

# Steering the Feedback Loop: Policy Gradients for Online Performative Optimization

Changsheng Chen<sup>1</sup>   Cheng Tang<sup>2</sup>   Lanran Fang<sup>1</sup>   Grani A. Hanasusanto<sup>1</sup>

<sup>1</sup>Department of Industrial and Enterprise Systems Engineering

<sup>2</sup>Siebel School of Computing and Data Science  
University of Illinois Urbana-Champaign

## Abstract

Performative decision-making arises in many real-world problems where decisions influence the populations that generate future data. A major line of existing research focuses on learning performatively stable solutions. We study average-reward performative optimization, where the learner observes an evolving population through sample batches of fixed size and learns from a single continuing trajectory with unknown dynamics and without resets. We propose a policy-gradient method combining randomized trajectory lengths and gradient tracking. Our analysis uses Poisson equations to control initialization and return-truncation bias, even when population-state distributions do not converge over time. Under regularity assumptions, a uniform Poisson inverse bound, and an exact gradient-domination condition, the algorithm achieves  $\tilde{O}(\sqrt{T})$  expected regret against the best fixed policy in the chosen class.

## 1 Introduction

When decisions affect the population generating future data, their value depends on both immediate and future rewards. An action that performs well today may change the population in ways that reduce the rewards available tomorrow. This feedback motivates performative prediction, in which deploying a model changes the data distribution (Perdomo et al., 2020). In stateful models, the population response also depends on its current state, so successive decisions shape its evolution (Brown et al., 2022).

We study long-run reward when policies choose actions using fixed-size sample batches from a hidden population. At each interaction, the learner uses the empirical distribution of the current batch to select an action, observes a reward, and continues from the population resulting from that action. The population dynamics are unknown, and learning proceeds along a single continuing trajectory without resets. Our benchmark is the best fixed policy in a prescribed observation-based class, evaluated with the same batch size. A fixed policy keeps its parameter unchanged while allowing its actions to vary with the observed batch. A fixed batch size can leave sampling noise in the policy’s inputs even after its parameter has been learned. Our goal is to minimize expected regret, the expected cumulative reward shortfall relative to earning the benchmark policy’s long-run average reward at every interaction. This comparison includes every interaction used to learn the policy.

Existing performative methods study stable deployments and optimize the long-term risk of a fixed deployed model (Perdomo et al., 2020; Izzo et al., 2022). Sequential-control approaches also give regret guarantees, with different information and planning requirements. Brown et al. (2024) study

systems with observed states and local controllability, meaning that actions can steer the state to nearby targets. Their main algorithm queries known dynamics, with an extension for learning structured unknown dynamics. [Chen and Hanasusanto \(2026\)](#) allow unknown population dynamics and finite-batch observations along one continuing trajectory. Their algorithms use a specified latent statistic and known feature maps to estimate reward and transition models, and use confidence sets in optimistic planning.

In this work, we study direct optimization within a prescribed observation-based policy class, without fitting reward or transition models or solving a dynamic program. Policy-gradient methods offer a way to do this: observed actions and rewards, together with derivatives of the known policy, allow us to estimate how its long-run reward changes without reconstructing the hidden population ([Baxter and Bartlett, 2001](#)). This requires accounting for an action’s effects on future rewards using the available continuing trajectory. We therefore ask:

*Can policy-gradient methods achieve sublinear regret with unknown population dynamics, fixed-size sample batches, and a single continuing trajectory?*

A gradient formula alone does not answer this question, because learning requires both accurate estimates and a connection between gradients and global performance. The first obstacle is to estimate the gradient from finite segments of the continuing trajectory. Each segment starts from the population left by earlier policy updates, and the law of this starting population need not match the current policy’s stationary law. A preliminary run under the current policy, called burn-in, is intended to reduce the resulting initialization error. A second error arises when a finite sum of rewards, called a return, omits the selected action’s effects beyond the observed segment. Increasing these durations does not by itself guarantee accurate estimates: in a periodic system, population-state laws can keep oscillating, and a longer fixed burn-in need not remove dependence on the initial phase.

To control both errors, we randomize the burn-in and return durations under a uniform Poisson inverse condition. The condition controls cumulative expected deviations of Lipschitz functions of the population from their stationary means, even when population-state laws do not converge at deterministic times. Expectations at a randomized endpoint average over several possible trajectory lengths. We show that composing independent blocks of random length makes expectations at these endpoints approach their stationary values geometrically in the number of blocks. Applied to both burn-in and return endpoints, this yields uniform control of initialization and return-truncation errors.

Each gradient estimate requires a trajectory segment and remains noisy. We use gradient tracking to retain information from earlier estimates across policy updates ([Ganesh et al., 2025](#)). Because those estimates refer to earlier policy parameters, we adjust the previous estimate using an estimate of the gradient change before combining it with a fresh estimate.

Even with accurate gradient estimates, we face a second obstacle: obtaining a globally good policy. The long-run average reward can be nonconcave, so gradient-based updates may stall at suboptimal stationary points. We give sufficient conditions under which the best feasible first-order improvement controls the reward gap to the best policy in the class. These conditions connect the estimator guarantees to global regret through a gradient-domination inequality.

Our contributions are as follows:

- **Observable estimation with exponential bias reduction.** We construct gradient and gradient-change estimators using observed rewards and derivatives of the policy log density. With  $b$  randomized blocks for each of the burn-in and return durations, the gradient estimator

has bias  $O(2^{-b})$ , uniformly over policy parameters and starting populations, while its second moment and interaction cost are  $O(b)$  for fixed structural constants. We also bound the bias and second moment of the gradient-change estimator. Uniformity over starting populations allows these guarantees to be applied conditionally on the history, so estimation calls can continue from the population left by earlier calls.

- **Gradient domination beyond concavity.** We give sufficient conditions for gradient domination based on improvements available at individual observations (Appendix E). We verify exact gradient domination in a model with noisy observations and a nonconcave stationary objective (Appendix F).
- **Regret along one continuing trajectory.** Combining these estimates with gradient tracking (Ganesh et al., 2025) and stochastic Frank–Wolfe updates, we obtain  $\tilde{O}(\sqrt{T})$  expected regret under regularity, the uniform Poisson inverse condition, and exact gradient domination (Theorem 7.4). The guarantee holds from every initial population, using a supplied horizon and uniform bounds on the structural constants. The budget  $T$  counts every interaction, including all burn-in and estimation steps, and every collected reward contributes to the regret comparison.

Experiments illustrate the roles of randomized durations and finite-window estimation, and demonstrate learning with observation-dependent policies of up to 100 parameters.

## 2 Related Work

**Performative prediction and stateful control.** The classical formulation models the data distribution as a function of the deployed model alone. A model is performatively stable if it minimizes prediction risk on the distribution it induces, holding that distribution fixed when comparing models (Perdomo et al., 2020). Work in this setting studies stochastic retraining toward stable solutions (Mendler-Dünner et al., 2020) and direct risk optimization (Miller et al., 2021; Izzo et al., 2021). Jagadeesan et al. (2022) obtain regret guarantees from samples of the deployment-induced distribution, using this feedback to assess models that have not been deployed. Hardt and Mendler-Dünner (2025) survey the broader literature and the distinction between adapting to a distribution and steering it.

Stateful formulations also let the response depend on the current population. Brown et al. (2022) analyze convergence of retraining to an equilibrium, while Izzo et al. (2022) optimize long-term risk under a fixed deployed model. Li and Wai (2022) analyze stochastic approximation with samples from a decision-dependent controlled Markov chain, proving convergence to a performatively stable solution. He et al. (2025) develop a stochastic-gradient method for steady-state optimization under contracting distribution dynamics, using steady-state response sensitivities to obtain stationarity guarantees. For data from a continuing controlled Markov chain, Liu et al. (2024) obtain stationary-point guarantees through derivative-free optimization under geometric mixing.

Prior work also studies sequential control of evolving populations. Brown et al. (2024) obtain  $O(\sqrt{T})$  regret for known locally controllable dynamics and convex state losses, as well as sublinear guarantees for unknown locally action-linear dynamics and, separately, for bandit feedback. Their applications include stateful performative prediction. Chen and Hanasusanto (2026) study latent population distributions with finite-batch observations along one continuing trajectory without resets. Their optimistic planning approach uses finite-dimensional latent statistics, linear-in-feature models, and recoverability, with regret measured against a full-information finite-horizon controller.

Our focus is stationary average reward over a prescribed observation-based policy class, with regret against the best fixed policy using the same observation budget.

**Performative reinforcement learning.** Mandal et al. (2023) study discounted Markov decision processes (MDPs) whose reward and transition primitives depend on the deployed policy and establish performative stability through regularized retraining. Rank et al. (2024) also allow dependence on the previous environment, capturing gradual adaptation, and analyze regularized retraining toward stable policies. Policy gradients (Basu et al., 2026) and zeroth-order Frank–Wolfe optimization (Chen and Huang, 2025) target performative optimality for discounted objectives with policy-dependent environment primitives. In our formulation, policy dependence enters through the actions chosen from population samples: the response map and reward function are fixed, and the evolving population is hidden.

**Mean-field control and learning.** Distribution-valued population states also arise in mean-field control. Pham and Wei (2016) develop dynamic programming on these states, while Carmona et al. (2023) formulate discounted mean-field MDPs and develop Q-learning methods. Pásztor et al. (2023) give model-based regret guarantees for episodic mean-field control, where the learner observes the mean-field distribution and the system resets to a fixed initial distribution after each episode. Our learner has only finite-batch observations of the hidden population and cannot reset its trajectory.

**Policy gradients and partial observation.** GPOMDP estimates average-reward gradients in partially observable Markov decision processes (POMDPs) using observable log-density gradients (Baxter and Bartlett, 2001). We build on this principle with randomized durations and finite-trajectory bias and moment bounds. For finite POMDPs, Russo et al. (2025) obtain regret guarantees using known observation models, transition estimation, and belief-based planning. For fully observed ergodic MDPs, Ganesh et al. (2025) obtain a  $\tilde{O}(\sqrt{T})$  regret term through Hessian-aided gradient tracking, which informs our optimization procedure. Ganesh and Aggarwal (2025) also obtain a  $\tilde{O}(\sqrt{T})$  regret term using batched natural actor–critic under a unichain assumption: each fixed policy induces a chain with one recurrent class, while transient states and periodicity are allowed. Both analyses include additional approximation terms. Nguyen and Cheng (2023) study reset-free RL with sublinear regret and a sublinear reset count, whereas our interaction model permits no resets. Our analysis constructs observable gradient and gradient-change estimates for compact hidden population states under a uniform Poisson inverse bound, without requiring deterministic-time mixing or a learned value function of the hidden population (a latent-state critic).

**Policy improvement and global optimization.** Bhandari and Russo (2024) develop guarantees based on one-step improvement and policy-class closure, while Azizzadenesheli et al. (2020) study surrogate geometry and distribution mismatch for observation-based policies in episodic partially observed systems. We formulate policy improvement at each empirical observation and distinguish changes in observation frequencies from changes in the stationary hidden-population posterior. We give sufficient conditions under which observation-level improvements imply exact gradient domination of the stationary policy-class objective. Miller et al. (2021) also give convexity conditions involving loss curvature and distributional response.

### 3 Problem Formulation

Consider a firm choosing discounts using a sample of customers’ willingness to pay. After selecting a discount, it observes profit from a fresh customer drawn from the same population independently of that sample. Discounts can also change future willingness to pay, affecting later rewards and the population generating new samples. The firm therefore learns a discount policy from samples and

realized profits without knowing the full population or its response. We specify this interaction and the available feedback, then define the long-run reward of a fixed policy and the regret benchmark.

**Interaction model.** Let  $(\mathcal{Z}, d_{\mathcal{Z}})$  and  $(\mathcal{A}, d_{\mathcal{A}})$  be compact metric spaces of individual outcomes and actions. The hidden population distribution  $\mu_t$  belongs to a nonempty set  $\mathcal{D} \subseteq \mathcal{P}(\mathcal{Z})$  that is compact under  $d_{\mathcal{D}} := W_{1,\mathcal{Z}}$ , where  $\mathcal{P}(\mathcal{Z})$  denotes the Borel probability measures on  $\mathcal{Z}$ . Here  $W_{1,\mathcal{Z}}$  is the Wasserstein distance with ground metric  $d_{\mathcal{Z}}$ : the minimum expected ground distance over couplings, which are joint laws with the two prescribed marginals. We abbreviate it as  $W_1$ . For action distributions,  $W_{1,\mathcal{A}}$  uses the ground metric  $d_{\mathcal{A}}$ . A fixed unknown map  $F : \mathcal{D} \times \mathcal{A} \rightarrow \mathcal{D}$  determines the next population from the current population  $\mu_t$  and action  $A_t \in \mathcal{A}$ :

$$\mu_{t+1} = F(\mu_t, A_t). \quad (1)$$

Fix a batch size  $m \geq 1$  and a stochastic policy family  $\{\pi_{\theta} : \theta \in \Theta\}$ , with  $\Theta \subset \mathbb{R}^p$  compact and convex. Before the current batch is drawn, the learner selects  $\psi_t \in \Theta$  using its observed history. We use  $\psi_t$  for the parameter executed at physical time  $t$ , and  $\theta$  when discussing a fixed policy. Let  $\delta_z$  denote the point mass at outcome  $z$ . Conditional on the population and past history, the learner draws a batch and forms its empirical distribution:

$$Z_{t,1}, \dots, Z_{t,m} \stackrel{\text{iid}}{\sim} \mu_t, \quad \hat{\mu}_t = \frac{1}{m} \sum_{i=1}^m \delta_{Z_{t,i}}, \quad (2)$$

and draws  $A_t \sim \pi_{\psi_t}(\cdot \mid \hat{\mu}_t)$ . Thus, the trajectory is random through the observations and actions, even though (1) determines the next population once the current population and action are fixed. The observation domain  $\mathcal{O}_m$  contains all possible empirical measures, including those outside  $\mathcal{D}$ . The empirical measure in (2) is supported on finitely many sample points even when the admissible population distributions are smooth.

**Rewards and feedback.** The learner observes  $R_t \in [0, 1]$ , with unknown mean reward function  $r : \mathcal{D} \times \mathcal{A} \rightarrow [0, 1]$ . For analysis, let  $\mathcal{H}_t$  be the full history before the current batch, including  $\mu_t$  and  $\psi_t$ . We assume

$$\mathbb{E}[R_t \mid \mathcal{H}_t, \hat{\mu}_t, A_t] = r(\mu_t, A_t). \quad (3)$$

In the illustration, profit evaluated on a fresh customer drawn from  $\mu_t$  independently of the decision batch satisfies (3). The learner observes only sample batches, actions, and rewards along one continuing trajectory without resets, as summarized in Figure 1.

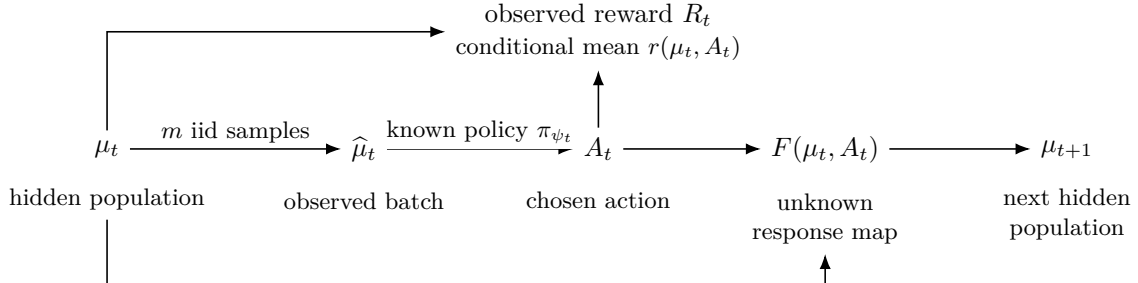


Figure 1: Interaction at time  $t$ . The policy samples  $A_t$  using the observed batch  $\hat{\mu}_t$ , while  $\mu_t$  remains hidden. The reward branch specifies only the conditional mean in (3). The unknown map  $F$  determines the next population through (1), and learning continues from that population without a reset.

**Dynamics under a fixed policy.** Let  $\mathbf{O}_m(d\xi \mid \mu)$  be the empirical observation law, and write  $\pi_\theta(a \mid \xi)$  for the policy density relative to a fixed reference measure  $da$ , with sums replacing integrals for discrete actions. The learner specifies the policy family and can evaluate its density and first and second parameter derivatives. The response map  $F$  and mean reward function  $r$  remain unknown. Since  $m$  is fixed throughout, we suppress it in the analytical quantities below. Averaging over observation noise gives

$$p_\theta(a \mid \mu) := \int_{\mathcal{O}_m} \pi_\theta(a \mid \xi) \mathbf{O}_m(d\xi \mid \mu). \quad (4)$$

The density  $p_\theta$  describes action selection from population  $\mu$  after averaging over the random observation batch. Under  $\pi_\theta$ , the next population's distribution depends on the past only through the current population, so the population process is Markov. Write  $\mathbb{E}^\theta$  for expectation under this fixed policy, with initialization specified by context, and  $\mathbb{E}_{\mu_0}^\theta$  when the trajectory starts from the fixed population  $\mu_0$ . For a bounded measurable function  $f : \mathcal{D} \rightarrow \mathbb{R}$ , we write the transition operator and expected reward as

$$(P_\theta f)(\mu) := \int_{\mathcal{A}} f(F(\mu, a)) p_\theta(a \mid \mu) da, \quad (5)$$

$$\bar{r}_\theta(\mu) := \int_{\mathcal{A}} r(\mu, a) p_\theta(a \mid \mu) da. \quad (6)$$

Thus  $(P_\theta f)(\mu)$  is the expected value of  $f$  at the next population. For example, when  $f$  equals one on a measurable set and zero elsewhere, the expectation is the probability of entering that set in one step. These probabilities determine the transition kernel. The quantity  $\bar{r}_\theta(\mu)$  is the expected immediate reward. These are analytical quantities and need not be computed by the learner.

**Objective and regret.** The regularity and Poisson assumptions in Section 4 ensure that each fixed policy induces a unique invariant population law  $\nu_\theta$  and has a well-defined long-run average reward. The law  $\nu_\theta \in \mathcal{P}(\mathcal{D})$  assigns probabilities to population distributions:  $\mu_t$  is a distribution over individual outcomes, whereas  $\nu_\theta$  is a law over such population states. Invariance means that one step under the fixed policy leaves this law unchanged. For a law  $\nu$  on  $\mathcal{D}$ , write  $\nu f := \int_{\mathcal{D}} f(\mu) \nu(d\mu)$  for the mean of  $f$  under  $\nu$ . Write  $\mathbb{E}_{\text{stat}}^\theta$  for expectation when  $\mu_0 \sim \nu_\theta$  and the trajectory follows  $\pi_\theta$ . Invariance gives  $\mu_t \sim \nu_\theta$  at every time. Averaging the expected immediate reward under  $\nu_\theta$  gives

$$J(\theta) := \nu_\theta \bar{r}_\theta. \quad (7)$$

The regularity and Poisson conditions also imply that  $J(\theta)$  is the limiting expected average reward from every initial population  $\mu_0 \in \mathcal{D}$ :

$$J(\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\mu_0}^\theta \left[ \sum_{t=0}^{N-1} R_t \right]. \quad (8)$$

This limit holds even when the probability law of  $\mu_t$  does not converge to  $\nu_\theta$ . Our benchmark is  $J^* := \max_{\theta \in \Theta} J(\theta)$ , using the same observation budget  $m$  at every step. For  $T$  interactions, define

$$\text{Reg}_T := T J^* - \sum_{t=0}^{T-1} R_t. \quad (9)$$

We seek sublinear expected regret from any initial population. Every interaction, including those used for estimation, counts toward  $T$  and contributes to the cumulative reward in (9).

## 4 Poisson Equations and Policy Gradients

Our goal in this section is to derive an expression for the gradient of (7). Changing the policy affects both immediate rewards and future population states, so the gradient must account for both effects.

We first state regularity assumptions and a uniform Poisson inverse condition. Under these assumptions, we establish a performance-difference identity and derive an average-reward policy-gradient formula. We then express the gradient using policy scores and relate it to finite observed returns. Section 5 constructs an estimator of this gradient from observations along a single continuing trajectory.

### 4.1 One-Step Regularity

We first control how a single interaction changes with the population, action, and policy parameter. The environment assumption bounds the sensitivity of the response map and reward, while the policy assumption controls dependence on the empirical observation and parameter. Together, they give regularity of the transition operator and expected reward after averaging over observation noise.

**Assumption 1** (Environment). *There are finite nonnegative constants  $q, L_a, L_r, L_u$  such that, for all  $\mu, \nu \in \mathcal{D}$  and  $a, b \in \mathcal{A}$ ,*

$$d_{\mathcal{D}}(F(\mu, a), F(\nu, b)) \leq q d_{\mathcal{D}}(\mu, \nu) + L_a d_{\mathcal{A}}(a, b), \quad (10)$$

$$|r(\mu, a) - r(\nu, b)| \leq L_r d_{\mathcal{D}}(\mu, \nu) + L_u d_{\mathcal{A}}(a, b). \quad (11)$$

The constants  $q, L_a$  bound next-population sensitivity to the current population and action, respectively, while  $L_r, L_u$  have the analogous roles for mean reward. Condition (10) bounds population-response sensitivity, as in the joint Wasserstein-sensitivity formulation of Brown et al. (2022, Definition 1), and holds for Lipschitz individual updates that preserve  $\mathcal{D}$ . Condition (11) similarly bounds mean-reward sensitivity and holds for expected bounded utilities that are jointly Lipschitz in individual outcomes and actions. These one-step regularity conditions allow atomic populations. Long-run control is supplied separately by Assumption 3.

For vectors,  $\|\cdot\|$  and  $\langle \cdot, \cdot \rangle$  denote the Euclidean norm and inner product, and  $\|\cdot\|_{\text{op}}$  is the induced operator norm. Unless another variable is indicated,  $\nabla$  and  $\nabla^2$  differentiate with respect to  $\theta$ , holding other arguments fixed. The space  $\mathcal{L}^1(da)$  consists of integrable functions modulo equality almost everywhere, with norm  $\|f\|_{\mathcal{L}^1} = \int_{\mathcal{A}} |f(a)| da$ . Differentiability in this norm controls changes in the whole action density. The policy score is the gradient of the log density, and the log-density Hessian is its derivative.

**Assumption 2** (Policy). *The policy family extends to an open neighborhood  $\mathcal{U}$  of  $\Theta$  and has a common positive support. For each observation  $\xi \in \mathcal{O}_m$ , the map  $\theta \mapsto \pi_\theta(\cdot \mid \xi)$  is twice continuously differentiable into  $\mathcal{L}^1(\mathrm{d}a)$ , with measurable derivative versions. There are finite nonnegative constants  $L_\pi, G, M, L_B$  such that, uniformly for  $\theta \in \mathcal{U}$  and  $\xi, \zeta \in \mathcal{O}_m$ ,*

$$W_{1,\mathcal{A}}(\pi_\theta(\cdot \mid \xi), \pi_\theta(\cdot \mid \zeta)) \leq L_\pi W_1(\xi, \zeta), \quad (12)$$

$$\|\nabla_\theta \log \pi_\theta(a \mid \xi)\| \leq G, \quad (13)$$

$$\left\| \nabla_\theta^2 \log \pi_\theta(a \mid \xi) \right\|_{\mathrm{op}} \leq M, \quad (14)$$

$$\int_{\mathcal{A}} \|\nabla_\theta \pi_\theta(a \mid \xi) - \nabla_\theta \pi_\theta(a \mid \zeta)\| \, \mathrm{d}a \leq L_B W_1(\xi, \zeta). \quad (15)$$

The log-density bounds hold for all actions in the common positive support.

The observation bounds (12) and (15) control how the action distribution and its parameter derivative change with the empirical observation. The score and Hessian bounds (13)–(14) control parameter sensitivity and the moments of the gradient and gradient-change estimates. Common positive support makes the log-density derivatives well-defined and permits likelihood comparisons across parameters. The extension to an open neighborhood gives these derivatives their usual meaning at boundary parameters.

Bound (12) is the observation-space analogue of the Wasserstein-Lipschitz policy condition in [Pirotta et al. \(2015, Assumption 2\(1\)\)](#). The score and log-density Hessian bounds (13)–(14) are also imposed by [Papini et al. \(2018, Assumption 4.1\)](#) and [Ganesh et al. \(2025, Assumption 2\)](#). Finite-action softmax policies  $\pi_\theta(a \mid \xi) \propto \exp(\theta^\top \phi_a(\xi))$  satisfy the full assumption, including common positive support and a smooth extension to a bounded neighborhood of  $\Theta$ , when the features  $\phi_a$  are uniformly bounded and Lipschitz in  $W_1$ . Appendix F verifies both regularity assumptions in a continuous-action model.

For  $f : \mathcal{D} \rightarrow \mathbb{R}$ , define its Lipschitz seminorm by  $\|f\|_{\mathrm{Lip}} := \sup_{\mu \neq \nu} |f(\mu) - f(\nu)| / W_1(\mu, \nu)$ . Coupling the sample batches and then the actions from two initial populations gives

$$\|\bar{r}_\theta\|_{\mathrm{Lip}} \leq L_r + L_u L_\pi, \quad (16)$$

and, for every Lipschitz  $f$ ,

$$\|P_\theta f\|_{\mathrm{Lip}} \leq (q + L_a L_\pi) \|f\|_{\mathrm{Lip}}. \quad (17)$$

The two terms in each bound account for the population’s direct effect and its effect through the sampled action. Thus the expected reward is Lipschitz, and  $P_\theta$  preserves Lipschitz regularity. Appendix A.1 gives the coupling argument. We next use this function space to control cumulative reward deviations along a trajectory.

## 4.2 Poisson Equations and Cumulative Deviations

To compare a finite trajectory with the stationary objective, we need to control cumulative deviations from stationary reward. More generally, consider a Lipschitz function  $f$  of the population. We seek a Lipschitz function  $h$  and a constant  $c$  satisfying the Poisson equation  $(I - P_\theta)h = f - c$ , where  $I$  is the identity operator. Along a trajectory under the fixed policy, summing expected deviations  $f(\mu_t) - c$  then gives the initial value of  $h$  minus its expected final value. A bound on  $h$  therefore controls the entire sum, regardless of the trajectory length.

The reward Poisson equation alone does not provide all the control needed for gradient estimation, which also requires controlling functions whose stationary averages give the policy gradient. We therefore impose a uniform Poisson inverse bound over the Lipschitz function space, which also allows repeated applications of the inverse in the randomized-block analysis.

Because  $P_\theta 1 = 1$ , adding a constant to  $h$  leaves  $h - P_\theta h$  unchanged. Let  $\text{Lip}(\mathcal{D})$  denote the Lipschitz functions and  $[f]$  the class of functions differing from  $f$  by a constant. The quotient space  $\dot{\text{Lip}}(\mathcal{D}) = \text{Lip}(\mathcal{D})/\mathbb{R}$ , with norm  $\|[f]\| = \|f\|_{\text{Lip}}$ , turns the Poisson equation into  $[(I - P_\theta)h] = [f]$ . It removes the additive-constant ambiguity so that an inverse can be defined. For a scalar function on  $\mathcal{D}$ , write  $\|f\|_\infty = \sup_{\mu \in \mathcal{D}} |f(\mu)|$ . For vector- and matrix-valued functions, sup norms and Lipschitz seminorms use the Euclidean and operator norms, respectively.

**Assumption 3** (Uniform bound on the Poisson inverse). *The induced operator  $I - P_\theta$  is invertible on  $\dot{\text{Lip}}(\mathcal{D})$ , and a known finite constant  $C_P$  satisfies*

$$\sup_{\theta \in \Theta} \|(I - P_\theta)^{-1}\|_{\dot{\text{Lip}} \rightarrow \dot{\text{Lip}}} \leq C_P. \quad (18)$$

A sufficient condition for Assumption 3 is  $q + L_a L_\pi < 1$ . In that case, (17) makes  $P_\theta$  a contraction on the quotient, and the Neumann series  $\sum_{k=0}^{\infty} P_\theta^k$  gives the valid upper bound  $C_P = 1/(1 - q - L_a L_\pi)$  (Rudin, 1991, Theorem 10.7). Here  $P_\theta^k$  denotes  $k$  successive transitions and  $P_\theta^0 = I$ .

Contraction is not necessary. On a deterministic two-state cycle,  $P = -I$  on the quotient and  $(I - P)^{-1} = \frac{1}{2}I$ . Starting from either state, the population law alternates forever, yet the Poisson inverse is bounded. This example explains why Assumption 3 can control cumulative deviations without deterministic-time mixing. On a finite state space, one can bound the quotient matrix inverse uniformly over  $\Theta$ ; Appendix F does so in (91).

The learner uses a supplied upper bound  $C_P$  to choose the random block lengths in Section 5. It does not need to compute the inverse. The following result establishes the invariant law and identifies the centering constant  $c$ .

**Proposition 4.1** (Invariant law and normalized Poisson solutions). *Under Assumptions 1–3, for every  $\theta \in \Theta$ , the kernel  $P_\theta$  has a unique invariant law  $\nu_\theta$ . Let  $D$  denote the diameter of  $\mathcal{D}$  under  $W_1$ . For every Lipschitz  $f$ , there is a unique Lipschitz solution  $h = \mathcal{R}_\theta f$  satisfying*

$$\begin{aligned} h - P_\theta h &= f - \nu_\theta f, & \nu_\theta h &= 0, \\ \|h\|_{\text{Lip}} &\leq C_P \|f\|_{\text{Lip}}, & \|h\|_\infty &\leq DC_P \|f\|_{\text{Lip}}. \end{aligned} \quad (19)$$

*The operator  $\mathcal{R}_\theta$  maps  $f$  to this normalized solution.*

The proof is in Appendix A.1.

Averaging both sides of the original equation  $h - P_\theta h = f - c$  under  $\nu_\theta$  gives  $c = \nu_\theta f$ , because invariance makes  $\nu_\theta(h - P_\theta h) = 0$ . The normalization  $\nu_\theta h = 0$  selects one solution from those differing by a constant. Its Lipschitz bound also gives a sup-norm bound through the diameter  $D$ , which is what the finite-trajectory comparison needs.

By (16), the expected reward lies in the domain of the Poisson solution operator. Applying Proposition 4.1 to  $\bar{r}_\theta$  yields the normalized differential value, or bias function,  $h_\theta := \mathcal{R}_\theta \bar{r}_\theta$ . It satisfies the average-reward Bellman equation (Sutton and Barto, 2018, Section 10.3)

$$h_\theta - P_\theta h_\theta = \bar{r}_\theta - J(\theta). \quad (20)$$

The normalization gives  $\nu_\theta h_\theta = 0$ . Taking expectations under the fixed policy and summing cancels the intermediate values of  $h_\theta$ , giving

$$\mathbb{E}_{\mu_0}^\theta \left[ \sum_{t=0}^{N-1} (R_t - J(\theta)) \right] = h_\theta(\mu_0) - \mathbb{E}_{\mu_0}^\theta h_\theta(\mu_N). \quad (21)$$

Thus the expected cumulative reward differs from  $NJ(\theta)$  by at most  $2\|h_\theta\|_\infty$ , independently of  $N$ . Combining the sup-norm bound in (19) with (21) gives

$$\left| \mathbb{E}_{\mu_0}^\theta \left[ \sum_{t=0}^{N-1} R_t \right] - NJ(\theta) \right| \leq 2DC_P(L_r + L_u L_\pi).$$

Dividing by  $N$  proves (8) from every initial population, with an  $O(N^{-1})$  error in expected average reward even when the population law at deterministic times does not converge. Appendix A.1 gives the cancellation calculation and the bounds. To improve this long-run performance, we now derive a formula for the policy gradient  $\nabla J(\theta)$ .

### 4.3 Performance Differences and the Policy Gradient

We first compare two fixed policies, using the differential value to account for their effects on future populations. Under Assumptions 1–3, the comparison takes the following form.

**Lemma 4.2** (A performance-difference identity). *For any  $\theta, \vartheta \in \Theta$ ,*

$$J(\vartheta) - J(\theta) = \nu_\vartheta[\bar{r}_\vartheta - \bar{r}_\theta + (P_\vartheta - P_\theta)h_\theta]. \quad (22)$$

To see the identity, substitute  $\bar{r}_\theta = J(\theta) + h_\theta - P_\theta h_\theta$  into the right-hand side and use  $\nu_\vartheta P_\vartheta h_\theta = \nu_\vartheta h_\theta$ . The full calculation is in Appendix A.2.

The term  $\bar{r}_\vartheta - \bar{r}_\theta$  compares immediate rewards, while  $(P_\vartheta - P_\theta)h_\theta$  compares the expected differential value of the next population under the two policies, using  $h_\theta$  from the reference policy  $\theta$ . Averaging these two terms under the new stationary law  $\nu_\vartheta$  gives the full performance difference. The choice of averaging law matters because changing the policy also changes how often different populations are visited.

To obtain the gradient from (22), we must also control how its averaging law changes as the new policy parameter approaches the reference parameter. For a Lipschitz  $f$ , set  $h = \mathcal{R}_\theta f$  in the Poisson equation (19). Averaging  $f - \nu_\theta f = h - P_\theta h$  under  $\nu_\vartheta$  and using invariance to cancel  $\nu_\vartheta(h - P_\vartheta h)$  leaves only the transition difference:

$$(\nu_\vartheta - \nu_\theta)f = \nu_\vartheta(P_\vartheta - P_\theta)\mathcal{R}_\theta f. \quad (23)$$

The Poisson bound and policy regularity control the right-hand side as  $\vartheta$  approaches  $\theta$ . This gives continuity of stationary expectations, allowing us to take the limit in (22).

With stationary expectations controlled, we can differentiate the two one-step quantities in (22). For a fixed function  $f$ , let  $(\nabla P_\theta)f$  denote the parameter derivative of  $P_\theta f$ , holding  $f$  fixed. In particular,  $(\nabla P_\theta)h_\theta$  differentiates the transition operator with the reference differential value held fixed.

**Theorem 4.3** (Average-reward policy gradient). *Under Assumptions 1–3, let*

$$b_\theta = \nabla \bar{r}_\theta + (\nabla P_\theta)h_\theta. \quad (24)$$

*Then*

$$\nabla J(\theta) = \nu_\theta b_\theta. \quad (25)$$

The proof, including continuity of stationary expectations, is in Appendix A.3.

The function  $b_\theta$  in (24) is the gradient integrand: averaging it under  $\nu_\theta$  gives  $\nabla J(\theta)$  in (25). Its two terms retain the immediate-reward and future-population effects from the performance comparison. Since  $r$  and  $F$  do not depend on  $\theta$  at fixed population and action, the derivatives in  $b_\theta$  act only on  $p_\theta$ . The formula therefore avoids computing  $\nabla \nu_\theta$  or differentiating the unknown response map  $F$ . The score and observation-derivative bounds (13) and (15), together with the regularity of the environment and  $h_\theta$ , also make  $b_\theta$  Lipschitz. Appendix B gives the derivation.

#### 4.4 Policy Scores and Finite Returns

To express the gradient using observed data, evaluate the policy score at the empirical distribution and action under the fixed policy:

$$S_t = \nabla \log \pi_\theta(A_t \mid \hat{\mu}_t). \quad (26)$$

The learner can compute the score in (26) even though the population is hidden. To differentiate the one-step averages in (5)–(6), substitute (4) and use the chain rule  $\nabla \pi_\theta = \pi_\theta \nabla \log \pi_\theta$ , as in likelihood-ratio gradient estimation under partial observation (Baxter and Bartlett, 2001). At fixed  $\mu$ , the observation law is independent of  $\theta$ , so differentiation acts only on the policy density. Conditioning on  $\mu_t = \mu$  fixes the current population, while the expectations average over the observation batch, sampled action, and any reward noise. Using (3) and holding  $h_\theta$  fixed gives

$$\nabla \bar{r}_\theta(\mu) = \mathbb{E}^\theta[S_t R_t \mid \mu_t = \mu], \quad (27)$$

$$((\nabla P_\theta)h_\theta)(\mu) = \mathbb{E}^\theta[S_t h_\theta(\mu_{t+1}) \mid \mu_t = \mu]. \quad (28)$$

Substituting these conditional score expectations into  $b_\theta$  and averaging over  $\mu_t \sim \nu_\theta$ , as prescribed by (25), gives

$$\nabla J(\theta) = \mathbb{E}_{\text{stat}}^\theta[S_t \{R_t + h_\theta(\mu_{t+1})\}]. \quad (29)$$

In (29), the reward  $R_t$  captures the immediate effect of action  $A_t$ . The term  $h_\theta(\mu_{t+1})$  evaluates the population reached after that action, with subsequent actions chosen under  $\pi_\theta$ . It accounts for the action’s future effect on rewards relative to the long-run average  $J(\theta)$ . The stationary expectation averages over the population, empirical observation, action, and reward. Although  $S_t$  and  $R_t$  are observed, the differential value  $h_\theta(\mu_{t+1})$  is unknown. The derivation is in Appendix A.4.

A useful property of the score is that its conditional mean is zero. Indeed, normalization gives

$$\mathbb{E}^\theta[S_t \mid \mu_t, \hat{\mu}_t] = \nabla \int_{\mathcal{A}} \pi_\theta(a \mid \hat{\mu}_t) \, da = 0. \quad (30)$$

The zero-mean property also holds conditional on the full pre-action history and current empirical observation. Consequently, subtracting a baseline fixed before the action is drawn does not change the expected score product. The baseline may depend on past observations and rewards or the current empirical observation, provided the product is integrable. For the analytical representation,

we subtract the baseline  $J(\theta) + h_\theta(\mu_t)$  from  $R_t + h_\theta(\mu_{t+1})$  in (29). Using (3) to replace  $R_t$  by  $r(\mu_t, A_t)$  inside the expected score product then yields the differential advantage

$$\mathcal{A}_\theta(\mu, a) = r(\mu, a) - J(\theta) + h_\theta(F(\mu, a)) - h_\theta(\mu). \quad (31)$$

The equivalent gradient representation is

$$\nabla J(\theta) = \mathbb{E}_{\text{stat}}^\theta [S_t \mathcal{A}_\theta(\mu_t, A_t)]. \quad (32)$$

The differential advantage in (31) measures the immediate reward above  $J(\theta)$  plus the change in differential value caused by the action.

**Finite observed returns.** The score formula (29) involves the unknown differential value  $h_\theta(\mu_{t+1})$ . To connect this formula to trajectory data, we iterate the reward Poisson equation (20), expressing this value as a finite sum of centered rewards plus an endpoint differential value while continuing under the same policy  $\pi_\theta$ .

Fix  $\theta$  and a starting population  $x$ , and use local indices with  $\mu_0 = x$ . The selected action  $A_0$  earns  $R_0$  and moves the population to  $\mu_1 = F(x, A_0)$ . Combining the definition (24) with the score identities (27)–(28) gives the quantity we aim to estimate at  $x$ :

$$b_\theta(x) = \mathbb{E}_x^\theta [S_0 \{R_0 + h_\theta(\mu_1)\}].$$

For an integer  $\ell \geq 0$ , collect the next  $\ell$  rewards. Applying (21) from  $\mu_1$  gives

$$h_\theta(\mu_1) = \mathbb{E}_x^\theta \left[ \sum_{j=1}^{\ell} (R_j - J(\theta)) + h_\theta(\mu_{\ell+1}) \middle| \hat{\mu}_0, A_0 \right].$$

The sum is empty when  $\ell = 0$ . Thus the differential value at  $\mu_1$  equals an expected sum of later centered rewards plus an expected endpoint value. Using only the finite reward sum omits that endpoint value and introduces return-truncation bias.

We use the observable score-weighted return  $S_0 \sum_{j=0}^{\ell} R_j$  as a finite-return estimate of  $b_\theta(x)$ . To identify its bias, multiply the preceding conditional identity by  $S_0$  and take expectations, which is valid because  $S_0$  is determined by  $\hat{\mu}_0, A_0$ . Substituting the result into the formula for  $b_\theta(x)$  above expresses this target through the observed return and the endpoint value. The centering term  $\ell J(\theta)$  contributes zero because  $\mathbb{E}_x^\theta S_0 = 0$  by (30). The estimator’s mean, denoted by  $g_{\theta, \ell}(x)$ , therefore satisfies

$$\begin{aligned} g_{\theta, \ell}(x) &:= \mathbb{E}_x^\theta \left[ S_0 \sum_{j=0}^{\ell} R_j \right] = b_\theta(x) - \mathbb{E}_x^\theta [S_0 h_\theta(\mu_{\ell+1})] \\ &= \underbrace{b_\theta(x)}_{\text{gradient integrand}} - \underbrace{[(\nabla P_\theta) P_\theta^\ell h_\theta](x)}_{\text{truncation term}}. \end{aligned} \quad (33)$$

Appendix A.5 derives this identity from (20). Here  $P_\theta^\ell h_\theta$  gives the expected endpoint differential value when the next  $\ell$  actions follow  $\pi_\theta$ . Applying  $\nabla P_\theta$  accounts for how the first action distribution affects this value, with those later transitions held fixed.

The full gradient is the stationary average  $\nabla J(\theta) = \nu_\theta b_\theta$ . Section 5 uses (33) to construct an estimator of this gradient from a finite segment of the continuing trajectory.

## 5 Gradient estimation from one trajectory

A finite-segment estimate must account for two sources of bias. First, the population at the start of the return need not have the stationary law  $\nu_\theta$ . We use a preliminary run of the fixed policy, called *burn-in*, to reduce this initialization error. Second, the finite return introduces the truncation term in (33).

Assumption 3 permits periodic population dynamics, so increasing fixed durations alone need not eliminate these errors. We randomize both durations, then control the estimator’s bias and second moment uniformly over its starting population. These uniform bounds apply to every call along the continuing trajectory, even though its starting population depends on the preceding interactions.

Algorithm 1 gives the complete estimation routine. Its inputs  $B, N$  are the numbers of randomized blocks in the burn-in and return durations, respectively, and  $s$  is the maximum length of each block. For each call, the routine saves its policy parameter, burn-in and return durations, and sequence of observations, actions, and rewards as a *trajectory record*. We first explain how to choose these lengths, then derive the estimate’s mean and control its bias and second moment.

---

### Algorithm 1 GRADIENTESTIMATE( $\theta; s, B, N$ )

---

**Require:** Continuing trajectory; fixed  $\theta$ ; integers  $s, B, N \geq 1$ .

**Ensure:** Estimate  $\hat{g}$  of  $\nabla J(\theta)$  and number of population interactions used; retain the complete call record.

1: Draw  $L_1, \dots, L_B, \tilde{L}_1, \dots, \tilde{L}_N$  independently from  $\text{Unif}\{1, \dots, s\}$ , independently of the trajectory.

2: Set  $U \leftarrow \sum_{i=1}^B L_i$  and  $V \leftarrow \sum_{i=1}^N \tilde{L}_i$ .

3: Record  $(\theta, U, V)$ .

*Burn-in and baseline.*

4: Run  $\pi_\theta$  for  $U$  decisions, recording  $(\hat{\mu}_t, A_t, R_t)$  for  $t = 0, \dots, U - 1$ .

5: Set  $\hat{J} \leftarrow \frac{1}{U} \sum_{t=0}^{U-1} R_t$ .

*Return and gradient estimate.*

6: Continue under  $\pi_\theta$  for  $V + 1$  decisions, recording  $(\hat{\mu}_t, A_t, R_t)$  for  $t = U, \dots, U + V$ .

7: Compute the first action’s score:  $S_U \leftarrow \nabla_\theta \log \pi_\theta(A_U \mid \hat{\mu}_U)$ .

8: Set  $Y \leftarrow \sum_{j=0}^V (R_{U+j} - \hat{J})$ ,  $\hat{g} \leftarrow S_U Y$ .

9: **return**  $(\hat{g}, U + V + 1)$ .

---

### 5.1 Randomized durations

For a process alternating between two states, averaging even and odd endpoints removes the oscillation in the population law at successive deterministic times. The following operator extends this averaging idea under the Poisson inverse bound. Using a supplied upper bound  $C_P$  from Assumption 3, set

$$s = \lceil 4 \max\{1, C_P\} \rceil, \quad Q_\theta = \frac{1}{s} \sum_{\ell=1}^s P_\theta^\ell. \quad (34)$$

The kernel  $P_\theta^\ell$  describes  $\ell$  physical steps, so averaging these kernels corresponds to drawing a duration uniformly from  $\{1, \dots, s\}$  and executing the fixed policy for that many steps. This constitutes one randomized block, with endpoint kernel  $Q_\theta$ . For a chosen integer  $b \geq 1$ , draw iid lengths

$L_i \sim \text{Unif}\{1, \dots, s\}$ ,  $i = 1, \dots, b$ , independently of the fixed-policy trajectory. Running these blocks consecutively uses between  $b$  and  $sb$  physical steps and has endpoint kernel  $Q_\theta^b$ .

Every physical step follows  $P_\theta$ , and all observations and rewards within a block are collected as usual. Randomization determines when a segment ends. Since  $\nu_\theta$  is invariant under  $P_\theta$  and  $Q_\theta$  is an average of powers of  $P_\theta$ , we also have  $\nu_\theta Q_\theta = \nu_\theta$ . Writing  $\alpha = \frac{2C_P}{s} \leq \frac{1}{2}$ , the endpoint error satisfies the following bound.

**Lemma 5.1** (Exponential accuracy at random times). *For every Lipschitz  $f$  and integer  $b \geq 1$ ,*

$$\left\| Q_\theta^b f - \nu_\theta f \right\|_\infty = \sup_{x \in \mathcal{D}} \left| (Q_\theta^b f)(x) - \nu_\theta f \right| \leq D \alpha^b \|f\|_{\text{Lip}} \leq D 2^{-b} \|f\|_{\text{Lip}}, \quad (35)$$

where  $D = \text{diam}_{W_1}(\mathcal{D})$ .

The proof is in Appendix B.1.

The expected value of a Lipschitz function at a randomized endpoint therefore approaches its stationary expectation geometrically in the number of blocks, uniformly over the initial population. This conclusion holds even when fixed-time population laws oscillate. The same bound holds in Euclidean norm for vector-valued functions, including the gradient integrand  $b_\theta$ . Below, we apply the lemma to  $b_\theta$  to control initialization error and to the differential value  $h_\theta$  to control the return tail.

## 5.2 Gradient estimation

The estimator in Algorithm 1 uses two consecutive segments: a randomized burn-in followed by a randomized return. Both segments use the same fixed parameter  $\theta$  and continue from the current population without a reset.

For integers  $B, N \geq 1$ , draw the segment lengths

$$U = \sum_{i=1}^B L_i, \quad V = \sum_{i=1}^N \tilde{L}_i, \quad L_i, \tilde{L}_i \stackrel{\text{iid}}{\sim} \text{Unif}\{1, \dots, s\}, \quad (36)$$

independently of the trajectory. Here  $B, N$  count randomized blocks, while  $U$  counts burn-in steps and  $V$  counts the steps after the first return action. We index the physical steps within this call from zero.

The first  $U$  steps form the burn-in segment and reduce initialization error. Centering subsequent rewards at  $J(\theta)$  removes an accumulated average-reward component whose product with the policy score has zero expectation but can add substantial noise. Since  $J(\theta)$  is unknown, we estimate it using the burn-in rewards:

$$\hat{J} = \frac{1}{U} \sum_{t=0}^{U-1} R_t. \quad (37)$$

Starting at time  $U$ , we continue under  $\pi_\theta$  for another  $V + 1$  steps and use  $\hat{J}$  as the baseline to form the centered return

$$Y = \sum_{j=0}^V (R_{U+j} - \hat{J}). \quad (38)$$

Multiplying this return by the score of its first action gives

$$\hat{g} = S_U Y, \quad S_U = \nabla_\theta \log \pi_\theta(A_U \mid \hat{\mu}_U). \quad (39)$$

The score  $S_U$  weights both the reward at time  $U$  and the subsequent rewards, accounting for the effect of  $A_U$  on future populations. Since  $\hat{J}$  is computed before  $A_U$  is drawn, the score-zero identity (30) ensures that subtracting this baseline does not change the estimator's mean. Every quantity in (39) is computed from observed data and derivatives of the known policy family.

Each call uses  $U + V + 1$  physical steps, with maximum length

$$n = s(B + N) + 1. \quad (40)$$

All of these steps count toward the interaction budget and regret. The routine also retains  $(\theta, U, V)$  and the full sequence of observations, actions, and rewards for the gradient-change procedure in Section 6. The following subsections analyze the output  $\hat{g}$  of Algorithm 1, first bounding its bias relative to  $\nabla J(\theta)$  and then its second moment.

### 5.3 Uniform bias control

For fixed  $\theta \in \Theta$  and starting population  $x \in \mathcal{D}$ , let

$$f_x(\theta) := \mathbb{E}_x^\theta[\hat{g}],$$

where  $x$  is the population before burn-in and the expectation averages over all randomness in the call. We first compute this mean, then bound its difference from  $\nabla J(\theta)$  uniformly over  $x$ .

**Lemma 5.2** (Mean of the gradient estimate). *The estimator in (39) satisfies*

$$f_x(\theta) = \left\{ Q_\theta^B \left[ b_\theta - (\nabla P_\theta) Q_\theta^N h_\theta \right] \right\} (x). \quad (41)$$

The proof is in Appendix C.1.

The baseline does not contribute to this mean. To see this, condition on  $U, V$  and the history before  $A_U$  is drawn, including  $\hat{\mu}_U$ . Both  $V$  and  $\hat{J}$  are then fixed, while  $S_U$  has conditional mean zero by (30). Hence

$$\mathbb{E}_x^\theta[(V + 1)S_U \hat{J}] = 0.$$

This cancellation holds even though the baseline and the subsequent rewards come from the same trajectory.

After removing the baseline from the expectation, (33) gives

$$f_x(\theta) = \mathbb{E}_x^\theta[g_{\theta,V}(\mu_U)] = \mathbb{E}_x^\theta[(P_\theta^U g_{\theta,V})(x)].$$

The second equality conditions on  $U, V$  and averages over the population reached after  $U$  burn-in transitions. Substituting (33) and averaging over the independent lengths replaces  $P_\theta^U$  and  $P_\theta^V$  by  $Q_\theta^B$  and  $Q_\theta^N$ , respectively, yielding (41).

Using  $\nabla J(\theta) = \nu_\theta b_\theta$ , we obtain

$$f_x(\theta) - \nabla J(\theta) = (Q_\theta^B b_\theta)(x) - \nu_\theta b_\theta - [Q_\theta^B (\nabla P_\theta) Q_\theta^N h_\theta](x).$$

The first difference is the initialization error remaining after burn-in. The final term is the error from truncating the return. To bound these two terms, define

$$L_0 = L_r + L_u L_\pi, \quad H_0 = D C_P L_0, \quad (42)$$

$$C_g = G L_r + L_B + (qG + D L_B) C_P L_0. \quad (43)$$

The regularity estimates established in Appendix B.2 give

$$\|\bar{r}_\theta\|_{\text{Lip}} \leq L_0, \quad \|h_\theta\|_{\text{Lip}} \leq C_P L_0, \quad \|b_\theta\|_{\text{Lip}} \leq C_g.$$

They also give  $\|h_\theta\|_\infty \leq H_0$ , which will be used in the second-moment analysis.

**Proposition 5.3** (Uniform gradient bias). *For every  $\theta \in \Theta$ ,*

$$\sup_{x \in \mathcal{D}} \|f_x(\theta) - \nabla J(\theta)\| \leq \varepsilon_{B,N} := \underbrace{DC_g \alpha^B}_{\text{burn-in bias}} + \underbrace{GDC_P L_0 \alpha^N}_{\text{return truncation}}. \quad (44)$$

The proof is in Appendix B.2.

For the initialization error, apply Lemma 5.1 to  $b_\theta$ . For the return-truncation error, apply the same lemma to  $h_\theta$ , using  $\nu_\theta h_\theta = 0$ . The score bound gives  $\|(\nabla P_\theta)v\|_\infty \leq G\|v\|_\infty$ , and  $Q_\theta^B$  does not increase the supremum norm. Together, these estimates give (44).

Because the bound is uniform in both  $\theta$  and  $x$ , it also holds conditional on the history before each call. This history may determine the chosen parameter and the distribution of the starting population. Averaging over that conditional distribution preserves the bound. Thus no stationary initialization or reset is required, provided each call draws fresh independent lengths and holds its chosen parameter fixed throughout.

## 5.4 Second-moment control

We next bound the estimator's second moment, which is needed for gradient tracking. By  $\hat{g} = S_U Y$  and the score bound  $\|S_U\| \leq G$ ,

$$\|\hat{g}\| = \|S_U\| |Y| \leq G|Y|.$$

It therefore suffices to bound  $\mathbb{E}_x^\theta Y^2$ .

We obtain the required reward-sum bounds from the reward Poisson equation (20). For any segment of length  $\ell \geq 1$  starting at time  $t$ , it gives the decomposition

$$\sum_{k=t}^{t+\ell-1} (R_k - J(\theta)) = h_\theta(\mu_t) - h_\theta(\mu_{t+\ell}) + \sum_{k=t}^{t+\ell-1} \Delta_{k+1},$$

where  $\Delta_{k+1} := R_k - J(\theta) + h_\theta(\mu_{k+1}) - h_\theta(\mu_k)$  has conditional mean zero given the full history before step  $k$  by the Poisson equation (Glynn and Meyn, 1996).

The endpoint difference is bounded in absolute value by  $2H_0$ , and each martingale difference satisfies  $|\Delta_{k+1}| \leq 1 + 2H_0$ . Distinct martingale differences have zero cross moments, so their sum has second moment at most  $\ell(1 + 2H_0)^2$ . Combining these bounds using  $(a + b)^2 \leq 2a^2 + 2b^2$  gives a second-moment bound of  $C_R \ell$ , uniformly over the starting population, where

$$C_R = 8H_0^2 + 2(1 + 2H_0)^2. \quad (45)$$

**Proposition 5.4** (Second moment of the gradient estimate). *For every  $\theta \in \Theta$  and starting population  $x \in \mathcal{D}$ ,*

$$\mathbb{E}_x^\theta Y^2 \leq \mathcal{U}_{B,N} := 2C_R \left[ sN + 1 + \frac{(sN + 1)^2}{B} \right]. \quad (46)$$

Consequently,

$$\mathbb{E}_x^\theta \|\hat{g}\|^2 \leq \mathcal{V} := G^2 \mathcal{U}_{B,N}. \quad (47)$$

The proof, including the conditional reward-sum bound, is in Appendix C.2.

The two contributions in (46) follow from

$$Y = \sum_{j=0}^V (R_{U+j} - J(\theta)) - (V+1)(\hat{J} - J(\theta)).$$

The first term is the return centered at the true average reward, whose second moment is at most  $C_R(sN+1)$ . For the second term, applying the same reward-sum bound to the burn-in segment gives

$$\mathbb{E}_x^\theta [(\hat{J} - J(\theta))^2 \mid U] \leq \frac{C_R}{U} \leq \frac{C_R}{B}.$$

The baseline error is multiplied by  $V+1$ , so its contribution to the second moment is at most  $C_R(sN+1)^2/B$ . Combining the two terms using  $(a-b)^2 \leq 2a^2 + 2b^2$  yields (46), without requiring independence between the two segments. In particular, the baseline cancellation used for bias control does not eliminate the baseline’s contribution to the second moment.

With  $B = N = b$  and fixed structural constants, both terms in (46) are  $O(b)$ : increasing the burn-in along with the return prevents baseline error from contributing a quadratic term. Together with (44), this gives geometrically decreasing bias  $O(2^{-b})$ , while the second moment and maximum call length are  $O(b)$ . Like the bias bound, the moment bound holds conditional on each call’s preceding history. These conditional guarantees will control estimation errors as the policy is updated along the continuing trajectory in Section 7.

## 6 Estimating Gradient Changes

Even when the estimates in Section 5 have small bias, each fresh estimate requires a trajectory segment and remains noisy. We use gradient tracking to retain information from earlier gradient estimates across policy updates. Since the previous estimate targets the gradient at the preceding policy parameter, we adjust it using an estimate of the gradient change before combining it with a fresh estimate in a weighted average. This section constructs that gradient-change estimate.

The correction and the fresh gradient estimate use separate trajectory segments. The correction’s second-moment bound is proportional to the squared distance between the two parameters, allowing small policy updates to keep its noise controlled. The construction follows the Hessian-aided policy-gradient approach of Ganesh et al. (2025), applied to the finite-trajectory estimator of Section 5.

### 6.1 Construction of the gradient-change estimator

Fix endpoints  $\vartheta, \theta \in \Theta$ , write  $\delta = \theta - \vartheta$ , and fix a starting population  $x$ . Recall that  $f_x(\theta) = \mathbb{E}_x^\theta[\hat{g}]$  is the mean of GRADIENTESTIMATE. With  $Df_x$  denoting its Jacobian with respect to the policy parameter, the fundamental theorem of calculus gives

$$f_x(\theta) - f_x(\vartheta) = \int_0^1 Df_x(\vartheta + u\delta) \delta \, du.$$

Thus we can estimate the difference by sampling  $u \sim \text{Unif}[0, 1]$  and estimating the Jacobian-vector product at  $\tilde{\theta} = \vartheta + u\delta$ . Convexity of  $\Theta$  keeps this parameter feasible. Both endpoint means in this

identity use the same starting population  $x$ , but the construction requires only one trajectory under  $\tilde{\theta}$ . In an actual call,  $x$  is the current hidden population; the procedure need not observe or reset it. To construct the Jacobian-vector estimate, consider a record collected under a fixed parameter  $\theta$ , and write

$$M_t = \nabla_{\theta}^2 \log \pi_{\theta}(A_t | \hat{\mu}_t), \quad S_{\text{traj}} = \sum_{t=0}^{U+V} S_t.$$

Here  $M_t$  is the derivative of the policy score, and  $S_{\text{traj}}$  is the score of the observation–action trajectory likelihood. Changing  $\theta$  affects both the score factor in  $\hat{g} = S_U Y$  and the law of the observation–action record. Holding the recorded observations, actions, rewards, and lengths fixed leaves  $Y$  with no explicit parameter dependence. Differentiating the score factor in  $S_U Y$  contributes  $Y M_U \delta$ , while differentiating the trajectory likelihood contributes  $Y S_U \langle S_{\text{traj}}, \delta \rangle$ . Together, these terms give

$$Df_x(\theta) \delta = \mathbb{E}_x^{\theta}[Y (M_U \delta + S_U \langle S_{\text{traj}}, \delta \rangle)].$$

The trajectory score includes the burn-in actions because they affect the population at which the return begins. The laws of  $U, V$  do not depend on the parameter, so the lengths contribute no additional score term.

The proof in Appendix D.1 first replaces rewards by their conditional means from (3) before differentiation, then uses conditional score identities to recover the displayed expression in observed rewards. The replacement is confined to the proof, and no derivatives of the full reward distribution are required. Algorithm 2 evaluates the resulting expression at the sampled intermediate parameter.

---

**Algorithm 2** GRADIENTCHANGE( $\vartheta, \theta; s, B, N$ )

---

**Require:** Continuing trajectory; endpoints  $\vartheta, \theta \in \Theta$ ; integers  $s, B, N \geq 1$ .

**Ensure:** Estimate  $\hat{v}$  of  $\nabla J(\theta) - \nabla J(\vartheta)$  and number of population interactions used.

- 1: Set  $\delta \leftarrow \theta - \vartheta$ .
  - 2: Draw  $u \sim \text{Unif}[0, 1]$ , independently of the past.
  - 3: Set  $\tilde{\theta} \leftarrow \vartheta + u\delta$ .  
*Collect a trajectory under the intermediate policy.*
  - 4: Run GRADIENTESTIMATE( $\tilde{\theta}; s, B, N$ ), retaining its call record.
  - 5: Retrieve  $U, V$  and the recorded observations, actions, and rewards from this call.
  - 6: Compute  $Y$  using (37)–(38).  
*Compute the gradient-change estimate using the saved trajectory data.*
  - 7: Compute  $S_t$  for  $t = 0, \dots, U + V$ , evaluated at  $\tilde{\theta}$ .
  - 8: Set  $S_{\text{traj}} \leftarrow \sum_{t=0}^{U+V} S_t$ .
  - 9: Compute the Hessian-vector product  $M_U \delta$ , evaluated at  $\tilde{\theta}$ .
  - 10: Set  $\hat{v} \leftarrow Y (M_U \delta + S_U \langle S_{\text{traj}}, \delta \rangle)$ .
  - 11: **return**  $(\hat{v}, U + V + 1)$ .
- 

The call to GRADIENTESTIMATE under  $\tilde{\theta}$  supplies the saved trajectory data used to compute  $\hat{v}$ . The gradient estimate returned by this call targets the intermediate policy and is unused. Computing  $\hat{v}$  from these data requires no further population interactions, so the maximum cost of Algorithm 2 is the same  $n = s(B + N) + 1$  as in (40). Its second-derivative term uses a Hessian-vector product, without forming or storing the full Hessian matrix.

## 6.2 Bias and second-moment bounds

To control the tracking error, we need to quantify the correction’s bias relative to the true gradient change and bound its second moment. We first analyze a Jacobian estimate at a fixed parameter, then apply the result at the single sampled intermediate parameter. For a record collected under  $\theta$ , define the random matrix  $\hat{A}$  that estimates the Jacobian  $Df_x(\theta)$ :

$$\hat{A} := Y(M_U + S_U S_{\text{traj}}^\top). \quad (48)$$

Algorithm 2 evaluates  $\hat{A}\delta$  at  $\tilde{\theta}$ , without forming the matrix. Here  $n$  and  $\mathcal{U}_{B,N}$  are the maximum record length and return-moment bound from (40) and (46).

**Lemma 6.1** (A Jacobian estimate). *For every fixed starting population  $x$  and parameter  $\theta$ ,*

$$\mathbb{E}_x^\theta \hat{A} = Df_x(\theta), \quad \mathbb{E}_x^\theta \|\hat{A}\|_{\text{op}}^2 \leq \mathcal{W} := (M + nG^2)^2 \mathcal{U}_{B,N}.$$

The proof is in Appendix D.1. By the lemma and the fundamental theorem of calculus, the product  $\hat{A}\delta$  computed at the single sampled intermediate parameter is an unbiased estimate of  $f_x(\theta) - f_x(\vartheta)$ . The expectation includes both the parameter draw and the trajectory data. The uniform gradient-bias bound (44) then controls its difference from the true stationary gradient change.

**Proposition 6.2** (Gradient-change bounds). *Fix the endpoints  $\vartheta, \theta \in \Theta$  before the call and a starting population  $x$ . Write  $\mathbb{E}_x$  for expectation over the sampled parameter and the trajectory collected under it. Then*

$$\mathbb{E}_x \hat{v} = f_x(\theta) - f_x(\vartheta). \quad (49)$$

With  $\mathcal{W}$  as in Lemma 6.1,

$$\|\mathbb{E}_x \hat{v} - (\nabla J(\theta) - \nabla J(\vartheta))\| \leq 2\varepsilon_{B,N}, \quad (50)$$

$$\mathbb{E}_x \|\hat{v}\|^2 \leq \mathcal{W} \|\theta - \vartheta\|^2. \quad (51)$$

The proof is in Appendix D.2.

Combining (49) with the endpoint bias bound (44) gives (50): each endpoint contributes at most  $\varepsilon_{B,N}$ . No derivative bound on the estimation bias is required.

For the second moment, the policy derivative bounds and the maximum record length give

$$\|\hat{v}\| \leq |Y|(\|M_U\|_{\text{op}} + \|S_U\| \|S_{\text{traj}}\|) \|\delta\| \leq |Y|(M + nG^2) \|\delta\|.$$

Applying (46) at the intermediate parameter yields (51). When  $B = N = b$  and the structural constants are fixed,  $\mathcal{W} = O(b^3)$ , but the moment bound is multiplied by  $\|\theta - \vartheta\|^2$ . Section 7 uses decreasing policy step sizes to exploit this factor.

Both bounds are uniform over endpoints and starting populations. They therefore remain valid conditional on the history before a call, when that history determines the endpoints and the distribution of the current population. Each call draws a fresh intermediate parameter and fresh lengths and holds the intermediate parameter fixed while collecting its record.

## 7 Policy Optimization and Regret

We combine the estimators of Sections 5 and 6 with gradient tracking and Frank–Wolfe updates. We first state the additional condition that connects first-order improvement to global reward, then analyze the updates and regret along the continuing trajectory. All interactions used for estimation count toward regret.

### 7.1 Gradient domination

Accurate gradients alone do not guarantee global optimization of a nonconcave objective. On the constrained domain  $\Theta$ , the Frank–Wolfe gap measures the largest gain predicted by the gradient over feasible parameter changes (Jaggi, 2013). We require this gain to control the remaining reward gap.

**Assumption 4** (Gradient domination by the Frank–Wolfe gap). *There are known constants  $\kappa \geq 1$  and  $\epsilon_{\text{geom}} \geq 0$  such that, for every  $\theta \in \Theta$ ,*

$$J^* - J(\theta) \leq \kappa \mathcal{G}(\theta) + \epsilon_{\text{geom}}, \quad \mathcal{G}(\theta) = \max_{v \in \Theta} \langle \nabla J(\theta), v - \theta \rangle. \quad (52)$$

*A measurable linear maximizer over  $\Theta$  is available.*

The gap  $\mathcal{G}(\theta)$  is nonnegative because  $v = \theta$  is feasible. If it vanishes, (52) places the policy within  $\epsilon_{\text{geom}}$  of the best reward in the same fixed- $m$  policy class. With zero error, this is the degree-one gradient-domination condition of Bhandari and Russo (2024, Definition 2), written for reward maximization.

Concavity of  $J$  implies (52) with  $\kappa = 1$  and  $\epsilon_{\text{geom}} = 0$ , but is not necessary. Appendix E gives sufficient conditions under partial observation, and Appendix F verifies them with  $\kappa = 3$  and zero error for a nonconcave objective with noisy observations. The regret analysis uses Assumption 4 directly; it does not require that particular set of sufficient conditions. A nonzero  $\epsilon_{\text{geom}}$  contributes  $T\epsilon_{\text{geom}}$  to the bound.

The linear maximization oracle is a separate computational requirement: it returns a maximizer of a supplied linear objective over  $\Theta$ . Boxes and simplices admit explicit maximizers with fixed tie-breaking.

### 7.2 Policy updates and gradient tracking

We call each iteration of estimation and policy updating an epoch. At epoch  $k$ , we first collect a correction  $\hat{v}_k$  for the move from  $\theta_{k-1}$  to  $\theta_k$ , then a fresh gradient estimate  $\hat{g}_k$  at  $\theta_k$ . Each call uses fresh randomization and continues from the population left by the preceding interactions.

Fix  $s$  as in (34) and block counts  $B, N \geq 1$ . Write

$$n = s(B + N) + 1, \quad \eta_k = \min\{1, 2\kappa/k\},$$

where  $\kappa$  is the constant in Assumption 4.

Starting from  $g_0 = 0$  and  $\theta_0 = \theta_1$ , set

$$g_k = (1 - 1/k)(g_{k-1} + \hat{v}_k) + (1/k)\hat{g}_k, \quad (53)$$

$$w_k \in \arg \max_{w \in \Theta} \langle g_k, w \rangle, \quad \theta_{k+1} = \theta_k + \eta_k(w_k - \theta_k). \quad (54)$$

The tracking update (53) combines the corrected estimate  $g_{k-1} + \hat{v}_k$ , which retains information from earlier epochs, with a fresh estimate weighted by  $1/k$ . At  $k = 1$ , the correction is zero and  $g_1 = \hat{g}_1$ . The Frank–Wolfe update (54) uses  $g_k$  to select a maximizer of the linear objective and takes a step toward it. Its convex-combination form keeps the iterates in  $\Theta$ .

Each call uses at most  $n$  physical steps. To give every epoch a fixed length, we continue  $\pi_{\theta_k}$  after the two calls until exactly  $2n$  steps have been used. This gives  $K = \lfloor T/(2n) \rfloor$  completed epochs under a horizon of  $T$  steps. Algorithm 3 also specifies the policy used for any remaining steps.

---

**Algorithm 3** Policy optimization on one continuing trajectory

---

**Require:** Integer horizon  $T \geq 1$ , initial  $\theta_1 \in \Theta$ , integers  $s, B, N \geq 1$ , step sizes  $\eta_k \in (0, 1]$ , and a linear maximizer over  $\Theta$ .

- 1: Set  $\theta_0 \leftarrow \theta_1$ ,  $g_0 \leftarrow 0$ ,  $n \leftarrow s(B + N) + 1$ ,  $K \leftarrow \lfloor T/(2n) \rfloor$ .
  - 2: **if**  $K = 0$  **then**
  - 3:   Execute  $\theta_1$  for all  $T$  steps and stop.
  - 4: **end if**
  - 5: **for**  $k = 1, \dots, K$  **do**
  - 6:    $(\hat{v}_k, \ell_k^v) \leftarrow \text{GRADIENTCHANGE}(\theta_{k-1}, \theta_k; s, B, N)$ .
  - 7:    $(\hat{g}_k, \ell_k^g) \leftarrow \text{GRADIENTESTIMATE}(\theta_k; s, B, N)$ .
  - 8:   Execute  $\theta_k$  for  $2n - \ell_k^v - \ell_k^g$  further steps.
  - 9:    $g_k \leftarrow (1 - 1/k)(g_{k-1} + \hat{v}_k) + (1/k)\hat{g}_k$ .
  - 10:   Choose  $w_k \in \arg \max_{w \in \Theta} \langle g_k, w \rangle$ .
  - 11:    $\theta_{k+1} \leftarrow \theta_k + \eta_k(w_k - \theta_k)$ .
  - 12: **end for**
  - 13: Execute  $\theta_K$  for the remaining  $T - 2nK$  steps.
- 

The two calls in an epoch generally start from different, dependent populations. Each satisfies its bias and moment bounds conditional on its own preceding history. Recall that  $\mathcal{V}$  bounds the second moment of a fresh estimate in (47), while  $\mathcal{W}\|\theta_k - \theta_{k-1}\|^2$  bounds that of the gradient-change estimate in (51). These conditional bounds give the following tracking guarantee.

**Lemma 7.1** (Gradient-tracking error). *Let  $D_\Theta = \text{diam}(\Theta)$ , and define  $\sigma = \sqrt{\mathcal{V} + 4\kappa^2 D_\Theta^2 \mathcal{W}}$ . For every completed epoch  $k$ ,*

$$\left(\mathbb{E}\|g_k - \nabla J(\theta_k)\|^2\right)^{1/2} \leq \frac{\sigma}{\sqrt{k}} + k\varepsilon_{B,N}. \quad (55)$$

The proof is in Appendix G.1.

The first term controls the accumulated noise, including noise in the gradient-change estimates. Multiplying (53) by  $k$  gives weight  $k - 1$  to the correction at epoch  $k$ . For  $k \geq 2$ , the policy update ensures  $\|\theta_k - \theta_{k-1}\| \leq 2\kappa D_\Theta / (k - 1)$ . Consequently, (51) bounds the second moment of this weighted correction by  $4\kappa^2 D_\Theta^2 \mathcal{W}$ , while the fresh estimate contributes at most  $\mathcal{V}$ . After subtracting each contribution's conditional mean given its preceding history, the cross moments vanish. Summing these noise contributions and dividing by  $k$  gives the term  $\sigma/\sqrt{k}$ .

The bias contributions accumulate to the second term,  $k\varepsilon_{B,N}$ . We therefore choose the block counts as a function of the horizon so that this term stays small throughout the run. The argument uses the conditional bounds for each call and requires neither independent trajectories nor stationary starting populations.

### 7.3 Expected regret

To obtain a regret bound, we must account for both the quality of the policies used during learning and the transient rewards collected while the population responds to them. Let  $\psi_t$  denote the parameter actually used at physical time  $t$ . We decompose expected regret as

$$\mathbb{E} \text{Reg}_T = \underbrace{\mathbb{E} \sum_{t=0}^{T-1} (J^* - J(\psi_t))}_{\text{stationary reward shortfall}} + \underbrace{\mathbb{E} \sum_{t=0}^{T-1} (J(\psi_t) - R_t)}_{\text{transient reward discrepancy}}. \quad (56)$$

We bound the first term using policy improvement and the second using the Poisson equation, then choose the trajectory lengths to control their sum. Both arguments require regularity across policy parameters.

**Regularity across policies.** To compare differential values across policies, fix one reference population  $\mu_{\text{ref}}$  and set

$$h_{\theta}^{\text{ref}}(\mu) = h_{\theta}(\mu) - h_{\theta}(\mu_{\text{ref}}).$$

This common normalization preserves the Poisson equation and gives  $\|h_{\theta}^{\text{ref}}\|_{\infty} \leq D \|h_{\theta}\|_{\text{Lip}} \leq H_0$ . It lets us control the transient cost of policy changes through differences in  $h_{\theta}^{\text{ref}}$ . These functions are used only in the analysis. The following comparison also uses  $C_g$ , the bound on  $\|b_{\theta}\|_{\text{Lip}}$  defined in (43).

**Lemma 7.2** (Regularity across policy parameters). *Define  $L_J = (M + G^2)(1 + H_0) + 2GD_C C_g$ . For all  $\theta, \vartheta \in \Theta$ ,*

$$\|h_{\vartheta}^{\text{ref}} - h_{\theta}^{\text{ref}}\|_{\infty} \leq DC_P C_g \|\vartheta - \theta\|, \quad (57)$$

$$|J(\vartheta) - J(\theta)| \leq G(1 + H_0) \|\vartheta - \theta\|, \quad (58)$$

$$\|\nabla J(\vartheta) - \nabla J(\theta)\| \leq L_J \|\vartheta - \theta\|. \quad (59)$$

The proof is in Appendix B.3. The first bound controls changes in differential values and hence transient rewards. The second controls the stationary loss from executing an intermediate parameter during a gradient-change call. The third says that  $\nabla J$  is  $L_J$ -Lipschitz, which controls the error in approximating an objective improvement by its first-order term. All three follow from the environment, policy, and Poisson assumptions.

**Bounding the stationary reward shortfall.** We first control the reward gaps of the iterates  $\theta_k$ . The choice of  $w_k$  gives

$$\langle \nabla J(\theta_k), w_k - \theta_k \rangle \geq \mathcal{G}(\theta_k) - D_{\Theta} \|g_k - \nabla J(\theta_k)\|.$$

Thus the tracking error reduces the available linear gain by at most its norm times  $D_{\Theta}$ . The smoothness bound (59) controls the additional error from taking a finite step. Combining these facts with Assumption 4 and Lemma 7.1, and writing  $a_k = \mathbb{E}[J^* - J(\theta_k)]$ , gives

$$\begin{aligned} a_{k+1} &\leq \left(1 - \frac{\eta_k}{\kappa}\right) a_k + \frac{\eta_k}{\kappa} \epsilon_{\text{geom}} \\ &\quad + \eta_k D_{\Theta} \left( \frac{\sigma}{\sqrt{k}} + k \varepsilon_{B,N} \right) + \frac{L_J D_{\Theta}^2}{2} \eta_k^2. \end{aligned}$$

For fixed structural constants, this recursion yields  $a_k \leq \epsilon_{\text{geom}} + O((1 + \sigma)/\sqrt{k} + k \varepsilon_{B,N})$ .

To bound the first term in (56), we must also account for the policies used to collect data. A gradient-change call executes an intermediate parameter on the segment from  $\theta_{k-1}$  to  $\theta_k$ . By (58), its stationary reward shortfall exceeds that of  $\theta_k$  by at most  $G(1+H_0)\|\theta_k - \theta_{k-1}\|$ . The call uses at most  $n$  interactions, and the entire epoch uses  $2n$ . The step sizes give  $\sum_{k=2}^K \|\theta_k - \theta_{k-1}\| \leq 2\kappa D_\Theta(1 + \log K)$  for  $K \geq 1$ . For  $K \geq 1$ , summing the epoch costs and bounding the fewer than  $2n$  remaining interactions by  $2n$ , we obtain

$$\begin{aligned} \mathbb{E} \sum_{t=0}^{T-1} (J^* - J(\psi_t)) &\leq 2n \sum_{k=1}^K a_k + O(n(1 + \log K)) \\ &\leq T\epsilon_{\text{geom}} + O\left(n(1 + \sigma)\sqrt{K} + n\varepsilon_{B,N}K^2 + n(1 + \log K)\right). \end{aligned} \quad (60)$$

The second inequality sums the iterate bounds and uses  $2nK \leq T$ . Appendix G.3 gives the recursion argument and explicit constants.

**Bounding the transient reward discrepancy.** The second term in (56) compares stationary rewards with those collected along the trajectory. Within a fixed-policy segment, the Poisson terms telescope in expectation to an endpoint difference. At a policy change, the sum also contains  $h_{\psi_{t-1}}^{\text{ref}}(\mu_t) - h_{\psi_t}^{\text{ref}}(\mu_t)$ . By (57), its magnitude is at most  $DC_P C_g \|\psi_t - \psi_{t-1}\|$ . Summing these contributions gives the following bound.

**Lemma 7.3** (Transient cost of changing policies). *Suppose  $\psi_t \in \Theta$  is selected before the batch at physical time  $t$ . Let  $\text{TV}_T = \sum_{t=1}^{T-1} \|\psi_t - \psi_{t-1}\|$  be the total parameter variation along the interaction. Then*

$$\left| \mathbb{E} \sum_{t=0}^{T-1} (J(\psi_t) - R_t) \right| \leq 2H_0 + DC_P C_g \mathbb{E} \text{TV}_T. \quad (61)$$

The proof is in Appendix G.2. This comparison applies along the actual trajectory, including the transient periods after policy updates. In our algorithm, the two parameter moves from  $\theta_{k-1}$  through the intermediate parameter to  $\theta_k$  have total length  $\|\theta_k - \theta_{k-1}\|$ . Padding and the remaining steps introduce no additional policy changes. For  $K \geq 1$ , the same step-size bound gives  $\text{TV}_T \leq 2\kappa D_\Theta(1 + \log K)$ . Substituting into (61) bounds the second term in (56) by

$$\left| \mathbb{E} \sum_{t=0}^{T-1} (J(\psi_t) - R_t) \right| \leq 2H_0 + 2\kappa D_\Theta DC_P C_g (1 + \log K) = O(1 + \log K). \quad (62)$$

**Combining the bounds and choosing trajectory lengths.** Substituting (60) and (62) into (56) controls the total regret. Longer records reduce  $\varepsilon_{B,N}$  but increase the interaction cost and estimator moments. A logarithmic choice of block counts makes accumulated estimation bias negligible while controlling these costs.

**Theorem 7.4** (Expected regret). *Suppose the model of Section 3 and Assumptions 1, 2, 3, and 4 hold, with known uniform upper bounds on their constants. Let  $C_\varepsilon = D(C_g + GC_P L_0)$  and set*

$$B = N = b_T := \max \left\{ 1, \left\lceil \log_2(\max\{1, C_\varepsilon\} T^3) \right\rceil \right\}. \quad (63)$$

*Use  $s$  from (34) and step sizes  $\eta_k = \min\{1, 2\kappa/k\}$ . For fixed structural constants, Algorithm 3 satisfies, from every initial hidden population,*

$$\mathbb{E} \text{Reg}_T \leq T\epsilon_{\text{geom}} + O(\sqrt{T} \log^2(2T)). \quad (64)$$

The bound includes all gradient-change calls, gradient calls, burn-in, return, padding, and remaining steps. If each physical step uses one independent evaluation sample, the total number of individual samples is  $(m + 1)T$ .

The full proof and explicit constants are in Appendix G.3. The choice  $B = N = b_T$  ensures  $\varepsilon_{B,N} \leq T^{-3}$ . For fixed structural constants,  $b_T = O(\log(2T))$ , and the estimator bounds give

$$n = O(b_T), \quad \mathcal{V} = O(b_T), \quad \mathcal{W} = O(b_T^3), \quad \sigma = O(b_T^{3/2}).$$

For  $K \geq 1$ , the leading term in (60) satisfies

$$n(1 + \sigma)\sqrt{K} \leq (1 + \sigma)\sqrt{nT/2} = O(\sqrt{T} b_T^2).$$

The two logarithmic factors therefore account jointly for the cost of collecting each record and the noise in tracking the gradient. The accumulated estimation-bias term is negligible because  $n\varepsilon_{B,N}K^2 \leq \varepsilon_{B,N}T^2/(4n)$ . The intermediate-policy and remainder costs in (60) are  $O(n(1 + \log K)) = O(\log^2(2T))$ , while the transient term in (62) is  $O(1 + \log K)$ . If  $K = 0$ , the direct bound  $\mathbb{E} \text{Reg}_T \leq T < 2n$  gives the same theorem without using the logarithmic epoch bounds.

When  $\epsilon_{\text{geom}} = 0$ , the theorem gives sublinear expected regret, with average regret bounded by  $O(T^{-1/2} \log^2(2T))$ . For nonzero  $\epsilon_{\text{geom}}$ , the average-regret bound is  $\epsilon_{\text{geom}}$  plus this vanishing term.

*Remark 7.5* (Information required by the schedule). The schedule uses a predetermined horizon and supplied upper bounds on the structural constants, including  $C_P$  and  $\kappa$ , to set the block counts and step sizes. Policy derivative bounds can be obtained from the chosen policy family, while the Poisson and gradient-domination bounds also require information about the environment.

## 8 Experiments

The experiments examine how estimator design affects learning along a single continuing trajectory. E1 isolates the role of randomized durations when population laws are periodic. E2 compares finite-window estimation with GPOMDP when rewards contain a delayed action effect. E3 evaluates the complete algorithm, including its sampled gradient-change correction, as population and policy dimensions increase. These runs assess finite-budget behavior under practical schedules.

**Experimental protocol.** Each run uses one continuing trajectory without resets, with  $m = 1$  and every interaction counted towards  $T$ . Each reward comes from a fresh evaluation sample that is independent of the policy observation conditional on the population. We report the policy gap  $\Delta(\theta) = J^* - J(\theta)$  and conditional-mean regret in all experiments. Conditional-mean regret replaces each sampled reward by its mean given the current population and deployed policy. Its expectation equals that of (9). Latent states and exact objectives are used only for evaluation. Appendix H gives full experimental details and results.

**E1: Randomized durations under periodic dynamics.** We consider a population with a contracting continuous coordinate and an alternating phase. We compare randomized burn-in and return durations with durations fixed at their respective means, keeping the remaining algorithm and interaction budget unchanged. At  $b = s = 2$ , fixing  $U = V = 3$  retains a phase-dependent estimation error that can reverse the gradient direction. Randomization cancels the phase-dependent term in the mean gradient estimate. The randomized method learns from both initial phases. Fixed durations perform better in the favorable phase; see Appendix H.2.

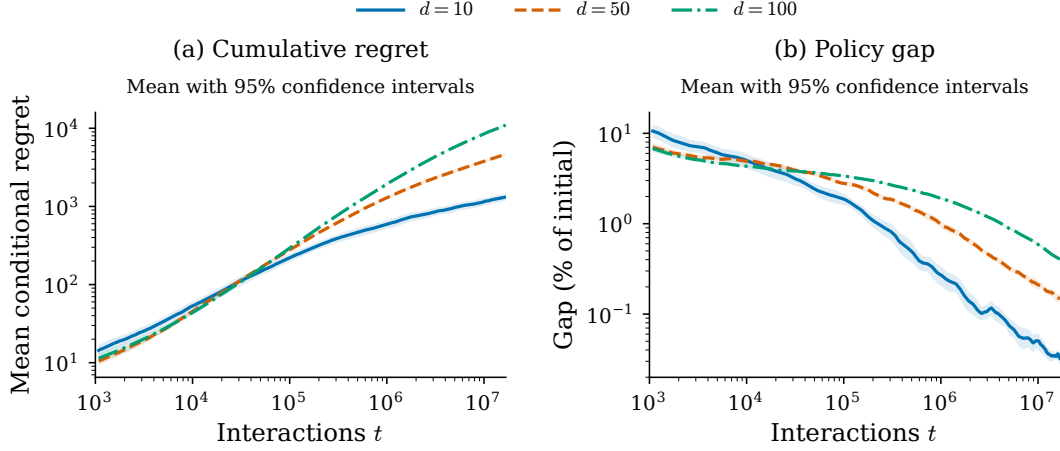


Figure 2: E3 with  $d$  latent population coordinates and  $d$  policy parameters, using  $b = s = 2$  and  $T = 2^{24}$ . Left: conditional-mean cumulative regret. Right: policy gap as a percentage of its initial value. Curves average 16 seeds; bands show pointwise 95% bootstrap confidence intervals.

**E2: Finite-window estimation versus GPOMDP.** In a model with an immediate cost and a slightly larger delayed benefit, discounting can reverse the mean gradient direction. GPOMDP (Baxter and Bartlett, 2001) therefore requires a long eligibility trace, a discounted sum of past policy scores, to preserve the correct direction, increasing sensitivity to reward noise. A short finite window captures the full delayed effect in this instance. We compare our complete algorithm with GPOMDP estimates paired with either projected gradient updates or gradient averaging followed by Frank–Wolfe updates. All three methods are tuned separately under the same interaction budget. Our method achieves lower mean regret and final policy gap than both GPOMDP implementations; see Table 3 in the appendix.

**E3: Scaling population and policy dimensions.** We study ring-coupled populations with  $d \in \{10, 50, 100\}$  latent coordinates and policy parameters. Policies use noisy observations, and every run retains the sampled gradient-change correction. Figure 2 shows substantial policy-gap reduction in all tested dimensions. For  $d = 100$  and  $b = s = 2$ , the log–log slope of mean conditional regret over  $[T/2, T]$  is 0.499 (95% bootstrap interval  $[0.486, 0.513]$ ), consistent with approximately square-root growth over this measured interval. These finite-budget runs use fixed block counts and pilot-tuned schedules. The guarantee (64) uses horizon-dependent block counts (63) and the step sizes specified in Section 7. Appendix H.4 gives the model, tuning protocol, and detailed results.

## 9 Conclusion

Our policy-gradient method learns along one continuing population trajectory using randomized durations. Under regularity, a uniform Poisson inverse bound, and exact gradient domination, it achieves  $\tilde{O}(\sqrt{T})$  expected regret against the best fixed policy in the chosen class, even when the population law under a fixed policy does not converge at deterministic times. This guarantee uses the calibrated schedule in Theorem 7.4, compares policies with the same observation budget, and counts every interaction. We further give sufficient conditions for gradient domination under partial observation and extend the bound to include  $T\epsilon_{\text{geom}}$  under approximate domination.

In the periodic example, randomized durations remove the phase-dependent bias in the gradient

estimate. With delayed rewards, our method achieves lower mean regret and final policy gap than both tested GPOMDP implementations. We also demonstrate learning with observation-dependent policies up to 100 dimensions. A natural next step is to develop adaptive window and correction schedules with guarantees requiring less prior knowledge of the structural constants.

## References

- K. Azizzadenesheli, Y. Yue, and A. Anandkumar. Policy gradient in partially observable environments: Approximation and convergence. *arXiv preprint arXiv:1810.07900*, 2020. URL <https://arxiv.org/abs/1810.07900>. Version 3.
- Debabrota Basu, Udvass Das, Brahim Driss, and Uddalak Mukherjee. Performative policy gradient: Optimality in performative reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 6939–6983, 2026. URL <https://proceedings.mlr.press/v306/basu26a.html>.
- J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001. URL <https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume15/baxter01a.pdf>.
- J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024. URL <https://doi.org/10.1287/opre.2021.0014>.
- Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, second edition, 1999.
- Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6045–6061, 2022. URL <https://proceedings.mlr.press/v151/brown22a.html>.
- William Brown, Christos Papadimitriou, and Tim Roughgarden. Online Stackelberg optimization via nonlinear control. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 697–749, 2024. URL <https://proceedings.mlr.press/v247/brown24a.html>.
- René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning. *The Annals of Applied Probability*, 33(6B):5334–5381, 2023. doi: 10.1214/23-AAP1949. URL <https://doi.org/10.1214/23-AAP1949>.
- Changsheng Chen and Grani A. Hanasusanto. Online performative decision making with latent distribution states. Preprint, Optimization Online, 2026. URL <https://optimization-online.org/?p=36162>.
- Ziyi Chen and Heng Huang. Achieve performatively optimal policy for performative reinforcement learning. *arXiv preprint arXiv:2510.04430*, 2025. URL <https://arxiv.org/abs/2510.04430>.
- S. Ganesh, W. U. Mondal, and V. Aggarwal. Order-optimal regret with novel policy gradient approaches in infinite-horizon average reward MDPs. In *Proceedings of AISTATS*, volume 258 of *PMLR*, pages 3421–3429, 2025. URL <https://proceedings.mlr.press/v258/ganesh25a.html>.

- Swetha Ganesh and Vaneet Aggarwal. Regret analysis of average-reward unichain MDPs via an actor-critic approach. In *Advances in Neural Information Processing Systems*, volume 38, pages 73709–73746, 2025. URL <https://arxiv.org/abs/2505.19986>.
- P. W. Glynn and S. P. Meyn. A Liapounov bound for solutions of the Poisson equation. *The Annals of Probability*, 24(2):916–931, 1996. URL <https://web.stanford.edu/~glynn/papers/1996/GM96.pdf>.
- Moritz Hardt and Celestine Mendler-Dünnér. Performative prediction: Past and future. *Statistical Science*, 40(3):417–436, 2025. doi: 10.1214/25-STS986. URL <https://doi.org/10.1214/25-STS986>.
- Zhiyu He, Saverio Bolognani, Florian Dörfler, and Michael Muehlebach. Decision-dependent stochastic optimization: The role of distribution dynamics. *arXiv preprint arXiv:2503.07324*, 2025. URL <https://arxiv.org/abs/2503.07324>.
- Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4641–4650, 2021. URL <https://proceedings.mlr.press/v139/izzo21a.html>.
- Zachary Izzo, James Zou, and Lexing Ying. How to learn when data gradually reacts to your model. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3998–4035, 2022. URL <https://proceedings.mlr.press/v151/izzo22a.html>.
- Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünnér. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9760–9785, 2022. URL <https://proceedings.mlr.press/v162/jagadeesan22a.html>.
- Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, 2013. URL <https://proceedings.mlr.press/v28/jaggi13.html>. Part 1.
- Olav Kallenberg. *Foundations of Modern Probability*. Springer, New York, second edition, 2002. doi: 10.1007/978-1-4757-4015-8. URL <https://doi.org/10.1007/978-1-4757-4015-8>.
- Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3164–3186, 2022. URL <https://proceedings.mlr.press/v151/li22c.html>.
- Haitong Liu, Qiang Li, and Hoi To Wai. Two-timescale derivative free optimization for performative prediction with Markovian data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 31425–31450, 2024. URL <https://proceedings.mlr.press/v235/liu24aj.html>.
- Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23642–23680, 2023. URL <https://proceedings.mlr.press/v202/mandal23a.html>.

- Celestine Mendler-Dünnér, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html).
- Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. URL <https://probability.ca/MT/>.
- J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of ICML*, volume 139 of *PMLR*, pages 7710–7720, 2021. URL <https://proceedings.mlr.press/v139/miller21a.html>.
- Hoai-An Nguyen and Ching-An Cheng. Provable reset-free reinforcement learning by no-regret reduction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25939–25955, 2023. URL <https://proceedings.mlr.press/v202/nguyen23b.html>.
- Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirodda, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4026–4035, 2018. URL <https://proceedings.mlr.press/v80/papini18a.html>.
- Barna Pásztor, Ilija Bogunovic, and Andreas Krause. Efficient model-based multi-agent mean-field reinforcement learning. *Transactions on Machine Learning Research*, 2023. URL <https://las.inf.ethz.ch/files/pasztor2023efficient.pdf>.
- Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünnér, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609, 2020. URL <https://proceedings.mlr.press/v119/perdomo20a.html>.
- Huyên Pham and Xiaoli Wei. Discrete time McKean–Vlasov control problem: A dynamic programming approach. *Applied Mathematics & Optimization*, 74(3):487–506, 2016. doi: 10.1007/s00245-016-9386-9. URL <https://doi.org/10.1007/s00245-016-9386-9>.
- Matteo Pirodda, Marcello Restelli, and Luca Bascetta. Policy gradient in Lipschitz Markov decision processes. *Machine Learning*, 100(2–3):255–283, 2015. doi: 10.1007/s10994-015-5484-1. URL <https://doi.org/10.1007/s10994-015-5484-1>.
- Ben Rank, Stelios Triantafyllou, Debmalaya Mandal, and Goran Radanovic. Performative reinforcement learning in gradually shifting environments. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 3041–3075, 2024. URL <https://proceedings.mlr.press/v244/rank24a.html>.
- Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, third edition, 1976.
- Walter Rudin. *Functional Analysis*. McGraw-Hill, New York, second edition, 1991.
- Alessio Russo, Alberto Maria Metelli, and Marcello Restelli. Achieving  $\tilde{O}(\sqrt{T})$  regret in average-reward POMDPs with known observation models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4168–4176, 2025. URL <https://proceedings.mlr.press/v258/russo25b.html>.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018. URL <https://mitpress.mit.edu/9780262039246/reinforcement-learning/>.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9. URL <https://doi.org/10.1007/978-3-540-71050-9>.

## A compact map of the model and the analysis

This overview collects the main objects for reference, connecting the physical interaction to the quantities used in the analysis and computed by the algorithm. The physical system always follows

$$\mu_t \longrightarrow (Z_{t,1}, \dots, Z_{t,m}) \longrightarrow \hat{\mu}_t \longrightarrow A_t \longrightarrow \mu_{t+1} = F(\mu_t, A_t).$$

At each step, the learner also observes the reward  $R_t$ . For a fixed policy parameter and the fixed batch size  $m$ , the table summarizes the quantities used throughout the paper.

| Object                        | Role                                                                                                                                       |
|-------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| $P_\theta$                    | Transition operator for the hidden population after averaging observation and action randomness.                                           |
| $\nu_\theta$                  | Invariant law over population states.                                                                                                      |
| $J(\theta)$                   | Average reward of policy $\pi_\theta$ at the fixed batch size $m$ .                                                                        |
| $h_\theta$                    | Differential value function, centered under $\nu_\theta$ , satisfying the average-reward Bellman equation.                                 |
| $h_\theta^{\text{ref}}$       | The same differential value function shifted to vanish at a fixed reference population, used to compare policies.                          |
| $\nabla J(\theta)$            | Immediate policy effect plus the future population-response effect.                                                                        |
| $Q_\theta$                    | Transition operator after a uniformly drawn number of physical steps.                                                                      |
| $M_\theta$                    | Stationary law of the observation actually used by the policy.                                                                             |
| $B_\vartheta(\cdot \mid \xi)$ | Posterior population distribution given observation $\xi$ under policy $\pi_\vartheta$ in stationarity.                                    |
| $Q_{\theta,\vartheta}$        | Conditional mean of $r(\mu, a) + h_\theta(F(\mu, a))$ given observation $\xi$ , using the posterior population law under $\pi_\vartheta$ . |
| $d_{\theta,\vartheta}$        | Best feasible improvement over the current action law at an observation, evaluated under comparator policy $\pi_\vartheta$ 's posterior.   |
| $\mathcal{G}(\theta)$         | Frank–Wolfe gap: best feasible first-order improvement in parameter space.                                                                 |
| $\hat{g}$                     | Trajectory estimate of the average-reward gradient.                                                                                        |
| $\hat{A}$                     | Estimate of the Jacobian of the finite-trajectory estimator's mean.                                                                        |

The algorithm computes  $\hat{g}$  and the required products of  $\hat{A}$  with parameter differences using observable rewards and policy derivatives. Primitive regularity and the Poisson inverse bound justify the gradient-bias bound (44) and the second-moment bound (47) used in the analysis, without requiring the learner to observe the population or its differential value.

## A Policy evaluation and the average-reward gradient

We first establish an invariant law using compactness and continuity, then use the Poisson inverse to prove uniqueness. The resulting policy comparisons yield the policy-gradient and finite-return identities of Section 4 without differentiating the invariant law.

### A.1 Invariant law and normalized Poisson solutions

*Proof of Proposition 4.1. Existence.* Fix  $\theta \in \Theta$ . We first show that  $P_\theta$  preserves continuous functions, which is the Feller property. Compactness of  $\mathcal{D}$  then guarantees the existence of an invariant probability.

We begin with a Lipschitz function  $f$  and two population states  $\mu, \nu \in \mathcal{D}$ . Draw  $m$  independent pairs  $(Z_i, Z'_i)$  from an optimal coupling of  $\mu$  and  $\nu$ , which exists by Villani (2009, Theorem 4.1), and let  $\xi, \zeta$  be the corresponding empirical measures. Pairing the corresponding samples gives

$$\mathbb{E}W_1(\xi, \zeta) \leq \frac{1}{m} \sum_{i=1}^m \mathbb{E}d_{\mathcal{Z}}(Z_i, Z'_i) = W_1(\mu, \nu).$$

Conditionally on these observations, optimally couple  $A \sim \pi_\theta(\cdot \mid \xi)$  and  $A' \sim \pi_\theta(\cdot \mid \zeta)$ . The coupling can be chosen measurably in  $(\xi, \zeta)$  by Villani (2009, Corollary 5.22). The policy regularity bound (12) then yields

$$\mathbb{E}[d_{\mathcal{A}}(A, A') \mid \xi, \zeta] \leq L_\pi W_1(\xi, \zeta), \quad \mathbb{E}d_{\mathcal{A}}(A, A') \leq L_\pi W_1(\mu, \nu).$$

Using the transition bound (10), we obtain

$$\begin{aligned} |(P_\theta f)(\mu) - (P_\theta f)(\nu)| &= |\mathbb{E}[f(F(\mu, A)) - f(F(\nu, A'))]| \\ &\leq \|f\|_{\text{Lip}} \mathbb{E}W_1(F(\mu, A), F(\nu, A')) \\ &\leq \|f\|_{\text{Lip}} [qW_1(\mu, \nu) + L_a \mathbb{E}d_{\mathcal{A}}(A, A')] \\ &\leq (q + L_a L_\pi) \|f\|_{\text{Lip}} W_1(\mu, \nu). \end{aligned}$$

Thus  $P_\theta f$  is Lipschitz, with  $\|P_\theta f\|_{\text{Lip}} \leq (q + L_a L_\pi) \|f\|_{\text{Lip}}$ .

Now let  $f$  be any continuous function on  $\mathcal{D}$ . Since  $\mathcal{D}$  is a compact metric space, there are Lipschitz functions  $f_n$  converging uniformly to  $f$ : the Lipschitz functions form an algebra containing constants and separating points, so the Stone–Weierstrass theorem applies (Rudin, 1976, Theorem 7.32). Averaging preserves the uniform approximation, since  $\|P_\theta f_n - P_\theta f\|_\infty \leq \|f_n - f\|_\infty \rightarrow 0$ . Each  $P_\theta f_n$  is Lipschitz by the preceding argument and hence continuous. Their uniform limit  $P_\theta f$  is therefore continuous as well. Since  $\mathcal{D}$  is compact, the Cesàro averages of the transition laws are tight. The Feller property therefore yields an invariant probability measure by Meyn and Tweedie (1993, Theorem 12.0.1(i)).

**Uniqueness of the invariant law.** We prove uniqueness by showing that any two invariant probabilities give the same expectation for each Lipschitz function. Fix a Lipschitz function  $f$ . By Assumption 3,  $I - P_\theta$  is invertible on  $\text{Lip}(\mathcal{D})$ . Hence there is a Lipschitz function  $h$  such that  $[h - P_\theta h] = [f]$ . Equality in this quotient space means equality up to an additive constant. Thus, for some  $c \in \mathbb{R}$ ,  $h - P_\theta h = f - c$ . Here,  $h$  and  $c$  are chosen before selecting an invariant probability.

Let  $\lambda$  be any invariant probability for  $P_\theta$ . Integrating the preceding identity against  $\lambda$  and using invariance, we obtain  $0 = \lambda h - \lambda(P_\theta h) = \lambda f - c$ . Therefore, every invariant probability has the same

expectation  $c$  of  $f$ . Since this holds for every Lipschitz function, uniform approximation extends the equality to all continuous functions. These determine probability measures on the compact metric space  $\mathcal{D}$  (Billingsley, 1999, Theorem 1.2), so the invariant probability is unique. We denote it by  $\nu_\theta$ . The same calculation also identifies  $c = \nu_\theta f$ , so  $h - P_\theta h = f - \nu_\theta f$ .

**Uniqueness of the normalized Poisson solution.** Adding a constant to  $h$  leaves the Poisson equation unchanged. We remove this ambiguity by imposing  $\nu_\theta h = 0$ , which can always be achieved by replacing  $h$  with  $h - \nu_\theta h$ .

To see why the normalized solution is unique, let  $h_1$  and  $h_2$  be two such solutions. Subtracting their equations gives  $(I - P_\theta)(h_1 - h_2) = 0$ . Injectivity of  $I - P_\theta$  on the quotient implies that  $h_1 - h_2$  is constant. Since both solutions have zero  $\nu_\theta$ -mean, this constant is zero, and hence  $h_1 = h_2$ .

The inverse bound (18) directly controls the Lipschitz seminorm, giving  $\|h\|_{\text{Lip}} \leq C_P \|f\|_{\text{Lip}}$ . To obtain an absolute bound on  $h$ , we additionally use its zero-mean normalization. For every  $\mu \in \mathcal{D}$ , this gives  $h(\mu) = \int_{\mathcal{D}} [h(\mu) - h(\mu')] \nu_\theta(d\mu')$ . Since  $|h(\mu) - h(\mu')| \leq \|h\|_{\text{Lip}} W_1(\mu, \mu')$  and  $W_1(\mu, \mu') \leq D$ , we obtain  $|h(\mu)| \leq D \|h\|_{\text{Lip}} \leq DC_P \|f\|_{\text{Lip}}$ . Taking the supremum over  $\mu$  completes the proof.  $\square$

For the reward solution  $h_\theta = \mathcal{R}_\theta \bar{r}_\theta$ , take expectations in the reward Poisson equation (20). Using the reward mean (6) and the transition identity (5) gives

$$\mathbb{E}_{\mu_0}^\theta [R_t - J(\theta)] = \mathbb{E}_{\mu_0}^\theta [\bar{r}_\theta(\mu_t) - J(\theta)] = \mathbb{E}_{\mu_0}^\theta [h_\theta(\mu_t) - h_\theta(\mu_{t+1})].$$

Summing over  $t = 0, \dots, N - 1$  makes the expected differential values telescope, leaving the two endpoints in (21).

## A.2 Comparing two fixed policies

*Proof of Lemma 4.2.* Integrate (20) under  $\nu_\vartheta$ . Since  $\nu_\vartheta P_\vartheta = \nu_\vartheta$ ,

$$\nu_\vartheta \bar{r}_\theta - J(\theta) = \nu_\vartheta (P_\vartheta - P_\theta) h_\theta.$$

Adding  $\nu_\vartheta(\bar{r}_\vartheta - \bar{r}_\theta)$  to both sides gives (22).  $\square$

The general Poisson equation (19) similarly expresses changes in stationary expectations through the transition operator, without differentiating the invariant law. Integrating  $(I - P_\theta)\mathcal{R}_\theta f = f - \nu_\theta f$  under  $\nu_\vartheta$  gives

$$(\nu_\vartheta - \nu_\theta)f = \nu_\vartheta(I - P_\theta)\mathcal{R}_\theta f = \nu_\vartheta(P_\vartheta - P_\theta)\mathcal{R}_\theta f,$$

where the last equality uses  $\nu_\vartheta P_\vartheta = \nu_\vartheta$ . This proves (23).

## A.3 The average-reward gradient

*Proof of Theorem 4.3.* We first control how the policy changes with its parameter. The score bound (13) gives, for every observation  $\xi$ ,

$$\int_{\mathcal{A}} \|\nabla \pi_\theta(a \mid \xi)\| \, da = \int_{\mathcal{A}} \|\nabla \log \pi_\theta(a \mid \xi)\| \pi_\theta(a \mid \xi) \, da \leq G,$$

since the policy density integrates to one.

For  $\theta, \vartheta \in \Theta$ , let  $\theta_u = \theta + u(\vartheta - \theta)$ ,  $u \in [0, 1]$ . By the continuous policy differentiability in Assumption 2, the fundamental theorem of calculus along this segment gives the following identity in  $\mathcal{L}^1(da)$ :

$$\pi_{\vartheta}(a \mid \xi) - \pi_{\theta}(a \mid \xi) = \int_0^1 \langle \nabla \pi_{\theta_u}(a \mid \xi), \vartheta - \theta \rangle \, du.$$

Taking absolute values and integrating over actions therefore yields

$$\int_{\mathcal{A}} |\pi_{\vartheta}(a \mid \xi) - \pi_{\theta}(a \mid \xi)| \, da \leq \|\vartheta - \theta\| \int_0^1 \int_{\mathcal{A}} \|\nabla \pi_{\theta_u}(a \mid \xi)\| \, da \, du \leq G \|\vartheta - \theta\|.$$

For a fixed population  $\mu$ , the observation law  $\mathbf{O}_m(\cdot \mid \mu)$  does not depend on the policy parameter. Averaging the preceding bound over observations gives

$$\int_{\mathcal{A}} |p_{\vartheta}(a \mid \mu) - p_{\theta}(a \mid \mu)| \, da \leq \int_{\mathcal{O}_m} \left[ \int_{\mathcal{A}} |\pi_{\vartheta}(a \mid \xi) - \pi_{\theta}(a \mid \xi)| \, da \right] \mathbf{O}_m(d\xi \mid \mu) \leq G \|\vartheta - \theta\|.$$

Finally, for any bounded function  $g$ ,

$$\begin{aligned} |((P_{\vartheta} - P_{\theta})g)(\mu)| &= \left| \int_{\mathcal{A}} g(F(\mu, a)) [p_{\vartheta}(a \mid \mu) - p_{\theta}(a \mid \mu)] \, da \right| \\ &\leq \|g\|_{\infty} \int_{\mathcal{A}} |p_{\vartheta}(a \mid \mu) - p_{\theta}(a \mid \mu)| \, da \leq G \|g\|_{\infty} \|\vartheta - \theta\|. \end{aligned}$$

Taking the supremum over  $\mu$  gives  $\|(P_{\vartheta} - P_{\theta})g\|_{\infty} \leq G \|g\|_{\infty} \|\vartheta - \theta\|$ . Combining it with (23) and the Poisson solution bound from (19) gives, for every Lipschitz  $f$ ,

$$|(\nu_{\vartheta} - \nu_{\theta})f| \leq G \|\mathcal{R}_{\theta} f\|_{\infty} \|\vartheta - \theta\| \leq GDC_P \|f\|_{\text{Lip}} \|\vartheta - \theta\|. \quad (65)$$

Now fix  $\theta$  and a feasible direction  $v \in \mathbb{R}^p$ . Let  $\varepsilon$  tend to zero through nonzero values for which  $\theta + \varepsilon v \in \Theta$ . Collect the one-step reward and transition difference quotients in

$$f_{\varepsilon} := \frac{\bar{r}_{\theta+\varepsilon v} - \bar{r}_{\theta}}{\varepsilon} + \frac{P_{\theta+\varepsilon v} - P_{\theta}}{\varepsilon} h_{\theta}.$$

Here  $h_{\theta}$  remains fixed as  $\varepsilon$  varies. Applying (22) with comparator parameter  $\theta + \varepsilon v$  and dividing by  $\varepsilon$  gives

$$\frac{J(\theta + \varepsilon v) - J(\theta)}{\varepsilon} = \nu_{\theta+\varepsilon v} f_{\varepsilon}.$$

We now justify the uniform first-order approximation of the policy density. Differentiating  $\nabla \pi_{\theta} = \pi_{\theta} \nabla \log \pi_{\theta}$  gives  $\nabla^2 \pi_{\theta} = \pi_{\theta} [\nabla^2 \log \pi_{\theta} + (\nabla \log \pi_{\theta})(\nabla \log \pi_{\theta})^{\top}]$ . Since the density integrates to one, the score and Hessian bounds (13)–(14) imply  $\int_{\mathcal{A}} \|\nabla^2 \pi_{\theta}(a \mid \xi)\|_{\text{op}} \, da \leq M + G^2$ , uniformly over parameters and observations. Taylor's formula with integral remainder in  $\mathcal{L}^1(da)$  therefore yields

$$\begin{aligned} \sup_{\xi} \int_{\mathcal{A}} \left| \frac{\pi_{\theta+\varepsilon v}(a \mid \xi) - \pi_{\theta}(a \mid \xi)}{\varepsilon} - \langle \nabla \pi_{\theta}(a \mid \xi), v \rangle \right| \, da \\ \leq |\varepsilon| \|v\|^2 \sup_{\xi} \int_0^1 (1-u) \int_{\mathcal{A}} \|\nabla^2 \pi_{\theta+u\varepsilon v}(a \mid \xi)\|_{\text{op}} \, da \, du \\ \leq \frac{1}{2} (M + G^2) |\varepsilon| \|v\|^2 \longrightarrow 0. \end{aligned}$$

Multiplying this error by the bounded function  $r(\mu, a) + h_{\theta}(F(\mu, a))$  and averaging over the parameter-independent observation law preserves uniform convergence. Hence  $\|f_{\varepsilon} - \langle b_{\theta}, v \rangle\|_{\infty} \longrightarrow 0$ .

To control the change in the invariant law, we need a Lipschitz test function. The state-regularity bounds (67), proved in Appendix B.2, imply that  $\langle b_\theta, v \rangle$  is Lipschitz. We can therefore apply (65) to this fixed function. Separating the change in the function from the change in its averaging law, we obtain

$$|\nu_{\theta+\varepsilon v} f_\varepsilon - \nu_\theta \langle b_\theta, v \rangle| \leq \|f_\varepsilon - \langle b_\theta, v \rangle\|_\infty + |(\nu_{\theta+\varepsilon v} - \nu_\theta) \langle b_\theta, v \rangle| \longrightarrow 0.$$

Thus the directional derivative is  $\nu_\theta \langle b_\theta, v \rangle$ . Uniformity over unit directions gives the gradient formula. Appendix B proves gradient Lipschitz continuity and extends the argument to boundary parameters.  $\square$

#### A.4 Representing the gradient with policy scores

The chain-rule identity in Section 4.4 converts density derivatives into observable scores. We give the intermediate integrations for (27)–(29).

At a fixed population  $\mu$ , the conditional observation law  $\mathbf{O}_m(\cdot \mid \mu)$  is independent of  $\theta$ . Since the next display averages only over observations drawn from this law, its expectations need no policy superscript. Applying the chain-rule identity to (4), with differentiation under the expectation justified by Assumption 2, gives

$$\begin{aligned} \nabla p_\theta(a \mid \mu) &= \nabla \mathbb{E}[\pi_\theta(a \mid \hat{\mu}_t) \mid \mu_t = \mu] = \mathbb{E}[\nabla \pi_\theta(a \mid \hat{\mu}_t) \mid \mu_t = \mu] \\ &= \mathbb{E}[\pi_\theta(a \mid \hat{\mu}_t) \nabla \log \pi_\theta(a \mid \hat{\mu}_t) \mid \mu_t = \mu]. \end{aligned}$$

Integrating against the policy density  $\pi_\theta(a \mid \xi)$  is equivalent to taking an expectation over an action drawn from that policy. For example, differentiating (6) gives

$$\nabla \bar{r}_\theta(\mu) = \mathbb{E} \left[ \int [\nabla \log \pi_\theta(a \mid \hat{\mu}_t) r(\mu, a)] \pi_\theta(a \mid \hat{\mu}_t) \, da \mid \mu_t = \mu \right] = \mathbb{E}^\theta [S_t r(\mu, A_t) \mid \mu_t = \mu].$$

The conditional reward model (3) lets us replace  $r(\mu, A_t)$  by  $R_t$  inside this expectation. Applying the same argument to  $h_\theta(F(\mu, a))$ , with  $h_\theta$  held fixed, gives the corresponding transition derivative. The two resulting expectations are precisely (27)–(28). Averaging their sum under  $\nu_\theta$  and using (25) gives (29). The zero-mean score identity (30) then allows subtraction of the action-independent baseline  $J(\theta) + h_\theta(\mu_t)$ , giving (32).

#### A.5 The finite-return representation

Fix  $\theta$ ,  $\mu_0 = x$ , and an integer  $\ell \geq 0$ . Recall from (27)–(28) that  $b_\theta(x) = \mathbb{E}_x^\theta [S_0 \{R_0 + h_\theta(\mu_1)\}]$ . We express  $h_\theta(\mu_1)$  through rewards over the next  $\ell$  decisions by unfolding its Poisson equation.

Conditionally on  $\hat{\mu}_0, A_0$ , the next population is  $\mu_1 = F(x, A_0)$ . For  $\ell \geq 1$ , applying the finite-horizon comparison (21) from  $\mu_1$ , together with the Markov property and conditional reward mean (3), gives

$$h_\theta(\mu_1) = \mathbb{E}_x^\theta \left[ \sum_{j=1}^{\ell} (R_j - J(\theta)) + h_\theta(\mu_{\ell+1}) \mid \hat{\mu}_0, A_0 \right].$$

For  $\ell = 0$ , the reward sum is empty and the identity reduces to  $h_\theta(\mu_1) = h_\theta(\mu_1)$ . The differential value at the next population thus equals the expected centered return plus the expected differential value at the endpoint.

Since  $S_0$  is determined by  $\hat{\mu}_0, A_0$ , we can multiply this identity by  $S_0$  and then take expectations. The term  $\ell J(\theta)$  disappears because  $\mathbb{E}_x^\theta S_0 = 0$  by (30). Adding the immediate-reward contribution from the expression for  $b_\theta(x)$  above gives

$$\mathbb{E}_x^\theta \left[ S_0 \sum_{j=0}^{\ell} R_j \right] = b_\theta(x) - \mathbb{E}_x^\theta [S_0 h_\theta(\mu_{\ell+1})].$$

For the operator form in (33), condition on the first observation and action: the expected endpoint differential value is  $(P_\theta^\ell h_\theta)(F(x, A_0))$ . Applying the score identity (28) differentiates the first action distribution while holding this continuation function fixed:

$$\mathbb{E}_x^\theta [S_0 h_\theta(\mu_{\ell+1})] = [(\nabla P_\theta) P_\theta^\ell h_\theta](x).$$

This proves (33).

The mean  $g_{\theta, \ell}(x)$  of the score-weighted return also has the finite-sum representation

$$g_{\theta, \ell}(x) = \nabla \bar{r}_\theta(x) + \left[ (\nabla P_\theta) \sum_{k=0}^{\ell-1} P_\theta^k \bar{r}_\theta \right](x).$$

Indeed, the sum inside the brackets is the expected reward over  $\ell$  decisions starting from the next population. Here  $\nabla P_\theta$  differentiates only the first transition, holding the continuation function fixed, because the estimator uses only the first score  $S_0$ . When  $\ell = 0$ , the future-reward sum is empty and the estimator reduces to  $S_0 R_0$ .

## B Regularity, randomized durations, and policy sensitivity

This appendix proves the randomized-endpoint and gradient-bias bounds of Section 5, together with the parameter regularity used in Section 7. Constants are defined in (42)–(43) and Lemma 7.2.

### B.1 Accuracy at randomized endpoints

The following proof of Lemma 5.1 separates the sup-norm bound for a difference of Markov operators from the Lipschitz bound for the Poisson inverse.

*Proof of Lemma 5.1.* Fix  $\theta$ . We first reduce the claim to centered functions. Replacing  $f$  by  $f - \nu_\theta f$  preserves its Lipschitz seminorm and gives  $Q_\theta^b(f - \nu_\theta f) = Q_\theta^b f - \nu_\theta f$ , since  $Q_\theta$  preserves constants. It therefore suffices to consider  $\nu_\theta f = 0$ . The operators  $P_\theta, Q_\theta, \mathcal{R}_\theta$  all preserve centered Lipschitz functions.

We use the Poisson equation  $(I - P_\theta)\mathcal{R}_\theta f = f$  to simplify the time average:

$$Q_\theta f = \frac{1}{s} \sum_{\ell=1}^s P_\theta^\ell (I - P_\theta) \mathcal{R}_\theta f = \frac{1}{s} \sum_{\ell=1}^s (P_\theta^\ell - P_\theta^{\ell+1}) \mathcal{R}_\theta f = \frac{P_\theta - P_\theta^{s+1}}{s} \mathcal{R}_\theta f.$$

The intermediate terms cancel, leaving only the two endpoints. Defining  $H := (P_\theta - P_\theta^{s+1})/s$ , we obtain  $Q_\theta = H \mathcal{R}_\theta$  on centered Lipschitz functions.

To control repeated randomized blocks, we would like to write  $Q_\theta^b = H^b \mathcal{R}_\theta^b$ , so that the two factors can be bounded separately. This requires showing that  $H$  and  $\mathcal{R}_\theta$  commute. First,  $P_\theta \mathcal{R}_\theta f$  satisfies

$$(I - P_\theta)P_\theta \mathcal{R}_\theta f = P_\theta(I - P_\theta)\mathcal{R}_\theta f = P_\theta f, \quad \nu_\theta P_\theta \mathcal{R}_\theta f = \nu_\theta \mathcal{R}_\theta f = 0.$$

Thus  $P_\theta \mathcal{R}_\theta f$  and  $\mathcal{R}_\theta P_\theta f$  solve the same normalized Poisson equation. Uniqueness from Proposition 4.1 gives  $P_\theta \mathcal{R}_\theta f = \mathcal{R}_\theta P_\theta f$ . Consequently,  $\mathcal{R}_\theta$  commutes with every power of  $P_\theta$ , and hence with  $H$ . We may therefore write  $Q_\theta^b f = H^b \mathcal{R}_\theta^b f$ .

We now control  $H$  in the sup norm and  $\mathcal{R}_\theta$  in the Lipschitz seminorm. For any bounded  $g$ , Markov operators satisfy  $\|P_\theta^k g\|_\infty \leq \|g\|_\infty$ , so

$$\|Hg\|_\infty \leq \frac{\|P_\theta g\|_\infty + \|P_\theta^{s+1} g\|_\infty}{s} \leq \frac{2}{s} \|g\|_\infty.$$

Iterating gives  $\|H^b g\|_\infty \leq (2/s)^b \|g\|_\infty$ . Repeated application of the Poisson inverse bound likewise gives  $\|\mathcal{R}_\theta^b f\|_{\text{Lip}} \leq C_P^b \|f\|_{\text{Lip}}$ .

To combine these bounds, we use the zero-mean normalization to pass from the Lipschitz seminorm to the sup norm. Since  $\nu_\theta \mathcal{R}_\theta^b f = 0$ ,

$$(\mathcal{R}_\theta^b f)(x) = \int [(\mathcal{R}_\theta^b f)(x) - (\mathcal{R}_\theta^b f)(y)] \nu_\theta(dy).$$

Using  $W_1(x, y) \leq D$  gives  $\|\mathcal{R}_\theta^b f\|_\infty \leq D \|\mathcal{R}_\theta^b f\|_{\text{Lip}}$ . Therefore,

$$\begin{aligned} \|Q_\theta^b f\|_\infty &= \|H^b \mathcal{R}_\theta^b f\|_\infty \\ &\leq (2/s)^b \|\mathcal{R}_\theta^b f\|_\infty \\ &\leq D(2C_P/s)^b \|f\|_{\text{Lip}} \\ &= D\alpha^b \|f\|_{\text{Lip}} \leq D2^{-b} \|f\|_{\text{Lip}}. \end{aligned}$$

Applying this result to  $f - \nu_\theta f$  proves (35).  $\square$

## B.2 State regularity and gradient-estimation bias

The coupling used in the proof of Proposition 4.1 gives the one-step transition bound (17). Applying the reward regularity bound (11) to the same coupled samples and actions gives the corresponding reward bound:

$$\|P_\theta f\|_{\text{Lip}} \leq (q + L_a L_\pi) \|f\|_{\text{Lip}}, \quad \|\bar{r}_\theta\|_{\text{Lip}} \leq L_0. \quad (66)$$

The proof below also establishes the parameter-derivative bounds

$$\|(\nabla P_\theta) f\|_{\text{Lip}} \leq (qG + DL_B) \|f\|_{\text{Lip}}, \quad \|\nabla \bar{r}_\theta\|_{\text{Lip}} \leq GL_r + L_B. \quad (67)$$

Bounds (66) and (67) control the Lipschitz constant of the gradient integrand and, together with Lemma 5.1, the two sources of estimator bias.

*Proof of Proposition 5.3.* The reward Lipschitz bound (16) and the Poisson solution bounds (19) give  $\|h_\theta\|_{\text{Lip}} \leq C_P L_0$  and  $\|h_\theta\|_\infty \leq DC_P L_0 = H_0$ . It remains to control the Lipschitz seminorm of  $b_\theta$ , which requires bounds on the parameter derivatives.

Averaging the score bound (13) over observations, using the derivative representation in Appendix A.4, gives  $\int_{\mathcal{A}} \|\nabla p_{\theta}(a \mid x)\| \, da \leq G$ . For populations  $x, y \in \mathcal{D}$ , use the empirical-observation coupling from Appendix A.1, which satisfies  $\mathbb{E}W_1(\xi, \zeta) \leq W_1(x, y)$ . The derivative bound (15) then gives

$$\begin{aligned} \int_{\mathcal{A}} \|\nabla p_{\theta}(a \mid x) - \nabla p_{\theta}(a \mid y)\| \, da &\leq \mathbb{E} \left[ \int_{\mathcal{A}} \|\nabla \pi_{\theta}(a \mid \xi) - \nabla \pi_{\theta}(a \mid \zeta)\| \, da \right] \\ &\leq L_B \mathbb{E}W_1(\xi, \zeta) \leq L_B W_1(x, y). \end{aligned}$$

These two estimates control how the transition derivative varies with the population. Let  $f$  be Lipschitz and fix a reference population  $x_{\text{ref}} \in \mathcal{D}$ . Because  $\int_{\mathcal{A}} \nabla p_{\theta}(a \mid x) \, da = 0$ , subtracting the constant  $f(x_{\text{ref}})$  does not change  $(\nabla P_{\theta})f$ . We can therefore write

$$\begin{aligned} ((\nabla P_{\theta})f)(x) - ((\nabla P_{\theta})f)(y) &= \int_{\mathcal{A}} [f(F(x, a)) - f(F(y, a))] \nabla p_{\theta}(a \mid x) \, da \\ &\quad + \int_{\mathcal{A}} [f(F(y, a)) - f(x_{\text{ref}})] [\nabla p_{\theta}(a \mid x) - \nabla p_{\theta}(a \mid y)] \, da. \end{aligned}$$

The first integral changes the population while keeping the action fixed. By (10) and the integrated score bound above, its norm is at most  $qG \|f\|_{\text{Lip}} W_1(x, y)$ . For the second integral, the diameter bound gives  $|f(F(y, a)) - f(x_{\text{ref}})| \leq D \|f\|_{\text{Lip}}$ . Its norm is consequently at most  $DL_B \|f\|_{\text{Lip}} W_1(x, y)$ . Together, these estimates give  $\|(\nabla P_{\theta})f\|_{\text{Lip}} \leq (qG + DL_B) \|f\|_{\text{Lip}}$ .

Applying the same splitting to  $\nabla \bar{r}_{\theta}(x) = \int_{\mathcal{A}} r(x, a) \nabla p_{\theta}(a \mid x) \, da$  gives

$$\|\nabla \bar{r}_{\theta}(x) - \nabla \bar{r}_{\theta}(y)\| \leq (GL_r + L_B)W_1(x, y).$$

Here the first term uses the population sensitivity in (11), and the second uses  $0 \leq r \leq 1$ . Thus

$$\|b_{\theta}\|_{\text{Lip}} \leq \|\nabla \bar{r}_{\theta}\|_{\text{Lip}} + \|(\nabla P_{\theta})h_{\theta}\|_{\text{Lip}} \leq GL_r + L_B + (qG + DL_B)C_P L_0 = C_g.$$

We can now bound the two bias terms. Applying Lemma 5.1 to  $b_{\theta}$  gives

$$\|Q_{\theta}^B b_{\theta} - \nu_{\theta} b_{\theta}\|_{\infty} \leq D\alpha^B \|b_{\theta}\|_{\text{Lip}} \leq DC_g \alpha^B.$$

For vector-valued  $b_{\theta}$ , this follows by applying the scalar lemma to every unit-vector projection.

For the return tail, the score bound (13) implies  $\|(\nabla P_{\theta})f\|_{\infty} \leq G\|f\|_{\infty}$  for every bounded scalar function  $f$ . Also,  $Q_{\theta}^B$ , being a Markov operator, does not increase the sup norm. Since  $\nu_{\theta} h_{\theta} = 0$ ,

$$\|Q_{\theta}^B (\nabla P_{\theta}) Q_{\theta}^N h_{\theta}\|_{\infty} \leq G \|Q_{\theta}^N h_{\theta}\|_{\infty} \leq GD\alpha^N \|h_{\theta}\|_{\text{Lip}} \leq GDC_P L_0 \alpha^N.$$

Adding the two bounds proves (44).  $\square$

### B.3 Regularity across policy parameters

Before proving Lemma 7.2, we record some consequences of the policy score and Hessian bounds. Differentiating  $\nabla \pi_{\theta} = \pi_{\theta} \nabla \log \pi_{\theta}$  once more and applying the product rule gives

$$\nabla^2 \pi_{\theta} = \pi_{\theta} \nabla^2 \log \pi_{\theta} + (\nabla \log \pi_{\theta})(\nabla \pi_{\theta})^{\top} = \pi_{\theta} \{\nabla^2 \log \pi_{\theta} + (\nabla \log \pi_{\theta})(\nabla \log \pi_{\theta})^{\top}\}.$$

Taking norms, using the score and Hessian bounds (13)–(14), and integrating the policy density, whose total mass is one, gives the following bounds. The reward bound also uses  $0 \leq r \leq 1$ :

$$\begin{aligned} \|(\nabla P_{\theta})f\|_{\infty} &\leq G \|f\|_{\infty}, \\ \|(\nabla^2 P_{\theta})f\|_{\infty} &\leq (M + G^2) \|f\|_{\infty}, \quad \|\nabla^2 \bar{r}_{\theta}\|_{\infty} \leq M + G^2. \end{aligned} \tag{68}$$

*Proof of Lemma 7.2.* We first compare the two differential value functions. Let  $\delta = \vartheta - \theta$ . Collect the reward and transition changes in the scalar function  $a : \mathcal{D} \rightarrow \mathbb{R}$ :

$$a = \bar{r}_\vartheta - \bar{r}_\theta + (P_\vartheta - P_\theta)h_\theta^{\text{ref}}.$$

Integrating the uniform derivative bounds (67) along the segment joining  $\theta$  and  $\vartheta$ , while holding  $h_\theta^{\text{ref}}$  fixed, yields

$$\|a\|_{\text{Lip}} \leq (GL_r + L_B + (qG + DL_B)C_P L_0)\|\delta\| = C_g\|\delta\|.$$

Subtracting the reward Poisson equations (20) gives

$$(I - P_\vartheta)(h_\vartheta^{\text{ref}} - h_\theta^{\text{ref}}) = a - (J(\vartheta) - J(\theta)).$$

The performance-difference identity (22) identifies the constant on the right as  $\nu_\vartheta a$ . Both differential value functions vanish at  $\mu_{\text{ref}}$ , so uniqueness of the Poisson solution up to a constant implies

$$h_\vartheta^{\text{ref}} - h_\theta^{\text{ref}} = \mathcal{R}_\vartheta a - (\mathcal{R}_\vartheta a)(\mu_{\text{ref}}).$$

Consequently,

$$\|h_\vartheta^{\text{ref}} - h_\theta^{\text{ref}}\|_\infty \leq DC_P \|a\|_{\text{Lip}} \leq DC_P C_g \|\delta\|.$$

This proves (57).

The score representation (29) and the bounds on rewards and the differential value function give  $\|\nabla J(\theta)\| \leq G(1 + H_0)$ . Integration along the parameter segment proves (58).

It remains to control changes in the gradient. The density Hessian identity above gives, after averaging over observations,

$$\int_{\mathcal{A}} \|\nabla_\theta^2 p_\theta(a | x)\|_{\text{op}} da \leq M + G^2.$$

Applying the fundamental theorem of calculus once more,

$$\sup_x \int_{\mathcal{A}} \|\nabla p_\vartheta(a | x) - \nabla p_\theta(a | x)\| da \leq (M + G^2)\|\delta\|.$$

Since  $(\nabla P_\theta)1 = 0$ , we may write the gradient integrand using the reference-normalized differential value function:

$$b_\theta = \nabla \bar{r}_\theta + (\nabla P_\theta)h_\theta^{\text{ref}}.$$

The preceding density bound,  $0 \leq r \leq 1$ , the transition derivative bound in (68), and (57) therefore give

$$\|b_\vartheta - b_\theta\|_\infty \leq ((M + G^2)(1 + H_0) + GDC_P C_g)\|\delta\|.$$

Finally, decompose the gradient difference as

$$\nabla J(\vartheta) - \nabla J(\theta) = \nu_\vartheta(b_\vartheta - b_\theta) + (\nu_\vartheta - \nu_\theta)b_\theta.$$

The first term is controlled by the preceding sup-norm bound. For the second, the stationary perturbation bound (65) and  $\|b_\theta\|_{\text{Lip}} \leq C_g$  give

$$\|(\nu_\vartheta - \nu_\theta)b_\theta\| \leq GDC_P C_g \|\delta\|.$$

For vector-valued functions this follows by applying the scalar bound to each unit-vector projection. Adding the two bounds proves (59).  $\square$

The same Poisson inverse argument also gives the Lipschitz-seminorm counterpart of (57):

$$\|h_{\vartheta}^{\text{ref}} - h_{\theta}^{\text{ref}}\|_{\text{Lip}} \leq C_P C_g \|\vartheta - \theta\|.$$

The bounds used above also give

$$\|b_{\theta}\|_{\text{Lip}} \leq C_g, \quad \|b_{\theta}\|_{\infty} \leq G(1 + H_0).$$

The uniform bounds (67) supply the state regularity used in the directional-limit argument in Appendix A.3. Finite-trajectory differentiation is justified separately in Appendix D.1.

For boundary parameters, integrating the first-derivative bounds along parameter segments gives continuity in the operator norm on  $\text{Lip}(\mathcal{D})$ , without requiring differentiability in that norm. The primitive conditions hold on a neighborhood of  $\Theta$ . For  $\vartheta$  sufficiently close to  $\theta \in \Theta$ ,  $\|(I - P_{\theta})^{-1}(P_{\vartheta} - P_{\theta})\| < 1$  in this operator norm. The Neumann-series perturbation argument (Rudin, 1991, Theorem 10.7) therefore extends invertibility to a smaller neighborhood of  $\Theta$ . The invariant-law construction and policy-gradient proof then apply on this neighborhood, giving ordinary parameter derivatives at boundary points of  $\Theta$ .

## C Mean and second moments of the gradient estimate

The mean calculation for (39) combines baseline cancellation with randomized endpoints. Poisson reward-sum bounds control the second moment, allowing dependence between the burn-in baseline and the return.

### C.1 Mean of the score-weighted return

*Proof of Lemma 5.2.* We first remove the baseline from the expectation. Let  $\mathcal{G}_U$  contain all sampled lengths, the interaction history before step  $U$ , and the empirical observation  $\hat{\mu}_U$ , but not the action  $A_U$ . Both  $\hat{J}$  and  $V$  are determined by this information. The zero-mean score argument in (30) gives  $\mathbb{E}_x^{\theta}[S_U | \mathcal{G}_U] = 0$ . Consequently,

$$\mathbb{E}_x^{\theta}[S_U(V+1)\hat{J}] = \mathbb{E}_x^{\theta}[(V+1)\hat{J}\mathbb{E}_x^{\theta}[S_U | \mathcal{G}_U]] = 0.$$

Thus subtracting the baseline does not change the estimator's mean, without assuming that  $\hat{J}$  is unbiased for  $J(\theta)$  or independent of the later rewards.

The cancellation also holds conditional on  $U = u$  and  $V = \ell$ . The first  $u$  decisions bring the population from  $x$  to  $\mu_u$ . Conditional on this population and the preceding history, the subsequent score-weighted return has mean  $g_{\theta,\ell}(\mu_u)$ , defined in (33). Averaging over  $\mu_u$  gives

$$\mathbb{E}_x^{\theta}[\hat{g} | U = u, V = \ell] = \mathbb{E}_x^{\theta}[g_{\theta,\ell}(\mu_u) | U = u, V = \ell] = (P_{\theta}^u g_{\theta,\ell})(x).$$

Using the finite-return identity (33), this becomes

$$\mathbb{E}_x^{\theta}[\hat{g} | U = u, V = \ell] = \left\{ P_{\theta}^u \left[ b_{\theta} - (\nabla P_{\theta}) P_{\theta}^{\ell} h_{\theta} \right] \right\} (x).$$

It remains to average over the lengths. By (36) and the definition of  $Q_{\theta}$  in (34),

$$\mathbb{E}[P_{\theta}^U] = Q_{\theta}^B, \quad \mathbb{E}[P_{\theta}^V] = Q_{\theta}^N.$$

Averaging the conditional-mean formula over the independent lengths  $U, V$  gives (41). The operator order follows the trajectory:  $Q_{\theta}^B$  accounts for burn-in,  $\nabla P_{\theta}$  for the action at time  $U$ , and  $Q_{\theta}^N$  for the subsequent return steps.  $\square$

## C.2 Reward sums and the estimator second moment

The reward Poisson equation turns the centered reward sum into a bounded endpoint difference plus a sum of martingale differences. This decomposition applies from every starting population.

*Proof of Proposition 5.4.* Fix the policy parameter  $\theta$  and starting population  $x$ . Let  $\mathcal{F}_t$  contain the population at time  $t$ , the preceding history, and all independently pre-drawn lengths, but not the current observation, action, or reward. This filtration is used only for analysis; the learner need not observe the population.

Define  $\Delta_{t+1} := R_t - J(\theta) + h_\theta(\mu_{t+1}) - h_\theta(\mu_t)$ . The reward Poisson equation (20) gives

$$\begin{aligned}\mathbb{E}_x^\theta[\Delta_{t+1} \mid \mathcal{F}_t] &= \bar{r}_\theta(\mu_t) - J(\theta) + (P_\theta h_\theta)(\mu_t) - h_\theta(\mu_t) = 0, \\ |\Delta_{t+1}| &\leq |R_t - J(\theta)| + |h_\theta(\mu_{t+1})| + |h_\theta(\mu_t)| \leq 1 + 2H_0.\end{aligned}$$

The second line uses  $R_t, J(\theta) \in [0, 1]$  and  $\|h_\theta\|_\infty \leq H_0$ . For  $t \leq i < j$ , the martingale-difference property implies  $\mathbb{E}_x^\theta[\Delta_{i+1}\Delta_{j+1} \mid \mathcal{F}_t] = \mathbb{E}_x^\theta[\Delta_{i+1}\mathbb{E}_x^\theta[\Delta_{j+1} \mid \mathcal{F}_j] \mid \mathcal{F}_t] = 0$ .

Consequently, for any integer  $\ell \geq 1$ ,

$$\begin{aligned}\mathbb{E}_x^\theta \left[ \left( \sum_{k=t}^{t+\ell-1} (R_k - J(\theta)) \right)^2 \middle| \mathcal{F}_t \right] &= \mathbb{E}_x^\theta \left[ \left( h_\theta(\mu_t) - h_\theta(\mu_{t+\ell}) + \sum_{k=t}^{t+\ell-1} \Delta_{k+1} \right)^2 \middle| \mathcal{F}_t \right] \\ &\leq 8H_0^2 + 2\mathbb{E}_x^\theta \left[ \left( \sum_{k=t}^{t+\ell-1} \Delta_{k+1} \right)^2 \middle| \mathcal{F}_t \right] \\ &= 8H_0^2 + 2 \sum_{k=t}^{t+\ell-1} \mathbb{E}_x^\theta[\Delta_{k+1}^2 \mid \mathcal{F}_t] \\ &\leq 8H_0^2 + 2\ell(1 + 2H_0)^2 \leq C_R \ell.\end{aligned}$$

Here the first equality follows by telescoping the differential values; the first inequality uses  $(a+b)^2 \leq 2a^2 + 2b^2$  and the endpoint bound  $2H_0$ ; the next equality uses the vanishing cross terms. The final inequalities use the bound on  $\Delta_{k+1}$ ,  $\ell \geq 1$ , and the definition of  $C_R$  in (45). This estimate is uniform over the population and preceding history.

Now condition on  $U = u, V = v$ . Set  $A := \sum_{t=u}^{u+v} (R_t - J(\theta))$  and  $C := \sum_{t=0}^{u-1} (R_t - J(\theta))$ . Since  $\hat{J} - J(\theta) = C/u$ , we have  $Y = A - (v+1)C/u$ , and hence

$$\begin{aligned}\mathbb{E}_x^\theta[Y^2 \mid U = u, V = v] &= \mathbb{E}_x^\theta \left[ \left( A - \frac{v+1}{u}C \right)^2 \middle| U = u, V = v \right] \\ &\leq 2\mathbb{E}_x^\theta[A^2 \mid U = u, V = v] + 2\frac{(v+1)^2}{u^2}\mathbb{E}_x^\theta[C^2 \mid U = u, V = v] \\ &\leq 2C_R(v+1) + 2\frac{(v+1)^2}{u^2}C_R u \\ &= 2C_R \left[ v+1 + \frac{(v+1)^2}{u} \right].\end{aligned}$$

The first inequality requires no independence between  $A$  and  $C$ . The second applies the preceding reward-sum bound to the two segments, averaging over the history at time  $u$  for the return segment. Conditioning on the lengths is valid because they are drawn independently of the trajectory.

Finally,

$$\begin{aligned}\mathbb{E}_x^\theta Y^2 &\leq 2C_R \mathbb{E}_x^\theta \left[ V + 1 + \frac{(V+1)^2}{U} \right] \\ &\leq 2C_R \left[ sN + 1 + \frac{(sN+1)^2}{B} \right] = \mathcal{U}_{B,N}, \\ \mathbb{E}_x^\theta \|\hat{g}\|^2 &= \mathbb{E}_x^\theta [\|S_U\|^2 Y^2] \leq G^2 \mathbb{E}_x^\theta Y^2 \leq G^2 \mathcal{U}_{B,N} = \mathcal{V},\end{aligned}$$

where we used  $U \geq B$ ,  $V + 1 \leq sN + 1$ , and  $\|S_U\| \leq G$ . These are the bounds (46)–(47).  $\square$

In particular, the conditional bound for the burn-in sum in this proof gives the baseline mean-squared error  $\mathbb{E}_x^\theta (\hat{J} - J(\theta))^2 \leq C_R/B$ . This estimate does not require stationary initialization.

## D Likelihood-ratio differentiation and gradient differences

Differentiating the finite-trajectory mean  $f_x$  gives an unbiased Jacobian estimator. We evaluate this estimator at a single uniformly sampled intermediate parameter and multiply it by  $\theta - \vartheta$  to estimate  $f_x(\theta) - f_x(\vartheta)$  without bias. Replacing rewards by their conditional means before differentiation avoids assumptions on derivatives of the full reward distribution.

### D.1 The observable Jacobian estimator

*Proof of Lemma 6.1.* For the observable estimate (48), fix the starting population  $x$  and the lengths  $U = u, V = v$ . Until we average over the lengths below, write  $\mathbb{E}$  for  $\mathbb{E}_x^\theta[\cdot \mid U = u, V = v]$ . Baseline cancellation and the conditional reward model give  $f_{x;u,v}(\theta) = \mathbb{E}[\hat{g}] = \sum_{j=u}^{u+v} \mathbb{E}[S_u r(\mu_j, A_j)]$ . We first differentiate these mean-reward contributions, then compare the result with the observable estimator.

For  $j \geq u$ , let  $\tau_j = (\hat{\mu}_0, A_0, \dots, \hat{\mu}_j, A_j)$ . Fixing this record and  $x$  fixes the populations through  $\mu_{t+1} = F(\mu_t, A_t)$ , so  $r_j(\tau_j) := r(\mu_j, A_j)$  has no further dependence on  $\theta$ . Choose a fixed reference parameter  $\theta_\circ$  in a sufficiently small convex neighborhood where the policy bounds hold, and let  $\lambda_j$  be its record law from  $x$ . Common positive support gives the likelihood ratio

$$\begin{aligned}q_{\theta,j}(\tau_j) &= \prod_{t=0}^j \frac{\pi_\theta(A_t \mid \hat{\mu}_t)}{\pi_{\theta_\circ}(A_t \mid \hat{\mu}_t)}, \\ \nabla_\theta \log q_{\theta,j}(\tau_j) &= \sum_{t=0}^j \nabla_\theta \log \pi_\theta(A_t \mid \hat{\mu}_t) = \sum_{t=0}^j S_t, \\ \nabla_\theta q_{\theta,j}(\tau_j) &= q_{\theta,j}(\tau_j) \sum_{t=0}^j S_t.\end{aligned}$$

The first equality follows from the sequential sampling law: the observation kernels are identical under both parameters when the preceding action record is fixed. The remaining equalities follow by differentiating the log likelihood.

With the record held fixed,  $D_\theta S_u = M_u$ . The product rule therefore yields

$$\begin{aligned}
D_\theta \mathbb{E}[S_u r_j] &= D_\theta \int S_u r_j q_{\theta,j} d\lambda_j \\
&= \int r_j \left[ M_u q_{\theta,j} + S_u (\nabla_\theta q_{\theta,j})^\top \right] d\lambda_j \\
&= \int r_j \left( M_u + S_u \sum_{t=0}^j S_t^\top \right) q_{\theta,j} d\lambda_j \\
&= \mathbb{E} \left[ r_j \left( M_u + S_u \sum_{t=0}^j S_t^\top \right) \right].
\end{aligned}$$

The second equality differentiates the explicit score and the record density separately; the third substitutes the likelihood derivative. The record length is finite, and the bounded policy scores bound its likelihood ratio locally. Together with the uniform derivative bounds, this justifies differentiation under the integral. Summing over rewards gives

$$Df_{x;u,v}(\theta) = \mathbb{E} \left[ \sum_{j=u}^{u+v} r_j \left( M_u + S_u \sum_{t=0}^j S_t^\top \right) \right]. \quad (69)$$

To recover this expression from  $\hat{A}$ , let  $\mathcal{G}_t$  contain the sampled lengths, the preceding history, and  $\hat{\mu}_t$ , but not  $A_t$ . Differentiating the action-density normalization once and twice gives

$$\begin{aligned}
\mathbb{E}[S_t \mid \mathcal{G}_t] &= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a \mid \hat{\mu}_t) da = \nabla_\theta 1 = 0, \\
\mathbb{E}[M_t + S_t S_t^\top \mid \mathcal{G}_t] &= \int_{\mathcal{A}} \nabla_\theta^2 \pi_\theta(a \mid \hat{\mu}_t) da = \nabla_\theta^2 1 = 0.
\end{aligned}$$

The second line uses  $\nabla^2 \pi_\theta = \pi_\theta [\nabla^2 \log \pi_\theta + (\nabla \log \pi_\theta)(\nabla \log \pi_\theta)^\top]$ . Assumption 2 justifies exchanging derivatives and integrals.

For each  $j \geq u$ , scores after time  $j$  contribute zero:  $\mathbb{E}[R_j S_u S_t^\top] = \mathbb{E}[R_j S_u \mathbb{E}[S_t^\top \mid \mathcal{G}_t]] = 0$  for  $t > j$ . Consequently,

$$\begin{aligned}
&\mathbb{E} \left[ R_j \left( M_u + S_u \sum_{t=0}^{u+v} S_t^\top \right) \right] \\
&= \mathbb{E} \left[ R_j \left( M_u + S_u \sum_{t=0}^j S_t^\top \right) \right] \\
&= \mathbb{E} \left[ r_j \left( M_u + S_u \sum_{t=0}^j S_t^\top \right) \right].
\end{aligned}$$

The first equality removes the future scores using their conditional zero means. In the second, the remaining coefficient is determined before  $R_j$  is observed, so the conditional reward model (3) replaces  $R_j$  by  $r_j$ .

The baseline also contributes zero. Since  $\hat{J}$  is known before  $A_u$ , the two conditional score identities imply

$$\begin{aligned}
\mathbb{E}[\hat{J} S_u S_t^\top] &= \mathbb{E}[\hat{J} \mathbb{E}[S_u \mid \mathcal{G}_u] S_t^\top] = 0, & t < u, \\
\mathbb{E}[\hat{J} S_u S_t^\top] &= \mathbb{E}[\hat{J} S_u \mathbb{E}[S_t^\top \mid \mathcal{G}_t]] = 0, & t > u, \\
\mathbb{E}[\hat{J}(M_u + S_u S_u^\top)] &= \mathbb{E}[\hat{J} \mathbb{E}[M_u + S_u S_u^\top \mid \mathcal{G}_u]] = 0.
\end{aligned}$$

The first two lines condition on the history before the later action; the final line combines the Hessian with the same-time score product. Thus, expanding  $Y = \sum_{j=u}^{u+v} R_j - (v+1)\hat{J}$  and using (69), we obtain

$$\begin{aligned}\mathbb{E}[\hat{A}] &= \mathbb{E}\left[\left(\sum_{j=u}^{u+v} R_j - (v+1)\hat{J}\right)\left(M_u + S_u \sum_{t=0}^{u+v} S_t^\top\right)\right] \\ &= \sum_{j=u}^{u+v} \mathbb{E}\left[r_j \left(M_u + S_u \sum_{t=0}^j S_t^\top\right)\right] = Df_{x;u,v}(\theta).\end{aligned}$$

Restoring the expectation over lengths gives

$$\mathbb{E}_x^\theta \hat{A} = \mathbb{E}_{U,V}[Df_{x;U,V}(\theta)] = D_\theta \mathbb{E}_{U,V}[f_{x;U,V}(\theta)] = Df_x(\theta),$$

where differentiation commutes with averaging because the length distribution has finite support and is independent of  $\theta$ .

Finally, the call contains at most  $n$  decisions, so  $\|S_{\text{traj}}\| \leq nG$ . Hence

$$\begin{aligned}\mathbb{E}_x^\theta \|\hat{A}\|_{\text{op}}^2 &\leq \mathbb{E}_x^\theta \left[ Y^2 (\|M_U\|_{\text{op}} + \|S_U\| \|S_{\text{traj}}\|)^2 \right] \\ &\leq (M + nG^2)^2 \mathbb{E}_x^\theta Y^2 \\ &\leq (M + nG^2)^2 \mathcal{U}_{B,N} = \mathcal{W}.\end{aligned}$$

The first inequality uses the triangle inequality and the operator norm of a rank-one matrix; the second uses the policy derivative bounds; the last applies Proposition 5.4.  $\square$

## D.2 Integrating along the policy-parameter segment

*Proof of Proposition 6.2.* Let  $\delta = \theta - \vartheta$ . Conditional on  $u$ , Lemma 6.1 gives  $\mathbb{E}_x[\hat{v} \mid u] = Df_x(\vartheta + u\delta)\delta$ . Taking expectation over the single uniformly sampled  $u$  and applying the fundamental theorem of calculus along the parameter segment,

$$\mathbb{E}_x \hat{v} = \int_0^1 Df_x(\vartheta + u\delta)\delta \, du = f_x(\theta) - f_x(\vartheta).$$

The same hypothetical starting population  $x$  is held fixed throughout this identity. It does not require collecting two trajectories from  $x$ .

Subtracting the desired gradient difference now gives

$$\mathbb{E}_x \hat{v} - (\nabla J(\theta) - \nabla J(\vartheta)) = (f_x(\theta) - \nabla J(\theta)) - (f_x(\vartheta) - \nabla J(\vartheta)).$$

Each endpoint error is bounded by  $\varepsilon_{B,N}$ , which proves (50). No derivative of the estimator bias is required.

Finally, conditional on  $u$ ,

$$\mathbb{E}_x[\|\hat{v}\|^2 \mid u] \leq \mathbb{E}_x[\|\hat{A}_\theta\|_{\text{op}}^2 \mid u] \|\delta\|^2 \leq \mathcal{W} \|\delta\|^2.$$

Averaging proves (51).  $\square$

The bias bound (50) is uniform in the endpoints. The trajectory lengths are chosen to make  $\varepsilon_{B,N}$  small enough to control its accumulation over repeated calls.

## E Gradient domination under partial observation

We give sufficient conditions for the gradient-domination inequality (52). The proof connects two comparisons. The performance-difference identity bounds the global reward gap by local improvements evaluated under the optimal policy. Surrogate concavity and policy-improvement closure bound local improvements under the current policy by the Frank–Wolfe gap. Posterior transfer and observation coverage connect these two evaluations: they control, respectively, which hidden populations an observation represents and how frequently that observation occurs.

The posterior laws and local improvement gaps below are used to verify the guarantee; the learner need not estimate them. Remark F.2 covers coarser observation summaries.

### E.1 Improvements available from an observation

Fix a current policy  $\pi_\theta$  and a comparator  $\pi_\vartheta$ . We first express  $J(\vartheta) - J(\theta)$  in terms of action changes at individual observations. Taking the comparator to be optimal will then bound the remaining reward gap.

The same batch  $\xi$  can arise from different hidden populations, but the policy must use the same action distribution whenever it sees that batch. We therefore condition the population on the observation before comparing actions. Let  $\mathbf{M}_\vartheta$  be the stationary observation law and  $B_\vartheta(\cdot \mid \xi)$  the conditional law of the population given  $\xi$  under  $\pi_\vartheta$ .<sup>1</sup> The joint law can then be written in either order:

$$\Gamma_\vartheta(d\mu, d\xi) = \nu_\vartheta(d\mu) \mathbf{O}_m(d\xi \mid \mu) = B_\vartheta(d\mu \mid \xi) \mathbf{M}_\vartheta(d\xi). \quad (70)$$

In the observation-first order,  $\mathbf{M}_\vartheta$  gives the frequency of each observation and  $B_\vartheta(\cdot \mid \xi)$  gives the weights on populations consistent with it. Changing the policy can change both laws through  $\nu_\vartheta$ , even though the sampling kernel  $\mathbf{O}_m$  is fixed. These stationary conditional laws are used only in the analysis. For finite observation spaces,  $\mathbf{M}_\vartheta(\xi)$  denotes the mass at  $\xi$ .

At a fixed observation  $\xi$  and action  $a$ , the density difference  $\pi_\vartheta(a \mid \xi) - \pi_\theta(a \mid \xi)$  does not depend on  $\mu$ . Averaging the immediate reward and the next population’s differential value under  $B_\vartheta(\cdot \mid \xi)$  therefore defines the conditional action value

$$\mathbf{Q}_{\theta, \vartheta}(\xi, a) = \int_{\mathcal{D}} [r(\mu, a) + h_\theta(F(\mu, a))] B_\vartheta(d\mu \mid \xi).$$

The two subscripts have different roles:  $\theta$  fixes the differential value used to evaluate future effects, while  $\vartheta$  selects the posterior weights on hidden populations.

**Lemma E.1** (Performance difference conditioned on the observation). *For every  $\theta, \vartheta \in \Theta$ ,*

$$J(\vartheta) - J(\theta) = \int_{\mathcal{O}_m} \mathbf{M}_\vartheta(d\xi) \int_{\mathcal{A}} \mathbf{Q}_{\theta, \vartheta}(\xi, a) [\pi_\vartheta(a \mid \xi) - \pi_\theta(a \mid \xi)] da. \quad (71)$$

*Proof.* At a fixed population  $\mu$ , combine the integrated reward and transition differences using (6) and (5), keeping  $h_\theta$  fixed:

$$\bar{r}_\vartheta(\mu) - \bar{r}_\theta(\mu) + [(P_\vartheta - P_\theta)h_\theta](\mu) = \int_{\mathcal{A}} [r(\mu, a) + h_\theta(F(\mu, a))] [p_\vartheta(a \mid \mu) - p_\theta(a \mid \mu)] da. \quad (72)$$

---

<sup>1</sup>Empirical observations take values in the compact image of  $\mathcal{Z}^m$  under the empirical-measure map in (2). This image and the compact population space are standard Borel, so a regular conditional law exists (Kallenberg, 2002, Theorem 6.3).

Expand the action densities in (72) using (4), then average under  $\nu_\vartheta$ . The performance-difference identity (22) gives

$$J(\vartheta) - J(\theta) = \int \nu_\vartheta(d\mu) \int \mathcal{O}_m(d\xi | \mu) \int_{\mathcal{A}} [r(\mu, a) + h_\theta(F(\mu, a))] [\pi_\vartheta(a | \xi) - \pi_\theta(a | \xi)] da.$$

Use (70) to condition this joint law on  $\xi$ . Both action laws depend only on  $\xi$ , so averaging over  $\mu$  replaces the reward and differential-value term by  $\mathbf{Q}_{\theta, \vartheta}(\xi, a)$ , proving the identity. Bounded rewards and the bounded Poisson solution justify exchanging the integrals.  $\square$

The inner integral in (71) is the gain from replacing the current action law by the comparator's law at the same observation. We upper-bound it by choosing the best action law separately at each observation. Let  $\mathcal{K}(\xi)$  contain every available law  $\pi_v(\cdot | \xi)$ ,  $v \in \Theta$ , on the common action support. With posterior weights selected by  $\vartheta$ , the resulting local improvement gap is

$$d_{\theta, \vartheta}(\xi) = \sup_{p \in \mathcal{K}(\xi)} \int_{\mathcal{A}} \mathbf{Q}_{\theta, \vartheta}(\xi, a) p(da) - \int_{\mathcal{A}} \mathbf{Q}_{\theta, \vartheta}(\xi, a) \pi_\theta(a | \xi) da. \quad (73)$$

Throughout this comparison,  $\theta$  fixes both the differential value and the action law being improved;  $\vartheta$  selects the posterior used to evaluate that improvement. We require these gaps to be measurable. They are nonnegative because the current action law is feasible, and no attainment of the supremum is required.

The comparator's action law  $\pi_\vartheta(\cdot | \xi)$  belongs to  $\mathcal{K}(\xi)$ , so the inner integral in (71) is at most  $d_{\theta, \vartheta}(\xi)$  by (73). Averaging these bounds gives the desired comparison:

$$J(\vartheta) - J(\theta) \leq \int d_{\theta, \vartheta}(\xi) \mathbf{M}_\vartheta(d\xi). \quad (74)$$

This bounds a comparator's reward advantage by its average local improvement gap. The bound uses the comparator's observation law and posterior, whereas the gradient in (25) uses the current policy's stationary law. We next control local gains under the current policy, then compare the two sets of averaging laws.

## E.2 Implementing local improvements with one policy

The bound in (74) allows the action law from  $\mathcal{K}(\xi)$  to be chosen separately at every observation. Those choices need not belong to a single policy in our class. For example, the best action law at one observation may require a value of a shared parameter that conflicts with the best law at another observation. We need to control the loss in average gain from requiring one parameter  $v$  to determine all the action laws.

To connect these gains to  $\nabla J(\theta)$ , we first evaluate them under the current policy. Hold its stationary population law  $\nu_\theta$  and differential value  $h_\theta$  fixed, and change only the action law to  $\pi_v$ . The resulting policy-improvement surrogate is

$$\begin{aligned} \mathcal{L}_\theta(v) &= \int \mathbf{M}_\theta(d\xi) \int_{\mathcal{A}} \mathbf{Q}_{\theta, \theta}(\xi, a) \pi_v(a | \xi) da = \nu_\theta[\bar{r}_v + P_v h_\theta], \\ \nabla_v \mathcal{L}_\theta(v)|_{v=\theta} &= \nabla J(\theta). \end{aligned} \quad (75)$$

Thus  $\mathcal{L}_\theta(v)$  evaluates the replacement action law using the populations and differential values of the current policy. Differentiating at  $v = \theta$ , with  $\nu_\theta$  and  $h_\theta$  fixed, gives  $\nu_\theta[\nabla \bar{r}_\theta + (\nabla P_\theta) h_\theta] = \nu_\theta b_\theta$ , which

equals  $\nabla J(\theta)$  by Theorem 4.3. Although  $\mathcal{L}_\theta(v)$  and  $J(v)$  need not agree away from the current parameter, this identity connects surrogate improvement to the true gradient.

The best average gain from choosing an action law separately at each observation is the left side of the following inequality. The supremum on the right is the best gain achievable by one parameter. We require their difference to be at most  $\epsilon_{\text{obs}}$ , a condition called approximate closure under policy improvement:

$$\int d_{\theta,\theta}(\xi) \mathbf{M}_\theta(d\xi) \leq \sup_{v \in \Theta} [\mathcal{L}_\theta(v) - \mathcal{L}_\theta(\theta)] + \epsilon_{\text{obs}}. \quad (76)$$

If  $\mathcal{K}(\xi)$  contains laws outside the policy family, the error also accounts for those additional choices. Exact closure means that restricting the choices to one parameter does not reduce the best average gain.

The next lemma uses concavity of this surrogate to turn an achievable policy improvement into a bound involving the true gradient.

**Lemma E.2** (From policy improvement to the Frank–Wolfe gap). *If  $v \mapsto \mathcal{L}_\theta(v)$  is concave on  $\Theta$  and (76) holds, then*

$$\int d_{\theta,\theta}(\xi) \mathbf{M}_\theta(d\xi) \leq \mathcal{G}(\theta) + \epsilon_{\text{obs}}. \quad (77)$$

*Proof.* For each feasible  $v$ , concavity and the gradient identity in (75) imply

$$\mathcal{L}_\theta(v) - \mathcal{L}_\theta(\theta) \leq \langle \nabla J(\theta), v - \theta \rangle \leq \mathcal{G}(\theta).$$

Take the supremum over  $v$  and substitute in (76). □

Only the surrogate must be concave: its population law and differential value are held fixed, whereas the stationary objective  $J$  may still be nonconcave (Bhandari and Russo, 2024). Both the posterior in the local gap and the observation law in this bound belong to the current policy. The remaining step connects them to the comparator’s laws.

### E.3 From local improvement to global reward

Equation (74) bounds the reward gap using the comparator’s posterior and observation law, whereas Lemma E.2 uses the current policy’s laws. We connect them in two steps.

First, the same observation can represent different hidden populations under the two stationary policies. This changes the conditional action values and the available local improvement. We require the local gap under the comparator’s posterior to be bounded by a multiple of the gap under the current posterior, plus an additive error. This posterior transfer condition compares the gains directly; it does not require the two posterior distributions to be close in a probability metric.

Second, the observation itself can occur at different frequencies. Observation coverage bounds the optimal policy’s probability of any observation event by a constant times its probability under the current policy. In particular, the current posterior is then defined almost surely at observations encountered by the optimal comparator.

**Proposition E.3** (Gradient domination under partial observation). *Under the model of Section 3 and Assumptions 1–3, suppose there is an optimal comparator  $\theta^* \in \arg \max_{v \in \Theta} J(v)$  such that the following hold uniformly over  $\theta \in \Theta$ :*

- (i) **Implementable improvement.** The surrogate  $\mathcal{L}_\theta$  is concave and (76) holds with a uniform  $\epsilon_{\text{obs}} \geq 0$ .
- (ii) **Observation coverage.** For some  $C_{\text{obs}} \geq 1$ ,

$$M_{\theta^*} \ll M_\theta, \quad \frac{dM_{\theta^*}}{dM_\theta} \leq C_{\text{obs}} \quad M_\theta\text{-almost surely.} \quad (78)$$

- (iii) **Transfer across posteriors.** For some  $\Lambda \geq 1$ ,  $\epsilon_{\text{tr}} \geq 0$ , and nonnegative measurable  $e_{\theta, \theta^*}$ ,

$$d_{\theta, \theta^*}(\xi) \leq \Lambda d_{\theta, \theta}(\xi) + e_{\theta, \theta^*}(\xi), \quad \int e_{\theta, \theta^*} dM_{\theta^*} \leq \epsilon_{\text{tr}}. \quad (79)$$

The first inequality holds  $M_{\theta^*}$ -almost surely.

Then (52) holds with

$$\kappa = C_{\text{obs}}\Lambda, \quad \epsilon_{\text{geom}} = C_{\text{obs}}\Lambda\epsilon_{\text{obs}} + \epsilon_{\text{tr}}. \quad (80)$$

The linear maximization oracle in Assumption 4 is a separate property of the parameter domain.

*Proof of Proposition E.3.* Use the optimal comparator  $\theta^*$  specified in the hypotheses. Its action law belongs to  $\mathcal{K}(\xi)$  at each observation, so Lemma E.1 gives

$$\begin{aligned} J^* - J(\theta) &\leq \int d_{\theta, \theta^*}(\xi) M_{\theta^*}(d\xi) \\ &\leq \Lambda \int d_{\theta, \theta}(\xi) M_{\theta^*}(d\xi) + \epsilon_{\text{tr}} \\ &\leq C_{\text{obs}}\Lambda \int d_{\theta, \theta}(\xi) M_\theta(d\xi) + \epsilon_{\text{tr}} \\ &\leq C_{\text{obs}}\Lambda[\mathcal{G}(\theta) + \epsilon_{\text{obs}}] + \epsilon_{\text{tr}}. \end{aligned}$$

The successive bounds use transfer across posteriors, observation coverage with nonnegativity of the local gap, and Lemma E.2, establishing (80). Since  $C_{\text{obs}}, \Lambda \geq 1$ , their product is at least one.  $\square$

The comparisons need hold with only one optimal policy, uniformly over all current parameters  $\theta$ . The factor  $C_{\text{obs}}\Lambda$  multiplies both the Frank–Wolfe gap and the closure error. The transfer error enters directly because it is already averaged under the comparator’s observation law.

## E.4 Checking the structural conditions

**Observation coverage.** Absolute continuity ensures that the current posterior is defined at observations encountered by the comparator. For finite  $\mathcal{O}_m$ , coverage bounds  $M_{\theta^*}(\xi)/M_\theta(\xi)$ . A uniform lower bound  $M_\theta(\xi) \geq c > 0$  for every policy and observation implies  $C_{\text{obs}} \leq 1/c$ . Fixed-size success counts, for example, have uniformly positive probabilities when every population’s success probability is bounded away from zero and one. Such a certificate can be conservative for rare counts. Positive action probabilities alone do not provide it, because observation frequencies also depend on the stationary population law. For general observation spaces, the density bound in (78) imposes the same comparison between observation laws without requiring a Lebesgue density.

**Affine policy classes.** The policy architecture can make surrogate concavity and closure easy to check. Affine dependence of the action law on its parameters makes  $\mathcal{L}_\theta$  affine, hence concave. If each observation has an independently constrained parameter block and  $\mathcal{K}(\xi)$  is exactly that block’s

family of action laws, one parameter can realize all local choices at once. This gives exact closure. The following corollary combines these choices. We then give a separate test for how posterior changes affect local improvements.

**Corollary E.4** (Affine policies with independent observation blocks). *Suppose  $\mathcal{O}_m = \{\xi_1, \dots, \xi_L\}$ , and label these empirical observations by  $o = 1, \dots, L$ . Within this corollary, observation arguments use the label  $o$  in place of  $\xi_o$ . Let  $\Theta = \prod_{o=1}^L \Theta_o$ , and suppose the action law at observation  $o$  depends affinely only on its block  $\theta_o \in \Theta_o$ , where each  $\Theta_o$  is compact and convex. Take  $\mathcal{K}(o) = \{\pi_{v_o}(\cdot | o) : v_o \in \Theta_o\}$ . Then the surrogate is affine,  $\epsilon_{\text{obs}} = 0$ , and*

$$\int d_{\theta, \theta}(o) \mathbf{M}_{\theta}(\text{d}o) = \mathcal{G}(\theta). \quad (81)$$

Under (79) with zero error and observation coverage,  $J^* - J(\theta) \leq C_{\text{obs}} \Lambda \mathcal{G}(\theta)$ .

*Proof.* Fix the current parameter  $\theta$ . For each observation  $o$ , define

$$\ell_o(v_o) := \int_{\mathcal{A}} \mathbf{Q}_{\theta, \theta}(o, a) \pi_{v_o}(a | o) \text{d}a.$$

The conditional action values  $\mathbf{Q}_{\theta, \theta}(o, a)$  and stationary observation probabilities  $\mathbf{M}_{\theta}(o)$  remain fixed as  $v$  varies. Since the action law is affine in  $v_o$ , each  $\ell_o$  is affine, and so is  $\mathcal{L}_{\theta}(v) = \sum_{o=1}^L \mathbf{M}_{\theta}(o) \ell_o(v_o)$ .

The local family  $\mathcal{K}(o)$  is exactly the family generated by  $v_o \in \Theta_o$ . Hence

$$\begin{aligned} \int d_{\theta, \theta}(o) \mathbf{M}_{\theta}(\text{d}o) &= \sum_{o=1}^L \mathbf{M}_{\theta}(o) \left[ \max_{v_o \in \Theta_o} \ell_o(v_o) - \ell_o(\theta_o) \right] \\ &= \max_{v \in \prod_{o=1}^L \Theta_o} \sum_{o=1}^L \mathbf{M}_{\theta}(o) [\ell_o(v_o) - \ell_o(\theta_o)] \\ &= \max_{v \in \Theta} [\mathcal{L}_{\theta}(v) - \mathcal{L}_{\theta}(\theta)]. \end{aligned}$$

The first equality uses the definition of the local gap. The second holds because each summand depends only on its own parameter block, so all local maximizers can be chosen simultaneously. These maximizers exist by continuity and compactness. The final equality is the definition of the surrogate. This proves exact closure in (76), with  $\epsilon_{\text{obs}} = 0$ .

Moreover,

$$\begin{aligned} \max_{v \in \Theta} [\mathcal{L}_{\theta}(v) - \mathcal{L}_{\theta}(\theta)] &= \max_{v \in \Theta} \langle \nabla_v \mathcal{L}_{\theta}(v) |_{v=\theta}, v - \theta \rangle \\ &= \max_{v \in \Theta} \langle \nabla J(\theta), v - \theta \rangle \\ &= \mathcal{G}(\theta). \end{aligned}$$

The first equality uses affineness of the surrogate, and the second uses the gradient identity in (75). Thus the average local improvement gap equals the Frank–Wolfe gap, proving (81). Under observation coverage and zero-error posterior transfer, Proposition E.3 gives  $J^* - J(\theta) \leq C_{\text{obs}} \Lambda \mathcal{G}(\theta)$ .  $\square$

For a concrete example, suppose the individual outcome space is finite. With a fixed batch size, only finitely many empirical count vectors are possible. Give each count vector  $o$  its own mixture-weight vector  $w_o \in \Delta_K$ , where  $\Delta_K = \{w \in \mathbb{R}^K : w_j \geq 0, \sum_{j=1}^K w_j = 1\}$ , and define

$$\pi_w(a | o) = \eta q_0(a) + (1 - \eta) \sum_{j=1}^K w_{o,j} q_j(a), \quad 0 < \eta < 1,$$

where  $q_0, \dots, q_K$  are fixed action densities. The action space may be continuous; only the mixture weights are learned. Each observation has an independent parameter block, and the action law is affine in that block. Thus, when  $\mathcal{K}(o)$  is this mixture family, Corollary E.4 gives surrogate concavity and exact closure.

This construction can also satisfy the policy regularity assumptions. If the component densities are bounded and  $q_0$  is bounded away from zero, the exploration term keeps the policy density uniformly positive and bounds its score and log-density Hessian. For the smooth extension required by Assumption 2, divide the mixture by its total mass  $\eta + (1 - \eta) \sum_j w_{o,j}$ . This leaves the policy unchanged on the simplex and defines a positive, smooth density on a sufficiently small neighborhood. The observation-Lipschitz bounds follow from boundedness of the policy derivatives and the positive minimum distance between distinct observations in this finite space, together with compactness of the action space. With infinitely many observations, a separate block for each observation generally exceeds a finite-dimensional parameterization; shared parameters instead require a separate argument for (76), possibly with nonzero error. Nonlinear policy parameterizations also require a separate argument for surrogate concavity.

**Posterior transfer.** We next consider changes in the posterior over hidden populations. An action-independent shift in conditional action values cancels when comparing action laws, while a nonnegative rescaling multiplies the local improvement gap by the same factor. The next lemma uses these facts to establish posterior transfer and quantify the effect of residual errors.

**Lemma E.5** (Positive affine changes of conditional action values). *Suppose that, for every relevant  $\theta, \vartheta, \xi$ ,*

$$\mathbf{Q}_{\theta, \vartheta}(\xi, a) = \alpha_{\theta, \vartheta}(\xi) \mathbf{Q}_{\theta, \theta}(\xi, a) + c_{\theta, \vartheta}(\xi), \quad 0 \leq \alpha_{\theta, \vartheta}(\xi) \leq \Lambda, \quad (82)$$

*for actions in the common support. Then (79) holds with zero error. More generally, if the equality has an additional residual  $w_{\theta, \vartheta}(\xi, a)$ , the bound on local gaps holds with*

$$e_{\theta, \vartheta}(\xi) = \text{span}_a w_{\theta, \vartheta}(\xi, a), \quad \text{span}_a w = \sup_a w - \inf_a w,$$

*provided this nonnegative error is measurable and integrable.*

*Proof.* Integrating a difference of probability laws against an action-independent constant gives zero. For every  $p \in \mathcal{K}(\xi)$ , the local improvement under  $\mathbf{Q}_{\theta, \vartheta}$  is therefore  $\alpha_{\theta, \vartheta}$  times that under  $\mathbf{Q}_{\theta, \theta}$ . Since  $\alpha_{\theta, \vartheta} \geq 0$ , taking the supremum gives  $d_{\theta, \vartheta} = \alpha_{\theta, \vartheta} d_{\theta, \theta} \leq \Lambda d_{\theta, \theta}$ . With a residual, its integral against the signed law  $p - \pi_\theta(\cdot \mid \xi)$  is bounded by  $\sup_a w - \inf_a w$ , yielding the approximate statement after taking the supremum. If every local law assigns zero probability to action sets of zero reference measure, almost-everywhere identities and essential extrema (which ignore those sets) suffice.  $\square$

Policy-independent posteriors give  $\mathbf{Q}_{\theta, \vartheta} = \mathbf{Q}_{\theta, \theta}$ , so the lemma applies with  $\Lambda = 1$ . It also permits posterior changes that act through an action-independent shift and a bounded nonnegative rescaling. Comparing only the local improvement gaps, as in (79), is weaker than either requirement on the full conditional values. Even so, a comparison with zero error is restrictive: if the current action law is already locally optimal under the current posterior, it must remain locally optimal under the comparator's posterior at the observations covered by the condition. Otherwise the current gap is zero and the comparator gap is positive, which no finite  $\Lambda$  can cover without an additive error.

**Remark E.6** (Fixed auxiliary action randomization). Some models include an auxiliary random choice within each action whose law is fixed while the policy is optimized. Write the physical action as  $(a, u)$ , and suppose every admissible policy has the form  $\pi_v(da \mid \xi)K(du \mid \xi, a)$ , with the same

parameter-independent conditional kernel  $K$ . Since changing  $v$  changes only the law of  $a$ , first average the conditional action value over the auxiliary choice  $u$ :

$$\tilde{Q}_{\theta,\vartheta}(\xi, a) = \int Q_{\theta,\vartheta}(\xi, a, u) K(du \mid \xi, a),$$

and apply the local improvement conditions and gap comparisons to this function of  $a$ . Every policy-difference integral has exactly this form, so this operation is an identity. The physical transition remains the original deterministic  $F(\mu, a, u)$ . In particular, the positive affine relation in (82) need only hold after this averaging.

## E.5 Posterior shifts and approximation error

Another way to account for posterior changes is to bound their effect on the performance-difference identity. Evaluate the comparator's action changes using the posterior population distribution under the current policy  $\pi_\theta$ , while retaining the comparator's observation frequencies. Whenever  $M_\vartheta \ll M_\theta$ , define

$$A_{\theta,\vartheta}(\xi) = \int Q_{\theta,\vartheta}(\xi, a) [\pi_\vartheta(a \mid \xi) - \pi_\theta(a \mid \xi)] da.$$

Add and subtract  $Q_{\theta,\vartheta}$  inside the action-value term of (71). This gives

$$\begin{aligned} J(\vartheta) - J(\theta) &= \int A_{\theta,\vartheta} dM_\vartheta + \mathcal{E}_\theta(\vartheta), \\ \mathcal{E}_\theta(\vartheta) &= \int M_\vartheta(d\xi) \int [Q_{\theta,\vartheta}(\xi, a) - Q_{\theta,\theta}(\xi, a)] [\pi_\vartheta(a \mid \xi) - \pi_\theta(a \mid \xi)] da. \end{aligned} \tag{83}$$

The first term values the action changes under the current posterior, weighted by the comparator's observation frequencies. The remainder corrects for the posterior mismatch by combining the change in action law with the change in conditional values. Both  $Q_{\theta,\vartheta}$  and  $Q_{\theta,\theta}$  use the same differential value function  $h_\theta$  under policy  $\pi_\theta$  and differ only in their conditional population distributions. For densities  $p$  and  $q$  with respect to the common reference measure  $da$ , write  $\text{TV}(p, q) = \frac{1}{2} \int |p(a) - q(a)| da$ . The bound  $|\int f(p - q)| \leq \text{TV}(p, q) \text{span}(f)$  gives

$$|\mathcal{E}_\theta(\vartheta)| \leq \int M_\vartheta(d\xi) \text{TV}(\pi_\vartheta(\cdot \mid \xi), \pi_\theta(\cdot \mid \xi)) \text{span}_a[Q_{\theta,\vartheta}(\xi, a) - Q_{\theta,\theta}(\xi, a)]. \tag{84}$$

To connect this remainder to gradient domination, the local-gap definition (73) gives  $A_{\theta,\vartheta}(\xi) \leq d_{\theta,\vartheta}(\xi)$ , because the comparator's action law belongs to  $\mathcal{K}(\xi)$ . For the optimal comparator  $\vartheta = \theta^*$ , observation coverage (78) and the surrogate bound (77) bound the first term in (83) by  $C_{\text{obs}}[\mathcal{G}(\theta) + \epsilon_{\text{obs}}]$ . A bound on  $|\mathcal{E}_\theta(\theta^*)|$  that is uniform over the current parameter  $\theta$ , such as one obtained from (84), therefore enters as an additive error in (52). If the invariant law is policy independent, the posterior remainder is zero even with a fixed noisy batch.

Alternatively, (79) with zero error bounds the comparator's local improvement by a multiple of the current policy's, even when the posteriors differ. Together with surrogate concavity, exact closure, and observation coverage, it gives  $\epsilon_{\text{geom}} = 0$ . These conclusions concern the same fixed- $m$  policy-class benchmark  $J^*$ . They do not identify it with a fully observed optimum. A fixed positive structural error instead contributes  $T\epsilon_{\text{geom}}$  to regret, so a sublinear guarantee from that bound requires further control of the error.

The posterior laws, differential values, and local gaps certify the conditions in Proposition E.3, while Algorithm 3 forms its updates from trajectory estimates and linear maximization over  $\Theta$ . The next appendix verifies these conditions in a model with noisy observations and a nonconcave objective.

## F An example with hidden populations and gradient domination

We construct an environment in which a finite sample batch never reveals the current population exactly, yet the Frank–Wolfe gap controls suboptimality in long-run reward. The model has two hidden populations and a continuous control that affects both current reward and future populations. For suitable parameters it satisfies Corollary E.4, even though the long-run objective is not concave. After checking regularity and the Poisson bound, we compute stationary probabilities and differential values and use Bayes’ rule to verify the observation conditions. We then show that no policy maximizes improvement at both hidden populations simultaneously and establish nonconcavity.

### F.1 The population, continuous action, and deterministic update

The population can be in one of two states. Each action combines a continuous control with a sampled update rule: keep the present population, move to the first population, or move to the second. We call this update rule the action’s mode. The control affects both the reward and the probabilities of these three modes. Fix  $0 < \delta < 1$ ,  $0 < \beta < 1$ ,  $0 < \lambda < 1$ ,  $0 < \epsilon < 1 - \lambda$ , and an odd integer  $m \geq 1$ . Set

$$\mathcal{Z} = \{0, 1\}, \quad \mu_\sigma = \text{Bernoulli}\left(\frac{1 + \delta\sigma}{2}\right), \quad \sigma \in \{-1, +1\}, \quad \mathcal{D} = \{\mu_{-1}, \mu_{+1}\}.$$

Under the usual absolute distance on  $\mathcal{Z}$ , the populations are separated by  $W_1(\mu_{-1}, \mu_{+1}) = \delta$ . The empirical observation from  $m$  fresh samples is determined by their success count  $C \in \{0, \dots, m\}$ : set  $\xi_k = \text{Bernoulli}(k/m)$ , so  $\hat{\mu} = \xi_C$  and  $\mathcal{O}_m = \{\xi_0, \dots, \xi_m\}$ . The correspondence between  $k$  and  $\xi_k$  is one-to-one and loses no observation information. Throughout this example,  $k$  in an observation argument is shorthand for  $\xi_k$ , as in  $\pi_\theta(\cdot \mid k)$ ,  $M_\theta(k)$ , and  $B_\theta(\cdot \mid k)$ . The count probabilities at population  $\mu_\sigma$  are

$$e_\sigma(k) = \binom{m}{k} \left(\frac{1 + \delta\sigma}{2}\right)^k \left(\frac{1 - \delta\sigma}{2}\right)^{m-k}.$$

The restriction  $0 < \delta < 1$  makes the populations distinct while keeping every  $e_\sigma(k)$  positive. Thus samples carry information about the hidden population, but no count identifies it with certainty.

Let  $\Theta = [-1, 1]^{m+1}$ , with coordinate  $[\theta]_k$  for observed count  $k$ , indexed from zero. Given  $C = k$ , draw a continuous control  $a \in [-1, 1]$  from a tilted uniform density:

$$\varphi_\theta(a \mid k) = \frac{1 + \beta[\theta]_k a}{2}. \tag{85}$$

The sign of  $[\theta]_k$  favors positive or negative controls. The restriction  $\beta < 1$  leaves the density at least  $(1 - \beta)/2$ , retaining exploration throughout the action interval across all feasible parameters. Separate coordinates for the counts let the policy change its response to one observation without constraining its response to another. Given  $a$ , draw a mode  $j \in \{0, +, -\}$  from the fixed, known probabilities

$$\omega_0(a) = \lambda, \quad \omega_+(a) = \frac{1 - \lambda + \epsilon a}{2}, \quad \omega_-(a) = \frac{1 - \lambda - \epsilon a}{2}.$$

The actual action is  $A = (j, a)$  in the compact space  $\mathcal{A} = \{0, +, -\} \times [-1, 1]$ , with density  $\pi_\theta(j, a \mid k) = \omega_j(a) \varphi_\theta(a \mid k)$  relative to counting measure times Lebesgue measure, with both factors

strictly positive. Along a trajectory, write  $\mu_t = \mu_{\sigma_t}$  and  $A_t = (j_t, a_t)$ , omitting time indices when discussing one step. The sampled mode determines the population update:

$$F(\mu_\sigma, (j, a)) = \begin{cases} \mu_\sigma, & j = 0, \\ \mu_{+1}, & j = +, \\ \mu_{-1}, & j = -. \end{cases} \quad (86)$$

The update remains deterministic given the population and action. Averaging over its random mode gives

$$\mathbb{E}[\sigma_{t+1} \mid \sigma_t = \sigma, a_t = a] = \lambda\sigma + \epsilon a. \quad (87)$$

Here  $\lambda$  controls persistence, while  $\epsilon$  controls how strongly the continuous action shifts the next population. The restriction  $\epsilon < 1 - \lambda$  keeps both modes  $j = +$  and  $j = -$  possible at every control value. This example thus assumes a specific intervention mechanism whose sampled update mode is recorded as part of the action. Choose the population reward

$$r(\mu_\sigma, (j, a)) = \frac{1 + \delta\sigma a}{2}. \quad (88)$$

Given the population  $\mu_\sigma$ , draw a fresh evaluation sample  $Z^{\text{ev}} \sim \mu_\sigma$  independently of the observation batch, action, and earlier history. The observed reward is  $R = \frac{1 + (2Z^{\text{ev}} - 1)a}{2} \in [0, 1]$ . Its conditional mean is (88), satisfying (3). Under this reward, the two hidden populations prefer opposite controls.

## F.2 Regularity and the uniform Poisson bound

We check that the construction satisfies the assumptions needed for gradient estimation before checking gradient domination. With  $d_{\mathcal{A}}((j, a), (j', a')) = \mathbf{1}_{\{j \neq j'\}} + |a - a'|$ , the map (86) satisfies

$$d_{\mathcal{D}}(F(\mu, A), F(\nu, A')) \leq d_{\mathcal{D}}(\mu, \nu) + \delta d_{\mathcal{A}}(A, A'),$$

and the reward satisfies

$$|r(\mu, A) - r(\nu, A')| \leq d_{\mathcal{D}}(\mu, \nu) + \frac{1}{2} d_{\mathcal{A}}(A, A').$$

The action metric assigns distance at least one to different modes, so even a mode change that switches the population satisfies the Lipschitz bound.

For count  $C = k$ , let  $\mathbf{e}_k$  be the corresponding coordinate vector. The observable derivatives are

$$\nabla \log \pi_\theta(j, a \mid k) = \frac{\beta a}{1 + \beta[\theta]_k a} \mathbf{e}_k, \quad \nabla^2 \log \pi_\theta(j, a \mid k) = -\frac{\beta^2 a^2}{(1 + \beta[\theta]_k a)^2} \mathbf{e}_k \mathbf{e}_k^\top.$$

The score's Euclidean norm and the Hessian's operator norm are bounded on  $\Theta$  by  $\beta/(1 - \beta)$  and  $\beta^2/(1 - \beta)^2$ , respectively. The bounds deteriorate as  $\beta$  approaches one, so retaining a fixed exploration floor also controls the derivative constants. Choose  $r_{\text{nbhd}} > 0$  such that  $\beta(1 + r_{\text{nbhd}}) < 1$ . On the open neighborhood  $\mathcal{U} = \{\theta : \|\theta\|_\infty < 1 + r_{\text{nbhd}}\}$ , the score denominator is at least  $1 - \beta(1 + r_{\text{nbhd}}) > 0$ , giving analogous uniform bounds. The affine densities are twice continuously differentiable in  $\mathcal{L}^1(\text{d}A)$ .

The policy is specified at the success fractions  $k/m$  that can occur in a sample batch. For convenience in checking observation regularity, extend it between these grid points to every Bernoulli( $x$ ),  $x \in [0, 1]$ , by linearly interpolating  $[\theta]_0, \dots, [\theta]_m$  in (85). The interpolant is  $2m$ -Lipschitz in  $x$  on  $\Theta$ , and

$2m(1 + r_{\text{nbhd}})$ -Lipschitz uniformly on  $\mathcal{U}$ . Its parameter gradient is also continuous, piecewise linear, and Lipschitz in  $x$ . Both the observation-policy Wasserstein bound (12) and the  $\mathcal{L}^1$  derivative bound (15) follow, with constants that may depend on the fixed batch size  $m$ .

To describe the population transition, first compute the mean control selected in each current population. At count  $k$ , integrating the tilted density gives  $\int_{-1}^1 a \varphi_\theta(a | k) da = \beta[\theta]_k/3$ . Let  $\bar{a}$  bound the magnitude of the mean control at any count. Averaging over the count gives the mean control  $c_\sigma(\theta)$  in population  $\mu_\sigma$ . We also define the common component  $u_\theta$  and difference component  $v_\theta$  of the two population means:

$$\begin{aligned} \bar{a} &= \frac{\beta}{3}, & c_\sigma(\theta) &= \mathbb{E}^\theta[a | \sigma] = \bar{a} \sum_{k=0}^m e_\sigma(k) [\theta]_k, \\ u_\theta &= \frac{c_{+1}(\theta) + c_{-1}(\theta)}{2}, & v_\theta &= \frac{c_{+1}(\theta) - c_{-1}(\theta)}{2}. \end{aligned} \tag{89}$$

These definitions give  $c_\sigma(\theta) = u_\theta + \sigma v_\theta$  and  $|c_\sigma(\theta)| \leq \bar{a}$ , hence

$$|u_\theta| + |v_\theta| \leq \bar{a}. \tag{90}$$

Averaging the conditional dynamics (87) under the policy, and using  $c_\sigma(\theta) = u_\theta + \sigma v_\theta$ , gives the expected next population sign:

$$\mathbb{E}^\theta[\sigma_{t+1} | \sigma_t = \sigma] = \epsilon u_\theta + (\lambda + \epsilon v_\theta) \sigma.$$

Every function on the two populations is affine in their sign  $\sigma$ . The displayed conditional mean therefore shows that  $P_\theta$  multiplies the difference between a function's two values by  $\lambda + \epsilon v_\theta$ . Thus, for a given function  $f$ , subtracting the two state equations in  $(I - P_\theta)h = f - c$  cancels  $c$  and determines the solution's value difference:

$$h(\mu_{+1}) - h(\mu_{-1}) = \frac{f(\mu_{+1}) - f(\mu_{-1})}{1 - \lambda - \epsilon v_\theta}.$$

The Lipschitz seminorm on two states is the absolute value of this difference divided by the distance between the states. The inverse scales that seminorm by the same factor. Using  $|v_\theta| \leq \bar{a}$  from (90) gives

$$\|(I - P_\theta)^{-1}\|_{\text{Lip} \rightarrow \text{Lip}} = \frac{1}{1 - \lambda - \epsilon v_\theta} \leq \frac{1}{1 - \lambda - \epsilon \bar{a}}. \tag{91}$$

Since  $\epsilon < 1 - \lambda$  and  $\bar{a} < 1$ , the denominator in (91) is positive and Assumption 3 holds uniformly. The uniformly positive probabilities of the modes  $j = +$  and  $j = -$  also ensure a unique invariant law. This particular construction is aperiodic, although the general assumption permits periodic dynamics.

### F.3 The stationary prior and differential values

To compare controls after observing a sample count, we need the populations' stationary probabilities and the effect of the next population on subsequent reward. Let  $z_\theta$  be the stationary probability of  $\mu_{+1}$  minus that of  $\mu_{-1}$ . We also use the difference  $\Delta h_\theta$  between their differential values and a uniform bound  $\bar{z}$  on the population imbalance:

$$z_\theta = \mathbb{E}_{\text{stat}}^\theta[\sigma], \quad \Delta h_\theta = h_\theta(\mu_{+1}) - h_\theta(\mu_{-1}), \quad \bar{z} = \frac{\epsilon \bar{a}}{1 - \lambda} < 1.$$

By (89), the mean control at population  $\mu_\sigma$  is  $u_\theta + v_\theta\sigma$ . Taking stationary expectations in (87) therefore gives  $z_\theta = (\lambda + \epsilon v_\theta)z_\theta + \epsilon u_\theta$ . Solving for  $z_\theta$  yields

$$z_\theta = \frac{\epsilon u_\theta}{1 - \lambda - \epsilon v_\theta}, \quad |z_\theta| \leq \bar{z}. \quad (92)$$

This determines the invariant weights as  $\nu_\theta(\{\mu_\sigma\}) = (1 + \sigma z_\theta)/2$ . For the bound, stationarity also gives  $(1 - \lambda)z_\theta = \epsilon \mathbb{E}_{\text{stat}}^\theta[a]$ , and the last expectation lies in  $[-\bar{a}, \bar{a}]$ .

Averaging the reward (88) over controls with mean  $c_\sigma(\theta) = u_\theta + \sigma v_\theta$  gives  $\bar{r}_\theta(\mu_\sigma) = \frac{1}{2} + \frac{\delta}{2}(\sigma u_\theta + v_\theta)$ . The common term cancels when we subtract the two populations' rewards, so  $\bar{r}_\theta(\mu_{+1}) - \bar{r}_\theta(\mu_{-1}) = \delta u_\theta$ . As shown above,  $P_\theta$  multiplies a two-state value difference by  $\lambda + \epsilon v_\theta$ . Subtracting the reward Poisson equation (20) at the two populations therefore gives  $(1 - \lambda - \epsilon v_\theta)\Delta h_\theta = \delta u_\theta$ . Solving this scalar equation and comparing with (92) yields

$$\Delta h_\theta = \frac{\delta u_\theta}{1 - \lambda - \epsilon v_\theta}, \quad \epsilon \Delta h_\theta = \delta z_\theta. \quad (93)$$

Thus the difference in differential values under policy  $\pi_\theta$  is proportional to the stationary imbalance. The centered solution is  $h_\theta(\mu_\sigma) = \frac{\Delta h_\theta}{2}(\sigma - z_\theta)$ .

#### F.4 Bayes' rule and policy improvement

The same sample count can arise from either population. Its posterior population probabilities depend on how often the policy visits each population. We now find a condition ensuring that these probabilities do not reverse the preferred control direction. Start with a prior assigning equal mass to both populations. After observing  $C = k$ , its posterior mean of  $\sigma$  is

$$t_k = \frac{e_{+1}(k) - e_{-1}(k)}{e_{+1}(k) + e_{-1}(k)} = \tanh((2k - m) \operatorname{arctanh} \delta).$$

An odd batch size rules out a tied count, giving the uniform margin

$$|t_k| \geq \delta, \quad \operatorname{sign}(t_k) = \operatorname{sign}(2k - m) =: \tau_k.$$

Under policy  $\pi_\vartheta$ , the stationary prior instead has mean  $z_\vartheta$ , so the joint probability of population  $\mu_\sigma$  and count  $k$  is  $(1 + \sigma z_\vartheta)e_\sigma(k)/2$ . Let  $\mathbf{M}_\vartheta(k)$  denote the stationary probability of count  $k$  under policy  $\pi_\vartheta$ . It is the sum of these two joint probabilities, while their difference divided by their sum gives the posterior mean. Factoring out  $e_{+1}(k) + e_{-1}(k)$  yields

$$\mathbf{M}_\vartheta(k) = \frac{e_{+1}(k) + e_{-1}(k)}{2}(1 + z_\vartheta t_k), \quad (94)$$

$$\bar{\sigma}_\vartheta(k) := \mathbb{E}_{\text{stat}}^\vartheta[\sigma \mid C = k] = \frac{t_k + z_\vartheta}{1 + t_k z_\vartheta}. \quad (95)$$

Every policy assigns positive probability to every sample count. Since  $|t_k| \leq 1$  and  $|z_\vartheta| \leq \bar{z}$ , the factor  $1 + z_\vartheta t_k$  lies between  $1 - \bar{z}$  and  $1 + \bar{z}$ . In (94), the common factor  $(e_{+1}(k) + e_{-1}(k))/2$  cancels when comparing count probabilities under two policies, giving

$$\mathbf{M}_\vartheta(k) \leq C_{\text{obs}} \mathbf{M}_\theta(k), \quad C_{\text{obs}} = \frac{1 + \bar{z}}{1 - \bar{z}}. \quad (96)$$

The policy parameter changes the control density, but the conditional mode probabilities  $\omega_j(a)$  are fixed. We can therefore compare controls by first averaging their reward and next-state differential value over the mode, as in Remark E.6:

$$\tilde{q}_\theta(\sigma, a) = \sum_j [r(\mu_\sigma, (j, a)) + h_\theta(F(\mu_\sigma, (j, a)))] \omega_j(a).$$

To compute this average, write  $h_\theta(\mu_\sigma) = \frac{1}{2}[h_\theta(\mu_{+1}) + h_\theta(\mu_{-1})] + \frac{1}{2}\Delta h_\theta \sigma$ . By the conditional dynamics (87), averaging over the mode gives the expected next sign  $\lambda\sigma + \epsilon a$ . Adding the reward (88) and using  $\epsilon\Delta h_\theta = \delta z_\theta$  from (93) then gives

$$\tilde{q}_\theta(\sigma, a) = \frac{1}{2} + \frac{h_\theta(\mu_{+1}) + h_\theta(\mu_{-1})}{2} + \frac{\lambda\Delta h_\theta}{2}\sigma + \frac{\delta(\sigma + z_\theta)}{2}a. \quad (97)$$

Averaging (97) over the posterior population distribution under policy  $\pi_\theta$ , given the sample count  $C = k$ , gives

$$\tilde{\mathbf{Q}}_{\theta, \vartheta}(k, a) := \mathbb{E}_{\text{stat}}^\vartheta[\tilde{q}_\theta(\sigma, a) \mid C = k] = \tilde{\mathbf{Q}}_{\theta, \vartheta}(k, 0) + \frac{\delta(\bar{\sigma}_\vartheta(k) + z_\theta)}{2}a, \quad (98)$$

where the value at zero control is the action-independent intercept. In (98), only the slope in  $a$  affects improvement. It combines the conditional mean  $\bar{\sigma}_\vartheta(k)$  of the population sign  $\sigma$  under policy  $\pi_\theta$ , given  $C = k$ , with  $z_\theta = \epsilon\Delta h_\theta/\delta$ , the contribution from the next population's differential value under policy  $\pi_\theta$ , by (93).

To keep posterior changes from reversing this slope, impose the margin condition

$$\delta > \frac{2\bar{z}}{1 + \bar{z}^2}. \quad (99)$$

The information in a sample count must outweigh both the shift in the population prior and the contribution from  $\Delta h_\theta$ . For fixed  $\delta, \beta, \lambda$ , the condition holds for sufficiently small positive  $\epsilon$ , while stronger feedback can violate it. To see how it controls the slope, define  $\gamma = \frac{\delta - \bar{z}}{1 - \delta\bar{z}}$  and rewrite (95) as

$$\tau_k \bar{\sigma}_\vartheta(k) = \frac{|t_k| + \tau_k z_\vartheta}{1 + |t_k| \tau_k z_\vartheta}.$$

The fraction  $(x + y)/(1 + xy)$  is increasing in each argument for  $x \in [\delta, 1)$  and  $y \in [-\bar{z}, \bar{z}]$ . Its minimum is therefore attained at  $x = \delta$ ,  $y = -\bar{z}$ , and equals  $\gamma$ . Adding  $\tau_k z_\theta \in [-\bar{z}, \bar{z}]$ , and using  $|\bar{\sigma}_\vartheta(k)| \leq 1$  for the upper bound, gives

$$\tau_k \bar{\sigma}_\vartheta(k) \geq \gamma, \quad \gamma - \bar{z} \leq \tau_k(\bar{\sigma}_\vartheta(k) + z_\theta) \leq 1 + \bar{z} \quad \text{for all } \theta, \vartheta, k. \quad (100)$$

The margin condition is equivalent to  $\gamma > \bar{z}$ , so even the smallest posterior margin exceeds the largest opposing contribution from  $\Delta h_\theta$ . The slope therefore has the same nonzero sign  $\tau_k$  under every stationary prior. For a candidate parameter block  $w \in [-1, 1]$ , replace the current density  $\varphi_\theta(a \mid k)$  by  $(1 + \beta wa)/2$ . Integrating their difference gives

$$\int_{-1}^1 \tilde{\mathbf{Q}}_{\theta, \vartheta}(k, a) \left[ \frac{1 + \beta wa}{2} - \varphi_\theta(a \mid k) \right] da = \frac{\delta\beta}{6}(w - [\theta]_k)(\bar{\sigma}_\vartheta(k) + z_\theta).$$

Maximizing over the feasible scalar parameter  $w$  gives the largest improvement available after observing count  $k$ :

$$d_{\theta, \vartheta}(k) = \frac{\delta\beta}{6} |\bar{\sigma}_\vartheta(k) + z_\theta| (1 - \tau_k [\theta]_k). \quad (101)$$

The same parameter block  $w = \tau_k$  maximizes the conditional improvement under every stationary prior. Applying the coefficient bounds (100) to the local-gap formula (101) proves

$$d_{\theta,\vartheta}(k) \leq \Lambda d_{\theta,\theta}(k), \quad \Lambda = \frac{1 + \bar{z}}{\gamma - \bar{z}}. \quad (102)$$

The bound also holds when  $1 - \tau_k[\theta]_k = 0$ , since both sides then vanish. Equivalently, the reduced conditional  $Q$ -functions differ by a positive scaling and an action-independent offset, as in Lemma E.5.

**Proposition F.1** (Gradient domination with fixed noisy observations). *Under (99), this model satisfies Assumptions 1–3, and*

$$J^* - J(\theta) \leq \kappa \mathcal{G}(\theta), \quad \kappa = C_{\text{obs}} \Lambda = \frac{(1 + \bar{z})^2}{(1 - \bar{z})(\gamma - \bar{z})} \quad \text{for every } \theta \in \Theta. \quad (103)$$

There is no additive error. Here  $J^*$  is the optimum over the specified observation-based policy class.

*Proof.* With regularity and the uniform Poisson bound (91) established, it remains to verify gradient domination. The affine conditional policy has an independent scalar parameter for each count  $k$ , so one parameter vector can achieve all the best feasible conditional improvements simultaneously. The policy-gradient formula (29) gives

$$[\nabla J(\theta)]_k = \mathbb{M}_\theta(k) \frac{\delta\beta}{6} (\bar{\sigma}_\theta(k) + z_\theta), \quad \mathcal{G}(\theta) = \sum_{k=0}^m \mathbb{M}_\theta(k) d_{\theta,\theta}(k).$$

Applying the performance-difference identity (71) to an optimal parameter  $\theta^*$ , then taking the best feasible conditional improvement, gives

$$J^* - J(\theta) \leq \sum_{k=0}^m \mathbb{M}_{\theta^*}(k) d_{\theta,\theta^*}(k) \leq C_{\text{obs}} \Lambda \sum_{k=0}^m \mathbb{M}_\theta(k) d_{\theta,\theta}(k).$$

Here the last inequality uses observation coverage (96) and the comparison of local gaps (102), completing the verification of Corollary E.4.  $\square$

For the concrete choice  $\delta = \frac{1}{2}$ ,  $\beta = \frac{1}{2}$ ,  $\lambda = \frac{1}{2}$ , and  $\epsilon = \frac{1}{10}$ , we have  $\bar{z} = 1/30$  and  $\gamma = 28/59$ . Substituting these values into (103) gives  $\kappa = \frac{56699}{22649} < 3$ . Hence  $J^* - J(\theta) \leq 3\mathcal{G}(\theta)$  for every fixed odd  $m$ , with a domination constant independent of  $m$ , even though the parameter dimension and some observation-regularity constants depend on it.

## F.5 Gradient domination despite nonconcavity

We now show that no policy parameter maximizes one-step improvement at both hidden populations, and that the long-run objective is not concave. For a reference parameter  $\theta$  and a replacement vector  $w \in \Theta$ , hold the differential value function  $h_\theta$  under policy  $\pi_\theta$  fixed. The improvement at population  $\mu_\sigma$  is

$$D_{\theta,w}(\mu_\sigma) = \sum_{k=0}^m e_\sigma(k) \int_{-1}^1 \tilde{q}_\theta(\sigma, a) (\varphi_w(a | k) - \varphi_\theta(a | k)) da.$$

Integrating the affine control densities gives

$$D_{\theta,w}(\mu_\sigma) = \frac{\delta\beta}{6} (\sigma + z_\theta) \sum_{k=0}^m e_\sigma(k) ([w]_k - [\theta]_k).$$

Since  $|z_\theta| < 1$  and every count probability is positive, the unique maximizers for  $\sigma = +1$  and  $\sigma = -1$  require, respectively,  $[w]_k = +1$  and  $[w]_k = -1$  for every  $k$ . No single feasible parameter can attain both. As both populations have positive stationary mass, compactness gives the strict inequality

$$\nu_\theta \left[ \sup_{w \in \Theta} D_{\theta,w} \right] > \sup_{w \in \Theta} \nu_\theta D_{\theta,w}.$$

The left side performs policy improvement separately at each hidden population; the right side must use one common parameter. A fully observed controller could choose different controls at the two populations, but each sample count is compatible with both. Closure in (76) instead asks for the best feasible decision at each observed count, which the policy implements through its separate count coordinates.

To check concavity, compute the stationary reward under the same policy. Since  $c_\sigma(\theta) = u_\theta + v_\theta \sigma$ , we have  $\mathbb{E}_{\text{stat}}^\theta[\sigma a] = v_\theta + u_\theta z_\theta$ . Averaging the reward (88) in stationarity and substituting (92) gives

$$J(\theta) = \frac{1}{2} + \frac{\delta}{2} \left( v_\theta + \frac{\epsilon u_\theta^2}{1 - \lambda - \epsilon v_\theta} \right). \quad (104)$$

Along the feasible line  $[\theta]_k = t$  for every  $k$ ,  $u_\theta = \bar{a}t$  and  $v_\theta = 0$ , so (104) reduces to

$$J(t\mathbf{1}) = \frac{1}{2} + \frac{\delta \bar{a}^2}{2(1 - \lambda)} t^2. \quad (105)$$

Expression (105) is strictly convex in  $t$ , so  $J$  is not concave on  $\Theta$ .

The coefficient margin (100) also identifies the optimizer: because  $[\nabla J(\theta)]_k$  has the fixed sign  $\tau_k$  throughout the parameter cube,

$$[\theta^*]_k = \tau_k = \text{sign}(2k - m)$$

is the unique optimum, a majority-rule controller within the chosen density family. Thus gradient domination holds even though the objective is nonconcave and no single policy maximizes improvement at both hidden populations. Extending this conclusion to policies with shared parameters requires separate verification.

*Remark F.2* (Policies using observation summaries). The analysis in Appendix E also applies to a measurable summary  $o = \mathsf{T}(\xi)$  in a standard Borel space  $\mathcal{O}$ , provided every policy factors through it:  $\pi_\theta(\cdot \mid \xi) = \tilde{\pi}_\theta(\cdot \mid \mathsf{T}(\xi))$ . Use the law of  $\mathsf{T}(\xi)$  when  $\xi \sim \mathcal{O}_m(\cdot \mid \mu)$  as the sampling kernel, and define the stationary observation law, posterior, conditional action values, and local action families on the summary space. The same proofs apply after verifying closure, surrogate concavity, coverage, and the comparison of local gaps on the summary space. These conditions can differ from those for the full empirical observation. Retain the original model and regularity assumptions for the composed policy as a function of  $\xi$ , since an arbitrary hard partition need not preserve this regularity. The integrated transition and reward, objective, and learning algorithm are unchanged because the implemented action law is the same.

## G Complete proof of the regret guarantee

The proof of Theorem 7.4 combines gradient tracking, the transient cost of changing policies, and gradient domination. It accounts for the policies used during data collection before substituting the horizon-dependent trajectory lengths.

## G.1 Gradient tracking on one trajectory

*Proof of Lemma 7.1.* We separate the tracking error into accumulated centered noise and conditional bias. Let  $\mathcal{F}_k$  be the history before the gradient-change call, including its endpoints but not its fresh randomization, and let  $\mathcal{H}_k$  be the history after that call and before the fresh gradient call. These analytical histories may include the hidden populations. Each estimate is centered using the history immediately before its own call.

Write  $\delta_k = \theta_k - \theta_{k-1}$  and  $\Delta_k = \nabla J(\theta_k) - \nabla J(\theta_{k-1})$ , and define

$$\begin{aligned} a_k^v &= \mathbb{E}[\hat{v}_k \mid \mathcal{F}_k] - \Delta_k, & \xi_k^v &= \hat{v}_k - \mathbb{E}[\hat{v}_k \mid \mathcal{F}_k], \\ a_k^g &= \mathbb{E}[\hat{g}_k \mid \mathcal{H}_k] - \nabla J(\theta_k), & \xi_k^g &= \hat{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{H}_k]. \end{aligned}$$

The conditional oracle guarantees give  $\|a_k^v\| \leq 2\varepsilon_{B,N}$ ,  $\|a_k^g\| \leq \varepsilon_{B,N}$ ,  $\mathbb{E}[\|\xi_k^v\|^2 \mid \mathcal{F}_k] \leq \mathcal{W}\|\delta_k\|^2$ , and  $\mathbb{E}[\|\xi_k^g\|^2 \mid \mathcal{H}_k] \leq \mathcal{V}$ . The noise bounds use the fact that centering cannot increase the second moment. All bounds hold at the actual starting population of each call, so the two starting populations need not coincide.

Set  $e_k = g_k - \nabla J(\theta_k)$ . Since  $g_{k-1} + \Delta_k = e_{k-1} + \nabla J(\theta_k)$ , the update (53) yields

$$\begin{aligned} ke_k &= (k-1)e_{k-1} + (k-1)(a_k^v + \xi_k^v) + a_k^g + \xi_k^g \\ &= \underbrace{\sum_{i=1}^k [(i-1)\xi_i^v + \xi_i^g]}_{\mathcal{N}_k} + \underbrace{\sum_{i=1}^k [(i-1)a_i^v + a_i^g]}_{\mathcal{B}_k}. \end{aligned}$$

The first equality cancels the change in the true gradient; the second unrolls the recursion, with the initial term multiplied by zero. Thus  $\mathcal{N}_k$  contains the accumulated centered noise and  $\mathcal{B}_k$  the accumulated conditional bias.

For the noise term, chronological conditioning eliminates all cross moments. In particular, within an epoch,  $\mathbb{E}\langle \xi_i^v, \xi_i^g \rangle = \mathbb{E}\langle \xi_i^v, \mathbb{E}[\xi_i^g \mid \mathcal{H}_i] \rangle = 0$ . The same argument applies across epochs: each earlier noise term is known before a later call, whose centered noise has conditional mean zero. For  $i \geq 2$ , the policy update (54) gives  $\|\delta_i\| \leq D_\Theta \eta_{i-1} \leq 2\kappa D_\Theta / (i-1)$ . Therefore,

$$\begin{aligned} \mathbb{E}\|\mathcal{N}_k\|^2 &= \sum_{i=1}^k \left[ (i-1)^2 \mathbb{E}\|\xi_i^v\|^2 + \mathbb{E}\|\xi_i^g\|^2 \right] \\ &\leq \mathcal{W} \sum_{i=2}^k (i-1)^2 \mathbb{E}\|\delta_i\|^2 + k\mathcal{V} \\ &\leq 4(k-1)\kappa^2 D_\Theta^2 \mathcal{W} + k\mathcal{V} \\ &\leq k(\mathcal{V} + 4\kappa^2 D_\Theta^2 \mathcal{W}) = k\sigma^2. \end{aligned}$$

Conditional centering removes the cross moments, while the shrinking policy steps offset the squared correction weights  $(i-1)^2$ . The correction at  $i=1$  is zero. No independence between calls is required.

For the bias term, the triangle inequality gives the pathwise bound

$$\begin{aligned} \|\mathcal{B}_k\| &\leq \sum_{i=1}^k [(i-1)\|a_i^v\| + \|a_i^g\|] \\ &\leq \sum_{i=1}^k [2(i-1)\varepsilon_{B,N} + \varepsilon_{B,N}] = k^2 \varepsilon_{B,N}. \end{aligned}$$

Unlike the centered noise, these biases need not cancel. Combining the two estimates gives

$$\begin{aligned}
\left(\mathbb{E}\|g_k - \nabla J(\theta_k)\|^2\right)^{1/2} &= \frac{1}{k} \left(\mathbb{E}\|\mathcal{N}_k + \mathcal{B}_k\|^2\right)^{1/2} \\
&\leq \frac{1}{k} \left[ \left(\mathbb{E}\|\mathcal{N}_k\|^2\right)^{1/2} + \left(\mathbb{E}\|\mathcal{B}_k\|^2\right)^{1/2} \right] \\
&\leq \frac{\sqrt{k}\sigma + k^2\varepsilon_{B,N}}{k} = \frac{\sigma}{\sqrt{k}} + k\varepsilon_{B,N},
\end{aligned}$$

where the first inequality is the triangle inequality in  $L^2$ . This proves (55).  $\square$

## G.2 Transient cost of policy changes

This comparison controls the difference between actual and stationary rewards along the whole continuing trajectory, including the estimation segments.

*Proof of Lemma 7.3.* Let  $\mathcal{F}_t$  contain the history, current population, and chosen parameter before the batch at time  $t$ . The reward and transition identities, together with the Poisson equation (20) for  $\psi_t$ , show that

$$Z_{t+1} := R_t - J(\psi_t) + h_{\psi_t}^{\text{ref}}(\mu_{t+1}) - h_{\psi_t}^{\text{ref}}(\mu_t)$$

has conditional mean zero given  $\mathcal{F}_t$ .

Rearranging this identity and summing over physical time gives

$$\sum_{t=0}^{T-1} (R_t - J(\psi_t)) = h_{\psi_0}^{\text{ref}}(\mu_0) - h_{\psi_{T-1}}^{\text{ref}}(\mu_T) + \sum_{t=1}^{T-1} \left( h_{\psi_t}^{\text{ref}}(\mu_t) - h_{\psi_{t-1}}^{\text{ref}}(\mu_t) \right) + \sum_{t=0}^{T-1} Z_{t+1}.$$

Within a fixed-policy segment, the intermediate differential value terms cancel exactly. At a policy change, only the difference between the two differential value functions remains.

Taking expectations removes the last sum. The endpoint terms have total absolute value at most  $2H_0$ . By (57), the remaining terms are bounded by  $DC_{\text{PC}} \sum_{t=1}^{T-1} \|\psi_t - \psi_{t-1}\|$ . This proves (61).  $\square$

## G.3 From stationary policy improvement to expected regret

*Proof of Theorem 7.4.* Recall that  $\sigma = \sqrt{\mathcal{V} + 4\kappa^2 D_{\Theta}^2 \mathcal{W}}$ . We first bound the cumulative stationary suboptimality of the iterates, and then compare it with the rewards actually collected. The tracking bound controls error in policy updates, while the adaptive reward comparison controls deviations from stationary rewards.

*Improvement in the stationary objective.* Fix an epoch  $k$ , and let  $w_k^* \in \arg \max_{w \in \Theta} \langle \nabla J(\theta_k), w \rangle$ . Since  $w_k$  maximizes the linear objective defined by  $g_k$ ,  $\langle g_k, w_k - w_k^* \rangle \geq 0$ . Consequently,

$$\begin{aligned}
\langle \nabla J(\theta_k), w_k - \theta_k \rangle &= \mathcal{G}(\theta_k) + \langle \nabla J(\theta_k), w_k - w_k^* \rangle \\
&\geq \mathcal{G}(\theta_k) + \langle \nabla J(\theta_k) - g_k, w_k - w_k^* \rangle \\
&\geq \mathcal{G}(\theta_k) - D_{\Theta} \|g_k - \nabla J(\theta_k)\|.
\end{aligned}$$

Thus using the estimated gradient loses at most  $D_{\Theta} \|g_k - \nabla J(\theta_k)\|$  in the linear gain.

By Lemma 7.2,  $J$  has an  $L_J$ -Lipschitz gradient. Using the update rule and the preceding bound on the selected direction, we obtain

$$\begin{aligned} J(\theta_{k+1}) &\geq J(\theta_k) + \eta_k \langle \nabla J(\theta_k), w_k - \theta_k \rangle - \frac{L_J}{2} \eta_k^2 \|w_k - \theta_k\|^2 \\ &\geq J(\theta_k) + \eta_k \mathcal{G}(\theta_k) - \eta_k D_\Theta \|g_k - \nabla J(\theta_k)\| - \frac{L_J D_\Theta^2}{2} \eta_k^2, \end{aligned}$$

where we used  $\|w_k - \theta_k\| \leq D_\Theta$ . Let  $a_k = \mathbb{E}[J^* - J(\theta_k)]$ . Subtracting the preceding improvement bound from  $J^*$  and taking expectations gives

$$a_{k+1} \leq a_k - \eta_k \mathbb{E}[\mathcal{G}(\theta_k)] + \eta_k D_\Theta \mathbb{E}\|g_k - \nabla J(\theta_k)\| + \frac{L_J D_\Theta^2}{2} \eta_k^2.$$

Gradient domination implies  $\mathbb{E}[\mathcal{G}(\theta_k)] \geq \frac{a_k - \epsilon_{\text{geom}}}{\kappa}$ . Moreover, Cauchy–Schwarz and Lemma 7.1 give

$$\mathbb{E}\|g_k - \nabla J(\theta_k)\| \leq (\mathbb{E}\|g_k - \nabla J(\theta_k)\|^2)^{1/2} \leq \frac{\sigma}{\sqrt{k}} + k\varepsilon_{B,N}.$$

Substituting these two bounds yields

$$a_{k+1} \leq \left(1 - \frac{\eta_k}{\kappa}\right) a_k + \frac{\eta_k}{\kappa} \epsilon_{\text{geom}} + \eta_k D_\Theta \left(\frac{\sigma}{\sqrt{k}} + k\varepsilon_{B,N}\right) + \frac{L_J D_\Theta^2}{2} \eta_k^2. \quad (106)$$

To bound this recurrence, isolate the gap above the approximation level  $\epsilon_{\text{geom}}$  by setting  $z_k = (a_k - \epsilon_{\text{geom}})_+$ , where  $x_+ = \max\{x, 0\}$ . Define  $k_0 = \lceil 2\kappa \rceil \geq 2$  and  $A = \sqrt{k_0} + 2\kappa D_\Theta \sigma + 2\kappa^2 L_J D_\Theta^2$ . For  $k \geq k_0$ , we have  $\eta_k = 2\kappa/k$  and  $1 - \eta_k/\kappa = 1 - 2/k \geq 0$ . Hence (106) implies

$$z_{k+1} \leq \left(1 - \frac{2}{k}\right) z_k + \frac{2\kappa D_\Theta \sigma}{k^{3/2}} + 2\kappa D_\Theta \varepsilon_{B,N} + \frac{2\kappa^2 L_J D_\Theta^2}{k^2}.$$

We claim that

$$z_k \leq \frac{A}{\sqrt{k}} + 2\kappa D_\Theta k\varepsilon_{B,N}. \quad (107)$$

For  $k \leq k_0$ , this follows from  $0 \leq J \leq 1$  and  $A \geq \sqrt{k_0}$ . If it holds at  $k \geq k_0$ , then  $A \geq 2\kappa D_\Theta \sigma + 2\kappa^2 L_J D_\Theta^2$  gives

$$z_{k+1} \leq \frac{A}{\sqrt{k}} - \frac{A}{k^{3/2}} + 2\kappa D_\Theta (k-1)\varepsilon_{B,N}.$$

Using  $\frac{1}{\sqrt{k}} - \frac{1}{k^{3/2}} \leq \frac{1}{\sqrt{k+1}}$  proves the claim at  $k+1$ .

Since  $a_k \leq \epsilon_{\text{geom}} + z_k$ , summing (107) gives

$$\sum_{k=1}^K a_k \leq K \epsilon_{\text{geom}} + 2A\sqrt{K} + \kappa D_\Theta \varepsilon_{B,N} K(K+1). \quad (108)$$

*The policies used to collect data.* The gradient-change call in epoch  $k$  uses  $\tilde{\theta}_k = \theta_{k-1} + u_k(\theta_k - \theta_{k-1})$ , with  $u_k \in [0, 1]$ , whereas the gradient call and padding use  $\theta_k$ . Let  $\delta_k = \theta_k - \theta_{k-1}$ . By (58),

$$J^* - J(\tilde{\theta}_k) \leq J^* - J(\theta_k) + G(1 + H_0)\|\delta_k\|.$$

The gradient-change call lasts at most  $n$  steps, and the whole epoch lasts  $2n$  steps. The epoch's cumulative stationary shortfall is therefore at most  $2n(J^* - J(\theta_k)) + nG(1 + H_0)\|\delta_k\|$ . There are fewer than  $2n$  remaining steps, whose total stationary shortfall is at most  $2n$ .

The same parameter differences also control the transient cost. Between epochs  $k - 1$  and  $k$ , the executed parameters move from  $\theta_{k-1}$  to  $\tilde{\theta}_k$ , and then to  $\theta_k$ . Both moves follow the same line segment, so their lengths add to  $\|\delta_k\|$ . There is no parameter change between the gradient call and padding, or during the final remaining steps. Since  $\delta_1 = 0$ ,

$$\text{TV}_T \leq \sum_{k=2}^K \|\delta_k\| \leq D_\Theta \sum_{k=1}^{K-1} \eta_k \leq 2\kappa D_\Theta (1 + \log K).$$

Let  $\psi_t$  denote the parameter executed at physical time  $t$ . Decompose regret as

$$\mathbb{E} \text{Reg}_T = \mathbb{E} \sum_{t=0}^{T-1} (J^* - J(\psi_t)) + \mathbb{E} \sum_{t=0}^{T-1} (J(\psi_t) - R_t).$$

The stationary shortfall bounds above control the first sum. Lemma 7.3 controls the second. For  $K \geq 1$ , this gives

$$\mathbb{E} \text{Reg}_T \leq 2n \sum_{k=1}^K a_k + 2\kappa D_\Theta (nG(1 + H_0) + DC_P C_g)(1 + \log K) + 2n + 2H_0.$$

Substituting (108) and using  $2nK \leq T$ , we obtain the explicit bound

$$\begin{aligned} \mathbb{E} \text{Reg}_T &\leq T\epsilon_{\text{geom}} + 4nA\sqrt{K} + 2n\kappa D_\Theta \varepsilon_{B,N} K(K+1) \\ &\quad + 2\kappa D_\Theta [nG(1 + H_0) + DC_P C_g](1 + \log K) + 2n + 2H_0. \end{aligned} \tag{109}$$

*Substitution of the trajectory lengths.* Because  $\alpha \leq 1/2$ , the choice (63) ensures

$$\varepsilon_{B,N} \leq C_\varepsilon 2^{-b_T} \leq T^{-3}.$$

With fixed structural constants,  $b_T = O(\log(2T))$  and  $\mathcal{U}_{B,N} = O(b_T)$ , so

$$n = O(b_T), \quad \mathcal{V} = O(b_T), \quad \mathcal{W} = O(b_T^3), \quad \sigma = O(b_T^{3/2}), \quad A = O(b_T^{3/2}).$$

Using  $K \leq T/(2n)$ , the leading term in (109) satisfies

$$nA\sqrt{K} \leq A\sqrt{\frac{nT}{2}} = O(\sqrt{T} b_T^2) = O(\sqrt{T} \log^2(2T)).$$

The accumulated oracle-bias term is bounded by

$$n\varepsilon_{B,N} K(K+1) \leq \varepsilon_{B,N} \left( \frac{T^2}{4n} + \frac{T}{2} \right),$$

which is negligible under  $\varepsilon_{B,N} \leq T^{-3}$ . The parameter-variation term is  $O(\log^2(2T))$ , and the remaining endpoint and leftover terms are  $O(\log(2T))$ . This proves (64) when  $K \geq 1$ .

If  $K = 0$ , bounded rewards give  $\mathbb{E} \text{Reg}_T \leq T < 2n$ . Since  $n = O(\log(2T))$ , the stated bound also covers this case.

Finally, each of the  $T$  physical steps uses  $m$  action-selection samples. If it also uses one independent evaluation sample, the total individual-sample cost is  $(m+1)T$ .  $\square$

## H Experimental details

### H.1 Common protocol and evaluation

**Observations.** Each run follows one continuing population trajectory without resets, using one population observation ( $m = 1$ ) and an independent evaluation sample for the reward at each interaction. Each observation is a vector, with dimension  $d$  in E3. The policies in E1 and E2 ignore the observations, whereas the action distribution in E3 depends on both  $\theta$  and  $\hat{\mu}_t$ . Latent states and exact stationary objectives are used only for evaluation.

**Implementation.** Our method follows Algorithm 3, with equal burn-in and return block counts  $B = N = b$  and independent block lengths uniformly distributed on  $\{1, \dots, s\}$ . We keep  $b$  and  $s$  fixed within each run. Throughout this appendix,  $g_k$  denotes the tracked gradient. We replace the averaging weight  $1/k$  by  $\alpha_k$  and use the step sizes  $\eta_k$  specified for each experiment. Initially,  $g_0 = 0$ ,  $\theta_0 = \theta_1$ , and  $\alpha_1 = 1$ . For the box linear oracle, a zero tracker coordinate leaves the corresponding parameter coordinate unchanged.

Each epoch collects a gradient-change record followed by a fresh-gradient record, then uses the current epoch policy to pad its total duration to  $4bs + 2$  interactions. All interactions, including estimation, padding, and the final incomplete-epoch remainder, count towards  $T$ . The remainder uses the last epoch parameter; the final policy gap is evaluated at the latest updated output. The GPOMDP baselines in E2 instead update after every interaction.

**Evaluation.** Let  $X_t$  be the latent population state,  $\psi_t$  the parameter actually deployed at time  $t$ , and  $\bar{r}_\theta(x) = \mathbb{E}[R_t \mid X_t = x, \psi_t = \theta]$ . For  $J^* = \max_{\theta \in \Theta} J(\theta)$ , we report the final policy gap  $\Delta(\theta_{\text{out}}) = J^* - J(\theta_{\text{out}})$  and conditional-mean regret, computed as

$$TJ^* - \sum_{t=0}^{T-1} \bar{r}_{\psi_t}(X_t). \quad (110)$$

The conditional reward averages over observation, action, and evaluation randomness. This regret has the same expectation as observed-reward regret, while the learner always uses sampled rewards. Results are averaged over independent seeds. Standard errors and bootstrap intervals are identified in the tables. Bootstrap intervals resample entire seeded trajectories, preserving dependence across time and, for paired comparisons, across methods.

### H.2 E1: Randomized durations under periodic dynamics

This experiment isolates the effect of randomizing estimation durations. We compare our algorithm with a version that fixes both durations at their respective means. Both variants use the same estimators, gradient tracking, policy updates, and interaction budget.

**Population and policy.** The hidden population is indexed by  $x_t \in [-1, 1]$  and a phase  $e_t \in \{-1, +1\}$ , with dynamics

$$x_{t+1} = \rho x_t + (1 - \rho)A_t, \quad e_{t+1} = -e_t, \quad \rho = \frac{1}{5}. \quad (111)$$

A sample from  $\mu_{x,e}$  is a pair  $(Z, Q) \in \{-1, +1\}^2$  with conditionally independent coordinates satisfying

$$\Pr(Z = 1 \mid x, e) = \frac{1+x}{2}, \quad \Pr(Q = 1 \mid x, e) = \frac{1}{2} + \frac{e}{4}. \quad (112)$$

The policy ignores the observation and samples

$$\pi_\theta(a) = \frac{1+\theta a}{2}, \quad a \in \{-1, +1\}, \quad \theta \in [-1/2, 1/2]. \quad (113)$$

Its score is  $S_\theta(a) = a/(1+\theta a)$ , with  $\partial_\theta S_\theta(a) = -S_\theta(a)^2$ . Before the transition, an independent evaluation pair produces

$$\begin{aligned} R_t &= \frac{1}{2} + Z_t^{\text{ev}} \left( \frac{1}{10} + \frac{2}{5} Q_t^{\text{ev}} \right) \in [0, 1], \\ r(x, e) &= \mathbb{E}[R_t \mid x_t = x, e_t = e] = \frac{1}{2} + \frac{1}{10}x + \frac{1}{5}ex. \end{aligned} \quad (114)$$

The interaction term  $ex$  in (114) makes an action's future reward contribution depend on the phase.

**Stationary objective and assumptions.** Under the fixed policy (113), the dynamics (111) have an invariant law equal to the product of the invariant  $x$ -law, whose mean is  $\theta$ , and a uniform phase. Hence

$$J(\theta) = \frac{1}{2} + \frac{1}{10}\theta, \quad J'(\theta) = \frac{1}{10}. \quad (115)$$

Equation (115) gives  $\theta^* = 1/2$  and  $J^* = 0.55$ . The Frank–Wolfe gap equals the objective gap, so gradient domination holds with  $\kappa = 1$  and zero residual.

The alternating phase prevents fixed-time population laws from converging, but the model satisfies the uniform Poisson inverse condition. Under the two-coordinate Hamming cost, the observation law (112) gives the population metric

$$d_{\mathcal{D}}((x, e), (x', e')) = \frac{|x - x'|}{2} + \frac{\mathbf{1}\{e \neq e'\}}{2}.$$

Let  $K_\theta$  be the  $x$ -transition kernel. Decomposing a function into its symmetric and antisymmetric phase components reduces the Poisson equation to operators  $I - K_\theta$  and  $I + K_\theta$ , respectively. Since  $K_\theta$  contracts Lipschitz seminorms by  $1/5$ , the centered components are controlled by convergent geometric series. On constants,  $I + K_\theta$  acts as multiplication by two. These bounds are uniform in  $\theta$ .

**Matched comparison.** Both variants follow Algorithm 3 with  $B = N = b = 2$  and  $s = 2$ . For the randomized variant, each of  $U$  and  $V$  is the sum of two independent  $\text{Unif}\{1, 2\}$  block lengths, with all four draws independent. The fixed-duration variant sets  $U = V = 3$ , matching the mean burn-in and return durations of the randomized variant.

Each epoch collects a gradient-change record followed by a fresh-gradient record and is padded to eighteen interactions. We initialize  $x_0 = 0$ ,  $\theta_0 = \theta_1 = 0$ , and  $g_0 = 0$ , test both  $e_0 = -1$  and  $e_0 = +1$ , and use  $\alpha_k = 1/k$  and  $\eta_k = \min\{1, 2/k\}$ . Since eighteen is even, every epoch begins in phase  $e_0$ . Each fixed-duration record has seven interactions, so its second, fresh-gradient record begins in phase  $-e_0$ .

**Why randomization removes the phase bias.** Consider a record starting from  $(x, e)$  under a fixed  $\theta$ . The burn-in baseline and first return reward  $R_U$  have zero expected product with the score  $S_U$ . For  $j \geq 1$ ,

$$\mathbb{E}[S_U R_{U+j} \mid x, e, \theta, U, V] = (1 - \rho)\rho^{j-1}[0.1 + 0.2e(-1)^{U+j}].$$

Summing over the return gives

$$\mathbb{E}[\hat{g} \mid x, e, \theta, U, V] = 0.1(1 - 0.2^V) - \frac{2}{15}e(-1)^U[1 - (-0.2)^V]. \quad (116)$$

The second term in (116) depends on the phase at time  $U$ ,  $e(-1)^U$ . For the randomized durations,  $\mathbb{E}[(-1)^U] = 0$  and  $\mathbb{E}[0.2^V] = 0.12^2$ . Independence of  $U$  and  $V$  therefore eliminates the phase term.

The fresh-gradient record used in each update therefore has mean  $0.0992 - 0.1344e_0$  with fixed durations and  $0.09856$  with randomized durations. The fixed estimator has negative mean when  $e_0 = +1$ , despite the positive true gradient  $J'(\theta) = 0.1$ . The randomized estimator has positive mean for either initial phase. The expected gradient-change correction is zero here because the record mean is independent of  $\theta$ ; both variants nevertheless retain the sampled correction.

**Results.** Each run uses  $T = 2^{18}$  interactions. Table 1 averages sixteen seeds, numbered 72000–72015, for each method and initial phase. We report conditional-mean regret as defined in (110) and the final policy gap. The gap intervals are 95% percentile bootstrap intervals for the mean, computed from 4,000 seed resamples.

Table 1: E1 with  $b = s = 2$  and  $T = 2^{18}$ . Each row averages sixteen seeds; intervals are 95% percentile bootstrap intervals for the mean final policy gap.

| $e_0$ | Durations | Conditional regret | Final gap              | Gap interval                                 |
|-------|-----------|--------------------|------------------------|----------------------------------------------|
| −1    | Fixed     | 141.83             | $2.777 \times 10^{-6}$ | $[2.64 \times 10^{-7}, 6.5 \times 10^{-6}]$  |
| −1    | Random    | 437.67             | $4.552 \times 10^{-5}$ | $[2.6 \times 10^{-6}, 1.08 \times 10^{-4}]$  |
| +1    | Fixed     | 21310.65           | 0.08856                | $[0.0721, 0.0997]$                           |
| +1    | Random    | 759.19             | $1.181 \times 10^{-4}$ | $[5.42 \times 10^{-6}, 2.84 \times 10^{-4}]$ |

The fixed schedule performs better in the favorable phase  $e_0 = -1$ , but incurs a large final gap when  $e_0 = +1$ . Randomization produces small final gaps from both phases, illustrating robustness to the initial phase under the matched algorithm and interaction budget.

### H.3 E2: Finite-window estimation versus GPOMDP

This experiment compares our method with two GPOMDP implementations in a model where an action incurs an immediate cost and a slightly larger benefit two interactions later. A finite return window captures the entire effect, whereas a discounted eligibility trace requires a discount factor close to one to preserve the gradient’s direction.

**Model and policy.** The latent state is  $x_t \in [-a, a]^2$ , with  $a = 0.8$  and  $x_0 = (0, 0)$ . An observation  $Z_t \in \{-1, +1\}^2$  has conditionally independent coordinates satisfying

$$\Pr(Z_{t,i} = +1 \mid x_t) = \frac{1 + x_{t,i}}{2}, \quad i = 1, 2. \quad (117)$$

The policy ignores the observation and samples a complete action  $(\sigma_t, I_t)$  according to

$$\Pr_{\theta}(\sigma_t = +1) = \frac{1+\theta}{2}, \quad I_t \sim \text{Unif}[1/2, 1], \quad \theta \in [-0.8, 0.8]. \quad (118)$$

The intensity  $I_t$  is independent of  $\sigma_t$  and has mean  $\mu = 3/4$ . All methods start at  $\theta_0 = -0.4$ . The transition is

$$x_{t+1,1} = a\sigma_t I_t, \quad x_{t+1,2} = x_{t,1}. \quad (119)$$

The transition (119) is deterministic given the complete action. Since the intensity distribution in (118) is parameter independent, the score is  $S_{\theta}(\sigma) = \sigma/(1 + \sigma\theta)$ , with derivative  $\partial_{\theta}S_{\theta}(\sigma) = -S_{\theta}(\sigma)^2$ .

Using an independent evaluation observation from (117), with  $\mathbb{E}[Z_{t,2}^{\text{ev}} | x_t] = x_{t,2}$ , the reward is

$$R_t = \frac{1}{2} + \frac{vZ_{t,2}^{\text{ev}}/(a\mu) - c\sigma_t}{C}, \quad v = 0.2, \quad c = 0.198, \quad (120)$$

where  $C = 2(v/(a\mu) + c) = 797/750$  ensures  $R_t \in [0, 1]$ . The reward (120) uses the pre-transition state, so the current action's cost appears in  $R_t$  and its benefit in  $R_{t+2}$ . For conditional-mean regret, the evaluator uses

$$\bar{r}_{\theta}(x) = \frac{1}{2} + \frac{vx_2/(a\mu) - c\theta}{C}. \quad (121)$$

**Objective and finite-window estimate.** Under a fixed policy, the state forgets its initial condition after two interactions and has stationary coordinate means  $a\mu\theta$ . Consequently,

$$J(\theta) = \frac{1}{2} + g\theta, \quad g = \frac{v-c}{C}, \quad (122)$$

$$\theta^* = 0.8, \quad \Delta(\theta) = g(0.8 - \theta). \quad (123)$$

Since the objective (122) is affine, the Frank–Wolfe gap equals the objective gap (123), giving gradient domination with constant one and zero residual. Coupling the same complete actions is nonexpansive in normalized  $\ell_1$  distance and coalesces after two steps. For a centered function  $f$ , meaning  $\nu_{\theta}f = 0$ , we therefore have  $P_{\theta}^2 f = 0$ . The Poisson solution is  $f + P_{\theta}f$ , giving a uniform inverse bound of two.

For the centered record in Algorithm 1, the score  $S_U$  captures the immediate cost and, when  $V \geq 2$ , the delayed benefit. Thus

$$\mathbb{E}[\hat{g} | x_{\text{start}}, \theta] = \frac{v \Pr(V \geq 2) - c}{C}. \quad (124)$$

Our selected configuration uses  $(b, s) = (2, 1)$ , so  $U = V = 2$ . Equation (124) therefore gives an exactly unbiased gradient estimate. The sampled gradient-change correction is retained, although its mean is zero because  $g$  is constant. These deterministic durations test finite-window estimation; E1 separately examines the role of randomization.

**GPOMDP estimator and updates.** We use the observable-score estimator with the control-dependent-reward timing of [Baxter and Bartlett \(2001\)](#). Let  $e_t$  denote the eligibility trace stored before interaction  $t$ . With  $S_t = S_{\theta_t}(\sigma_t)$ , the estimate and trace update are

$$h_t = (R_t - 0.5)(e_t + S_t), \quad e_{t+1} = \beta e_t + S_t, \quad e_0 = 0. \quad (125)$$

The current score in (125) accounts for the immediate action cost. The trace persists across policy updates, and each score is evaluated at the parameter that generated its action.

Under the stationary law of a fixed policy, the reward multiplier is  $S_t + S_{t-1} + \beta S_{t-2} + \dots$ . The immediate cost is therefore weighted by one and the delayed benefit by  $\beta$ , giving

$$\mathbb{E}[h_t] = \frac{v\beta - c}{C}. \quad (126)$$

A positive mean estimate requires  $\beta > 0.99$ . The small net benefit  $v - c = 0.002$  means that even slight discounting can reverse the gradient’s sign. For example, the selected projected GPOMDP uses  $\beta = 0.995$ , for which (126) gives  $\mathbb{E}[h_t] = g/2$  and trace scale  $1/(1 - \beta) = 200$ , despite the two-step causal delay. Because the trace weights decay geometrically, choosing  $\beta$  this close to one also retains scores from much earlier actions. The stationary fixed-policy trace satisfies

$$\mathbb{E}[(e_t + S_t)^2] = \frac{1 + (1 - \beta^2)^{-1}}{1 - \theta^2}. \quad (127)$$

Equation (127) shows how retaining the delayed signal increases the trace’s second moment.

Projected GPOMDP updates after every interaction according to

$$\theta_{t+1} = \Pi_{\Theta}(\theta_t + \eta_{t+1}h_t).$$

GPOMDP-EMA/FW instead averages the estimates,

$$d_{t+1} = (1 - \alpha_{t+1})d_t + \alpha_{t+1}h_t,$$

and applies the box Frank–Wolfe update using  $d_{t+1}$  and step size  $\eta_{t+1}$ . Both GPOMDP implementations use reward baseline 0.5. Our method uses the sampled burn-in baseline and the two-record tracking update in Algorithm 3, with the averaging weights specified below.

**Configurations and tuning.** Table 2 gives the selected configurations. Each method was tuned separately by mean conditional regret.

Table 2: Selected E2 configurations. For our method,  $k$  counts two-record epochs; for GPOMDP, it counts interactions. The first averaging weight is one.

| Method                        | $\beta$ | $\eta_k$              | $\alpha_k$       |
|-------------------------------|---------|-----------------------|------------------|
| Our method, $(b, s) = (2, 1)$ | —       | $\min\{0.25, 0.5/k\}$ | $\min\{1, 2/k\}$ |
| GPOMDP projected              | 0.995   | $64/k$                | —                |
| GPOMDP EMA/FW                 | 0.99999 | $\min\{0.25, 8/k\}$   | $1/k$            |

For our method, we searched  $(b, s) \in \{(2, 1), (2, 2)\}$ . The step sizes were  $\eta_k = \min\{0.25, c_\eta/k\}$  with  $c_\eta \in \{0.125, 0.5, 2, 8, 32\}$ , and the averaging weights were  $\alpha_k = \min\{1, c_\alpha/k^q\}$  with

$$(c_\alpha, q) \in \{(1, 1), (2, 1), (4, 1), (8, 1), (16, 1), (64, 1), (256, 1), (0.25, 2/3)\}.$$

Writing  $\varepsilon = (v - c)/v$ , both GPOMDP implementations searched

$$\beta \in \{0, 1 - 2\varepsilon, 1 - \varepsilon, 1 - \varepsilon/2, 1 - \varepsilon/3, 1 - \varepsilon/5, 0.9999, 0.99999\}.$$

Projected GPOMDP searched  $\eta_k = \min\{1024, c_\eta/k^q\}$  with  $c_\eta \in \{1, 16, 64, 256, 1024\}$  and  $q \in \{1/2, 3/4, 1\}$ , together with constant steps 0.001, 0.01, and 0.1. GPOMDP-EMA/FW used the same five Frank-Wolfe coefficients and eight averaging schedules as our method.

Screening used  $T = 2^{20}$  and seeds 93000–93003. For each method and cost coefficient  $c$ , the four candidates with lowest mean regret advanced, with ties resolved by configuration identifier. The finalist pool additionally included four settings for our method with  $(b, s) = (2, 1)$ ,  $q = 1$ , and  $(c_\eta, c_\alpha) \in \{(0.5, 4), (2, 16), (0.5, 2), (2, 8)\}$ , and one GPOMDP-EMA/FW setting with  $\beta = 1 - \varepsilon/3$ ,  $c_\eta = 32$ , and  $\alpha_k = 1/k$ . Finalists were evaluated at  $T = 2^{24}$  on seeds 93000–93007, and the lowest mean regret determined each method’s configuration. The exploratory search considered  $c \in \{0.198, 0.199\}$ , with  $c = 0.19$  control runs; the reported model uses  $c = 0.198$ . The selected EMA/FW discount is the largest value in its search grid.

**Evaluation and results.** The selected configurations were evaluated on 32 new seeds, 94100–94131, each with  $T = 2^{24}$  interactions. Conditional-mean regret uses the reward mean in (121). Our method uses ten interactions per update, including its correction record, whereas both GPOMDP implementations update every interaction. All interactions count towards the same budget. Terminal interactions follow the common protocol. Methods use the same environmental seed construction but follow their own sampled actions and state trajectories.

Table 3: E2 at  $T = 2^{24}$  over 32 evaluation seeds. Entries are mean  $\pm$  standard error; lower is better. The final-gap column is multiplied by  $10^3$ .

| Method           | Conditional regret     | $10^3$ final gap        |
|------------------|------------------------|-------------------------|
| Our method       | $13038.79 \pm 1453.56$ | $0.581253 \pm 0.074088$ |
| GPOMDP projected | $19287.17 \pm 2157.20$ | $1.047338 \pm 0.127388$ |
| GPOMDP EMA/FW    | $24958.49 \pm 2669.04$ | $1.340015 \pm 0.227878$ |

Standard errors are sample standard deviations divided by  $\sqrt{32}$ . Our method achieves lower mean conditional regret and final policy gap than both tuned GPOMDP implementations at the same interaction budget. This comparison evaluates the complete methods, including their centering and update rules, on the specified delayed-benefit instance.

#### H.4 E3: Scaling population and policy dimensions

We evaluate the complete algorithm with  $d \in \{10, 50, 100\}$  latent population coordinates and  $d$  policy parameters. The policy uses noisy observations, and every run retains the sampled gradient-change correction. This subsection details the  $b = s = 2$  configurations shown in Figure 2.

**Model and policy.** Let  $P$  denote the cyclic shift matrix and set

$$W = 0.8I + 0.1P + 0.1P^\top, \quad x_{t+1} = \frac{1}{2}Wx_t + \frac{1}{2}A_t, \quad (128)$$

where  $x_t \in [-1, 1]^d$ ,  $A_t \in \{-1, 1\}^d$ , and  $x_0 = 0$ . In (128), each coordinate interacts with its two ring neighbors. The population  $\mu_x$  is the law of  $x + \xi$ , where the coordinates  $\xi_i \sim \text{Unif}[-1/2, 1/2]$  are independent. At each interaction, the learner receives one vector observation  $O_t = x_t + \xi_t$ . An

independent evaluation vector  $O_t^{\text{ev}} = x_t + \xi_t^{\text{ev}}$  generates the reward

$$R_t = 1 - \frac{\|O_t^{\text{ev}} - z\|_2^2}{2.56d}, \quad (129)$$

$$r(x) = \mathbb{E}[R_t \mid x_t = x] = 1 - \frac{\|x - z\|_2^2 + d/12}{2.56d}. \quad (130)$$

The target  $z$  is drawn once, with independent coordinates uniform on  $[-0.1, 0.1]$ , using NumPy’s `default_rng` with seed 20261006. Smaller dimensions use prefixes of the same 100-dimensional draw, and the target remains fixed during tuning and evaluation. Since  $|O_{t,i}^{\text{ev}} - z_i| \leq 1.6$  for every coordinate, the reward (129) lies in  $[0, 1]$ . Only the scalar reward is supplied to the learner.

For  $\Theta = [-1/2, 1/2]^d$ , the policy is

$$\pi_\theta(a \mid o) = \prod_{i=1}^d \frac{1 + a_i(\theta_i + o_i/4)}{2}, \quad \theta_0 = 0.41. \quad (131)$$

Thus each action coordinate depends on both its policy parameter and its observed coordinate. The coordinate probabilities lie in  $[1/16, 15/16]$ . The score and log-policy Hessian used by the algorithm are

$$S_i = \frac{a_i}{1 + a_i(\theta_i + o_i/4)}, \quad (132)$$

$$\nabla_\theta^2 \log \pi_\theta(a \mid o) = -\text{Diag}(S_i^2)_{i=1}^d. \quad (133)$$

The score (132) and Hessian (133) are computable from the observed action and observation. The sampled correction therefore requires neither latent states nor exact objective derivatives, and its Hessian–vector product can be computed coordinatewise.

**Exact objective and assumptions.** Averaging over observation noise gives  $\mathbb{E}[A_t \mid x_t] = \theta + x_t/4$ . Define

$$M = \frac{1}{2}W + \frac{1}{8}I, \quad K = \frac{1}{2}(I - M)^{-1}, \quad D = I + \frac{1}{4}K, \quad (134)$$

$$q = \frac{1}{d} \text{tr}((I - M^2)^{-1}), \quad w = \frac{q/4}{1 + q/64}, \quad Q = K^\top K - wD^\top D. \quad (135)$$

The matrices in (134) give stationary state and action means  $K\theta$  and  $D\theta$ , respectively. For the stationary state covariance  $\Sigma_\theta$ , the moment equations give

$$\Sigma_\theta = M\Sigma_\theta M^\top + \frac{1}{4} \text{Diag} \left( 1 - (D\theta)_i^2 - \frac{(\Sigma_\theta)_{ii}}{16} \right)_{i=1}^d.$$

Since  $M$  is symmetric circulant, every diagonal entry of  $(I - M^2)^{-1}$  equals  $q$ . Taking the trace of the covariance series and rearranging yields

$$\text{tr} \Sigma_\theta = w(d - \|D\theta\|_2^2).$$

Substitution into the mean reward gives the exact objective

$$J(\theta) = 1 - \frac{wd + \|z\|_2^2 + d/12 + \theta^\top Q\theta - 2\theta^\top Kz}{2.56d}. \quad (136)$$

For each tested dimension, the matrix  $Q$  in (135) is positive definite and the unconstrained maximizer  $\theta^* = Q^{-1}Kz$  lies inside  $\Theta$ . Hence

$$\Delta(\theta) = \frac{(\theta - \theta^*)^\top Q(\theta - \theta^*)}{2.56d}. \quad (137)$$

The quadratic objective (136) is concave, giving gradient domination with constant one and zero residual. These exact quantities are used only for evaluation, with the benchmark restricted to the policy class in (131).

We use the normalized  $\ell_1$  metric  $\rho(x, y) = d^{-1}\|x - y\|_1$  on individual outcomes and actions. For the translated population laws,  $W_1(\mu_x, \mu_y) = \rho(x, y)$ : coupling identical noise gives the upper bound, and the difference between the means gives the matching lower bound. Coupling the observation noise and action uniforms gives expected action distance at most  $\rho(x, y)/4$ . The population transition is therefore a  $5/8$  contraction, yielding the uniform Poisson inverse bound  $C_P \leq \sum_{k \geq 0} (5/8)^k = 8/3$ .

**Configurations and tuning.** For the configurations reported here,  $B = N = b = 2$  and  $s = 2$ , so each padded epoch costs eighteen interactions. We use the complete algorithm with schedules  $\alpha_k = \min\{1, c_\alpha/k\}$  and  $\eta_k = \min\{1/4, c_\eta/k\}$ . For each dimension, we tested all combinations  $c_\eta \in \{0.125, 0.25, 0.5, 1\}$  and  $c_\alpha \in \{4, 16, 64, 256\}$  at  $T_{\text{pilot}} = 2^{18}$  using seeds 121000–121003. We selected the pair with the lowest mean conditional regret, breaking ties by configuration identifier. The selected pairs appear in Table 4.

Evaluation uses sixteen new seeds, 121100–121115. After examining results at  $T_0 = 2^{22}$ , we extended these same seeded runs to  $T = 2^{24}$  without changing the schedules. The extension horizon and fitting interval were fixed before running the extension. All runs follow one continuing trajectory without resets.

**Evaluation and slope estimation.** Conditional-mean regret uses the reward mean in (130). Policy gaps are computed from (137); the main figure expresses them as percentages of  $\Delta(\theta_0)$ . All interactions are counted, including estimation, padding, and the final remainder, as specified in the common protocol.

We fit an ordinary least-squares slope of the logarithm of the seed-mean cumulative regret against  $\log t$  over  $[T/2, T]$ . The cumulative regret is measured from the start of the run, without resetting at  $T/2$ . For this fit, checkpoint epochs include 128 geometrically spaced indices from 1 to  $\lfloor T/18 \rfloor$ , rounded up, together with 32 evenly spaced time targets in  $[T/2, T]$  and 16 in  $[3T/4, T]$ , rounded down to complete epochs. Duplicate checkpoints are removed, and the terminal epoch is represented by the exact horizon  $T$ .

All reported 95% intervals use 20,000 percentile bootstrap resamples of entire seeded trajectories. For each slope resample, we recompute the mean regret curve and repeat the fit. The shaded bands in Figure 2 are pointwise intervals for the mean curves.

**Results.** Table 4 reports the endpoint results for the three configurations shown in the main figure. For  $d = 10, 50, 100$ , the mean final policy gap is respectively 99.968%, 99.850%, and 99.596% below its initial value.

Table 4: E3 with  $b = s = 2$  and  $T = 2^{24}$ . Entries are means over sixteen seeds with 95% bootstrap intervals. The last column reports  $10^4 \Delta(\theta_{\text{out}})$ .

| $d$ | $c_\eta$ | $c_\alpha$ | Conditional-mean regret       | Final gap $\times 10^4$ |
|-----|----------|------------|-------------------------------|-------------------------|
| 10  | 0.25     | 4          | 1315.76 [1184.27, 1464.20]    | 0.2263 [0.1894, 0.2607] |
| 50  | 0.5      | 16         | 4654.55 [4509.62, 4785.83]    | 1.0984 [0.9973, 1.2140] |
| 100 | 0.5      | 64         | 10967.44 [10707.71, 11189.58] | 3.0670 [2.9337, 3.1996] |

For  $d = 100$ , the log-log slope of mean conditional regret over  $[T/2, T]$  is 0.499, with 95% bootstrap interval  $[0.486, 0.513]$ . This is consistent with approximately square-root growth over the measured interval. These experiments use fixed block counts and pilot-tuned schedules; the fitted slope is a finite-horizon diagnostic rather than an asymptotic guarantee.

The results demonstrate effective learning with up to one hundred latent population coordinates and one hundred policy parameters. Online transitions, policy sampling, and the sampled Hessian-vector computation require  $O(d)$  arithmetic per interaction.