative evaluations against finite differences.

E.6 Ascent from tie points

Starting from the engineered tie points of Appendix E.5, we run 100 outer iterations with 10 seeds and compare Algorithm 2, using the four structural candidates of Table 9, with the vector specialization under fixed and random active selections (Table 10, Figure 9). These candidate-direction runs are distinct from the tests of Algorithm 1 in Appendix E.1.2.

All four candidate directions are certified at every iteration, with no false certificates, and all three actors are monotone in the exact objective in all 10 runs. The certified actor attains a slightly higher final objective: -3.51 ± 0.06 versus -3.65 ± 0.06 and -3.62 ± 0.06 on the coin toss, and -88.25 ± 0.11 versus -88.31 ± 0.11 and -88.34 ± 0.12 on the supply chain.

Thus, the certificate guarantees the sign of the chosen direction rather than faster convergence. Along these trajectories, the fixed and random selections also happen to be ascent directions, although Appendix E.5 shows that this need not hold in general.

Table 10: Ascent from engineered tie points (100 iterations, 10 seeds): final exact J, fraction of non-ascent steps, false certificates, and Δ_π, Δ_V against exact DP.

env	actor	J_0	J_{100}	$J_{100} - J_0$	J^{DP}	monotone runs	non-ascent steps	$D^+J[r] \le 0$ steps	skipped	#certified	Δ_π	$\Delta_V^{\mathrm{greedy}}$
coin	(i) Algorithm 1, certified candidates	-6.48227 ± 0.08902	-3.50684 ± 0.06226	2.97543 ± 0.07261	0.16129	10/10	0.0000 ± 0.0000	0.0000 ± 0.0000	0.00 ± 0.00	4.00 ± 0.00	0.3982 ± 0.0101	4.0520 ± 0.1520
coin	(ii) vector, first-index selection	-6.48227 ± 0.08902	-3.64519 ± 0.06374	2.83708 ± 0.06787	0.16129	10/10	0.0000 ± 0.0000	0.0000 ± 0.0000	0.00 ± 0.00	–	0.4145 ± 0.0095	4.3067 ± 0.1748
coin	(iii) vector, random active selection	-6.48227 ± 0.08902	-3.62349 ± 0.06278	2.85878 ± 0.06865	0.16129	10/10	0.0000 ± 0.0000	0.0000 ± 0.0000	0.00 ± 0.00	–	0.4118 ± 0.0095	4.2858 ± 0.1551
supply	(i) Algorithm 1, certified candidates	-97.25285 ± 0.24501	-88.25253 ± 0.11429	9.00032 ± 0.28281	-69.85289	10/10	0.0000 ± 0.0000	0.0000 ± 0.0000	0.00 ± 0.00	4.00 ± 0.00	0.7736 ± 0.0136	20.0955 ± 0.2608
supply	(ii) vector, first-index selection	-97.25285 ± 0.24501	-88.31057 ± 0.11420	8.94228 ± 0.27981	-69.85289	10/10	0.0000 ± 0.0000	0.0000 ± 0.0000	0.00 ± 0.00	–	0.7809 ± 0.0127	20.1375 ± 0.2327
supply	(iii) vector, random active selection	-97.25285 ± 0.24501	-88.33554 ± 0.11780	8.91731 ± 0.27949	-69.85289	10/10	0.0000 ± 0.0000	0.0000 ± 0.0000	0.00 ± 0.00	–	0.7745 ± 0.0140	20.1493 ± 0.2766

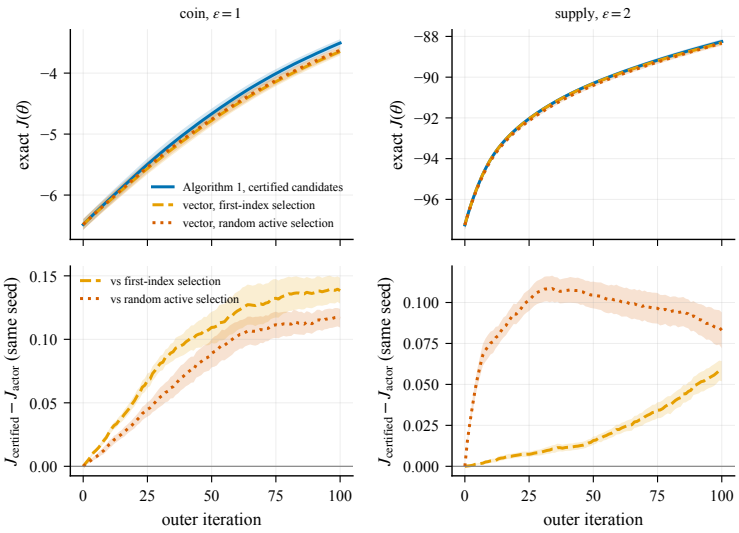


Figure 9: Exact objective over the iterations for the three actors (mean ± s.e., 10 seeds).

E.7 Evaluation under Distribution Shifts: Scalar LQ and Inverted Pendulum

Method. Each learned policy is evaluated with 10^4 common-random-number rollouts under mean shifts $w \sim \mathcal{N}(\delta, \sigma_0^2)$, $\delta \in \{-0.3, \dots, 0.3\}$, scale shifts $w \sim \mathcal{N}(0, (s\sigma_0)^2)$, $s \in \{0.5, \dots, 2\}$, heavy-tailed disturbances $w = \sigma_0 t_\nu / \sqrt{\nu/(\nu-2)}$, $\nu \in \{3, 5\}$, and the reference law. Domain randomisation (DR) uses the same networks with $\varepsilon = 0$ and resamples the disturbance law at each sweep.

For the inverted pendulum, we use $m = \ell = 1$, $g = 10$, $\mathrm{d}t = 0.05$, torque bounds $[-2, 2]$, horizon $T = 20$, and cost $\vartheta^2 + 0.1\dot{\vartheta}^2 + 0.001u^2$. The reference disturbance law is the empirical measure of $N = 30$ draws from $\mathcal{N}(0, 0.5^2)$, and the adversary reweights these atoms within a W_2 ball. Network settings are given in Table 19.

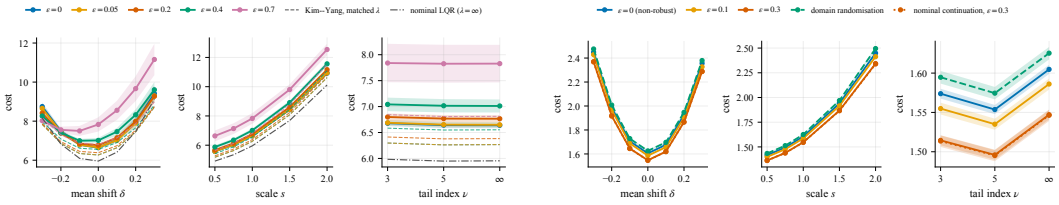


Figure 10: Cost under shifted disturbance laws (10^4 common-random-number rollouts, 10 seeds). Left: scalar LQ, including matched closed-form controllers of Kim and Yang (2023). Right: inverted pendulum, including DR.

Scalar LQ. Robust training trades nominal performance for protection against the matched adversary. As ε increases, nominal cost rises, and performance under arbitrary shift laws is not uniformly improved. Mean shifts can favor larger ε, while scale and heavy-tail shifts generally do not. The matched closed-form controllers of Kim and Yang (2023) show the same pattern.

Under a common Wasserstein adversary, however, policies trained with larger ε perform substantially better. At $\varepsilon_{\text{eval}} = 0.7$, the policy trained with $\varepsilon = 0.7$ has cost 13.46 ± 0.21, compared with 15.86 ± 0.52 for $\varepsilon = 0.2$ and 18.75 ± 0.94 for $\varepsilon = 0$. Thus the algorithm improves the robust objective it is trained for, while robustness need not transfer to unmatched scale or tail shifts.

The full and nominal-continuation targets are nearly indistinguishable under the common adversary at $\varepsilon = 0.7$ (13.46 versus 13.58), consistent with the tabular fixed-point discussion of Appendix E.4.

Inverted pendulum. On the pendulum, robust training improves both nominal and shifted performance over the tested grid. At $\varepsilon = 0.3$, the nominal cost is 1.484 ± 0.006, compared with 1.540 ± 0.006 for $\varepsilon = 0$ and 1.560 ± 0.008 for DR. The corresponding worst shifted cost is 2.370 ± 0.008, compared with 2.460 ± 0.010 and 2.503 ± 0.017.

Under common adversaries, the $\varepsilon = 0.3$ policy is also better at $\varepsilon_{\text{eval}} \in \{0, 0.1\}$ and statistically indistinguishable from $\varepsilon = 0$ at $\varepsilon_{\text{eval}} = 0.3$. DR is worse in all tested cases. The nominal-continuation target again gives nearly the same performance as the full recursion.

E.8 Lost-sales inventory

We consider a lost-sales inventory problem with holding, shortage, and fixed ordering costs, continuous demand, an empirical reference law with $N = 30$ atoms, and horizon $T = 20$. We use the networks of Table 19, $\varepsilon \in \{0, 0.3, 1\}$, and domain randomisation as in Appendix E.7.

With the 150-iteration budget of Table 19, the robust policies are worse than the nominal policy under all tested shifts and remain highly stochastic. Increasing the budget to 600 iterations removes this effect and restores the robustness trade-off: at $\varepsilon = 1$, the nominal cost is 95.5 ± 1.8 versus 93.4 ± 1.2 for $\varepsilon = 0$, while the worst-shift cost improves to 131.1 ± 1.1 versus 132.8 ± 1.0. The $\varepsilon = 0.3$ policy is intermediate, while DR gives the largest worst-shift cost.

Unlike in LQ and the pendulum, the transported continuation also outperforms the nominal continuation at $\varepsilon = 1$, with worst-shift cost 131.1 versus 136.2 and nominal cost 95.5 versus 98.7. Thus, the shorter training budget is insufficient for this higher-cost, non-smooth problem, and we report it as a budget-sensitivity result.

Table 11: Lost-sales inventory: 150 outer iterations (10 seeds, top) and 600 outer iterations (5 seeds, bottom).

method	C_{nom}	C_{worst}	argmax shift	ΔC_{worst} vs $\varepsilon = 0$	ΔC_{worst} vs DR	σ	unique argmin	wall (min)
eps=0 (non-robust)	102.358 ± 1.288	142.398 ± 1.760	scale:2	–	1.309 ± 2.392 (4/10 < 0)	1.752 ± 0.160	1.000 ± 0.000	25.2 ± 1.4
eps=0.3	105.480 ± 2.489	144.918 ± 4.163	scale:2	2.520 ± 2.655 (6/10 < 0)	3.828 ± 4.381 (4/10 < 0)	1.717 ± 0.141	1.000 ± 0.000	69.4 ± 1.1
eps=1	125.540 ± 10.734	173.279 ± 15.993	scale:2	30.881 ± 14.778 (2/10 < 0)	32.189 ± 16.042 (1/10 < 0)	1.369 ± 0.071	1.000 ± 0.000	82.0 ± 2.4
DR	105.260 ± 2.659	141.089 ± 1.450	scale:2	-1.309 ± 2.392 (6/10 < 0)	–	1.864 ± 0.239	1.000 ± 0.000	35.0 ± 0.4

method	C_{nom}	C_{worst}	argmax shift	ΔC_{worst} vs $\varepsilon = 0$	ΔC_{worst} vs DR	σ	unique argmin	wall (min)
eps=0 (non-robust)	93.433 ± 1.191	132.772 ± 1.036	scale:2	–	-4.129 ± 4.852 (2/5 < 0)	0.559 ± 0.046	1.000 ± 0.000	128.0 ± 0.7
eps=0.3	93.764 ± 0.961	132.136 ± 1.242	scale:2	-0.637 ± 0.773 (4/5 < 0)	-4.765 ± 4.191 (3/5 < 0)	0.601 ± 0.052	1.000 ± 0.000	309.3 ± 0.9
eps=1	95.456 ± 1.846	131.140 ± 1.113	scale:2	-1.632 ± 0.940 (4/5 < 0)	-5.761 ± 5.596 (3/5 < 0)	0.642 ± 0.063	1.000 ± 0.000	244.3 ± 1.9
DR	95.842 ± 2.124	136.901 ± 5.225	scale:2	4.129 ± 4.852 (3/5 < 0)	–	0.488 ± 0.037	1.000 ± 0.000	58.4 ± 1.4
nominal-continuation eps=1	98.717 ± 1.827	136.203 ± 4.353	scale:2	3.430 ± 4.037 (2/5 < 0)	-0.698 ± 1.280 (3/5 < 0)	0.610 ± 0.073	1.000 ± 0.000	162.5 ± 1.7

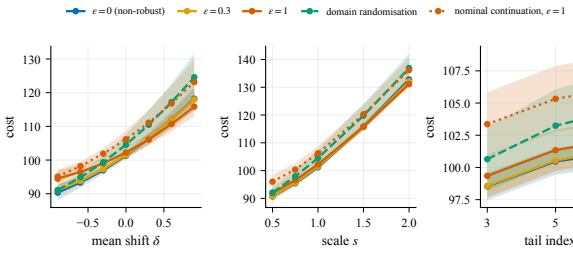


Figure 11: Inventory problem after 600 outer iterations: cost under shifted demand laws.

E.9 Robust Q-learning baseline

On the coin toss, we compare with the robust Q-iteration of Neufeld and Sester (2024). Its greedy policy agrees with our exact-DP time-0 policy on all 11 states for $\varepsilon \in \{0.5, 1, 2\}$, including at horizon $T = 100$. It also agrees with the policies learned by Algorithm 2 on all states for $\varepsilon \in \{1, 2\}$ and on $9/11$ states for $\varepsilon = 0.5$. The corresponding robust values agree with our V_0 to within 0.003 in sup norm.

The sample-based scheme converges to the fixed point of the sampled operator rather than the exact operator. Its sup-norm error to $Q^\star$ decreases from $0.32/0.058/0.070$ at $K = 10$ to $0.029/0.0095/0.015$

at $K = 100$ and $0.0028/0.0022/0.0018$ at $K = 1000$ for $\varepsilon = 0.5/1/2$. Increasing the number of updates at fixed K does not remove this error. Algorithm 2 avoids this empirical-measure bias in the tabular exact-reference setting; its sampled version has the same dependence on the sample size.

Table 12: Robust Q-iteration baseline on the coin toss: policy and value agreement with exact DP and sup-norm error of the sampled scheme.

ε	agree DP (T=10)	agree learned	agree DP (T=100)	$\overline{V^Q}$	$\overline{V_0}$	max $\lvert V^Q - V_0\rvert$	max per-step gap (T=10)	max per-step gap (T=100)	Q-iter its	sampled err
0.5	1.000	0.818 ± 0.000	1.000	0.366	0.367	0.003	0.0723	0.0000	45	3.176e-01
1.0	1.000	1.000 ± 0.000	1.000	0.160	0.161	0.003	0.0502	0.0000	25	5.823e-02
2.0	1.000	1.000 ± 0.000	1.000	0.000	0.000	0.000	0.0000	0.0000	2	7.028e-02

E.10 Uniqueness of the worst-case transport

We record the fraction of transport cells whose argmin in line 5 is unique up to 10^{-8}. With the λ-grids used during training, the argmin is unique on 92–94% of coin-toss cells, 99.6–100% of supply-chain cells, and more than 99.9% of LQ and inventory cells. These are grid-point argmin frequencies, not a verification of Proposition B.3 for the exact dual optimum. Appendix E.1.1 distinguishes coupling, law, and directional-sensitivity nonuniqueness.

At the exact maximiser $\lambda^\star$, however, the unique fraction falls to 0.65–0.87 on the coin toss and 0.79–0.81 on the supply chain. Hence transport ties are common at the exact dual optimum, which motivates the optimal-coupling treatment of Appendix E.5. On the pendulum, the unique fraction is 0.98 on average but falls to 0.83–0.86 near the end of training, showing that ties can also arise in the sample-supported setting.

Table 13: Fraction of transport cells with a unique argmin: mean over training, final iteration, and minimum over iterations.

exp.	env	dual solver	ε	unique (mean over training)	unique (final)	min
E1	coin	grid $n_\lambda = 101$ (Algorithm 1)	0.5	0.9362 ± 0.0001	0.9364 ± 0.0000	0.9074
E1	coin	grid $n_\lambda = 101$ (Algorithm 1)	1.0	0.9299 ± 0.0002	0.9303 ± 0.0000	0.8636
E1	coin	grid $n_\lambda = 101$ (Algorithm 1)	2.0	0.9242 ± 0.0000	0.9242 ± 0.0000	0.9152
E1	coin	exact $\lambda^\star$	0.5	0.6899 ± 0.0010	0.6694 ± 0.0053	0.6372
E1	coin	exact $\lambda^\star$	1.0	0.6684 ± 0.0012	0.6501 ± 0.0059	0.5939
E1	coin	exact $\lambda^\star$	2.0	0.8678 ± 0.0031	0.8555 ± 0.0065	0.7499
E1	supply	grid $n_\lambda = 101$ (Algorithm 1)	0.5	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000
E1	supply	grid $n_\lambda = 101$ (Algorithm 1)	1.0	0.9956 ± 0.0000	0.9956 ± 0.0000	0.9956
E1	supply	grid $n_\lambda = 101$ (Algorithm 1)	2.0	0.9956 ± 0.0000	0.9956 ± 0.0000	0.9718
E1	supply	exact $\lambda^\star$	0.5	0.8106 ± 0.0001	0.8114 ± 0.0014	0.7870
E1	supply	exact $\lambda^\star$	1.0	0.8112 ± 0.0006	0.8093 ± 0.0026	0.7489
E1	supply	exact $\lambda^\star$	2.0	0.7956 ± 0.0002	0.7938 ± 0.0067	0.6653
E2	scalar LQ (paper code)	grid $n_\lambda = 50$	0.05	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9997
E2	scalar LQ (paper code)	grid $n_\lambda = 50$	0.2	0.9999 ± 0.0000	0.9999 ± 0.0000	0.9995
E2	scalar LQ (paper code)	grid $n_\lambda = 50$	0.4	0.9999 ± 0.0000	1.0000 ± 0.0000	0.9996
E2	scalar LQ (paper code)	grid $n_\lambda = 50$	0.7	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9999
E6	pendulum	grid $n_\lambda = 50$	0.1	0.9821 ± 0.0020	0.8321 ± 0.0173	0.7342
E6	pendulum	grid $n_\lambda = 50$	0.3	0.9865 ± 0.0019	0.8626 ± 0.0179	0.7540
E6b	inventory	grid $n_\lambda = 50$	0.3	0.9999 ± 0.0000	0.9998 ± 0.0000	0.9990
E6b	inventory	grid $n_\lambda = 50$	1.0	0.9998 ± 0.0000	0.9995 ± 0.0002	0.9973
E6-lq	scalar LQ (generic)	grid $n_\lambda = 50$	0.2	0.9999 ± 0.0000	0.9999 ± 0.0000	0.9995
E6-lq	scalar LQ (generic)	grid $n_\lambda = 50$	0.7	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000

E.11 Sample sensitivity

We study the sampled form of Algorithm 2, using B on-policy trajectories per iteration and N independent successor samples per visited state–action pair (5 seeds, $\varepsilon = 1$). With one successor sample, the method performs poorly on both tasks.

On the supply chain, the error is mainly controlled by N: Δ_π decreases from approximately 0.47 at $N = 1$ to 0.28, 0.06, and 0.007 at $N = 4, 16, 64$, respectively, and reaches zero at $N = 64$, $B = 256$. The result is comparatively insensitive to B. On the coin toss, the error is more sensitive to B because small trajectory batches rarely visit the binomial tail states and leave their critic values stale.

Thus, N controls approximation of the transition law, whereas B controls state coverage. Appendix E.1.3 separately measures the error of exact empirical-model solutions relative to the population model. These finite-sample checks do not establish convergence of the fitted actor–critic.

Table 14: Sampled Algorithm 2 with N successor samples and B trajectories per iteration ($\varepsilon = 1$, 5 seeds).

env	N	B	Δ_π	Δ_V	wall (s)
coin	1	16	0.1600 ± 0.0214	0.6812 ± 0.0256	4.3 ± 0.0
coin	1	64	0.1327 ± 0.0055	0.4348 ± 0.0821	4.3 ± 0.0
coin	1	256	0.1073 ± 0.0093	0.0850 ± 0.0016	5.3 ± 0.1
coin	4	16	0.1291 ± 0.0139	0.3085 ± 0.0146	3.5 ± 0.0
coin	4	64	0.0855 ± 0.0179	0.1722 ± 0.0602	4.4 ± 0.0
coin	4	256	0.0509 ± 0.0046	0.0704 ± 0.0006	5.3 ± 0.1
coin	16	16	0.1455 ± 0.0155	0.4951 ± 0.0564	4.0 ± 0.0
coin	16	64	0.0800 ± 0.0078	0.1465 ± 0.0167	4.5 ± 0.0
coin	16	256	0.0382 ± 0.0053	0.0951 ± 0.0122	5.5 ± 0.1
coin	64	16	0.1564 ± 0.0045	0.4567 ± 0.0601	3.8 ± 0.0
coin	64	64	0.0927 ± 0.0156	0.3898 ± 0.1112	5.0 ± 0.0
coin	64	256	0.0345 ± 0.0101	0.1131 ± 0.0217	5.7 ± 0.1
coin	exact	exact	0.0000 ± 0.0000	0.0142 ± 0.0002	1.8 ± 0.1
supply	1	16	0.4655 ± 0.0123	2.6132 ± 0.1050	38.4 ± 0.2
supply	1	64	0.4709 ± 0.0088	2.1123 ± 0.1666	50.9 ± 0.1
supply	1	256	0.4618 ± 0.0127	2.0209 ± 0.1342	56.3 ± 0.2
supply	4	16	0.2436 ± 0.0238	0.4956 ± 0.1096	40.1 ± 0.1
supply	4	64	0.3055 ± 0.0114	0.5383 ± 0.0233	49.1 ± 0.1
supply	4	256	0.2764 ± 0.0094	0.4432 ± 0.0290	52.4 ± 0.1
supply	16	16	0.0618 ± 0.0136	0.0994 ± 0.0046	38.5 ± 0.0
supply	16	64	0.0491 ± 0.0055	0.0950 ± 0.0097	48.5 ± 0.1
supply	16	256	0.0582 ± 0.0062	0.0963 ± 0.0052	49.1 ± 0.2
supply	64	16	0.0073 ± 0.0018	0.0978 ± 0.0432	35.7 ± 0.2
supply	64	64	0.0073 ± 0.0034	0.0246 ± 0.0005	47.6 ± 0.2
supply	64	256	0.0000 ± 0.0000	0.0242 ± 0.0004	50.3 ± 0.2
supply	exact	exact	0.0000 ± 0.0000	0.0808 ± 0.0006	19.6 ± 0.2

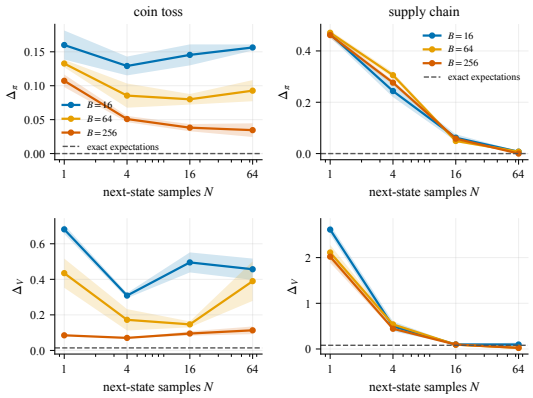


Figure 12: Policy and value errors as functions of N, for different trajectory batch sizes B.

E.12 Some additional plots

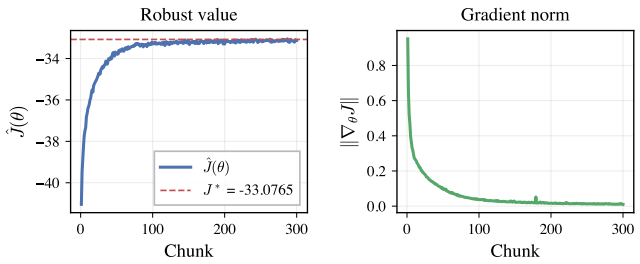


Figure 13: Learning curves of the supply chain example over training for $\varepsilon = 1$.

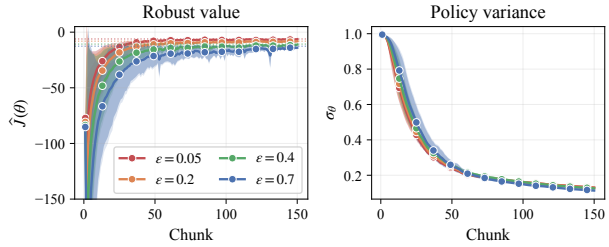
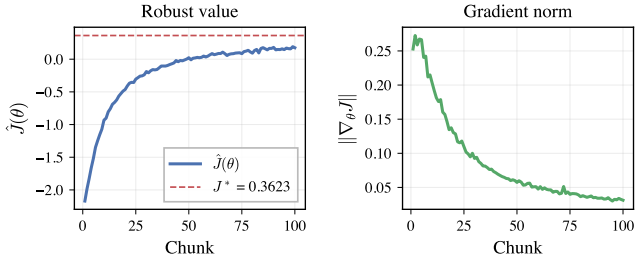
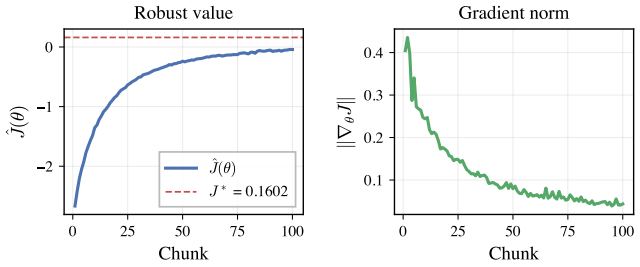


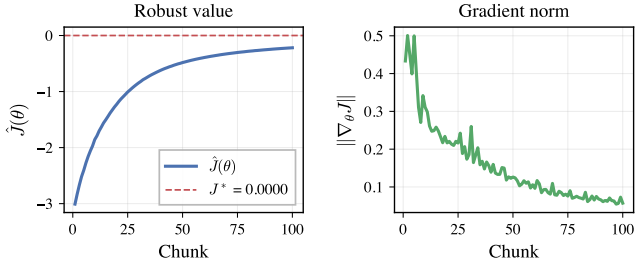
Figure 14: Learning curves of the robust linear-quadratic control example over training (robust value, left; policy standard deviation σ_θ, right).



(a) $\varepsilon = 0.5$



(b) $\varepsilon = 1$



(c) $\varepsilon = 2$

Figure 15: Learning curves of the coin toss example over training for different values of ε.

F TRAINING SETUP DETAILS

The sweep costs and timings below concern the original implementations of Algorithm 2; they exclude the vertex and branch enumeration of Algorithm 1. The exact finite-model costs and failure counts are reported separately in Appendix E.1.

Computational cost. Let B denote the number of sampled states, S the number of action or next-state samples per state, n_λ the number of dual grid points, and $C_\mathcal{F}$ the cost of evaluating the c-transform or worst-case transport. A backward sweep costs

$$O(TBSn_\lambda C_\mathcal{F}).$$

In tabular problems, $C_\mathcal{F} = O(|\mathcal{X}|)$, giving

$$O(TBSn_\lambda|\mathcal{X}|).$$

Robust overhead. With the defaults of Table 18, one robust outer iteration takes $5.3\,\text{ms}$ on the coin toss and $18\,\text{ms}$ on the supply chain. About 86–88% of this time is spent in the λ-grid search, 6–7% in the transport step, and 5–6% in the critic and actor updates. The corresponding non-robust iterations take 0.57 and $1.38\,\text{ms}$, respectively, giving overheads of approximately $9.4\times$ and $13\times$. The grid search also explains the discretisation error studied in Appendix E.5; an exact scalar dual search avoids this error.

Table 15: Wall-clock time per outer iteration and fraction of time spent in the dual search, transport step, and critic/actor updates (single CPU thread).

env	ε	dual	ms / iter	dual	transport	critic	$\times$ vs $\varepsilon = 0$
coin	0.0	short-circuit	0.57 ± 0.01	0.082	0.277	0.515	1.00
coin	0.0	grid	5.35 ± 0.01	0.857	0.070	0.058	9.40
coin	0.5	grid	5.32 ± 0.01	0.860	0.068	0.058	9.34
coin	1.0	grid	5.29 ± 0.01	0.861	0.068	0.057	9.28
coin	2.0	grid	5.31 ± 0.01	0.860	0.068	0.058	9.33
supply	0.0	short-circuit	1.38 ± 0.01	0.045	0.207	0.680	1.00
supply	0.0	grid	18.22 ± 0.05	0.880	0.057	0.056	13.18
supply	0.5	grid	18.02 ± 0.02	0.881	0.057	0.054	13.04
supply	1.0	grid	17.95 ± 0.02	0.882	0.056	0.054	12.99
supply	2.0	grid	17.99 ± 0.02	0.883	0.056	0.054	13.01

Scaling. In the tabular implementation, the dominant computation scales with the number of state–action–next-state triples. On the coin toss at $T = 10$, the adjoint implementation takes approximately $5\,\text{ms}$, $0.46\,\text{s}$, and $66\,\text{s}$ per iteration for $|\mathcal{X}| = 11, 101, 1001$, respectively. On the supply chain, where $|A| = |\mathcal{X}|$, the corresponding times are $17\,\text{ms}$, $1.7\,\text{s}$, and $13\,\text{s}$ for $|\mathcal{X}| = 11, 51, 101$.

The dense sensitivity critic of Algorithm 2 requires

$$O(T^2|\mathcal{X}|^2|A|)$$

memory. The algebraically equivalent adjoint formulation avoids this storage and is used for the larger problems.

For the coin toss, raw policy mismatch can remain large on rarely visited tail states, while the occupancy-weighted mismatch

$$\Delta_\pi^w = \frac{1}{T}\sum_{t,x} d_t(x)\,\mathbf{1}\{\hat{a}_t(x) \neq a_t^{\mathrm{DP}}(x)\}$$

falls below 10^{-3} rapidly and reaches 10^{-15}–10^{-11} at the end of training. On the supply chain, the fixed iteration budget becomes insufficient as the state and action spaces grow, indicating that the iteration budget must increase with problem size.

Table 16: Scaling at $\varepsilon = 1$ (5 seeds): cost per iteration for the dense and adjoint implementations, exact-DP time, iterations to policy agreement, and final errors.

env	$\lvert\mathcal{X}\rvert$	T	ms/it (dense)	ms/it (adjoint)	exact DP (s)	iters	iters to $\Delta_\pi = 0$	iters to $\Delta_\pi^w < 10^{-3}$	Δ_π	Δ_π^w	Δ_V
coin	11	10	5.3	5.2	0.01	100	15.2 (5/5)	15.2 (5/5)	0.0000 ± 0.0000	0.00e+00 ± 0.0e+00	0.0142 ± 0.0002
coin	101	10	483.4	456.1	0.20	100	not reached (100 its)	5.4 (5/5)	0.0655 ± 0.0018	7.59e-16 ± 2.1e-16	0.0085 ± 0.0002
coin	1001	10	90397.9	66451.9	26.98	71	not reached (70 its)	34.2 (5/5)	0.4318 ± 0.0015	1.48e-11 ± 8.7e-13	0.0096 ± 0.0001
coin	11	50	29.1	37.8	0.07	100	16.4 (5/5)	16.2 (5/5)	0.0000 ± 0.0000	0.00e+00 ± 0.0e+00	0.0788 ± 0.0002
coin	101	50	4430.3	3836.8	1.82	100	not reached (100 its)	6.0 (5/5)	0.1601 ± 0.0019	9.00e-15 ± 9.9e-16	0.0282 ± 0.0001
coin	1001	50	–	376341.3	178.83	9	not reached (9 its)	not reached (9 its)	0.5199 ± 0.0007	3.77e-02 ± 9.6e-04	3.6177 ± 0.1551
supply	11	10	18.3	17.1	0.02	300	30.8 (5/5)	30.8 (5/5)	0.0000 ± 0.0000	0.00e+00 ± 0.0e+00	0.0808 ± 0.0006
supply	51	10	2813.0	1662.9	0.82	300	not reached (300 its)	not reached (300 its)	0.0737 ± 0.0010	7.37e-02 ± 1.0e-03	0.0422 ± 0.0006
supply	101	10	–	13310.4	9.12	202	not reached (203 its)	not reached (203 its)	0.0800 ± 0.0021	8.00e-02 ± 2.1e-03	0.1637 ± 0.0008

Directional sweeps with neural policies. On the scalar LQ and pendulum tasks, we compare the vector sensitivity critic with $K = 4$ directional sensitivity critics for policy widths 128, 512, and 2048. The directional implementation uses either the exact dual with face linear programs or the grid dual with a single argmin selection.

At tie iterations, face-based and argmin-based predictions can differ in sign, although critic noise dominates their difference in prediction accuracy. Their main distinction is computational. The directional implementation uses approximately constant additional memory (36–56 MB and a 0.2 MB target tensor), whereas the vector critic materializes a target tensor proportional to the policy dimension. Consequently, the vector-critic runtime grows strongly with network width, while the directional sweeps remain nearly constant.

Table 17: Cost per sweep of the vector critic and $K = 4$ directional sweeps as a function of policy width, together with wrong-sign predictions at tie iterations.

env	width	method	d_θ	ms / sweep (median; mean ± s.e.)	peak RSS (MB)	ΔRSS (MB)	$BSN\, d_{\text{out}}$ tensor (MB)	wrong sign at ties
lq	128	(a) vector critic, d_θ outputs (paper)	514	1272 (1274 ± 3)	799	337	30.1	0.362 ± 0.045
lq	128	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face	514	1062 (1062 ± 3)	503	36	0.2	0.337 ± 0.047
lq	128	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin	514	1045 (1042 ± 4)	676	209	0.2	0.332 ± 0.048
lq	512	(a) vector critic, d_θ outputs (paper)	2050	2679 (2679 ± 8)	852	385	120.1	–
lq	512	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face	2050	1086 (1087 ± 1)	509	38	0.2	–
lq	512	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin	2050	1072 (1071 ± 3)	679	209	0.2	–
lq	2048	(a) vector critic, d_θ outputs (paper)	8194	8108 (8109 ± 7)	1732	1251	480.1	–
lq	2048	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face	8194	1360 (1360 ± 1)	531	45	0.2	–
lq	2048	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin	8194	1377 (1378 ± 7)	696	210	0.2	–
pendulum	128	(a) vector critic, d_θ outputs (paper)	642	3122 (3128 ± 5)	734	270	37.6	0.105 ± 0.012
pendulum	128	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face	642	6745 (6701 ± 207)	508	39	0.2	0.190 ± 0.008
pendulum	128	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin	642	2041 (2041 ± 4)	688	218	0.2	0.173 ± 0.014
pendulum	512	(a) vector critic, d_θ outputs (paper)	2562	6325 (6323 ± 8)	942	468	150.1	–
pendulum	512	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face	2562	7498 (7391 ± 234)	517	38	0.2	–
pendulum	512	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin	2562	2165 (2163 ± 5)	693	215	0.2	–
pendulum	2048	(a) vector critic, d_θ outputs (paper)	10242	19392 (19437 ± 66)	2136	1623	600.1	–
pendulum	2048	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face	10242	5884 (5886 ± 112)	573	56	0.2	–
pendulum	2048	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin	10242	2760 (2748 ± 8)	725	208	0.2	–
lq	crossover	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face		< 128				
lq	crossover	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin		< 128				
pendulum	crossover	(b) $K = 4$ directional, scalar critics + JVP, exact dual + face		609				
pendulum	crossover	(b′) $K = 4$ directional, scalar critics + JVP, grid dual + argmin		< 128				

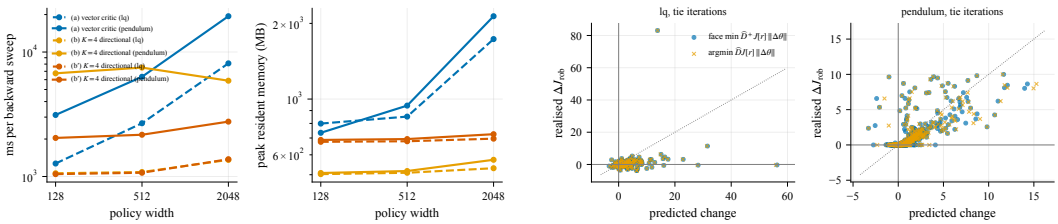


Figure 16: Left: cost per sweep versus policy width for the vector critic and $K = 4$ directional sweeps. Right: predicted versus realized change of the robust objective at tie iterations.

Compute resources. The original coin-toss, supply-chain, and continuous-control experiments were run on a single CPU thread on an Apple M-series machine without GPU acceleration. The exact finite-model validation batch used one c7i.16xlarge instance (64 vCPUs), with up to 62 concurrent processes. Its timings are per process and are not directly comparable with the Apple-machine measurements. At $|\mathcal{X}| = 11$, a tabular training run takes approximately 0.5 s for the coin toss (100 iterations) and 5.4 s for the supply chain (300 iterations). The sample-supported continuous experiments take approximately 35–130 minutes per seed for 150 outer iterations and 245–310 minutes for the 600-iteration inventory runs. Reproducing the continuous LQ experiments of Section 4.2 takes approximately 20 minutes.

Exact finite-model numerical settings. The finite-model batch uses face and active-set tolerances of 10^{-10} and HiGHS defaults. Depending on the experiment, the stopping tolerance is 10^{-3} or 5×10^{-4}, the iteration cap is 20,000, wall caps are 900–1,200 seconds, and selection caps are 10^5–2×10^5. Cap hits and missing outputs are retained in failure accounting. The scaling runs use $\zeta_0 = 0.01$ for $n \geq 4$ and the 10^5 selection cap. In the shared-parameter comparison, the corrected near-active and candidate-actor cells use $\zeta_0 = 0.01$ rather than the initial 0.5 runs that exceeded the selection cap. These settings are separate from the original defaults below.

Hyperparameters and experimental details We report the main hyperparameters used in the tabular and continuous experiments.

Table 18: Hyperparameters for the tabular experiments.

Hyperparameter	Coin toss	Supply chain
Actor learning rate η_θ	0.2	0.15
Dual grid resolution n_λ	101	101
Dual upper bound $\Lambda_{\max}$	50	50
Outer iterations	100	300
Seed	0	0

Table 19: Hyperparameters for the continuous linear–quadratic experiment.

Hyperparameter	Default
Outer iterations n_{outer}	150
Critic SGD steps per sweep	15
State batch size B	64
Action samples per state S	8
Dual grid resolution n_λ	50
State sampling support L	3.0
Actor hidden dimension	128
Critic hidden dimension	64
Actor learning rate	0.05
Critic learning rate	10^{-2}
Initial $\log \sigma_\theta$	0.0
Optimizer	Adam
Activation	ReLU
Discount	none
Seed	0