Online Performative Decision Making with Latent Distribution States

Changsheng Chen and Grani A. Hanasusanto
Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. {cc81, gah}@illinois.edu

Many operational decisions reshape the populations they act on: routing policies alter traffic, care interventions affect health outcomes, and public programs change participation. We study online control of such decision-dependent populations when the primitive state is a distribution, actions determine both current reward and the next distribution, and the reward and transition laws are unknown. Existing performative-prediction models largely ask whether repeated deployment converges to a stable decision-distribution pair. In many operational settings, however, the goal is sequential control of one evolving population, which cannot be reset for offline exploration. Exploration is therefore risky: an informative action can move the population to a state from which high reward is difficult to recover. We propose a latent-state formulation in which a finite-dimensional statistic summarizes decision-relevant information in the population, is estimated from finite batches, and evolves under unknown linear-in-feature reward and transition models. The key structural condition is recoverability: actions taken from one latent state can be matched at nearby latent states with comparable rewards and contracted next-state discrepancies. This converts transition-estimation error into controlled value error in continuous latent dynamics. Under robust recovery of the confidence class, full-model optimistic planning achieves high-probability sublinear regret. When only the true dynamics are recoverable, we introduce recoverable-core extended optimism, which optimizes over locally plausible reward-transition triples while pruning nonrecoverable ones. We give a finite-grid implementation using a Lipschitz majorant. The regret bounds separate reward learning, transition learning, epoch switching, finite-batch state estimation, and finite-grid approximation. A lower bound shows that without recovery-type structure, irreversible latent dynamics can force linear regret for every single-trajectory algorithm.

Key words: performative decision making; online learning; latent states; regret; decision-dependent uncertainty; reinforcement learning

1. Introduction

Modern decision systems often operate in environments that react to deployed decisions. Routing recommendations alter traffic, healthcare interventions change population outcomes, public programs affect participation, and prices reshape future demand. In such settings, the data distribution is part of the controlled system: today's decision affects both current reward and the population faced tomorrow. Ignoring this feedback can lead to systematically poor decisions.

This feedback is the central object of *performative prediction*, a literature initiated by Perdomo et al. (2020). Much of that literature studies repeated risk minimization and convergence to perfor-

matively stable decisions of the form $a^\star \in \arg\min_a R(a, d(a^\star))$, where $R(a,d)$ is the risk of decision a under distribution d (Perdomo et al. 2020, Drusvyatskiy and Xiao 2023, Cutler et al. 2024, Li and Wai 2022). The basic framework is stateless: the induced distribution is determined by the deployed decision. Stateful performative prediction allows the next distribution to depend on both the current distribution and the decision (Brown et al. 2022). This dynamic distributional perspective is central to our setting, but the objective in that stateful line of work remains equilibrium behavior under retraining rather than finite-horizon regret minimization.

From an operations-research perspective, the regret objective is closer to optimization under decision-dependent uncertainty, where a decision affects the uncertainty distribution on which its cost is evaluated. Classical stochastic, robust, and distributionally robust optimization provide the standard decision-focused language for such problems (Shapiro et al. 2009, Ben-Tal et al. 2009, Rahimian and Mehrotra 2019, Kuhn et al. 2025). Much of this literature studies static models in which a decision induces a distribution $d(a)$, or assumes that the response map is known, available through an oracle, or estimated offline before optimization (Jonsbraten et al. 1998, Goel and Grossmann 2006, Nohadani and Sharma 2018, Luo and Mehrotra 2020, Basciftci et al. 2021, Noyan et al. 2022, Jin et al. 2024, Yu and Shen 2022, Zhu et al. 2026, Yu and Basciftci 2026). In a stateful problem, however, offline learning is limited by reachability: estimating the transition requires distribution-action-next-distribution triples, yet without a simulator or broad exploratory data one cannot reset the population to arbitrary distributions and test arbitrary actions. Without additional extrapolative structure, historical data identify transitions only at the triples they cover.

Together, these perspectives motivate a single-trajectory control problem. At round t, the primitive state is a population distribution $d_t \in \mathcal{P}(\mathcal{Z})$, the decision maker chooses $a_t \in \mathcal{A}$, and the population evolves as $d_{t+1} = \mathcal{T}(d_t, a_t)$. Both the reward and $\mathcal{T}$ are unknown, so exploration concerns not only immediate rewards but also the future population dynamics created by the chosen actions.

Directly learning a full distributional transition and planning over probability measures can be statistically and computationally difficult without additional structure. Our modeling step is to summarize the information needed for decision making by a finite-dimensional latent statistic and to impose linear structure on rewards and latent transitions:

$$x_t = x(d_t) = \mathbb{E}_{Z\sim d_t}[\psi(Z)], \qquad r_\star(x,a) = \theta_\star^\top \Psi(x,a), \qquad T_\star(x,a) = W_\star \Phi(x,a).$$

This latent process preserves the performative feedback loop while providing a statistically tractable abstraction. The decision maker does not observe x_t directly. Instead, at each round it receives a finite batch of samples from the current population and forms an empirical estimate of the latent state. The goal is to learn both $r_\star$ and $T_\star$ online while minimizing finite-horizon regret.

This formulation is neither a contextual bandit nor a standard observed-state Markov decision process (Li et al. 2010, Chu et al. 2011, Lattimore and Szepesvári 2020, Puterman 1994). The primitive state is a population distribution, its finite-dimensional statistic is observed only through samples, and the context is endogenous: today's action changes tomorrow's latent state. Exploration can therefore be dangerous. An informative action may move the population into a low-value region from which high reward is hard to recover, and raw optimism can amplify this risk by exploiting statistically plausible but dynamically nonrecoverable transition predictions.

This leads to the central question: when is reliable online decision making possible? Our algorithms combine confidence-set optimism with recoverability. A lazy epoch structure freezes optimistic value functions within epochs, allowing Bellman potentials to telescope while determinant doubling controls the number of switches.

The key analytical bottleneck is that transition error appears as a displacement in a continuous latent state. If an optimistic planner predicts z, while the true next state is $T_\star(x,a)$, the regret analysis must control $V(z)-V(T_\star(x,a))$. A span bound of the kind often used in reinforcement-learning regret analyses gives only a constant-size bound. We therefore use *recoverability*: informally, actions taken from one latent state can be matched at nearby latent states with comparable immediate reward and contracted next-state discrepancy. This property yields Lipschitz value regularity, so small transition-estimation errors translate into small regret.

We first analyze a *robust recovery* setting, where every statistically plausible model is recoverable and full-model optimism is valid. This setting captures relevant examples, including stabilizable affine control, but can be restrictive. To relax it, we introduce *recoverable-core extended optimism*, which requires recoverability only of the true latent dynamics. The planner optimizes over locally plausible action-reward-next-state triples, but restricts attention to the largest recoverable subset. We also provide a finite-grid implementation, using a Lipschitz majorant to prevent grid-induced recovery errors from accumulating through the horizon.

Informal main result. Under either robust recovery of the confidence class or true-model recovery combined with recoverable-core planning, the decision maker achieves a high-probability regret bound of the form

$$\begin{aligned}\mathrm{Reg}(H) \lesssim\ & \text{reward-learning error} + \text{transition-learning error} + \text{epoch-switching cost} \\ & + \text{finite-batch state-estimation cost} + \text{finite-grid approximation cost}.\end{aligned}$$

The first two terms are controlled by elliptical-potential arguments for the reward and transition regressions. The epoch-switching term is controlled by determinant doubling. The finite-batch term reflects that the latent state is estimated from samples rather than observed directly, and the finite-grid term appears only in the implementable recoverable-core planner. In particular, when

the batch size and grid resolution are chosen so that the last two terms are lower order, the regret is sublinear in H up to logarithmic factors.

Our results concern statistical regret. Because the feature maps Ψ and Φ may be nonlinear in states and actions, the induced planning problems can be nonconvex even though the unknown parameters enter linearly. Following oracle-based regret analyses in reinforcement learning (Russo and Van Roy 2013, Osband and Van Roy 2014, Ayoub et al. 2020), we assume access to the required planning oracles. A complementary lower bound shows that without recovery-type structure, irreversible latent dynamics can force linear single-trajectory regret.

To the best of our knowledge, existing work has not studied this finite-horizon online control problem for stateful performative systems with latent population states, finite-batch state observations, and unknown reward and transition laws. Our contributions are as follows.

- We formulate stateful performative decision making as online control of a finite-dimensional population statistic $x(d) = \mathbb{E}_{Z\sim d}[\psi(Z)]$, estimated from finite batches along a single evolving trajectory with unknown reward and transition laws.
- We develop optimistic decision algorithms based on recoverability. Under robust recovery, full-model optimism is sufficient. For settings where only the true latent dynamics are recoverable, we propose recoverable-core extended optimism, which preserves optimism without requiring every plausible model to be recoverable. We also develop a finite-grid implementation of the recoverable-core planner, replacing the continuous recoverable core by a greatest fixed point of a pruning operator over grid states and grid actions.
- We establish high-probability regret bounds separating reward and transition learning, epoch switching, finite-batch state estimation, and finite-grid approximation. A Lipschitz majorant prevents grid-induced recovery errors from accumulating and yields the continuous-planning statistical terms plus an explicit discretization cost.
- We prove a lower bound showing why some recovery-type structure is needed for sublinear regret from a single evolving trajectory. Two models can be observationally indistinguishable before the first action, while any action that resolves the uncertainty irreversibly determines the future latent state, forcing linear regret for every algorithm.

1.1. Related Work

Performative prediction. Performative prediction studies learning problems in which the deployed decision affects the future data distribution. The framework of Perdomo et al. (2020) formalizes this feedback effect and introduces performative stability as an equilibrium notion for predictors evaluated on the distributions they induce. Subsequent work studies risk optimization, stochastic approximation, retraining dynamics, regret, and feedback models under performativity

(Miller et al. 2021, Mendler-Duenner et al. 2020, Izzo et al. 2021, 2022, Drusvyatskiy and Xiao 2023, Cutler et al. 2024, Li and Wai 2022, Jagadeesan et al. 2022, Park et al. 2024, Chen et al. 2024); related extensions address richer model classes, response-map learning, robust formulations, generalization, and partially performative shifts (Mofakhami et al. 2023, Bracale et al. 2024, Jia et al. 2025, Xue and Sun 2024, Wang et al. 2026, Rodemann et al. 2026, Lee and Zrnic 2026). See Hardt and Mendler-Duenner (2023), Kehrenberg et al. (2026) for recent surveys.

Two dynamic-distribution papers are especially relevant. In the stateful performative framework of Brown et al. (2022), the next distribution depends on the current distribution and deployed classifier, but the objective is convergence under retraining to an equilibrium pair. He et al. (2025) study nonlinear distribution dynamics using sensitivity-based online stochastic gradient. We instead consider finite-horizon control with unknown reward and transition laws, finite-batch latent observations, and single-trajectory regret against a dynamic benchmark. An informative action can therefore move the population to a region from which high reward is hard to recover, a safe-exploration issue absent from equilibrium and stochastic-gradient convergence guarantees.

Online nonlinear control and Stackelberg optimization. Brown et al. (2024) study interactions with adaptive agents as online nonlinear control along one trajectory. Under local controllability, inverse dynamics steer the system toward target states selected by online convex optimization, with performance measured against the best fixed target or stable policy. Our benchmark is instead the full-information finite-horizon dynamic optimum. We learn reward and transition models from sample-based latent states and use recoverability to couple nearby trajectories and exclude plausible but nonrecoverable transitions. Thus our problem combines a dynamic benchmark with potentially irreversible exploration rather than a convex fixed-state reduction.

Decision-dependent uncertainty in stochastic and robust optimization. Operations research has a parallel literature on decision-dependent, or endogenous, uncertainty. Decisions may affect random elements, probabilities, uncertainty sets, or ambiguity sets (Jonsbraten et al. 1998, Goel and Grossmann 2006, Nohadani and Sharma 2018, Luo and Mehrotra 2020, Basciftci et al. 2021, Noyan et al. 2022, Yu and Shen 2022, Yu and Basciftci 2026). Recent contextual DRO models learn decision- and covariate-dependent nominal distributions from logged observations before solving an ambiguity model (Zhu et al. 2026). These static, two-stage, multistage, and contextual models specify or estimate the uncertainty structure before optimization. In our setting, the affected distribution is itself the state for future decisions, and reward and transition laws are learned along the single trajectory altered by exploration.

Performative reinforcement learning. Recent work has extended performativity to reinforcement learning, where the deployed policy can affect the reward and transition dynamics of the environment. Mandal et al. (2023) introduce performative reinforcement learning and study stable

policies under repeated optimization; subsequent work considers gradual shifts and linear MDP structure (Rank et al. 2024, Mandal and Radanovic 2025). In contrast, the evolving object here is a population distribution summarized by a sample-observed latent statistic, and the decision maker learns its reward and transition models along one trajectory.

Mean-field control and learning. Mean-field control also treats population distributions as states and studies dynamic programming over probability measures (Pham and Wei 2016, Bäuerle 2023). Learning variants develop Q-learning, distributional dynamic programming, and model-based methods (Carmona et al. 2023, Gu et al. 2022, Bayraktar and Kara 2025). Most closely, Pásztor et al. (2023) give an optimistic model-based algorithm for episodic mean-field control. Their planner selects individual-agent policies and learns across episodes; our population-level actions are chosen along one nonresettable trajectory from finite-sample latent observations. Recoverable-core planning addresses the resulting irreversible exploration. Related reset-free RL controls regret and the number of resets (Nguyen and Cheng 2023), whereas our population cannot be reset.

Optimism, function approximation, and stability. Our algorithmic analysis is related to optimism-based reinforcement learning. UCRL2 (Jaksch et al. 2010) plans in an extended MDP whose actions include locally plausible rewards and transitions; because it can imitate true reaching policies on the good event, not every plausible MDP must have small diameter. Similarly, our local extended confidence model contains the true one-step graph, while recoverable-core pruning is specific to continuous deterministic latent dynamics.

We also draw on linear-function-approximation RL and adaptive control (Jin et al. 2020, Yang and Wang 2020, Zanette et al. 2020, Abbasi-Yadkori and Szepesvári 2011, Dean et al. 2018, Simchowitz and Foster 2020, Cohen et al. 2019). In adaptive nonlinear control, Boffi et al. (2021) use Lyapunov stability and contraction to obtain regret guarantees under matched uncertainty. Here contraction instead matches actions across nearby latent states, establishing Bellman-value regularity even when plausible transition models need not inherit the true system's recovery structure.

Discretization and Lipschitz value regularity. Our finite-grid planner is related to discretization methods for continuous-state dynamic programming and reinforcement learning (Munos and Moore 2002, Powell 2007, Auer and Ortner 2007, Ortner and Ryabko 2012, Song and Sun 2019). Because projection preserves recovery only up to additive error, we use a Lipschitz majorant to prevent error accumulation and obtain an $O(H\Delta_\eta)$ discretization term.

1.2. Paper Organization and Notation

Paper Organization. The rest of the paper is organized as follows. Section 2 formalizes the model, observation protocol, and regret benchmark, and Section 3 develops the confidence sets. Section 4 studies full-model optimism when the confidence class is robustly recoverable. Section 5

develops recoverable-core optimism when only the true dynamics need be recoverable and gives the finite-grid regret guarantee. Section 6 presents the lower-bound implications of finite-batch observation and irreversible exploration. Section 7 reports synthetic and data-calibrated pricing experiments, and Section 8 concludes. Proofs, finite-grid details, and additional numerical results are in the appendices, except for the value-regularity proof in Proposition 1.

Notation. We write $\|\cdot\|_2$ for the Euclidean norm, $\|\cdot\|_{\mathrm{op}}$ for the operator norm, and $\|\cdot\|_F$ for the Frobenius norm. For a positive definite matrix M, $\|v\|_M = (v^\top M v)^{1/2}$. The Loewner order is denoted by $\preceq$. For a real-valued function f on a set $\mathcal{S}$, define $\mathrm{span}(f) := \sup_{x\in\mathcal{S}} f(x) - \inf_{x\in\mathcal{S}} f(x)$. For a subset $\mathcal{S}$ of a Euclidean space, write $\mathrm{diam}(\mathcal{S}) := \sup_{x,y\in\mathcal{S}} \|x-y\|_2$. The notation $j(t)$ denotes the epoch containing period t, and τ_j denotes the start of epoch j.

2. Model and Online Decision Problem

In this section, we formalize the latent-state performative decision problem. We first describe the distributional state model and the latent sufficiency conditions that reduce the problem to a finite-dimensional state. We then introduce the linear structure on rewards and transitions, the observation protocol, and the regret benchmark.

We consider a decision maker who acts over a finite horizon H. The primitive state is a population distribution. Let $\mathcal{Z}$ be the sample space and let $\mathcal{P}(\mathcal{Z})$ denote the set of probability measures on $\mathcal{Z}$. At period $t \in \{1,\dots,H\}$, the population state is $d_t \in \mathcal{P}(\mathcal{Z})$. After receiving sample information about the current population, the decision maker chooses an action $a_t \in \mathcal{A}$, where $\mathcal{A} \subset \mathbb{R}^{d_a}$ is compact. The population then evolves according to an unknown performative transition map

$$\mathcal{T} : \mathcal{P}(\mathcal{Z}) \times \mathcal{A} \to \mathcal{P}(\mathcal{Z}), \qquad d_{t+1} = \mathcal{T}(d_t, a_t).$$

For each action $a \in \mathcal{A}$, let $r(\cdot, a) : \mathcal{Z} \to \mathbb{R}$ denote the sample-level reward.

2.1. Latent Sufficiency

The distribution-state formulation captures settings in which a decision rule is repeatedly deployed on a population and changes its future distribution. Our central modeling step replaces the full distribution d_t by a finite-dimensional statistic

$$x(d) = \mathbb{E}_{Z\sim d}[\psi(Z)],$$

where $\psi : \mathcal{Z} \to \mathbb{R}^m$ is a known feature map. Let $\mathcal{X} \subset \mathbb{R}^m$ denote the compact latent-state domain considered by the algorithms, and assume $x(d) \in \mathcal{X}$ for all population states under consideration. The statistic may collect aggregate demand, traffic, health, or labor-market features, or parameters of a structured distribution family. We write

$$x_t := x(d_t).$$

The latent statistic must be sufficient for both rewards and future latent states. For rewards, we assume

$$\mathbb{E}_{Z\sim d}[r(Z,a)] = r_\star(x(d),a). \tag{1}$$

For transitions, we assume that the next statistic needed for decision making is determined by the current latent state and action:

$$x(d_{t+1}) = T_\star(x(d_t),a_t). \tag{2}$$

ASSUMPTION 1 **(Forward-invariant latent domain)**. *The latent planning domain $\mathcal{X}$ is forward invariant under the true latent transition:*

$$T_\star(x,a)\in\mathcal{X} \qquad \textit{for every } x\in\mathcal{X} \textit{ and } a\in\mathcal{A}.$$

Assumption 1 keeps the benchmark and realized latent trajectories inside the planning domain. Equation (1) does not require individual rewards to depend only on $x(d)$; only their population average must do so. This is necessary for the exact latent dynamic program: otherwise, two distributions with the same statistic could have different expected rewards under the same action.

For example, if $r(Z,a)$ is polynomial in Z, including the corresponding monomials in ψ makes the relevant moments sufficient for expected reward. Equation (2) is analogous: the full next distribution need not be determined by x_t, only the information needed for future rewards and latent transitions.

We give two examples of this modeling step in practice.

Healthcare and public-health resource allocation. A health system may repeatedly allocate outreach, screening, reminders, or treatment capacity. Here d_t describes patient risk, disease, adherence, access, and utilization; a_t is an intervention policy; and reward measures health or capacity outcomes. Because intervention changes future participation, adherence, and risk, the population evolves endogenously. A latent state may collect prevalence, risk moments, adherence, utilization, and capacity pressure, with these summaries determining expected reward and the next latent state.

Routing and traffic control. A transportation agency may repeatedly recommend routes or deploy congestion warnings and tolls. The distribution d_t describes trips, routes, travel times, locations, and road conditions; a_t is a routing, pricing, or information intervention; and reward measures congestion, travel time, throughput, or welfare. Aggregate link flows, speeds, demand moments, and congestion levels can form x_t, whose transition records the intervention's effect on future traffic. The dependence on x_t matters because the same toll or recommendation can have different effects under different congestion patterns.

2.2. Latent Linear Model and Observations

The latent sufficiency conditions reduce the population problem to a finite-dimensional state x_t, but they do not specify how the reward and transition functions are learned. For the algorithmic and regret analysis, we assume that both functions are linear in unknown parameters after applying known feature maps:

$$r_\star(x,a) = \theta_\star^\top \Psi(x,a), \qquad T_\star(x,a) = W_\star \Phi(x,a).$$

Here $\Psi : \mathbb{R}^m \times \mathcal{A} \to \mathbb{R}^p$ and $\Phi : \mathbb{R}^m \times \mathcal{A} \to \mathbb{R}^q$ are known feature maps, while $\theta_\star \in \mathbb{R}^p$ and $W_\star \in \mathbb{R}^{m\times q}$ are unknown. This is a linear-function-approximation model in the unknown parameters, but it may be nonlinear in states and actions through the feature maps Ψ and Φ. Along the realized trajectory, the latent state evolves as $x_{t+1} = W_\star \Phi(x_t, a_t)$. The decision maker must learn both unknown parameters from data collected along this single controlled trajectory.

The standard assumptions below are imposed on the sample space, the action set, and the latent planning domain $\mathcal{X}$. We choose $\mathcal{X}$ large enough to contain the true latent states and the empirical latent averages $\hat{x}_t$ used by the algorithms; for example, it is enough for $\mathcal{X}$ to contain the closed convex hull of $\psi(\mathcal{Z})$. Feature bounds and Lipschitz constants are required only on this domain.

Assumption 2 **(Boundedness of features and parameters)**. *For all $z \in \mathcal{Z}$, $x \in \mathcal{X}$, and $a \in \mathcal{A}$,*

$$\|\psi(z)\|_2 \le B_\psi, \qquad \|\Psi(x,a)\|_2 \le B_\Psi, \qquad \|\Phi(x,a)\|_2 \le B_\Phi,$$

$$\|\theta_\star\|_2 \le S_\theta, \qquad \|W_\star\|_{\mathrm{op}} \le S_W.$$

Assumption 3 **(Lipschitzness of features)**. *For all $x, x' \in \mathcal{X}$ and $a, a' \in \mathcal{A}$,*

$$\|\Psi(x,a) - \Psi(x',a)\|_2 \le L_\Psi \|x - x'\|_2, \qquad \|\Psi(x,a) - \Psi(x,a')\|_2 \le L_{\Psi,a} \|a - a'\|_2,$$

$$\|\Phi(x,a) - \Phi(x',a)\|_2 \le L_\Phi \|x - x'\|_2, \qquad \|\Phi(x,a) - \Phi(x,a')\|_2 \le L_{\Phi,a} \|a - a'\|_2.$$

We next specify the observations. The decision maker does not observe d_t or x_t directly, but has sample access to the current population. We use sample splitting to separate the samples used to estimate the current latent state from the samples used to evaluate the chosen action.

At the beginning of period t, before choosing an action, the decision maker observes a state-estimation batch $Z^x_{t,1}, \dots, Z^x_{t,n} \overset{\text{i.i.d.}}{\sim} d_t$ and forms

$$\hat{x}_t := \frac{1}{n} \sum_{i=1}^{n} \psi(Z^x_{t,i}).$$

The action $a_t \in \mathcal{A}$ is chosen after observing this state batch. The period-t reward is then evaluated on the same population d_t, before the transition to d_{t+1}, using a separate reward batch $Z^R_{t,1}, \dots, Z^R_{t,n} \overset{\text{i.i.d.}}{\sim} d_t$, where $Y_{t,i} = r(Z^R_{t,i}, a_t) + \xi_{t,i}$ and $\xi_{t,i}$ is observation noise. Let $\mathcal{F}_{t,0}$ denote the

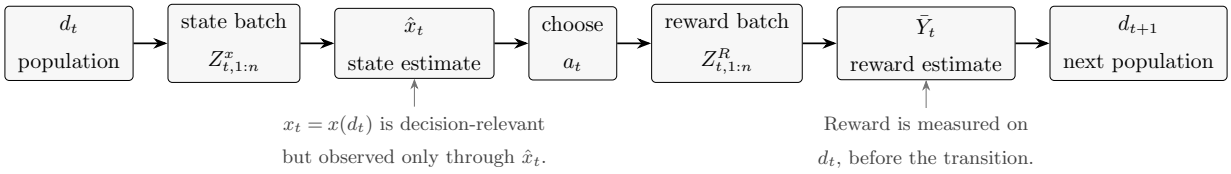


Figure 1 Period-t timing. The decision maker observes the current population only through a finite state-estimation batch, chooses a_t using $\hat{x}_t$, forms the reward estimate $\bar{Y}_t$ from an independent reward batch, and then the action changes the next population state.

information available after the state-estimation batch is observed and the action a_t is chosen, but before the reward batch is observed. Conditional on $\mathcal{F}_{t,0}$ and the current distribution d_t, the reward batch is sampled independently of the state-estimation batch. We write the average reward in this batch as

$$\bar{Y}_t := \frac{1}{n}\sum_{i=1}^{n} Y_{t,i}.$$

After the reward is evaluated, the transition map determines the next population state d_{t+1}. For transition estimation, we also observe a terminal state-estimation batch at time $H+1$, which is used to form $\hat{x}_{H+1}$ but not to choose an action.

Assumption 4 **(Reward fluctuation)**. *Let*

$$\zeta^R_t := \bar{Y}_t - r_\star(x_t, a_t)$$

be the fluctuation of the average reward around its true value. Conditional on $\mathcal{F}_{t,0}$, ζ^R_t is mean-zero and $\sigma_R/\sqrt{n}$-sub-Gaussian: $\mathbb{E}[\zeta^R_t \mid \mathcal{F}_{t,0}] = 0$, and, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(\lambda \zeta^R_t) \mid \mathcal{F}_{t,0}\big] \le \exp\left(\frac{\lambda^2 \sigma_R^2}{2n}\right).$$

Here σ_R controls both the sampling fluctuation of $r(Z^R_{t,i}, a_t)$ under $Z^R_{t,i} \sim d_t$ and the observation noise $\xi_{t,i}$.

This timing matches settings in which state information is available before deployment, while outcomes are measured afterward under the deployed action. It also gives the conditional mean-zero property used in the reward confidence bound.

Remark 1 (Sample splitting assumption). The independent reward batch ensures that, conditional on the information used to choose a_t, the reward average $\bar{Y}_t$ is an unbiased estimate of $r_\star(x_t, a_t) = \mathbb{E}_{Z\sim d_t}[r(Z, a_t)]$, with sub-Gaussian fluctuations of order $1/\sqrt{n}$, making the self-normalized confidence argument directly applicable. If the same batch were used both to estimate $\hat{x}_t$ and to form $\bar{Y}_t$, then a_t would be data-dependent and $\bar{Y}_t - r_\star(x_t, a_t)$ need not be conditionally mean-zero; the analysis would instead require uniform concentration over the action class. We therefore state the regret guarantees for the split-sample protocol. Analyzing data reuse or cross-fitting requires a separate argument.

2.3. Benchmark and Regret

Performance is measured against a full-information benchmark that knows the true reward and transition and observes the latent state exactly. The benchmark value functions are

$$V^\star_{H+1}(x) := 0, \qquad V^\star_h(x) := \max_{a\in\mathcal{A}} \left\{ r_\star(x,a) + V^\star_{h+1}(T_\star(x,a)) \right\}, \quad h = H,\dots,1.$$

The finite-horizon pseudo-regret is

$$\mathrm{Reg}(H) := V^\star_1(x_1) - \sum_{t=1}^{H} r_\star(x_t, a_t).$$

Here, x_1 is the initial latent state induced by d_1, while $x_2,\dots,x_H$ are the subsequent latent states induced by the transition dynamics and the chosen actions $a_1,\dots,a_H$. The word "pseudo" indicates that regret is evaluated using population mean rewards, although the decision maker observes only samples.

3. Confidence Sets and Optimistic Models

We construct a confidence ball for each latent state and then use empirical states as covariates in reward and transition regressions. On a joint high-probability event, the resulting sets contain all true quantities used by the optimistic planners. We then define the lazy epoch convention shared by both planners. The finite-grid implementation later replaces $\hat{x}_t$ by projected states and accounts separately for discretization error.

3.1. State Confidence Set

We first control the error from estimating the latent state. Having observed $\hat{x}_t$, the decision maker constructs a confidence set $\mathcal{X}_t$ that contains the true state x_t with high probability. The radius is chosen from concentration bounds for bounded vector-valued averages. For a state-confidence level $\delta_x \in (0,1)$, set

$$\beta^x := c_x B_\psi \sqrt{\frac{m + \log((H+1)/\delta_x)}{n}}, \tag{3}$$

where c_x is a universal constant. The dependence on $1/\sqrt{n}$ comes from averaging the bounded feature vectors $\psi(Z)$, while the logarithmic term accounts for uniform concentration over the $H+1$ state-estimation batches. The confidence set at time t is the following Euclidean ball around $\hat{x}_t$:

$$\mathcal{X}_t := \{x \in \mathbb{R}^m : \|x - \hat{x}_t\|_2 \le \beta^x\}.$$

By Lemma 1 in Appendix C, with probability at least $1-\delta_x$, we have $x_t \in \mathcal{X}_t$ for all $t = 1,\dots,H+1$. We denote this state-confidence event by

$$\mathcal{E}_x := \bigcap_{t=1}^{H+1} \{x_t \in \mathcal{X}_t\}.$$

All subsequent confidence bounds in this section are stated conditional on $\mathcal{E}_x$.

3.2. Reward Confidence Set

Using the observations collected before period t, the reward parameter is estimated by ridge regression on the observed state-action pairs:

$$\hat{\theta}_t := \arg\min_{\theta\in\mathbb{R}^p}\left\{\lambda_R\|\theta\|_2^2+\sum_{s<t}(\bar{Y}_s-\theta^\top\Psi(\hat{x}_s,a_s))^2\right\}=(\Lambda_t^R)^{-1}\sum_{s<t}\Psi(\hat{x}_s,a_s)\bar{Y}_s,$$

where

$$\Lambda_t^R := \lambda_R I_p+\sum_{s<t}\Psi(\hat{x}_s,a_s)\Psi(\hat{x}_s,a_s)^\top$$

is the reward design matrix. The parameter $\lambda_R>0$ is the ridge regularization parameter; it controls the usual bias–variance tradeoff in reward estimation. Note that the covariates are based on the estimated states $\hat{x}_s$. For a confidence level $\delta_R\in(0,1)$, the reward confidence set is the following ellipsoid around $\hat{\theta}_t$, with norm induced by Λ_t^R:

$$\mathcal{C}_t^R := \left\{\theta\in\mathbb{R}^p : \|\theta-\hat{\theta}_t\|_{\Lambda_t^R}\le\beta_t^R,\ \|\theta\|_2\le S_\theta\right\},$$

where

$$\beta_t^R := \sqrt{\lambda_R}S_\theta+\frac{\sigma_R}{\sqrt{n}}\left[2\log\left(\frac{\det(\Lambda_t^R)^{1/2}}{\det(\lambda_R I_p)^{1/2}\delta_R}\right)\right]^{1/2}+S_\theta L_\Psi\beta^x\sqrt{(t-1)p}. \tag{4}$$

The radius separates three effects. The first term is the ridge regularization cost, the second is the self-normalized martingale term for reward fluctuations, and the third is the deterministic approximation error from using $\hat{x}_s$ rather than the unobserved state x_s. This last term propagates latent-state estimation error through the Lipschitz feature map Ψ; the factor $\sqrt{(t-1)p}$ is obtained by applying Cauchy–Schwarz to the accumulated bias in the $(\Lambda_t^R)^{-1}$-norm.

By Proposition 4 in Appendix D, with probability at least $1-\delta_R$, simultaneously for all $t=1,\dots,H+1$, on $\mathcal{E}_x$, $\theta_\star\in\mathcal{C}_t^R$. We also define the reward prediction bonus

$$b_t^R(x,a) := \beta_t^R\|\Psi(x,a)\|_{(\Lambda_t^R)^{-1}}. \tag{5}$$

By Cauchy–Schwarz, every $\theta\in\mathcal{C}_t^R$ satisfies $|(\theta-\hat{\theta}_t)^\top\Psi(x,a)|\le b_t^R(x,a)$. Thus, whenever $\theta_\star\in\mathcal{C}_t^R$, b_t^R bounds the prediction error of the true reward model around the estimate. More generally, it bounds the diameter of plausible reward predictions: if $\theta,\theta'\in\mathcal{C}_t^R$, then $|(\theta-\theta')^\top\Psi(x,a)|\le 2b_t^R(x,a)$.

3.3. Transition Confidence Set

Using the same pre-period-t data, the transition parameter $W_\star$ is estimated by ridge regression on realized state-action pairs and next-state estimates:

$$\hat{W}_t := \arg\min_{W\in\mathbb{R}^{m\times q}}\left\{\lambda_T\|W\|_F^2+\sum_{s<t}\|W\Phi(\hat{x}_s,a_s)-\hat{x}_{s+1}\|_2^2\right\}=\left(\sum_{s<t}\hat{x}_{s+1}\Phi(\hat{x}_s,a_s)^\top\right)(\Lambda_t^T)^{-1},$$

where

$$\Lambda_t^T := \lambda_T I_q + \sum_{s<t} \Phi(\hat{x}_s, a_s)\Phi(\hat{x}_s, a_s)^\top$$

is the transition design matrix. The parameter $\lambda_T > 0$ is the transition ridge regularization parameter; it controls the usual bias–variance tradeoff in transition estimation. The transition confidence set is the corresponding matrix-valued ellipsoid:

$$\mathcal{C}_t^T := \left\{ W \in \mathbb{R}^{m\times q} : \|W\|_{\mathrm{op}} \le S_W,\ \|(W - \hat{W}_t)(\Lambda_t^T)^{1/2}\|_F \le \beta_t^T \right\},$$

where

$$\beta_t^T := \sqrt{\lambda_T m}\, S_W + (1 + S_W L_\Phi)\beta^x \sqrt{(t-1)q}. \tag{6}$$

The radius contains a regularization term for the unknown transition matrix and a deterministic term from finite-batch latent-state estimation. There is no martingale term analogous to the reward fluctuation term: conditional on the latent states, the transition is deterministic, and on $\mathcal{E}_x$ the regression error is bounded by the current and next state-estimation errors.

By Proposition 5 in Appendix D, on $\mathcal{E}_x$, for all $t = 1, \ldots, H+1$, $W_\star \in \mathcal{C}_t^T$. We similarly define the transition prediction bonus

$$b_t^T(x, a) := \beta_t^T \|\Phi(x, a)\|_{(\Lambda_t^T)^{-1}}. \tag{7}$$

By Cauchy–Schwarz and the Frobenius norm bound, every $W \in \mathcal{C}_t^T$ satisfies $\|(W - \hat{W}_t)\Phi(x, a)\|_2 \le b_t^T(x, a)$. Thus, whenever $W_\star \in \mathcal{C}_t^T$, b_t^T bounds the prediction error of the true transition model around the estimate. More generally, it bounds the diameter of plausible transition predictions: if $W, W' \in \mathcal{C}_t^T$, then $\|(W - W')\Phi(x, a)\|_2 \le 2b_t^T(x, a)$.

3.4. Global Good Event

The algorithms and regret analysis are stated on the event that all confidence sets contain their respective true quantities:

$$\mathcal{E} := \left\{ x_t \in \mathcal{X}_t,\ \theta_\star \in \mathcal{C}_t^R,\ W_\star \in \mathcal{C}_t^T \text{ for all } t = 1, \ldots, H+1 \right\}.$$

By the results in the preceding subsections, if the overall failure probability δ satisfies $\delta_x + \delta_R \le \delta$, then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$. No separate transition failure probability is needed because the transition confidence bound holds deterministically on the state-confidence event $\mathcal{E}_x$. On the *good event* $\mathcal{E}$, the empirical state, reward, and transition confidence sets contain the true quantities needed for optimistic planning.

3.5. Lazy Epoch Convention

The optimistic planners below use the confidence sets from this section through a lazy epoch convention. The running estimates and design matrices are updated as observations arrive, but the objects used to compute optimistic value functions are refreshed only at epoch boundaries. Let $\tau_1 = 1$ be the start of the first epoch. At the start of epoch j, the algorithm takes a snapshot of the current parameter estimates, design matrices, and confidence sets at time τ_j. The optimistic value functions for epoch j are computed from this snapshot. Thus planning during epoch j uses the fixed objects $\hat{\theta}_{\tau_j}$, $\hat{W}_{\tau_j}$, $\Lambda^R_{\tau_j}$, $\Lambda^T_{\tau_j}$, $\mathcal{C}^R_{\tau_j}$, and $\mathcal{C}^T_{\tau_j}$. Equivalently, the epoch-j plausible model class is

$$\mathcal{M}_j := \mathcal{C}^R_{\tau_j} \times \mathcal{C}^T_{\tau_j}.$$

The current empirical state $\hat{x}_t$ is still observed in every period; only the confidence snapshot and the associated optimistic value functions are held fixed within the epoch. The decision maker chooses an action by evaluating the epoch-j optimistic Bellman backup at the current empirical state.

Data collected within an epoch update the running estimates and design matrices but enter planning only at the next snapshot. A new epoch starts when either running design determinant has doubled relative to its value at the current epoch start:

$$\det(\Lambda^R_t) \geq 2\det(\Lambda^R_{\tau_j}) \qquad \text{or} \qquad \det(\Lambda^T_t) \geq 2\det(\Lambda^T_{\tau_j}).$$

Let $j(t)$ denote the epoch containing period t. Freezing keeps Bellman values fixed so regret potentials telescope, while determinant doubling limits recomputation. The number of epochs is controlled by the logarithms of the final-to-initial design-determinant ratios, which grow only polynomially in H under the feature bounds.

4. Full-Model Optimistic Planning Under Robust Recovery

We first record what happens under the most direct form of optimism. If every reward-transition model in the statistical confidence class satisfies a common recoverability property, then the decision maker can optimize over the full confidence class, as in classical optimistic model-based methods (Kearns and Singh 2002, Brafman and Tennenholtz 2002, Jaksch et al. 2010, Azar et al. 2017, Osband and Van Roy 2014, Ayoub et al. 2020). This benchmark is useful for two reasons. It shows how recoverability converts transition uncertainty into value regularity, and it identifies precisely which part of full-model optimism is too strong for the general setting considered below.

4.1. Robust Recovery of the Confidence Class

The robust-recovery condition is imposed on the entire epoch-j confidence class $\mathcal{M}_j$ over the latent domain $\mathcal{X}$, not only on the true model. This is the condition needed for full-model optimism, since

the planner may optimize with respect to any model in the confidence class. The condition is stated for latent states x, y; errors from acting at empirical states $\hat{x}_t$ are controlled separately by the state confidence set. For readability, define the reward and transition induced by $(\theta, W) \in \mathcal{M}_j$ as

$$r_\theta(x,a) := \theta^\top \Psi(x,a), \qquad T_W(x,a) := W\Phi(x,a).$$

ASSUMPTION 5 **(Robust recovery and invariance of the admissible optimistic class).** *For every epoch j, every $W \in \mathcal{C}^T_{\tau_j}$, every $x \in \mathcal{X}$, and every $a \in \mathcal{A}$, the admissible optimistic transition satisfies $T_W(x,a) \in \mathcal{X}$. Moreover, there exist constants $L_r \geq 0$ and $\alpha \in [0,1)$ such that, for every epoch j, every $(\theta, W) \in \mathcal{M}_j$, every $x, y \in \mathcal{X}$, and every $a \in \mathcal{A}$, there exists $a' \in \mathcal{A}$ satisfying*

$$|r_\theta(x,a) - r_\theta(y,a')| \leq L_r \|x-y\|_2, \qquad \|T_W(x,a) - T_W(y,a')\|_2 \leq \alpha \|x-y\|_2.$$

The forward-invariance part ensures that the full-model optimistic Bellman recursion below is well defined on $\mathcal{X}$. This condition is specific to full-model optimism; in the recoverable-core construction, optimistic next states are restricted to $\mathcal{X}$ directly. The recovery part states that, under every model in the confidence class, an action taken from state x can be matched from state y with comparable reward and contracted next-state discrepancy. For nearby y, the matching errors are small because they scale with $\|x-y\|_2$. Although the condition is imposed on the entire confidence class, it is not as stringent as it may first appear; the following example gives a sufficient condition.

EXAMPLE 1 (UNIFORMLY RECOVERABLE AFFINE CLASS WITH BALL-CONSTRAINED ACTIONS). Consider an admissible optimistic class in which each transition model has the affine form $T_W(x,a) = Ax + Ba + c$, and let the action space be the Euclidean ball $\mathcal{A} = R\mathbb{B} := \{a \in \mathbb{R}^k : \|a\|_2 \leq R\}$. Assume that every admissible transition model is forward invariant on $\mathcal{X}$. Suppose that for every admissible pair (A,B), there exists a feedback matrix $K_{A,B}$ satisfying

$$\|K_{A,B}\|_{\mathrm{op}} \leq K_{\max}, \qquad \|K_{A,B}\|_{\mathrm{op}} \operatorname{diam}(\mathcal{X}) \leq R, \qquad \|A - BK_{A,B}\|_{\mathrm{op}} + \|B\|_{\mathrm{op}} \|K_{A,B}\|_{\mathrm{op}} \leq \alpha < 1.$$

Fix $x, y \in \mathcal{X}$ and $a \in \mathcal{A}$, write $\delta := x - y$, and set $\lambda := \|K_{A,B}\delta\|_2 / R$. Then $0 \leq \lambda \leq 1$. Choose the recovery action $a' := (1-\lambda)a + K_{A,B}\delta$. If $\lambda = 0$, then $a' = a$; otherwise, $a' = (1-\lambda)a + \lambda(K_{A,B}\delta/\lambda)$, where $\|K_{A,B}\delta/\lambda\|_2 = R$. Thus a' is a convex combination of two points in $R\mathbb{B}$, so $a' \in \mathcal{A}$.

The transition difference satisfies

$$T_W(x,a) - T_W(y,a') = (A - BK_{A,B})\delta + \lambda Ba,$$

$$\|T_W(x,a) - T_W(y,a')\|_2 \leq \big(\|A - BK_{A,B}\|_{\mathrm{op}} + \|B\|_{\mathrm{op}}\|K_{A,B}\|_{\mathrm{op}}\big)\|\delta\|_2 \leq \alpha \|x-y\|_2.$$

The recovery action also satisfies $\|a - a'\|_2 \leq \lambda R + \|K_{A,B}\delta\|_2 = 2\|K_{A,B}\delta\|_2 \leq 2K_{\max}\|x-y\|_2$. Therefore, if uniformly over the admissible reward class, $|r_\theta(x,a) - r_\theta(y,a')| \leq L_x\|x-y\|_2 + L_a\|a-a'\|_2$, then Assumption 5 holds with $L_r = L_x + 2L_aK_{\max}$.

The construction may be viewed as constrained incremental feedback: the nominal action is first contracted toward the center of the action ball to create sufficient room for the feedback correction $K_{A,B}(x-y)$.

Remark 2 (Relation to standard feedback stabilization). The shrinkage factor in Example 1 is needed only to maintain feasibility under the compact action constraint. If the action space were unconstrained, $\mathcal{A}=\mathbb{R}^k$, one could instead use the standard incremental feedback action $a'=a+K_{A,B}(x-y)$, which gives $T_W(x,a)-T_W(y,a')=(A-BK_{A,B})(x-y)$. Thus, in the unconstrained setting, recovery reduces to the usual feedback contraction requirement $\|A-BK_{A,B}\|_{\mathrm{op}}<1$. This unconstrained case lies outside our standing compact-action assumptions but clarifies the control-theoretic origin of the construction. Moreover, if the open-loop dynamics are already contractive, one may take $K_{A,B}=0$. Then $\lambda=0$, $a'=a$, and the example reduces to the mean-reverting case.

4.2. Optimistic Planning Under Robust Recovery

Under robust recovery, the decision maker optimizes over the confidence class. At epoch j, set $\bar{V}^j_{H+1}(x):=0$, and, for $h=H,H-1,\ldots,\tau_j$,

$$\bar{V}^j_h(x):=\sup_{\substack{a\in\mathcal{A}\\(\theta,W)\in\mathcal{M}_j}}\left[\theta^\top\Psi(x,a)+\bar{V}^j_{h+1}(W\Phi(x,a))\right]. \tag{8}$$

The Bellman recursion is written for a generic latent state x. At period t in epoch j, after observing the empirical state $\hat{x}_t$, the decision maker solves the stage-t problem in (8) with input $x=\hat{x}_t$. The planner may use an optimistic model $(\theta,W)\in\mathcal{M}_j$ to evaluate the action, but only the real action a_t is executed. Here and below, *full-model optimism* means that each Bellman backup optimizes over the unpruned parameter confidence class. The optimizing pair (θ,W) may vary across states and stages.

The following proposition shows that, under robust recovery, the model-optimistic values $\bar{V}^j_h$ are optimistic and satisfy the Lipschitz regularity needed for the regret analysis. For this purpose, define

$$L_{\mathrm{lat}}:=\frac{L_r}{1-\alpha},\qquad B_{\mathrm{lat}}:=L_{\mathrm{lat}}\,\mathrm{diam}(\mathcal{X}). \tag{9}$$

Proposition 1 **(Robust recovery implies value regularity)**. *Suppose Assumption 5 holds. On the good event $\mathcal{E}$, for every epoch j and stage h, the model-optimistic values $\bar{V}^j_h$ satisfy*

$$\bar{V}^j_h(x)\ge V^\star_h(x)\qquad \textit{for all } x\in\mathcal{X},$$
$$|\bar{V}^j_h(x)-\bar{V}^j_h(y)|\le L_{\mathrm{lat}}\|x-y\|_2\qquad \textit{for all } x,y\in\mathcal{X},$$
$$\mathrm{span}(\bar{V}^j_h)\le B_{\mathrm{lat}}.$$

Proof. On the good event, $(\theta_\star, W_\star) \in \mathcal{M}_j$, so the true model is feasible in the model-optimistic Bellman recursion. The standard backward induction argument therefore gives optimism.

For Lipschitz regularity, the claim is trivial at $H+1$. Suppose $\bar{V}^j_{h+1}$ is L_{lat}-Lipschitz. Fix $x, y \in \mathcal{X}$, and take an ϵ-optimal triple (a, θ, W) in the Bellman recursion at x. By Assumption 5, there exists $a' \in \mathcal{A}$ such that $r_\theta(x,a) - r_\theta(y,a') \le L_r\|x-y\|_2$ and $\|T_W(x,a) - T_W(y,a')\|_2 \le \alpha\|x-y\|_2$. Using the same model (θ, W), the action a' is feasible in the Bellman recursion at y. Hence

$$\begin{aligned}\bar{V}^j_h(x) - \bar{V}^j_h(y) &\le r_\theta(x,a) - r_\theta(y,a') + \bar{V}^j_{h+1}(T_W(x,a)) - \bar{V}^j_{h+1}(T_W(y,a')) + \epsilon \\ &\le L_r\|x-y\|_2 + L_{\text{lat}}\alpha\|x-y\|_2 \\ &= L_{\text{lat}}\|x-y\|_2.\end{aligned}$$

Letting $\epsilon \downarrow 0$ gives the desired one-sided inequality. Swapping x and y gives the reverse inequality. The span bound follows from compactness of $\mathcal{X}$. □

4.3. Regret Bound Under Robust Recovery

The preceding proposition provides the optimism and horizon-uniform value regularity used in the regret analysis. The structure of the bound is easiest to see before introducing the final compact notation. Under robust recovery, regret is controlled by four quantities: the number of times the optimistic value function is refreshed, the accumulated reward uncertainty, the accumulated transition uncertainty, and the error from acting at empirical rather than true latent states.

Set $C_x := S_\theta L_\Psi + L_{\text{lat}}(S_W L_\Phi + 1)$. Here L_{lat} is the Lipschitz constant of the optimistic value functions and B_{lat} is the corresponding span bound. Recall the time-t prediction bonuses b^R_t and b^T_t from (5) and (7). For an epoch j, we use their frozen versions computed from the epoch snapshot:

$$b^R_j(x,a) := \beta^R_{\tau_j}\|\Psi(x,a)\|_{(\Lambda^R_{\tau_j})^{-1}}, \qquad b^T_j(x,a) := \beta^T_{\tau_j}\|\Phi(x,a)\|_{(\Lambda^T_{\tau_j})^{-1}}. \tag{10}$$

These widths are local prediction bounds: $b^R_j(x,a)$ measures reward uncertainty at (x,a), while $b^T_j(x,a)$ measures transition uncertainty at (x,a). With this notation, the regret proof first establishes the accounting bound

$$\text{Reg}(H) \le 2B_{\text{lat}}N_H + 2\sum_{t=1}^{H} b^R_{j(t)}(\hat{x}_t, a_t) + 2L_{\text{lat}}\sum_{t=1}^{H} b^T_{j(t)}(\hat{x}_t, a_t) + C_x H\beta^x,$$

where N_H is the number of determinant-doubling epochs up to time H. The first term is the price of changing optimistic value functions across epochs. The second and third terms are the cumulative statistical uncertainty along the realized trajectory. Transition uncertainty is multiplied by L_{lat} because it affects regret through the continuation value. The final term reflects the fact that the planner acts at $\hat{x}_t$, while regret is measured against the true latent state x_t.

To bound these quantities, define

$$\gamma_R(H) := \log \frac{\det(\Lambda^R_{H+1})}{\det(\lambda_R I_p)}, \qquad \gamma_T(H) := \log \frac{\det(\Lambda^T_{H+1})}{\det(\lambda_T I_q)},$$
$$\bar{\beta}^R_H := \max_{t \le H+1} \beta^R_t = \beta^R_{H+1}, \qquad \bar{\beta}^T_H := \max_{t \le H+1} \beta^T_t = \beta^T_{H+1},$$
$$\kappa_R := \frac{B^2_\Psi/\lambda_R}{\log(1 + B^2_\Psi/\lambda_R)}, \qquad \kappa_T := \frac{B^2_\Phi/\lambda_T}{\log(1 + B^2_\Phi/\lambda_T)}.$$

The equalities for $\bar{\beta}^R_H$ and $\bar{\beta}^T_H$ follow from the monotonicity of their defining terms. Determinant doubling and Lemmas 7–8 in Appendix I give

$$N_H \le 1 + \frac{\gamma_R(H)}{\log 2} + \frac{\gamma_T(H)}{\log 2},$$
$$\sum_{t=1}^H b^R_{j(t)}(\hat{x}_t, a_t) \le \bar{\beta}^R_H \sqrt{2H\kappa_R \gamma_R(H)}, \qquad \sum_{t=1}^H b^T_{j(t)}(\hat{x}_t, a_t) \le \bar{\beta}^T_H \sqrt{2H\kappa_T \gamma_T(H)}.$$

The following corollary collects these bounds.

Corollary 1 **(Regret under robust recovery)**. *Suppose Assumptions 2, 3, 4, and 5 hold. Let $\delta_x + \delta_R \le \delta$. Then the lazy model-optimistic planner satisfies, with probability at least $1 - \delta$,*

$$\mathrm{Reg}(H) \le 2B_{\mathrm{lat}} \left(1 + \frac{\gamma_R(H)}{\log 2} + \frac{\gamma_T(H)}{\log 2}\right) + 2\bar{\beta}^R_H \sqrt{2H\kappa_R \gamma_R(H)} + 2L_{\mathrm{lat}} \bar{\beta}^T_H \sqrt{2H\kappa_T \gamma_T(H)} + C_x H \beta^x.$$

Proof sketch. Proposition 1 supplies optimism and the constants L_{lat} and B_{lat}. Substitute the three estimates above into the accounting inequality proved in Lemma 3. Unlike the recoverable-core proof, no local pruning or finite-grid discretization is needed. □

By Lemma 6, $\gamma_R(H) \le p \log(1 + HB^2_\Psi/(\lambda_R p))$, and the analogous bound holds for $\gamma_T(H)$. Thus, when the feature dimensions are fixed, the epoch-switching term is $O(\log H)$, while the pure parameter-learning terms are $\widetilde{O}(\sqrt{H})$. The finite-batch state estimates enter in two ways: directly through $C_x H \beta^x$, and through the state-error components of $\bar{\beta}^R_H$ and $\bar{\beta}^T_H$. After the elliptical-potential multipliers are applied, these latter components contribute on the order of $H\beta^x\sqrt{\gamma_R(H)}$ and $H\beta^x\sqrt{\gamma_T(H)}$, respectively. Consequently, the robust-recovery regret is sublinear whenever $\beta^x\sqrt{\gamma_R(H) + \gamma_T(H)} = o(1)$. Since $\gamma_R(H) + \gamma_T(H) = O(\log H)$ for fixed feature dimensions, the definition of β^x in (3) shows that it suffices to take a per-period batch size n satisfying $n \gg (m + \log H)\log H$, up to logarithmic dependence on the confidence level.

5. Recoverable-Core Optimism

The benchmark in Section 4 requires every statistically plausible model to be recoverable. This is stronger than what is needed for optimism. The true latent dynamics may be recoverable even though the confidence sets contain other plausible, but dynamically nonrecoverable, transition

predictions. The challenge is therefore to keep enough optimism for learning while preventing the planner from exploiting such nonrecoverable predictions.

The recoverable-core construction addresses this problem locally. Rather than optimizing each Bellman backup directly over parameter pairs in the confidence class, the planner works with the one-step objects that enter the Bellman recursion: a real action, a plausible immediate reward, and a plausible next latent state. These triples form a raw extended confidence model. It is optimistic, but too permissive: some locally plausible next states can create value functions with uncontrolled Lipschitz behavior. We therefore extract the largest subcorrespondence of extended actions that satisfies the recovery property. Because the true reward-transition graph belongs to this subcorrespondence on the good event, the resulting *recoverable core* preserves optimism while discarding nonrecoverable one-step predictions. Thus the core is not merely a conservative restriction of the raw planner; it targets the optimistic one-step predictions whose downstream value cannot be controlled through recovery.

5.1. True Latent Recovery

The structural condition is imposed only on the true latent model.

Assumption 6 **(True latent recovery)**. *There exist constants $L_r \geq 0$ and $\alpha \in [0,1)$, independent of H, such that for every $x, y \in \mathcal{X}$ and every $a \in \mathcal{A}$, there exists $a' \in \mathcal{A}$ satisfying*

$$|r_\star(x,a) - r_\star(y,a')| \leq L_r \|x-y\|_2, \quad \text{and} \quad \|T_\star(x,a) - T_\star(y,a')\|_2 \leq \alpha \|x-y\|_2.$$

This condition says that, under the true latent model, any action at x can be matched from y with comparable reward and contracted next-state discrepancy, uniformly over pairs x, y. Thus nearby latent states admit recovery actions with small one-step errors. It is a finite-horizon, metric analogue of the reachability condition that controls bias spans in finite-state communicating MDPs.

Example 2 (Mean-reverting latent dynamics). Suppose $T_\star(x,a) = Ax + g(a)$ with $\|A\|_{\mathrm{op}} \leq \alpha < 1$. Choose $a' = a$. Then $T_\star(x,a) - T_\star(y,a') = A(x-y)$, and hence $\|T_\star(x,a) - T_\star(y,a')\|_2 \leq \alpha\|x-y\|_2$. If $r_\star$ is L_x-Lipschitz in x uniformly over a, then Assumption 6 holds with $L_r = L_x$.

Example 3 (Recoverable true affine dynamics). Suppose $T_\star(x,a) = Ax + Ba + c$, $\mathcal{A} = R\mathbb{B}$, and $T_\star$ is forward invariant on $\mathcal{X}$. If a feedback matrix K satisfies

$$\|K\|_{\mathrm{op}} \operatorname{diam}(\mathcal{X}) \leq R, \qquad \|A - BK\|_{\mathrm{op}} + \|B\|_{\mathrm{op}} \|K\|_{\mathrm{op}} \leq \alpha < 1.$$

then the recovery action in Example 1 gives, for every $x, y \in \mathcal{X}$ and $a \in \mathcal{A}$, an $a' \in \mathcal{A}$ satisfying $\|T_\star(x,a) - T_\star(y,a')\|_2 \leq \alpha\|x-y\|_2$ and $\|a - a'\|_2 \leq 2\|K\|_{\mathrm{op}}\|x-y\|_2$. If the true reward is L_x-Lipschitz in the state and L_a-Lipschitz in the action, then Assumption 6 holds with $L_r = L_x + 2L_a\|K\|_{\mathrm{op}}$.

These conditions apply only to the true pair (A, B), not uniformly over the optimistic confidence class as robust recovery requires.

5.2. Local Extended Confidence Actions

We first define the local extended actions used by the recoverable-core planner. Each extended action records a real action together with a statistically plausible reward and next latent state. Fix an epoch j and let τ_j be its start time. Within this epoch the algorithm freezes the confidence objects at time τ_j, as in the lazy epoch convention. The frozen confidence sets induce, at each latent state and action, a set of plausible one-step consequences. We write the frozen estimates as

$$\hat{r}_j(x,a) := \hat{\theta}_{\tau_j}^\top \Psi(x,a), \qquad \hat{T}_j(x,a) := \hat{W}_{\tau_j} \Phi(x,a). \tag{11}$$

The estimates in (11), together with the epoch-snapshot widths b_j^R and b_j^T in (10), define the local confidence intervals used below. On the good event, for all latent states $x \in \mathcal{X}$ and actions $a \in \mathcal{A}$,

$$|r_\star(x,a) - \hat{r}_j(x,a)| \le b_j^R(x,a), \tag{12}$$

$$\|T_\star(x,a) - \hat{T}_j(x,a)\|_2 \le b_j^T(x,a). \tag{13}$$

These inequalities are simply the prediction form of $\theta_\star \in \mathcal{C}_{\tau_j}^R$ and $W_\star \in \mathcal{C}_{\tau_j}^T$.

An extended action at state x is a triple $e = (a, \rho, z)$, where a is the real action to be executed, ρ is a plausible immediate reward, and z is a plausible next latent state. The *raw extended confidence* correspondence is

$$\mathfrak{E}_j(x) := \left\{ \begin{array}{c} a \in \mathcal{A},\ z \in \mathcal{X}, \\ (a,\rho,z) : |\rho - \hat{r}_j(x,a)| \le b_j^R(x,a), \\ \|z - \hat{T}_j(x,a)\|_2 \le b_j^T(x,a) \end{array} \right\}. \tag{14}$$

Thus, $\mathfrak{E}_j(x)$ contains all locally plausible one-step consequences of actions at state x. Under Assumption 1, on the good event the true consequence of every real action is included: for every $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$(a, r_\star(x,a), T_\star(x,a)) \in \mathfrak{E}_j(x). \tag{15}$$

Indeed, Assumption 1 gives $T_\star(x,a) \in \mathcal{X}$, while (12)–(13) verify the two confidence constraints defining $\mathfrak{E}_j(x)$.

5.3. Recoverable Correspondences and the Recoverable Core

The raw extended correspondence $\mathfrak{E}_j$ is the closest analogue of extended value iteration in UCRL2 (Jaksch et al. 2010): it lets the planner optimize over locally plausible one-step rewards and next states. In our continuous deterministic latent model, however, raw local optimism is not enough. The regret analysis must compare the value at an optimistic next state z with the value at the true next state $T_\star(x,a)$. A span bound controls this comparison only by a constant. To turn small transition-prediction error into small regret, the optimistic value functions must be Lipschitz. We therefore keep only the part of $\mathfrak{E}_j$ that preserves the recovery property responsible for this Lipschitz regularity.

Definition 1 (Recoverable core). Fix an epoch j. A correspondence $\mathfrak{K}$ with

$$\mathfrak{K}(x) \subseteq \mathfrak{E}_j(x) \qquad \text{for all } x \in \mathcal{X}$$

is (L_r, α)-recoverable if, for every $x, y \in \mathcal{X}$ and every $(a, \rho, z) \in \mathfrak{K}(x)$, there exists $(a', \rho', z') \in \mathfrak{K}(y)$ such that

$$\rho - \rho' \le L_r \|x - y\|_2 \quad \text{and} \quad \|z - z'\|_2 \le \alpha \|x - y\|_2.$$

Let $\mathfrak{R}_j$ denote the collection of all (L_r, α)-recoverable subcorrespondences of $\mathfrak{E}_j$. The recoverable core of $\mathfrak{E}_j$ is

$$\mathfrak{K}_j^{\max}(x) := \bigcup_{\mathfrak{K} \in \mathfrak{R}_j} \mathfrak{K}(x), \qquad x \in \mathcal{X}.$$

This condition says that if an extended action (a, ρ, z) is available at latent state x, then every latent state $y \in \mathcal{X}$ has a matching extended action whose reward is no more than $L_r \|x - y\|_2$ below ρ and whose next state is within $\alpha \|x - y\|_2$ of z. Because these tolerances scale with $\|x - y\|_2$, small state perturbations lead to small changes in the recovered one-step outcome. This is what yields Lipschitz regularity of the optimistic value functions. The reward condition is one-sided because the later value-regularity proof compares values one direction at a time; applying the same definition with x and y interchanged gives the reverse comparison when needed.

Define the true extended-action graph

$$\mathfrak{K}_\star(x) := \{(a, r_\star(x, a), T_\star(x, a)) : a \in \mathcal{A}\}.$$

On the good event, $\mathfrak{K}_\star(x) \subseteq \mathfrak{E}_j(x)$ for all x, by (15). Moreover, Assumption 6 implies that $\mathfrak{K}_\star$ is (L_r, α)-recoverable. The next proposition separates the two roles of the core: its recoverability is deterministic from the definition, while its containment of the true graph uses the good event.

Proposition 2 **(Recoverable core contains the true graph)**. *The correspondence $\mathfrak{K}_j^{\max}$ is (L_r, α)-recoverable and satisfies*

$$\mathfrak{K}_j^{\max}(x) \subseteq \mathfrak{E}_j(x) \qquad \textit{for every } x \in \mathcal{X}.$$

If Assumptions 1 and 6 hold, then, on the good event,

$$\mathfrak{K}_\star(x) \subseteq \mathfrak{K}_j^{\max}(x) \qquad \textit{for all } x \in \mathcal{X}.$$

The key point is that the union of recoverable subcorrespondences is recoverable: if an extended action belongs to the union, it belongs to one recoverable correspondence, and that correspondence provides the required recovery witness.

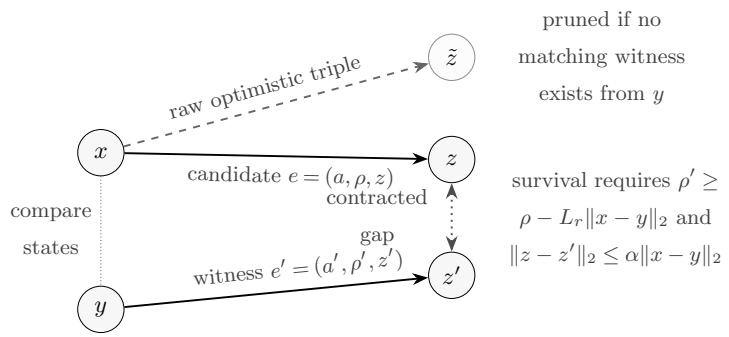


Figure 2 Recoverable-core intuition. The raw extended confidence model may contain optimistic one-step predictions with attractive next states. A triple survives in the recoverable core only if, for every comparison state y, a matching witness triple has comparable reward and a contracted next-state discrepancy. Raw triples without such witnesses are pruned.

Remark 3 (Finite-state implementation). When the latent state space and the extended-action sets are finite, the recoverable core has a concrete greatest-fixed-point interpretation. Starting from the raw extended correspondence, one repeatedly removes extended actions that lack recovery witnesses at some comparison state; Figure 2 illustrates this witness requirement. The terminal correspondence is the largest recoverable subcorrespondence. Section 5.5 uses this pruning construction as its implementable planning primitive. A direct implementation is polynomial in the number of grid states and candidate extended actions; see Remark 4 in Appendix F.

5.4. Continuous-Oracle Planning

We first state the idealized planner that has access to the continuous recoverable core and can solve the corresponding Bellman recursion.

Assumption 7 **(Recoverable-core planning oracle)**. *For each epoch j, the decision maker has access to an oracle that, given the raw extended correspondence $\mathfrak{E}_j$, returns its recoverable core $\mathfrak{K}_j^{\max}$ and solves the Bellman recursion below. We assume these Bellman optimization problems have finite optimal values and attain their suprema. Equivalently, all results may be stated for ϵ-optimal oracle calls, which add only an $O(H\epsilon)$ term to the regret.*

At epoch j, the optimistic planner computes $\widetilde{V}_{H+1}^j(x) := 0$, and, for $h = H, H-1, \ldots, \tau_j$,

$$\widetilde{V}_h^j(x) := \sup_{(a,\rho,z)\in\mathfrak{K}_j^{\max}(x)} \left[\rho + \widetilde{V}_{h+1}^j(z)\right]. \tag{16}$$

When the decision maker is at empirical state $\hat{x}_t$ in epoch j, it chooses an extended action (a_t, ρ_t, z_t) that attains the supremum in (16) and executes only the real action a_t. The pair (ρ_t, z_t) is used only for planning and analysis.

The following proposition records the properties of the continuous-oracle planner needed for regret analysis: optimism relative to the true dynamic program, Lipschitz regularity, and a span bound. Here L_{lat} and B_{lat} are the constants in (9), with L_r and α taken from Assumption 6.

PROPOSITION 3 **(Optimism and value regularity from the core)**. *Suppose Assumptions 1, 6, and 7 hold. On the good event, for every epoch j and stage h, the recoverable-core optimistic value functions satisfy*

$$\widetilde{V}_h^j(x) \geq V_h^\star(x) \qquad \textit{for all } x \in \mathcal{X},$$
$$|\widetilde{V}_h^j(x) - \widetilde{V}_h^j(y)| \leq L_{\mathrm{lat}}\|x-y\|_2 \qquad \textit{for all } x, y \in \mathcal{X},$$
$$\mathrm{span}(\widetilde{V}_h^j) \leq B_{\mathrm{lat}}.$$

This result isolates the role of the recoverable core: it gives optimism, Lipschitz regularity, and a horizon-uniform span bound while requiring recoverability only of the true latent dynamics. Consequently, if the continuous oracle were used directly with the lazy epoch convention, the same accounting as Corollary 1 would give the same statistical terms, with Assumption 5 replaced by Assumptions 6 and 7. We therefore do not state a separate continuous-oracle regret theorem. The implementable finite-grid version below adds the discretization term explicitly.

5.5. Finite-Grid Implementation

To approximate the continuous core, the implementable planner projects empirical states, restricts actions, rewards, and optimistic next states to finite grids, inflates confidence constraints for projection error, and prunes the resulting correspondence. The regret theorem below makes the discretization cost explicit.

Let $\mathcal{X}_\eta \subset \mathcal{X}$ and $\mathcal{A}_\eta \subset \mathcal{A}$ be finite state and action grids, let $\mathcal{R}_\eta$ be a finite reward grid, and let $\Pi_X : \mathbb{R}^m \to \mathcal{X}_\eta$ be a nearest-grid projection. For resolutions $\eta = (\eta_x, \eta_a, \eta_\rho, \eta_z)$, set

$$\bar{x}_t := \Pi_X(\hat{x}_t), \quad \beta_\eta^x := \beta^x + \eta_x, \quad \Delta_\eta := \eta_x + \eta_a + \eta_\rho + \eta_z, \quad \eta_z = \eta_x, \quad \varepsilon_R^\eta := \eta_\rho, \quad \varepsilon_T^\eta := \eta_x.$$

On the state-confidence event, $\|x_t - \bar{x}_t\|_2 \leq \beta_\eta^x$. The radii ε_R^η and ε_T^η, both bounded by Δ_η, absorb reward and next-state projection errors.

At the start of each epoch, the algorithm forms the inflated finite raw correspondence $\mathfrak{E}_{j,\eta}$, prunes it to its largest additively recoverable core $\mathfrak{K}_{j,\eta}^{\max}$, and computes finite-grid optimistic values $\widetilde{V}_{h,\eta}^j : \mathcal{X}_\eta \to \mathbb{R}$. Starting from $\widetilde{V}_{H+1,\eta}^j \equiv 0$, the algorithm computes

$$Q_{h,\eta}^j(x) := \max_{(a,\rho,z)\in\mathfrak{K}_{j,\eta}^{\max}(x)} \left\{\rho + \widetilde{V}_{h+1,\eta}^j(z)\right\},$$
$$\widetilde{V}_{h,\eta}^j(x) := \max_{y\in\mathcal{X}_\eta} \left\{Q_{h,\eta}^j(y) - L_{\mathrm{lat}}\|x-y\|_2\right\}, \qquad x \in \mathcal{X}_\eta.$$

The first line is the raw finite-grid backup; the second is the Lipschitz envelope, which preserves the value regularity needed to convert transition prediction error into regret. We call the resulting method *Recoverable-Core Latent UCB*. Algorithm 1 summarizes its finite-grid implementation. Appendix F gives the exact finite correspondence, pruning operator, envelope recursion, and full proof.

The finite-grid analysis uses the following additive grid-approximation condition. Informally, it says that a discretized approximation of the true reward-transition graph remains feasible for the inflated finite raw correspondence and remains recoverable up to $O(\Delta_\eta)$ additive error.

ASSUMPTION 8 **(Additive finite-grid approximation of the true recoverable graph)**. *For each epoch j, on the good event, there exists a finite-grid correspondence $\mathfrak{K}_{\star,j,\eta}$ that is additively recoverable on $\mathcal{X}_\eta$ with parameters $(L_r, \alpha, \varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T})$, such that $\mathfrak{K}_{\star,j,\eta}(x) \subseteq \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$. The additive slacks satisfy*

$$\varepsilon^\eta_{\mathrm{rec},R} \le c^\eta_{\mathrm{rec},R}\Delta_\eta, \qquad \varepsilon^\eta_{\mathrm{rec},T} \le c^\eta_{\mathrm{rec},T}\Delta_\eta,$$

for constants independent of H. Moreover, there is a constant C^η_{app}, independent of H, such that for every continuous state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, if $\bar{x} = \Pi_X x$, then there exists $(\bar{a}, \bar{\rho}, \bar{z}) \in \mathfrak{K}_{\star,j,\eta}(\bar{x})$ satisfying

$$\bar{\rho} \ge r_\star(x,a) - C^\eta_{\mathrm{app}}\Delta_\eta, \qquad \|\bar{z} - \Pi_X T_\star(x,a)\|_2 \le C^\eta_{\mathrm{app}}\Delta_\eta.$$

Algorithm 1 Recoverable-Core Latent UCB: Finite-Grid Summary

Require: Horizon H, confidence parameters, grids $\mathcal{X}_\eta, \mathcal{A}_\eta, \mathcal{R}_\eta$, recovery constants L_r, α.

1: Initialize reward and transition regressions and start epoch $j = 1$.

2: At the start of epoch j, form the finite raw correspondence $\mathfrak{E}_{j,\eta}$, compute its recoverable core $\mathfrak{K}^{\max}_{j,\eta}$ by Algorithm 2, and compute the finite-grid optimistic values.

3: **for** $t = 1, \dots, H$ **do**

4: Observe a state-estimation batch, compute $\hat{x}_t$, and project $\bar{x}_t = \Pi_X(\hat{x}_t)$.

5: If a reward or transition determinant has doubled, start a new epoch and recompute $\mathfrak{E}_{j,\eta}$, $\mathfrak{K}^{\max}_{j,\eta}$, and the optimistic values.

6: Choose

$$(a_t, \rho_t, z_t) \in \arg\max_{(a,\rho,z) \in \mathfrak{K}^{\max}_{j,\eta}(\bar{x}_t)} \big[\rho + \widetilde{V}^j_{t+1,\eta}(z)\big],$$

and execute only the real action a_t.

7: Observe the reward batch and update the reward regression; when the next projected state becomes available, update the transition regression.

8: **end for**

A simple sufficient condition is available under the feature Lipschitzness already imposed in Section 2. Since the reward and transition are linear in Ψ and Φ, Assumptions 2 and 3 imply that $r_\star$ and $T_\star$ are Lipschitz in both state and action. If the finite raw correspondence is inflated enough to absorb reward and next-state projection error, for example $\varepsilon_R^\eta \geq \eta_\rho$ and $\varepsilon_T^\eta \geq \eta_x$, then Assumption 8 holds with additive recovery and approximation errors of order Δ_η. Proposition 7 in Appendix F gives the formal statement.

The regret bound uses the same quantities $\gamma_R(H), \gamma_T(H), \kappa_R, \kappa_T, \bar{\beta}_H^R, \bar{\beta}_H^T$, and C_x introduced in Section 4. In the finite-grid analysis, the confidence radii are computed with β^x replaced by β_η^x. Thus the state-projection radius η_x enters the statistical width terms through $\bar{\beta}_H^R$ and $\bar{\beta}_H^T$, while the final $C_{\text{grid}}^\eta H\Delta_\eta$ term below is the additional Bellman-approximation cost of finite-grid planning.

THEOREM 1 **(Regret of Recoverable-Core Latent UCB with finite-grid planning).** *Suppose Assumptions 1, 2, 3, 4, 6, and 8 hold. Let $\delta_x + \delta_R \leq \delta$. Then, with probability at least $1-\delta$, the finite-grid implementation of Recoverable-Core Latent UCB summarized in Algorithm 1 satisfies*

$$\begin{aligned}\mathrm{Reg}(H) \leq\ & 2B_{\text{lat}}\left(1 + \frac{\gamma_R(H)}{\log 2} + \frac{\gamma_T(H)}{\log 2}\right) \\ & + 2\bar{\beta}_H^R\sqrt{2H\kappa_R\gamma_R(H)} + 2L_{\text{lat}}\bar{\beta}_H^T\sqrt{2H\kappa_T\gamma_T(H)} + C_x H\beta^x + C_{\text{grid}}^\eta H\Delta_\eta.\end{aligned}$$

Here C_{grid}^η is a finite constant depending on the discretization approximation constants and fixed problem parameters, but not on H. Moreover,

$$\gamma_R(H) \leq p\log\left(1 + \frac{HB_\Psi^2}{\lambda_R p}\right), \qquad \gamma_T(H) \leq q\log\left(1 + \frac{HB_\Phi^2}{\lambda_T q}\right).$$

As in the continuous-oracle analysis, elliptical potentials control reward and transition uncertainty, determinant doubling controls value-function switches, and finite-batch state estimation contributes $H\beta^x$. Finite-grid planning adds only $C_{\text{grid}}^\eta H\Delta_\eta$; see Appendix G. Thus, suppressing logarithmic factors and fixed constants,

$$\mathrm{Reg}(H) = \widetilde{O}\left(\sqrt{pH} + p\sqrt{\frac{H}{n}} + L_{\text{lat}}\sqrt{mqH} + B_{\text{lat}}(p+q) + H\beta^x\left[p + L_{\text{lat}}q + 1\right] + H\Delta_\eta\right).$$

COROLLARY 2 **(Sublinear regret under growing batch sizes and refining grids).** *Suppose the assumptions of Theorem 1 hold along a sequence of grids, with the constants in the grid-approximation condition uniformly bounded in H. If the per-period batch size n and grid resolution η, both allowed to scale with H, satisfy*

$$n \gg (m + \log H)\log H \qquad \textit{and} \qquad \Delta_\eta\sqrt{\log H} = o(1),$$

then, for any fixed $\delta \in (0,1)$, there is a deterministic sequence $\varepsilon_H(\delta) \to 0$ such that, for every sufficiently large H,

$$\Pr\left(\frac{\mathrm{Reg}(H)}{H} \leq \varepsilon_H(\delta)\right) \geq 1 - \delta.$$

For example, if all grid resolutions are of order $1/\log H$, then $\Delta_\eta\sqrt{\log H} = o(1)$. For fixed dimensions, a uniform grid with this resolution has size polynomial in $\log H$.

6. Lower Bounds

The principal result below concerns structural necessity within the controlled latent-dynamics framework. For comparison, Proposition 9 in Appendix H isolates the effect of finite-batch observation. In a sequence of independent one-step latent population problems, it gives an $\Omega(H/\sqrt{n})$ lower bound against a state-observing benchmark. This auxiliary result explains how an $H\beta^x$-type cost can arise when every period requires a fresh state estimate; it is not a minimax lower bound for the deterministic controlled-transition model, where past observations and the learned transition can provide additional information.

6.1. No Safe Exploration Without Recoverability

We show that without a recovery-type structural condition preventing irreversible dynamics, sublinear single-trajectory regret is impossible in general. This result is not a necessity theorem for the exact recoverable-core algorithm above. Rather, it shows that the unrestricted class contains instances in which an early action can irreversibly determine the future population state. In such instances, the decision maker cannot both explore the model and preserve the ability to recover, and linear regret is unavoidable.

THEOREM 2 **(No safe exploration without recoverability)**. *There exists a class of linear latent performative control instances with continuous action space $\mathcal{A} = [-1, 1]$ such that every algorithm suffers linear regret:*

$$\sup_M \mathbb{E}_M[\mathrm{Reg}(H)] \geq \frac{H-1}{2}.$$

Here the supremum is over instances M in the constructed class.

The lower bound has the following safe-exploration interpretation: before the decision maker acts, two models are observationally indistinguishable, and any action that attempts to resolve the model uncertainty also irreversibly determines the future latent state. Thus the decision maker cannot both learn the model and preserve the ability to recover.

7. Numerical Experiments

We report two complementary numerical studies. The first is a controlled synthetic experiment that isolates the value of recoverability-aware optimism relative to plug-in learning and raw optimism. The second uses a reference-price pricing environment calibrated from retail data to evaluate the operational value of planning for an endogenous, contractive reference-price state.

7.1. Synthetic Evaluation of Recoverable-Core Optimism

The synthetic experiment compares CE / plug-in, Raw UCB, and recoverable-core UCB in a one-dimensional instance where exploration is useful but unconstrained transition optimism can be misleading. The true system is

$$x_{t+1} = 0.7x_t + 0.1a_t, \qquad r_\star(x,a) = x + 1.2a - 2a^2, \qquad a_t \in [-1,1].$$

The action has a relatively weak effect on the next state but a stronger immediate reward effect, making the problem sensitive to reward-action excitation. A plug-in method initialized with an uninformative action-reward coefficient tends to choose actions near zero, which gives little information about that coefficient. In contrast, UCB methods choose nonzero actions because of optimism.

The algorithms observe the latent state directly, so the comparison isolates reward learning, transition optimism, and recoverability-aware planning rather than finite-batch state estimation. The reward model is correctly specified, whereas the transition feature map contains a localized nuisance direction whose true coefficient is zero. This direction does not change the true dynamics or the full-information benchmark, but it creates a plausible optimistic transition early in learning. CE / plug-in plans using the current point estimates, Raw UCB optimizes over plausible reward-transition triples, and recoverable-core UCB uses the same confidence information while removing triples that cannot be matched recoverably across states.

Regret is measured against the closed-form full-information finite-horizon benchmark. We consider ten horizons from 400 to 12,800, with 15 seed offsets and seven shifted action grids at each horizon. Independent random seeds are used across shifts, yielding 105 independent runs per algorithm–horizon pair. Appendix A.1 gives the benchmark derivation, exact feature maps, and grid construction. Appendix A also reports a diagnostic separating statistical learning error from finite-grid approximation error.

Figure 3 reports median cumulative regret with interquartile bands. CE / plug-in suffers from insufficient action excitation. Raw UCB explores aggressively but pays for unconstrained transition optimism. Recoverable-core UCB has lower regret by preserving useful optimism while imposing recoverability-aware restrictions.

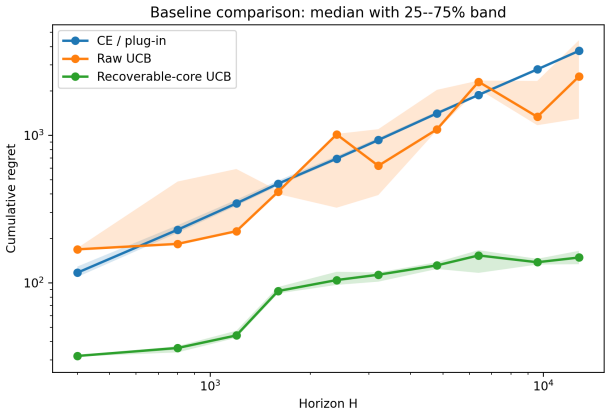


Figure 3 Median cumulative regret with interquartile bands in the exploration-sensitive synthetic instance. Recoverable-core UCB preserves reward optimism while excluding nonrecoverable transition predictions.

7.2. Data-Calibrated Reference-Price Experiment

We use Dominick's scanner data, widely used to study retail demand and store-level price sensitivity (Kilts Center for Marketing 2026, Hoch et al. 1995, Montgomery 1997), to calibrate a dynamic-pricing environment. Unlike the synthetic study, which isolates recoverable-core pruning, this experiment asks whether planning through endogenous reference-price dynamics improves on certainty-equivalent and myopic optimistic pricing. The design is semi-synthetic because counterfactual prices generate their own state and demand trajectories. Appendix B provides implementation details.

Environment and calibration. For product i, store s, and week t, let $p_{i,s,t}$ denote price, $r_{i,s,t}$ the consumer reference price, and $c_{i,s,t}$ a predictable context capturing seasonal, temporal, and store-level demand variation. Before applying a negligible numerical floor, conditional mean demand is

$$\widetilde{\mu}_{i,s,t} = c_{i,s,t}^{\top}\beta_i + b_{0,i} - a_i p_{i,s,t} - q_i p_{i,s,t}^2 + \eta_i(r_{i,s,t} - p_{i,s,t}).$$

The direct price terms allow nonlinear demand over the historically supported price range (Montgomery and Bradlow 1999). When $\eta_i > 0$, demand increases when price is below the reference price and decreases when it is above it. Observed demand adds a mean-zero Gaussian shock to the floored mean, and expected profit equals price-cost margin times that mean. The floor binds in less than 3% of test periods for every learned policy; Appendix B.2 reports its construction and the realized-profit model.

Reference prices evolve according to

$$r_{i,s,t+1} = \rho_i r_{i,s,t} + (1-\rho_i)p_{i,s,t}, \qquad 0 < \rho_i < 1,$$

a standard memory-based specification (Kalyanaram and Winer 1995, Briesch et al. 1997, Popescu and Wu 2007, Chen et al. 2017). Thus $x_t = r_t$ and $a_t = p_t$, while the calendar-driven context is known. The state is observed directly, so the study isolates reward and transition learning from state-estimation error. Contractivity makes current price affect both current demand and the reference state entering future demand. Off the floor, expected profit is a known context-dependent offset plus a term linear in $\vartheta_i = (b_{0,i}, a_i, q_i, \eta_i)^\top$; Appendix B.2 gives the explicit reward and transition feature maps.

For each UPC, profile least squares fits a pooled product-level model across stores. Weeks 1–239, 240–299, and 300–399 form the training, validation, and test periods, respectively. The training fit calibrates the simulator. The contextual coefficient is treated as known, whereas every learned policy re-estimates ϑ_i online and the dynamic policies also re-estimate ρ_i along their endogenous trajectories.

Mechanism-relevant instances. Data-support requirements leave 211 of the original 369 UPCs; validation fit, positive reference-price effects, decreasing fitted demand on the feasible price interval, and limited floor activation leave 39 qualified UPCs. Appendix B.5 gives the complete screening sequence. For each UPC, the validation-only procedure in Appendix B.6 selects one eligible store with operationally relevant reference-price feedback. The resulting instances are frozen before testing, and no test outcome is used for qualification, store selection, tuning, or removal. Reported performance is therefore conditional on this validation-selected population; it does not estimate gains across all Dominick's stores or the prevalence of consequential reference-price feedback.

Policies and evaluation. Four learned policies form a two-by-two comparison between dynamic and myopic planning and between certainty equivalence and optimism. *Myopic CE* maximizes estimated current profit. *Myopic OFU* adds reward optimism, as in linear contextual bandits (Abbasi-Yadkori et al. 2011), but assigns no continuation value to the state induced by the current price. *Dynamic CE* plans over the remaining horizon using point estimates of the reward and transition parameters. *Dynamic Raw UCB* (abbreviated *Dynamic Raw*) uses the same observation protocol and estimators but plans optimistically over both confidence sets. It is the pricing analogue of the full-model optimistic planner in Section 4. A *Dynamic Oracle* knows the calibrated parameters and provides a full-information benchmark.

All policies face the same context path, historical initial reference state, price set, warm-start observations, and matched demand-noise realizations. After the warm start, each learned policy updates from its own endogenous trajectory. Each policy is evaluated on the 39 frozen instances over the 100-week test period using 20 matched noise seeds. Expected profit is averaged across

seeds within each UPC, and paired Student-t intervals are then computed across the 39 UPC-level differences.

Results. Table 1 reports the principal comparisons. W/T/L gives product-level wins, ties, and losses for the first policy, and relative gain is normalized by the second policy's mean expected profit.

Table 1 Principal paired expected-profit comparisons across the 39 validation-selected product–store instances. Expected profit is first averaged over the 20 matched seeds within each instance.

Comparison	Mean gain	95% CI	W/T/L	Relative gain
Dynamic CE–Myopic CE	355.262	[246.325, 464.199]	36/1/2	10.812%
Dynamic Raw–Dynamic CE	61.745	[24.528, 98.962]	34/0/5	1.696%
Dynamic Raw–Myopic OFU	347.658	[230.099, 465.218]	37/0/2	10.361%

Dynamic CE outperforms Myopic CE by 10.812%, showing the value of planning for future reference prices without optimism. Dynamic Raw improves on Dynamic CE by a further 1.696% and outperforms the reward-optimistic myopic benchmark by 10.361%, with wins on 37 of the 39 instances. All three paired confidence intervals exclude zero. Appendix B.10 reports policy-level profits, floor activation, and boundary-action diagnostics.

Thus, within the validation-selected population, endogenous reference-price dynamics make dynamic planning operationally important; the synthetic study separately identifies the value of recoverability-aware pruning.

8. Conclusion

This paper develops a regret-based framework for online decisions along one evolving population, using a finite-dimensional latent statistic estimated from finite batches while unknown reward and transition laws are learned. Full-model optimism yields sublinear regret under robust recovery. Recoverable-core optimism requires only the true dynamics to be recoverable, preserves their reward-transition graph, and yields an implementable finite-grid planner with the same statistical learning terms plus discretization error.

A perhaps counterintuitive implication is that a larger optimistic set need not produce better decisions: plausible but nonrecoverable next states can destroy value regularity and move the population to hard-to-recover regions. Core pruning excludes such states without sacrificing optimism. Conversely, the lower bound shows that irreversible dynamics can force linear regret absent recovery-type structure. The numerical studies separately isolate the benefits of core pruning and of planning through endogenous population dynamics.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, volume 24, 2312–2320.

Abbasi-Yadkori Y, Szepesvári C (2011) Regret bounds for the adaptive control of linear quadratic systems. *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research* (PMLR), 1–26.

Agarwal A, Jiang N, Kakade SM, Sun W (2019) Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep* 32:96.

Auer P, Ortner R (2007) Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, volume 19 (MIT Press), 49–56.

Ayoub A, Jia Z, Szepesvári C, Wang M, Yang LF (2020) Model-based reinforcement learning with value-targeted regression. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (PMLR), 463–474.

Azar MG, Osband I, Munos R (2017) Minimax regret bounds for reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research* (PMLR), 263–272.

Basciftci B, Ahmed S, Shen S (2021) Distributionally robust facility location problem under decision-dependent stochastic demand. *European Journal of Operational Research* 292(2):548–561.

Bäuerle N (2023) Mean field markov decision processes. *Applied Mathematics & Optimization* 88(1):12.

Bayraktar E, Kara AD (2025) Learning with linear function approximations in mean-field control. *Journal of Machine Learning Research* 26:1–53.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press).

Boffi NM, Tu S, Slotine JJE (2021) Regret bounds for adaptive nonlinear control. *Proceedings of the Third Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research* (PMLR), 471–483, URL https://proceedings.mlr.press/v144/boffi21a.html.

Bracale D, Maity S, Banerjee M, Sun Y (2024) Learning the distribution map in reverse causal performative prediction. *arXiv preprint arXiv:2405.15172* .

Brafman RI, Tennenholtz M (2002) R-MAX: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3:213–231.

Briesch RA, Krishnamurthi L, Mazumdar T, Raj SP (1997) A comparative analysis of reference price models. *Journal of Consumer Research* 24(2):202–214, URL http://dx.doi.org/10.1086/209505.

Brown G, Hod S, Kalemaj I (2022) Performative prediction in a stateful world. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research* (PMLR), 6045–6061.

Brown W, Papadimitriou C, Roughgarden T (2024) Online stackelberg optimization via nonlinear control. *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research* (PMLR), 697–749, URL https://proceedings.mlr.press/v247/brown24a.html.

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

Chen X, Hu P, Hu Z (2017) Efficient algorithms for the dynamic pricing problem with reference price effect. *Management Science* 63(12):4389–4408, URL http://dx.doi.org/10.1287/mnsc.2016.2554.

Chen Y, Tang W, Ho CJ, Liu Y (2024) Performative prediction with bandit feedback: Learning through reparameterization. *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research* (PMLR), 7298–7324.

Chu W, Li L, Reyzin L, Schapire RE (2011) Contextual bandits with linear payoff functions. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research* (PMLR), 208–214.

Cohen A, Koren T, Mansour Y (2019) Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research* (PMLR), 1300–1309.

Cutler J, Díaz M, Drusvyatskiy D (2024) Stochastic approximation with decision-dependent distributions: Asymptotic normality and optimality. *Journal of Machine Learning Research* 25(90):1–49, arXiv:2207.04173.

Dean S, Mania H, Matni N, Recht B, Tu S (2018) Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, volume 31.

Drusvyatskiy D, Xiao L (2023) Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research* 48(2):954–998, URL http://dx.doi.org/10.1287/moor.2022.1287.

Goel V, Grossmann IE (2006) A class of stochastic programs with decision dependent uncertainty. *Mathematical Programming* 108(2):355–394.

Gu H, Guo X, Wei X, Xu R (2022) Dynamic programming principles for mean-field controls with learning. *arXiv preprint arXiv:1911.07314* .

Hardt M, Mendler-Duenner C (2023) Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608* .

He Z, Bolognani S, Dörfler F, Muehlebach M (2025) Decision-dependent stochastic optimization: The role of distribution dynamics. *arXiv preprint arXiv:2503.07324* .

Hoch SJ, Kim BD, Montgomery AL, Rossi PE (1995) Determinants of store-level price elasticity. *Journal of Marketing Research* 32(1):17–29, URL http://dx.doi.org/10.1177/002224379503200104.

Horn RA, Johnson CR (2012) *Matrix Analysis*, 2 ed. (Cambridge University Press).

Izzo Z, Ying L, Zou J (2021) How to learn when data reacts to your model: Performative gradient descent. *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (PMLR).

Izzo Z, Ying L, Zou J (2022) How to learn when data gradually reacts to your model. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research* (PMLR), 3998–4035.

Jagadeesan M, Zrnic T, Mendler-Duenner C (2022) Regret minimization with performative feedback. *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research* (PMLR), 9760–9785.

Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11:1563–1600.

Jia Z, Wang Y, Dong R, Hanasusanto GA (2025) Distributionally robust performative optimization. *arXiv preprint arXiv:2407.01344* .

Jin C, Yang Z, Wang Z, Jordan MI (2020) Provably efficient reinforcement learning with linear function approximation. *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research* (PMLR), 2137–2143.

Jin Q, Georghiou A, Vayanos P, Hanasusanto GA (2024) Distributionally robust optimization with decision-dependent information discovery. *arXiv preprint arXiv:2404.05900* .

Jonsbraten TW, Wets RJ, Woodruff DL (1998) A class of stochastic programs with decision dependent random elements. *Annals of Operations Research* 82:83–106.

Kalyanaram G, Winer RS (1995) Empirical generalizations from reference price research. *Marketing Science* 14(3, Supplement):G161–G169, URL http://dx.doi.org/10.1287/mksc.14.3.G161.

Kearns M, Singh S (2002) Near-optimal reinforcement learning in polynomial time. *Machine Learning* 49(2–3):209–232.

Kehrenberg T, Sanguino J, Lozano JA, Quadrianto N (2026) Dissecting performative prediction: A comprehensive survey. *arXiv preprint arXiv:2602.10176* .

Kilts Center for Marketing (2026) Dominick's dataset. University of Chicago Booth School of Business, URL https://www.chicagobooth.edu/research/kilts/research-data/dominicks, accessed August 5, 2026.

Kuhn D, Shafiee S, Wiesemann W (2025) Distributionally robust optimization. *Acta Numerica* 34:579–804.

Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press).

Lee J, Zrnic T (2026) Partially performative prediction. *arXiv preprint arXiv:2606.07890* .

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web* (ACM), 661–670.

Li Q, Wai HT (2022) State dependent performative prediction with stochastic approximation. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research* (PMLR), 3164–3186.

Luo F, Mehrotra S (2020) Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Letters* 14(8):2565–2594.

Mandal D, Radanovic G (2025) Performative reinforcement learning with linear markov decision process. *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research* (PMLR), 3232–3240, arXiv:2411.05234.

Mandal D, Triantafyllou S, Radanovic G (2023) Performative reinforcement learning. *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research* (PMLR), 23642–23680.

Mendler-Duenner C, Perdomo J, Zrnic T, Hardt M (2020) Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, volume 33, 4929–4939.

Miller JP, Perdomo JC, Zrnic T (2021) Outside the echo chamber: Optimizing the performative risk. *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research* (PMLR), 7710–7720.

Mofakhami M, Mitliagkas I, Gidel G (2023) Performative prediction with neural networks. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research* (PMLR).

Montgomery AL (1997) Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science* 16(4):315–337, URL http://dx.doi.org/10.1287/mksc.16.4.315.

Montgomery AL, Bradlow ET (1999) Why analyst overconfidence about the functional form of demand can lead to overpricing. *Marketing Science* 18(4):569–583, URL http://dx.doi.org/10.1287/mksc.18.4.569.

Munos R, Moore A (2002) Variable resolution discretization in optimal control. *Machine Learning* 49:291–323.

Nguyen HA, Cheng CA (2023) Provable reset-free reinforcement learning by no-regret reduction. *Proceedings of the Fortieth International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research* (PMLR), 25939–25955, URL https://proceedings.mlr.press/v202/nguyen23b.html.

Nohadani O, Sharma K (2018) Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization* 28(2):1773–1795.

Noyan N, Rudolf G, Lejeune M (2022) Distributionally robust optimization under a decision-dependent ambiguity set with applications to machine scheduling and humanitarian logistics. *INFORMS Journal on Computing* 34(2):729–751, URL http://dx.doi.org/10.1287/ijoc.2021.1096.

Ortner R, Ryabko D (2012) Online regret bounds for undiscounted continuous reinforcement learning. *Advances in Neural Information Processing Systems*, volume 25, 1772–1780.

Osband I, Van Roy B (2014) Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, volume 27.

Park S, Kwon J, Kim B, Chae S, Lee J, Lee D (2024) Parameter-free algorithms for performative regret minimization under decision-dependent distributions. *arXiv preprint arXiv:2402.15188* .

Pásztor B, Bogunovic I, Krause A (2023) Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050* .

Perdomo J, Zrnic T, Mendler-Duenner C, Hardt M (2020) Performative prediction. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (PMLR), 7599–7609.

Pham H, Wei X (2016) Discrete time McKean–Vlasov control problem: A dynamic programming approach. *Applied Mathematics & Optimization* 74(3):487–506.

Popescu I, Wu Y (2007) Dynamic pricing strategies with reference effects. *Operations Research* 55(3):413–429, URL http://dx.doi.org/10.1287/opre.1070.0393.

Powell WB (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (Wiley).

Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley).

Rahimian H, Mehrotra S (2019) Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659* .

Rank B, Triantafyllou S, Mandal D, Radanovic G (2024) Performative reinforcement learning in gradually shifting environments. *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research* (PMLR), 3041–3075.

Rodemann J, Fischer-Abaigar U, Bailie J, Muandet K (2026) Performative learning theory. *arXiv preprint arXiv:2602.04402* ICML 2026.

Russo D, Van Roy B (2013) Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, volume 26.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM).

Simchowitz M, Foster DJ (2020) Naive exploration is optimal for online LQR. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (PMLR), 8937–8948.

Song Z, Sun W (2019) Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475* .

Tsybakov AB (2009) *Introduction to Nonparametric Estimation*. Springer Series in Statistics (Springer, New York), URL http://dx.doi.org/10.1007/b13794.

Wang S, Wang Z, Johansson KH (2026) Wasserstein distributionally robust performative prediction. *arXiv preprint arXiv:2602.06730* .

Xue S, Sun Y (2024) Distributionally robust performative prediction. *Advances in Neural Information Processing Systems* 37:55030–55052.

Yang LF, Wang M (2020) Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (PMLR), 10746–10756.

Yu X, Basciftci B (2026) Distributionally robust optimization with multimodal decision-dependent ambiguity sets. *Mathematical Programming* URL http://dx.doi.org/10.1007/s10107-026-02337-1.

Yu X, Shen S (2022) Multistage distributionally robust mixed-integer programming with decision-dependent moment-based ambiguity sets. *Mathematical Programming* 196(1):1025–1064.

Zanette A, Lazaric A, Kochenderfer MJ, Brunskill E (2020) Learning near optimal policies with low inherent bellman error. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (PMLR), 10978–10989.

Zhu Q, Yu X, Bayraksan G (2026) Residuals-based contextual distributionally robust optimization with decision-dependent uncertainty: Theoretical guarantees and decomposition algorithm. *arXiv preprint arXiv:2406.20004* .

Appendix A: Numerical Experiment Details and Additional Diagnostics

This appendix gives implementation details for the synthetic experiment in Section 7.1 and reports an additional diagnostic examining the interaction between statistical learning error and finite-grid approximation error.

A.1. Main Synthetic Experiment

The synthetic experiments use the benchmark family $x_{t+1} = \gamma x_t + \kappa a_t$, $a_t \in [-1, 1]$, with reward $r_\star(x, a) = qx + \mu a - \lambda a^2$. This family admits a closed-form full-information finite-horizon benchmark. Let $V^\star_{H+1}(x) = 0$. Because the transition is affine and the reward is linear in x, the value function is affine: $V^\star_h(x) = p_h x + c_h$. Its Bellman recursion is

$$\begin{aligned} V^\star_h(x) &= \max_{a\in[-1,1]} \left\{ qx + \mu a - \lambda a^2 + V^\star_{h+1}(\gamma x + \kappa a) \right\} \\ &= \max_{a\in[-1,1]} \left\{ qx + \mu a - \lambda a^2 + p_{h+1}(\gamma x + \kappa a) + c_{h+1} \right\}. \end{aligned}$$

Therefore $p_h = q + \gamma p_{h+1}$, with $p_{H+1} = 0$, and hence $p_h = q(1 - \gamma^{H-h+1})/(1 - \gamma)$. The optimal action at stage h is

$$a^\star_h = \mathrm{clip}_{[-1,1]}\left(\frac{\mu + \kappa p_{h+1}}{2\lambda} \right),$$

where $\mathrm{clip}_{[-1,1]}(u) := \min\{1, \max\{-1, u\}\}$. The constant term satisfies $c_h = c_{h+1} + \mu a^\star_h - \lambda (a^\star_h)^2 + \kappa p_{h+1} a^\star_h$. We compute regret as $\mathrm{Reg}(H) = V^\star_1(x_1) - \sum_{t=1}^H r_\star(x_t, a_t)$ along the true trajectory generated by each decision rule.

To create statistically plausible transition directions that are absent from the true system, the decision maker's feature map may include localized nuisance features

$$\xi_k(x, a) = \exp\left\{ -\frac{1}{2}\left(\frac{(x - \bar{x}_k)^2}{s_x^2} + \frac{(a - \bar{a}_k)^2}{s_a^2} \right) \right\}, \qquad k = 1, \ldots, K.$$

Their true coefficients are zero, so they do not alter the true reward, transition, or benchmark. They nevertheless create uncertainty in directions that the decision maker has not yet ruled out.

The main experiment uses $\gamma = 0.7$, $\kappa = 0.1$, $q = 1$, $\mu = 1.2$, and $\lambda = 2$. Its reward feature map is correctly specified as $\Psi(x, a) = (x, a, a^2)$. The transition feature map and true coefficient are

$$\Phi(x, a) = (x, a, \xi(x, a)), \qquad W_\star = (0.7, 0.1, 0),$$

where the nuisance feature is centered near $(0, 1)$:

$$\xi(x, a) = \exp\left\{ -\frac{1}{2}\left(\frac{x^2}{0.20^2} + \frac{(a-1)^2}{0.20^2} \right) \right\}.$$

We use horizons $H \in \{400, 800, 1200, 1600, 2400, 3200, 4800, 6400, 9600, 12800\}$, 15 seed offsets, and shifted action grids with shifts $0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75$. Independent random seeds

are used across shifts, yielding 105 independent runs per algorithm–horizon pair. The grid size follows $N(H) \asymp H^{1/2}$, consistent with the diagnostic below.

A.2. Grid/Statistical Decomposition

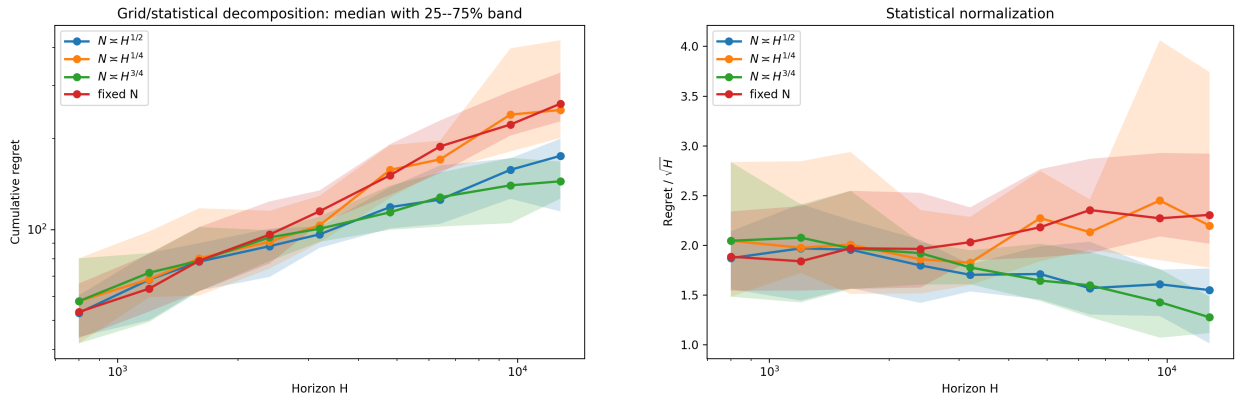
We test the finite-grid recoverable-core method on the one-dimensional system $x_{t+1} = 0.7x_t + 0.3a_t$, $a_t \in [-1, 1]$, with reward $r_\star(x, a) = x - 2a^2$. Thus $\gamma = 0.7$, $\kappa = 0.3$, $q = 1$, $\mu = 0$, and $\lambda = 2$. The absence of an immediate linear action reward means that nonzero actions are costly in the current period and are useful only through their effect on future latent states, making the instance a clean diagnostic for dynamic planning and grid approximation. The decision maker uses an overparameterized model: the reward feature map includes 12 localized bump features with zero true coefficients, and the transition feature map includes the same localized nuisance features with zero true transition coefficients.

In one dimension, an action grid of size N suggests the decomposition $\mathrm{Reg}(H, N) \lesssim C_{\mathrm{stat}}\sqrt{H} + C_{\mathrm{grid}}H/N$, so the balancing schedule is $N(H) \asymp H^{1/2}$. We compare fixed N_0, $H^{1/4}$, $H^{1/2}$, and $H^{3/4}$ action-grid schedules over horizons 800, 1200, 1600, 2400, 3200, 4800, 6400, 9600, and 12800, using 15 independent reward-noise seeds and shifted action grids with shifts 0, 0.25, 0.5, and 0.75. The same confidence-set construction, lazy epoch rule, and recoverable-core planning rule are used for all schedules.

Figure 4 reports cumulative regret and regret normalized by $\sqrt{H}$. Fixed and slowly growing grids incur larger regret at large horizons. After normalization by $\sqrt{H}$, the fixed and $H^{1/4}$ schedules continue to grow, while the $H^{1/2}$ and $H^{3/4}$ schedules are approximately stabilized. This is consistent with

$$\frac{\mathrm{Reg}(H, N)}{\sqrt{H}} \approx C_{\mathrm{stat}} + C_{\mathrm{grid}}\frac{\sqrt{H}}{N},$$

and supports the balancing choice $N(H) \asymp H^{1/2}$.



(a) Cumulative regret. (b) Regret normalized by $\sqrt{H}$.

Figure 4 Grid/statistical decomposition for recoverable-core UCB. Panel (a) shows median cumulative regret with interquartile bands. Panel (b) normalizes regret by $\sqrt{H}$. Under $\mathrm{Reg}(H,N) \approx C_{\mathrm{stat}}\sqrt{H} + C_{\mathrm{grid}}H/N$, fixed and $H^{1/4}$-level grids continue to grow after normalization, whereas $N(H) \asymp H^{1/2}$ and $N(H) \asymp H^{3/4}$ enter the statistical regime.

Appendix B: Additional Details for the Dominick's Pricing Experiment

This appendix provides additional details for the reference-price pricing experiment in Section 7.2. We describe the data processing, offline calibration, product and store selection, policy implementation, and statistical evaluation.

B.1. Data Processing

The Dominick's movement files contain weekly observations for individual products, identified by UPC, across stores.

We retain observations satisfying $\texttt{OK} = 1$, $\texttt{PRICE} > 0$, $\texttt{QTY} > 0$, and $\texttt{MOVE} \geq 0$.

Here, $\texttt{OK} = 1$ means that the observation is not marked as unreliable by the data provider. Demand is measured by `MOVE`.

The variable `PRICE` records the posted price for `QTY` units. We therefore define the price per unit as $p_{i,s,t} = \texttt{PRICE}_{i,s,t}/\texttt{QTY}_{i,s,t}$.

For example, an offer of two units for \$5 has $\texttt{PRICE} = 5$, $\texttt{QTY} = 2$, and unit price 2.5. Using unit price makes the price variable consistent with `MOVE`, which records the number of individual units sold.

For product i, store s, and week t, the context vector $c_{i,s,t}$ contains the annual terms $\sin(2\pi t/52)$ and $\cos(2\pi t/52)$, the semiannual terms $\sin(4\pi t/52)$ and $\cos(4\pi t/52)$, the normalized time trend $(t-120)/120$, and store indicators. The Fourier terms capture annual and semiannual seasonality, while the store indicators capture persistent differences in demand levels across stores.

Promotion variables, realized profits, category-level summaries, and future prices or demands are not included in the online context.

The temporal split uses weeks 1–239 for training, weeks 240–299 for validation, and weeks 300–399 for testing.

B.2. Demand and Reference-Price Model

Before applying the numerical demand floor, mean demand is

$$\widetilde{\mu}_{i,s,t} = c_{i,s,t}^{\top}\beta_i + b_{0,i} - a_i p_{i,s,t} - q_i p_{i,s,t}^2 + \eta_i \left(r_{i,s,t} - p_{i,s,t}\right).$$

The reference price follows

$$r_{i,s,t+1} = \rho_i r_{i,s,t} + (1-\rho_i) p_{i,s,t}.$$

The implemented mean demand is

$$\mu_{i,s,t} = \max\left\{\widetilde{\mu}_{i,s,t}, D_{\mathrm{floor},i}\right\}.$$

Observed demand is generated as

$$Y_{i,s,t} = \mu_{i,s,t} + \varepsilon_{i,s,t}, \qquad \varepsilon_{i,s,t} \sim N(0, \sigma_{D,i}^2).$$

Expected and realized weekly profits are $\overline{\Pi}_{i,s,t} = (p_{i,s,t} - w_i)\mu_{i,s,t}$ and $\Pi_{i,s,t}^{\mathrm{real}} = (p_{i,s,t} - w_i)Y_{i,s,t}$, respectively.

To connect this environment to Section 2, fix a product–store instance and suppress its indices. The latent state and action are $x_t = r_t$ and $a_t = p_t$; the predictable context path c_t is known to every policy and does not enter the latent transition. The transition feature representation is

$$\Phi_{\mathrm{price}}(x,p) = (x,p)^{\top}, \qquad W_{\star,i}^{\mathrm{price}} = (\rho_i, 1-\rho_i),$$

so the transition parameter has one unknown degree of freedom. For the reward model, define $\vartheta_i = (b_{0,i}, a_i, q_i, \eta_i)^{\top}$ and

$$\Psi_i^{\mathrm{price}}(x,p) := (p - w_i)(1, -p, -p^2, x - p)^{\top}.$$

Whenever the floor is inactive, expected profit equals the known offset $(p_{i,s,t} - w_i)c_{i,s,t}^{\top}\beta_i$ plus $\vartheta_i^{\top}\Psi_i^{\mathrm{price}}(r_{i,s,t}, p_{i,s,t})$. After offline calibration, the context coefficient is fixed and the known time-varying offset is included directly in the planning reward. Subtracting it from observed profit leaves a reward regression linear in ϑ_i.

B.3. Offline Calibration

For each UPC, we estimate one product-level model using all valid training observations from its available stores. Each store keeps its own price history and reference-price path, while the demand and reference-price parameters are shared across stores for the same product.

For a fixed value of ρ, the reference-price path is constructed recursively as

$$r_{i,s,t+1}(\rho) = \rho r_{i,s,t}(\rho) + (1-\rho) p_{i,s,t}.$$

Once ρ is fixed, the demand model is linear in $\gamma_i = (\beta_i, b_{0,i}, a_i, q_i, \eta_i)$.

We therefore solve

$$\widehat{\gamma}_i(\rho) \in \arg\min_{\gamma_i} \sum_{s,t \in \mathcal{T}_{\text{train}}} \left[Y_{i,s,t} - c_{i,s,t}^\top \beta_i - b_{0,i} + a_i p_{i,s,t} + q_i p_{i,s,t}^2 - \eta_i \left(r_{i,s,t}(\rho) - p_{i,s,t} \right) \right]^2.$$

The resulting profile loss is $L_i^{\text{prof}}(\rho) = L_i(\rho, \widehat{\gamma}_i(\rho))$. We evaluate this loss over $\mathcal{R} = \{0.01, 0.02, \ldots, 0.99\}$ and choose $\widehat{\rho}_i \in \arg\min_{\rho \in \mathcal{R}} L_i^{\text{prof}}(\rho)$.

The remaining coefficients are then $\widehat{\gamma}_i = \widehat{\gamma}_i(\widehat{\rho}_i)$.

This procedure solves the same least-squares problem as joint optimization over the same parameter set, while using the fact that the remaining coefficients can be estimated by linear regression once ρ is fixed.

B.4. Price Range, Cost, Noise, and Demand Floor

The feasible price interval for product i is $p_{i,s,t} \in [p_{L,i}, p_{H,i}]$, where

$$p_{L,i} = Q_{0.05}\left(p_{i,\text{train}}\right), \qquad p_{H,i} = Q_{0.95}\left(p_{i,\text{train}}\right).$$

Using the central historical price range limits how far the online policies can move outside the region supported by the training data.

Unit cost is set to $w_i = 0.65\,\text{median}(p_{i,\text{train}})$.

The product-specific demand-noise scale is the validation RMSE,

$$\sigma_{D,i} = \sqrt{\frac{1}{n_{i,\text{val}}} \sum_{s,t \in \mathcal{T}_{\text{val}}} \left(Y_{i,s,t} - \widehat{Y}_{i,s,t} \right)^2}.$$

The numerical demand floor is

$$D_{\text{floor},i} = 10^{-6} \max\left\{1, \text{median}\left(Y_{i,\text{train}} \mid Y_{i,\text{train}} > 0\right)\right\}.$$

This is a very small product-specific positive value. Its only purpose is to prevent the fitted quadratic demand curve from producing negative mean demand. The floor is applied to mean demand but does not truncate the Gaussian noise.

The fitted offline model defines the true parameters of the semi-synthetic environment. The contextual coefficient $\widehat{\beta}_i$ is fixed and known to all policies. The reward parameter $\vartheta_i = (b_{0,i}, a_i, q_i, \eta_i)^\top$ is re-estimated online by every learned policy. Dynamic CE and Dynamic Raw also estimate the transition parameter ρ_i. Each policy uses the observations generated by its own prices, and only the Oracle knows the calibrated parameter values.

B.5. Product Qualification

The processed data contain 369 UPCs. We first require enough training data, store coverage, and price variation to estimate a product-level model.

A product passes the training-support screen if $n_{i,\text{train}} \geq 2000$, $n_{i,\text{stores}} \geq 10$, $1.25 \leq \overline{p}_{i,\text{train}} \leq 8.00$, and $\text{sd}(p_{i,\text{train}})/\overline{p}_{i,\text{train}} \geq 0.05$.

The first two conditions require enough observations across enough stores. The price-variation condition excludes products whose historical prices change too little to estimate price response reliably. The average-price condition removes products with unusually small or large price scales.

These requirements leave 211 support-eligible UPCs.

We then require sufficient validation support and a minimum level of out-of-time predictive fit: $n_{i,\text{val}} \geq 500$ and $R^2_{i,\text{val}} \geq 0.30$.

Because the persistence parameter is estimated over $\mathcal{R} = \{0.01, 0.02, \ldots, 0.99\}$, the condition $0 < \widehat{\rho}_i < 1$ holds by construction rather than as an additional qualification screen. We separately require a positive reference-price effect, $\widehat{\eta}_i > 0$.

Fitted demand must be decreasing in price throughout the feasible interval. Since $\partial\widetilde{\mu}_i/\partial p = -(\widehat{a}_i + 2\widehat{q}_i p + \widehat{\eta}_i)$, we require $\partial\widetilde{\mu}_i/\partial p < 0$ for every $p \in [p_{L,i}, p_{H,i}]$.

Because the derivative is linear in p, it is enough to check the two endpoints of the interval.

We do not impose $\widehat{q}_i \geq 0$. A negative quadratic coefficient changes the curvature of demand but remains admissible when fitted demand is decreasing on the entire feasible interval.

We also exclude products for which the fitted demand floor is active in more than 1% of the training observations or for which the feasible price interval is inconsistent with the calibrated cost:

$$\frac{\#\left\{(s,t) : \widehat{\mu}_{i,s,t} \leq D_{\text{floor},i}\right\}}{n_{i,\text{train}}} \leq 0.01,$$

and $p_{L,i} > w_i$.

All estimated parameters, losses, price bounds, validation statistics, and demand derivatives must be finite.

The qualification process reduces the sample from 369 UPCs to 211 support-eligible UPCs and then to 39 qualified UPCs.

Product qualification uses only training and validation data. No test-period policy outcome is used in this process.

B.6. Validation-Based Construction of Mechanism-Relevant Instances

For each qualified UPC, we consider stores with enough historical data and complete context paths over the validation and test periods. This gives 3,351 eligible product–store candidates.

We evaluate each candidate over weeks 240–299 using Dynamic CE, Dynamic Raw, Myopic OFU, and the Oracle under matched demand-noise realizations and use these validation-period comparisons to select one store for each UPC. The resulting instances form a conditional test population for studying dynamic versus myopic pricing; they are not used to estimate average policy performance or the prevalence of consequential reference-price effects across stores.

The selected store is then fixed for that UPC. This procedure produces one frozen product–store instance for each of the 39 qualified UPCs.

Store selection uses validation data only. No observation or policy result from weeks 300–399 is used to select, tune, or remove any instance.

B.7. Initial Reference States and Warm Start

Each store has its own historical price path and therefore its own historical reference-price path. The reference state is constructed recursively using

$$r_{i,s,t+1} = \widehat{\rho}_i r_{i,s,t} + (1 - \widehat{\rho}_i) p^{\text{hist}}_{i,s,t}.$$

The validation and test experiments begin from the corresponding historically constructed reference states. Thus, stores are not initialized from a common artificial reference price.

All learning policies receive the same $n_0 = 2$ pre-horizon warm-start observations. They also receive the same initial reference state, context sequence, feasible price interval, and matched demand-noise realization.

After the common warm start, each policy selects its own prices. It therefore generates its own demand observations and its own reference-price trajectory.

B.8. Policy Implementation

Dynamic CE, Dynamic Raw, Myopic CE, and Myopic OFU use the same online reward estimator for $\vartheta_i = (b_{0,i}, a_i, q_i, \eta_i)^\top$. The online parameter set allows $q \in [-150, 25]$, which contains the calibrated values of q_i for all 39 qualified products. The estimate of η_i is constrained to remain positive.

Dynamic CE and Dynamic Raw also use the same online transition estimator, based on $r_{t+1} - p_t = \rho_i(r_t - p_t)$.

Dynamic CE plans over the remaining horizon using the current point estimates of the reward and transition models.

Dynamic Raw uses the same observation protocol and estimator forms as Dynamic CE, but plans optimistically over the reward and transition confidence sets.

Myopic CE uses the current point estimate of the reward model and chooses

$$p_t \in \arg\max_{p\in[p_{L,i},p_{H,i}]} \widehat{\Pi}_t(p;r_t),$$

where $\widehat{\Pi}_t(p;r_t)$ is the estimated current expected profit at the observed reference state. It uses no optimism bonus, continuation value, or transition-uncertainty contribution when selecting the current price.

Myopic OFU maximizes an optimistic estimate of current profit but assigns no continuation value to the future reference state induced by the current price.

The Oracle knows the calibrated reward and transition parameters and plans over the full remaining horizon.

For every qualified product, $|T_{\rho_i}(r,p) - T_{\rho_i}(r',p)| = \rho_i|r-r'|$, where $0<\rho_i<1$. The transition is therefore contractive. Core is not included in this experiment because the purpose is to compare dynamic optimism with certainty equivalence and myopic optimism, rather than to study recoverability-based pruning.

B.9. Final Evaluation and Statistical Inference

For each of the 39 frozen product–store instances, we evaluate all five policies over the 100-week test period using 20 matched demand-noise seeds.

The final experiment therefore contains $39\times5\times20 = 3{,}900$ policy trajectories and $3{,}900\times100 = 390{,}000$ weekly policy records.

For policy π, let $\overline{J}_i^{\pi}$ be cumulative profit averaged over the 20 seeds for UPC i. For each ordered policy pair (π,π'), we form the UPC-level differences $\Delta_i^{\pi,\pi'} = \overline{J}_i^{\pi} - \overline{J}_i^{\pi'}$ and report their mean across the 39 UPCs. The 95% confidence interval is the paired Student-t interval $\overline{\Delta}^{\pi,\pi'} \pm t_{0.975,38} s_{\pi,\pi'}/\sqrt{39}$, where $s_{\pi,\pi'}$ is the sample standard deviation of the paired differences. Win, tie, and loss counts use their signs. Relative gain divides the mean paired difference by the second policy's mean expected profit.

B.10. Additional Numerical Checks

All 39 frozen instances are included in the final results, including instances for which Dynamic Raw underperforms a baseline.

Table 2 Policy-level test performance and additional diagnostics for the pricing experiment. Profits are first averaged over the 20 matched seeds within each instance and then across the 39 instances.

Policy	Expected profit	Realized profit	Floor rate	Boundary rate
Myopic CE	3285.950	3294.304	2.631%	53.872%
Myopic OFU	3355.299	3365.384	2.685%	55.740%
Dynamic CE	3641.212	3650.382	2.886%	66.359%
Dynamic Raw	3702.958	3713.325	2.628%	72.685%
Dynamic Oracle	3968.714	3979.244	1.795%	91.385%

Realized profit produces the same policy ranking as expected profit. The demand floor is active in less than 3% of test periods for every learned policy. Boundary pricing is common across all policy classes and is most frequent under the Dynamic Oracle, indicating that boundary actions are primarily a feature of the calibrated pricing environments rather than an artifact of optimistic learning.

Appendix C: State Concentration

LEMMA 1 **(Concentration of empirical latent states)**. *Suppose* $\|\psi(Z)\|_2 \le B_\psi$ *almost surely. For each* $t = 1, \dots, H+1$, *let* $x_t = \mathbb{E}_{Z\sim d_t}[\psi(Z)]$ *and* $\hat{x}_t = n^{-1}\sum_{i=1}^n \psi(Z_{t,i}^x)$, *where* $\mathcal{F}_{t,-}$ *denotes the information available immediately before the period-t state batch is drawn, and* $Z_{t,1}^x, \dots, Z_{t,n}^x \overset{\text{i.i.d.}}{\sim} d_t$ *conditional on* $\mathcal{F}_{t,-}$. *Then there exists a universal constant* $c_x > 0$ *such that, with probability at least* $1-\delta_x$,

$$\|\hat{x}_t - x_t\|_2 \le c_x B_\psi \sqrt{\frac{m + \log((H+1)/\delta_x)}{n}} \qquad \textit{for all } t = 1, \dots, H+1.$$

Proof. Conditional on the pre-sampling history, the adaptively chosen distribution d_t is fixed, so the empirical latent state concentrates exactly as in the i.i.d. case; the only remaining step is a union bound over time. Fix $t \in \{1, \dots, H+1\}$. Conditional on $\mathcal{F}_{t,-}$, the distribution d_t is fixed, and hence $X_{t,i} := \psi(Z_{t,i}^x) - x_t$, $i = 1, \dots, n$, are conditionally independent mean-zero random vectors in $\mathbb{R}^m$. Moreover, by Jensen's inequality, $\|x_t\|_2 = \|\mathbb{E}_{Z\sim d_t}[\psi(Z)]\|_2 \le \mathbb{E}_{Z\sim d_t}\|\psi(Z)\|_2 \le B_\psi$. Therefore, almost surely, $\|X_{t,i}\|_2 \le \|\psi(Z_{t,i}^x)\|_2 + \|x_t\|_2 \le 2B_\psi$. Let $\mathcal{N}$ be a 1/2-net of $\mathbb{S}^{m-1}$. By the standard volumetric bound, we may choose $\mathcal{N}$ such that $|\mathcal{N}| \le 5^m$. Furthermore, for every $v \in \mathbb{R}^m$, $\|v\|_2 \le 2\max_{u\in\mathcal{N}} |u^\top v|$. Applying this inequality to $v = \hat{x}_t - x_t = n^{-1}\sum_{i=1}^n X_{t,i}$, we obtain

$$\|\hat{x}_t - x_t\|_2 \le 2 \max_{u\in\mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n u^\top X_{t,i} \right|.$$

Now fix $u \in \mathcal{N}$. Conditional on $\mathcal{F}_{t,-}$, the random variables $u^\top X_{t,i}$ are independent, mean zero, and satisfy $|u^\top X_{t,i}| \le \|u\|_2 \|X_{t,i}\|_2 \le 2B_\psi$ almost surely. Hence, by Hoeffding's inequality, for every $s > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} u^\top X_{t,i} \right| \ge s \,\middle|\, \mathcal{F}_{t,-} \right) \le 2 \exp\left(-\frac{n s^2}{8 B_\psi^2} \right).$$

Taking a union bound over $\mathcal{N}$, we get

$$\mathbb{P}\left(\max_{u \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^{n} u^\top X_{t,i} \right| \ge s \,\middle|\, \mathcal{F}_{t,-} \right) \le 2 \cdot 5^m \exp\left(-\frac{n s^2}{8 B_\psi^2} \right).$$

Using the net comparison with $s = \epsilon/2$, this implies

$$\mathbb{P}\left(\|\hat{x}_t - x_t\|_2 \ge \epsilon \,|\, \mathcal{F}_{t,-} \right) \le 2 \cdot 5^m \exp\left(-\frac{n \epsilon^2}{32 B_\psi^2} \right).$$

Choose $\epsilon = c_x B_\psi \sqrt{(m + \log((H+1)/\delta_x))/n}$, where $c_x > 0$ is a sufficiently large universal constant. Then $2 \cdot 5^m \exp(-n\epsilon^2/(32 B_\psi^2)) \le \delta_x/(H+1)$. Since this conditional probability bound holds almost surely, taking expectations yields $\mathbb{P}(\|\hat{x}_t - x_t\|_2 \ge \epsilon) \le \delta_x/(H+1)$. Finally, a union bound over $t = 1, \ldots, H+1$ gives

$$\mathbb{P}\left(\exists t \in \{1, \ldots, H+1\} : \|\hat{x}_t - x_t\|_2 \ge \epsilon \right) \le \delta_x.$$

This is exactly the asserted simultaneous bound, completing the proof. □

Appendix D: Confidence Sets

LEMMA 2 **(Self-normalized martingale inequality)**. *The following statement is Theorem 1 of Abbasi-Yadkori et al. (2011). Let $(\mathcal{F}_t)_{t \ge 0}$ be a filtration. Suppose $z_t \in \mathbb{R}^d$ is $\mathcal{F}_{t-1}$-measurable and η_t is conditionally R-sub-Gaussian:*

$$\mathbb{E}[\exp(\lambda \eta_t) \,|\, \mathcal{F}_{t-1}] \le \exp(\lambda^2 R^2/2).$$

Let

$$V_t = \lambda_0 I_d + \sum_{s<t} z_s z_s^\top, \qquad S_t = \sum_{s<t} z_s \eta_s.$$

Then with probability at least $1 - \delta$, simultaneously for all t,

$$\|S_t\|_{V_t^{-1}} \le R \sqrt{2 \log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda_0 I_d)^{1/2} \delta} \right)}.$$

PROPOSITION 4 **(Reward confidence)**. *With probability at least $1 - \delta_R$, simultaneously for all $t = 1, \ldots, H+1$, on $\mathcal{E}_x$, $\theta_\star \in \mathcal{C}_t^R$.*

Proof. The proof separates the reward regression error into three parts: the ridge regularization error, the martingale error from the reward fluctuations, and the deterministic bias from using $\bar{x}_s$ instead of x_s in the reward features. It is written for the projected-state regressions used by

the finite-grid algorithm, where $\bar{x}_s = \Pi_X(\hat{x}_s)$ and $\beta_\eta^x = \beta^x + \eta_x$. The unprojected confidence sets in Section 3 are recovered by setting $\bar{x}_s = \hat{x}_s$, $\eta_x = 0$, and $\beta_\eta^x = \beta^x$.

For $s < t$, $\bar{Y}_s = \theta_\star^\top \Psi(\bar{x}_s, a_s) + \varepsilon_s^R$, where $\varepsilon_s^R = \zeta_s^R + \theta_\star^\top (\Psi(x_s, a_s) - \Psi(\bar{x}_s, a_s))$.

Ridge decomposition. By definition, $\hat{\theta}_t = (\Lambda_t^R)^{-1} \sum_{s<t} \Psi(\bar{x}_s, a_s) \bar{Y}_s$. Substituting the preceding regression equation gives

$$\hat{\theta}_t = (\Lambda_t^R)^{-1} \sum_{s<t} \Psi(\bar{x}_s, a_s) \Psi(\bar{x}_s, a_s)^\top \theta_\star + (\Lambda_t^R)^{-1} \sum_{s<t} \Psi(\bar{x}_s, a_s) \varepsilon_s^R.$$

Since $\sum_{s<t} \Psi(\bar{x}_s, a_s) \Psi(\bar{x}_s, a_s)^\top = \Lambda_t^R - \lambda_R I_p$, we obtain

$$\hat{\theta}_t - \theta_\star = -(\Lambda_t^R)^{-1} \lambda_R \theta_\star + (\Lambda_t^R)^{-1} \sum_{s<t} \Psi(\bar{x}_s, a_s) \varepsilon_s^R.$$

Taking the Λ_t^R-norm yields

$$\begin{aligned} \|\theta_\star - \hat{\theta}_t\|_{\Lambda_t^R} &\le \lambda_R \|(\Lambda_t^R)^{-1} \theta_\star\|_{\Lambda_t^R} + \left\| \sum_{s<t} \Psi(\bar{x}_s, a_s) \varepsilon_s^R \right\|_{(\Lambda_t^R)^{-1}} \\ &\le \sqrt{\lambda_R} \|\theta_\star\|_2 + \left\| \sum_{s<t} \Psi(\bar{x}_s, a_s) \varepsilon_s^R \right\|_{(\Lambda_t^R)^{-1}}, \end{aligned}$$

where the last inequality uses $\Lambda_t^R \succeq \lambda_R I_p$.

We now split the second term into the martingale reward fluctuation and the deterministic state-estimation bias:

$$\sum_{s<t} \Psi(\bar{x}_s, a_s) \varepsilon_s^R = \sum_{s<t} \Psi(\bar{x}_s, a_s) \zeta_s^R + \sum_{s<t} \Psi(\bar{x}_s, a_s) \theta_\star^\top \left(\Psi(x_s, a_s) - \Psi(\bar{x}_s, a_s) \right).$$

Martingale reward fluctuation. The covariate $\Psi(\bar{x}_s, a_s)$ is $\mathcal{F}_{s,0}$-measurable, and ζ_s^R is conditionally $\sigma_R/\sqrt{n}$-sub-Gaussian given $\mathcal{F}_{s,0}$. Therefore, by Lemma 2, with probability at least $1 - \delta_R$, simultaneously for all t,

$$\left\| \sum_{s<t} \Psi(\bar{x}_s, a_s) \zeta_s^R \right\|_{(\Lambda_t^R)^{-1}} \le \frac{\sigma_R}{\sqrt{n}} \sqrt{2 \log \left(\frac{\det(\Lambda_t^R)^{1/2}}{\det(\lambda_R I_p)^{1/2} \delta_R} \right)}.$$

State-estimation and projection bias. It remains to bound the deterministic bias term on $\mathcal{E}_x$. On this event, $\|x_s - \bar{x}_s\|_2 \le \beta_\eta^x$ for all s. Thus, for every s,

$$\left| \theta_\star^\top \left(\Psi(x_s, a_s) - \Psi(\bar{x}_s, a_s) \right) \right| \le \|\theta_\star\|_2 \left\| \Psi(x_s, a_s) - \Psi(\bar{x}_s, a_s) \right\|_2 \le S_\theta L_\Psi \beta_\eta^x = \alpha_R^\eta.$$

Hence, by Cauchy–Schwarz,

$$\begin{aligned} \left\| \sum_{s<t} \Psi(\bar{x}_s, a_s) \theta_\star^\top \left(\Psi(x_s, a_s) - \Psi(\bar{x}_s, a_s) \right) \right\|_{(\Lambda_t^R)^{-1}} &\le \alpha_R^\eta \sum_{s<t} \|\Psi(\bar{x}_s, a_s)\|_{(\Lambda_t^R)^{-1}} \\ &\le \alpha_R^\eta \sqrt{t-1} \left(\sum_{s<t} \|\Psi(\bar{x}_s, a_s)\|_{(\Lambda_t^R)^{-1}}^2 \right)^{1/2}. \end{aligned}$$

The remaining sum is at most the feature dimension:

$$\begin{aligned}\sum_{s<t}\|\Psi(\bar{x}_s,a_s)\|^2_{(\Lambda_t^R)^{-1}} &= \operatorname{tr}\left((\Lambda_t^R)^{-1}\sum_{s<t}\Psi(\bar{x}_s,a_s)\Psi(\bar{x}_s,a_s)^\top\right)\\ &= \operatorname{tr}\left((\Lambda_t^R)^{-1}(\Lambda_t^R-\lambda_R I_p)\right)\\ &= p-\lambda_R\operatorname{tr}\left((\Lambda_t^R)^{-1}\right)\le p.\end{aligned}$$

Therefore, on $\mathcal{E}_x$,

$$\left\|\sum_{s<t}\Psi(\bar{x}_s,a_s)\theta_\star^\top\left(\Psi(x_s,a_s)-\Psi(\bar{x}_s,a_s)\right)\right\|_{(\Lambda_t^R)^{-1}}\le\alpha_R^\eta\sqrt{(t-1)p}.$$

Combining the three terms. Combining the regularization term, the martingale term, and the deterministic bias term gives $\|\theta_\star-\hat{\theta}_t\|_{\Lambda_t^R}\le\beta_t^R$ simultaneously for all t on $\mathcal{E}_x$, with probability at least $1-\delta_R$ over the reward fluctuations. Finally, $\|\theta_\star\|_2\le S_\theta$, so $\theta_\star\in\mathcal{C}_t^R$ for all t. □

PROPOSITION 5 **(Transition confidence)**. *On $\mathcal{E}_x$, for all $t=1,\dots,H+1$, $W_\star\in\mathcal{C}_t^T$.*

Proof. The proof is deterministic on the state-confidence event $\mathcal{E}_x$. The transition regression error comes only from using projected empirical latent states in place of the true latent states. We first bound this perturbation, then decompose the ridge estimator into a regularization term and a cumulative perturbation term. As in the reward proof, the unprojected confidence sets in Section 3 are obtained by setting $\bar{x}_s=\hat{x}_s$, $\eta_x=0$, and $\beta_\eta^x=\beta^x$.

Transition regression error. For $s<t$, the projected transition regression can be written as $\bar{x}_{s+1}=W_\star\Phi(\bar{x}_s,a_s)+\varepsilon_s^T$, where $\varepsilon_s^T=(\bar{x}_{s+1}-x_{s+1})+W_\star(\Phi(x_s,a_s)-\Phi(\bar{x}_s,a_s))$. On $\mathcal{E}_x$, we have $\|x_s-\bar{x}_s\|_2\le\beta_\eta^x$ and $\|x_{s+1}-\bar{x}_{s+1}\|_2\le\beta_\eta^x$. Therefore,

$$\begin{aligned}\|\varepsilon_s^T\|_2 &\le \|\bar{x}_{s+1}-x_{s+1}\|_2+\|W_\star\|_{\mathrm{op}}\|\Phi(x_s,a_s)-\Phi(\bar{x}_s,a_s)\|_2\\ &\le \beta_\eta^x+S_W L_\Phi\beta_\eta^x\\ &= (1+S_W L_\Phi)\beta_\eta^x=\alpha_T^\eta.\end{aligned}$$

Ridge decomposition. By definition, $\hat{W}_t=\left(\sum_{s<t}\bar{x}_{s+1}\Phi(\bar{x}_s,a_s)^\top\right)(\Lambda_t^T)^{-1}$. Substituting the transition regression equation gives

$$\begin{aligned}\hat{W}_t &= \left(\sum_{s<t}\left[W_\star\Phi(\bar{x}_s,a_s)+\varepsilon_s^T\right]\Phi(\bar{x}_s,a_s)^\top\right)(\Lambda_t^T)^{-1}\\ &= W_\star\left(\sum_{s<t}\Phi(\bar{x}_s,a_s)\Phi(\bar{x}_s,a_s)^\top\right)(\Lambda_t^T)^{-1}+\left(\sum_{s<t}\varepsilon_s^T\Phi(\bar{x}_s,a_s)^\top\right)(\Lambda_t^T)^{-1}.\end{aligned}$$

Since $\sum_{s<t}\Phi(\bar{x}_s,a_s)\Phi(\bar{x}_s,a_s)^\top=\Lambda_t^T-\lambda_T I_q$, we obtain

$$\hat{W}_t-W_\star=-\lambda_T W_\star(\Lambda_t^T)^{-1}+\left(\sum_{s<t}\varepsilon_s^T\Phi(\bar{x}_s,a_s)^\top\right)(\Lambda_t^T)^{-1}.$$

Multiplying on the right by $(\Lambda_t^T)^{1/2}$ and applying the triangle inequality yields

$$\|(W_\star - \hat{W}_t)(\Lambda_t^T)^{1/2}\|_F \le \lambda_T \|W_\star(\Lambda_t^T)^{-1/2}\|_F + \left\| \sum_{s<t} \varepsilon_s^T \Phi(\bar{x}_s, a_s)^\top (\Lambda_t^T)^{-1/2} \right\|_F .$$

Regularization term. Because $\Lambda_t^T \succeq \lambda_T I_q$, $\|(\Lambda_t^T)^{-1/2}\|_{\mathrm{op}} \le 1/\sqrt{\lambda_T}$. Hence $\lambda_T \|W_\star(\Lambda_t^T)^{-1/2}\|_F \le \sqrt{\lambda_T}\|W_\star\|_F$. By Assumption 2, $\|W_\star\|_F \le \sqrt{m}\|W_\star\|_{\mathrm{op}} \le \sqrt{m} S_W$. Thus the regularization term is at most $\sqrt{\lambda_T m}\, S_W$.

Perturbation term. It remains to control the cumulative perturbation term. Using the triangle inequality and the identity $\|uv^\top\|_F = \|u\|_2 \|v\|_2$,

$$\begin{aligned} \left\| \sum_{s<t} \varepsilon_s^T \Phi(\bar{x}_s, a_s)^\top (\Lambda_t^T)^{-1/2} \right\|_F &\le \sum_{s<t} \left\| \varepsilon_s^T \Phi(\bar{x}_s, a_s)^\top (\Lambda_t^T)^{-1/2} \right\|_F \\ &= \sum_{s<t} \|\varepsilon_s^T\|_2 \left\| (\Lambda_t^T)^{-1/2} \Phi(\bar{x}_s, a_s) \right\|_2 \\ &= \sum_{s<t} \|\varepsilon_s^T\|_2 \|\Phi(\bar{x}_s, a_s)\|_{(\Lambda_t^T)^{-1}}. \end{aligned}$$

On $\mathcal{E}_x$, $\|\varepsilon_s^T\|_2 \le \alpha_T^\eta$. Therefore,

$$\begin{aligned} \left\| \sum_{s<t} \varepsilon_s^T \Phi(\bar{x}_s, a_s)^\top (\Lambda_t^T)^{-1/2} \right\|_F &\le \alpha_T^\eta \sum_{s<t} \|\Phi(\bar{x}_s, a_s)\|_{(\Lambda_t^T)^{-1}} \\ &\le \alpha_T^\eta \sqrt{t-1} \left(\sum_{s<t} \|\Phi(\bar{x}_s, a_s)\|^2_{(\Lambda_t^T)^{-1}} \right)^{1/2}. \end{aligned}$$

The remaining sum is at most q:

$$\begin{aligned} \sum_{s<t} \|\Phi(\bar{x}_s, a_s)\|^2_{(\Lambda_t^T)^{-1}} &= \mathrm{tr}\left((\Lambda_t^T)^{-1} \sum_{s<t} \Phi(\bar{x}_s, a_s)\Phi(\bar{x}_s, a_s)^\top \right) \\ &= \mathrm{tr}\left((\Lambda_t^T)^{-1} (\Lambda_t^T - \lambda_T I_q) \right) \\ &= q - \lambda_T \,\mathrm{tr}\left((\Lambda_t^T)^{-1} \right) \le q. \end{aligned}$$

Thus,

$$\left\| \sum_{s<t} \varepsilon_s^T \Phi(\bar{x}_s, a_s)^\top (\Lambda_t^T)^{-1/2} \right\|_F \le \alpha_T^\eta \sqrt{(t-1)q}.$$

Combining the bounds. Combining the regularization and perturbation bounds gives

$$\|(W_\star - \hat{W}_t)(\Lambda_t^T)^{1/2}\|_F \le \sqrt{\lambda_T m}\, S_W + \alpha_T^\eta \sqrt{(t-1)q} = \beta_t^T.$$

Since $\|W_\star\|_{\mathrm{op}} \le S_W$, both constraints defining $\mathcal{C}_t^T$ are satisfied. Hence $W_\star \in \mathcal{C}_t^T$ for all $t = 1, \ldots, H+1$ on $\mathcal{E}_x$. $\square$

Appendix E: Recoverable-Core Proofs

Proof of Proposition 2. The proof separates the deterministic properties of the core from the good-event statement about the true graph.

The recoverable core is contained in the raw correspondence. Every $\mathfrak{K} \in \mathfrak{R}_j$ satisfies $\mathfrak{K}(x) \subseteq \mathfrak{E}_j(x)$. Hence the union also satisfies $\mathfrak{K}_j^{\max}(x) \subseteq \mathfrak{E}_j(x)$ for all $x \in \mathcal{X}$.

The recoverable core is recoverable. Fix $x, y \in \mathcal{X}$ and $(a, \rho, z) \in \mathfrak{K}_j^{\max}(x)$. By the definition of the union, there exists some $\mathfrak{K} \in \mathfrak{R}_j$ such that $(a, \rho, z) \in \mathfrak{K}(x)$. Since $\mathfrak{K}$ is (L_r, α)-recoverable, there exists $(a', \rho', z') \in \mathfrak{K}(y)$ satisfying $\rho - \rho' \leq L_r \|x - y\|_2$ and $\|z - z'\|_2 \leq \alpha \|x - y\|_2$. But $\mathfrak{K}(y) \subseteq \mathfrak{K}_j^{\max}(y)$. Therefore the same triple (a', ρ', z') is a valid recovery witness in $\mathfrak{K}_j^{\max}(y)$. Since x, y and (a, ρ, z) were arbitrary, $\mathfrak{K}_j^{\max}$ is (L_r, α)-recoverable.

The true graph is feasible. Fix $x \in \mathcal{X}$, $a \in \mathcal{A}$, and consider $(a, r_\star(x,a), T_\star(x,a)) \in \mathfrak{K}_\star(x)$. Assumption 1 ensures that $T_\star(x,a) \in \mathcal{X}$. On the good event, the reward and transition bounds (12)–(13) give $|r_\star(x,a) - \hat{r}_j(x,a)| \leq b_j^R(x,a)$, $\|T_\star(x,a) - \hat{T}_j(x,a)\|_2 \leq b_j^T(x,a)$. Therefore, $(a, r_\star(x,a), T_\star(x,a)) \in \mathfrak{E}_j(x)$. Hence, $\mathfrak{K}_\star(x) \subseteq \mathfrak{E}_j(x)$ for all $x \in \mathcal{X}$.

The true graph is recoverable. We next show that $\mathfrak{K}_\star$ is (L_r, α)-recoverable. Fix $x, y \in \mathcal{X}$ and a triple $(a, r_\star(x,a), T_\star(x,a)) \in \mathfrak{K}_\star(x)$. By Assumption 6, there exists $a' \in \mathcal{A}$ such that $|r_\star(x,a) - r_\star(y,a')| \leq L_r \|x - y\|_2$ and $\|T_\star(x,a) - T_\star(y,a')\|_2 \leq \alpha \|x - y\|_2$. Define the corresponding true triple at state y by $(a', \rho', z') := (a', r_\star(y,a'), T_\star(y,a'))$. Then $(a', \rho', z') \in \mathfrak{K}_\star(y)$. Moreover, the reward inequality above implies the one-sided condition $r_\star(x,a) - \rho' = r_\star(x,a) - r_\star(y,a') \leq L_r \|x - y\|_2$, and the transition inequality gives $\|T_\star(x,a) - z'\|_2 = \|T_\star(x,a) - T_\star(y,a')\|_2 \leq \alpha \|x - y\|_2$. Thus $\mathfrak{K}_\star$ is (L_r, α)-recoverable.

The recoverable core contains the true graph. The previous two paragraphs show that, on the good event, $\mathfrak{K}_\star$ is a recoverable subcorrespondence of $\mathfrak{E}_j$. Therefore $\mathfrak{K}_\star \in \mathfrak{R}_j$. By the definition of the recoverable core, $\mathfrak{K}_j^{\max}(x) = \bigcup_{\mathfrak{K} \in \mathfrak{R}_j} \mathfrak{K}(x)$, so $\mathfrak{K}_\star(x) \subseteq \mathfrak{K}_j^{\max}(x)$ for all $x \in \mathcal{X}$. □

Proof of Proposition 3. The proof has two parts. Optimism follows because the recoverable core contains the true reward-transition graph. Lipschitz regularity follows because every extended action in the core has a recovery witness at nearby states, and the contraction factor $\alpha < 1$ closes the backward induction.

Optimism. By Proposition 2, on the good event, $(a, r_\star(x,a), T_\star(x,a)) \in \mathfrak{K}_j^{\max}(x)$ for every $x \in \mathcal{X}$ and $a \in \mathcal{A}$. We prove by backward induction that

$$\widetilde{V}_h^j(x) \geq V_h^\star(x) \qquad \text{for all } x \in \mathcal{X}.$$

At $h = H+1$, both value functions are equal to zero. Suppose the claim holds at stage $h+1$. Then, for every $x \in \mathcal{X}$,

$$\begin{aligned}\widetilde{V}_h^j(x) &= \sup_{(a,\rho,z)\in\mathfrak{K}_j^{\max}(x)} \Big[\rho + \widetilde{V}_{h+1}^j(z)\Big] \\ &\geq \sup_{a\in\mathcal{A}} \Big[r_\star(x,a) + \widetilde{V}_{h+1}^j(T_\star(x,a))\Big] \\ &\geq \sup_{a\in\mathcal{A}} \big[r_\star(x,a) + V_{h+1}^\star(T_\star(x,a))\big] \\ &= V_h^\star(x).\end{aligned}$$

The first inequality uses the fact that all true triples are feasible in the core, and the second uses the induction hypothesis. This proves optimism.

Lipschitz regularity. We prove by backward induction that $\widetilde{V}_h^j$ is L_{lat}-Lipschitz. At $h = H+1$, $\widetilde{V}_{H+1}^j \equiv 0$, so the claim is immediate.

Suppose $\widetilde{V}_{h+1}^j$ is L_{lat}-Lipschitz. Fix $x, y \in \mathcal{X}$, and let $(a,\rho,z) \in \mathfrak{K}_j^{\max}(x)$ attain the supremum in (16) at state x. Thus $\widetilde{V}_h^j(x) = \rho + \widetilde{V}_{h+1}^j(z)$. Since $\mathfrak{K}_j^{\max}$ is recoverable by Proposition 2, there exists $(a',\rho',z') \in \mathfrak{K}_j^{\max}(y)$ such that $\rho - \rho' \leq L_r\|x-y\|_2$ and $\|z-z'\|_2 \leq \alpha\|x-y\|_2$. Because (a',ρ',z') is feasible in the Bellman recursion at state y, $\widetilde{V}_h^j(y) \geq \rho' + \widetilde{V}_{h+1}^j(z')$. Therefore

$$\begin{aligned}\widetilde{V}_h^j(x) - \widetilde{V}_h^j(y) &\leq \rho - \rho' + \widetilde{V}_{h+1}^j(z) - \widetilde{V}_{h+1}^j(z') \\ &\leq L_r\|x-y\|_2 + L_{\mathrm{lat}}\|z-z'\|_2 \\ &\leq L_r\|x-y\|_2 + \alpha L_{\mathrm{lat}}\|x-y\|_2 \\ &= (L_r + \alpha L_{\mathrm{lat}})\,\|x-y\|_2.\end{aligned}$$

By the choice $L_{\mathrm{lat}} = L_r/(1-\alpha)$, we have $L_r + \alpha L_{\mathrm{lat}} = L_{\mathrm{lat}}$. Hence $\widetilde{V}_h^j(x) - \widetilde{V}_h^j(y) \leq L_{\mathrm{lat}}\|x-y\|_2$. Repeating the same argument with x and y interchanged gives the reverse inequality. Therefore $|\widetilde{V}_h^j(x) - \widetilde{V}_h^j(y)| \leq L_{\mathrm{lat}}\|x-y\|_2$. This closes the induction.

Span bound. Since $\widetilde{V}_h^j$ is L_{lat}-Lipschitz on $\mathcal{X}$,

$$\operatorname{span}(\widetilde{V}_h^j) = \sup_{x,y\in\mathcal{X}} \Big(\widetilde{V}_h^j(x) - \widetilde{V}_h^j(y)\Big) \leq L_{\mathrm{lat}}\operatorname{diam}(\mathcal{X}) = B_{\mathrm{lat}}.$$

This completes the proof. □

Appendix F: Finite-Grid Recoverable Core

This appendix collects the algorithmic details for the finite-grid recoverable-core planner: the finite raw correspondence, the pruning procedure for computing the finite-grid core, the grid-approximation sufficient condition, the full lazy algorithm, and finite-grid value regularity.

Let $\mathcal{X}_\eta \subset \mathcal{X}$ be a finite η_x-net of the latent state domain, with nearest-grid projection $\Pi_X : \mathcal{X} \to \mathcal{X}_\eta$. Let $\mathcal{A}_\eta \subset \mathcal{A}$ be a finite η_a-net of the action space, with nearest-grid projection $\Pi_A : \mathcal{A} \to \mathcal{A}_\eta$.

Define the compact reward interval $\mathcal{I}_R := [-S_\theta B_\Psi, S_\theta B_\Psi]$. By Assumption 2, $|r_\star(x,a)| = |\theta_\star^\top \Psi(x,a)| \leq S_\theta B_\Psi$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, so $r_\star(\mathcal{X},\mathcal{A}) \subseteq \mathcal{I}_R$. Let $\mathcal{R}_\eta \subset \mathcal{I}_R$ be a finite η_ρ-net of

$\mathcal{I}_R$, with nearest-grid projection $\Pi_R : \mathcal{I}_R \to \mathcal{R}_\eta$. The next latent state grid is $\mathcal{X}_\eta$, so $\eta_z := \eta_x$, and $\Delta_\eta := \eta_x + \eta_a + \eta_z + \eta_\rho$. In the finite-grid analysis, the regressions use projected covariates and responses; equivalently, the confidence radii in Section 3 replace β^x by $\beta^x_\eta := \beta^x + \eta_x$.

At epoch j, the finite-grid raw extended correspondence is defined on grid states $x \in \mathcal{X}_\eta$ by

$$\begin{aligned}\mathfrak{E}_{j,\eta}(x) := \{(a,\rho,z) : a \in \mathcal{A}_\eta,\ \rho \in \mathcal{R}_\eta,\ z \in \mathcal{X}_\eta,\\ |\rho - \hat{r}_j(x,a)| \le b^R_j(x,a) + \varepsilon^\eta_R,\\ \|z - \hat{T}_j(x,a)\|_2 \le b^T_j(x,a) + \varepsilon^\eta_T\}.\end{aligned} \tag{17}$$

Set the deterministic inflation radii to $\varepsilon^\eta_R := \eta_\rho$ and $\varepsilon^\eta_T := \eta_x$. These inflations compensate, respectively, for projecting the true reward onto $\mathcal{R}_\eta$ and the true next latent state onto $\mathcal{X}_\eta$. In particular, $\varepsilon^\eta_R \le \Delta_\eta$ and $\varepsilon^\eta_T \le \Delta_\eta$.

Definition 2 (Additive recoverability on a finite grid). A finite-grid correspondence $\mathfrak{K}_\eta$ with $\mathfrak{K}_\eta(x) \subseteq \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$ is additively $(L_r, \alpha, \varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T})$-recoverable on $\mathcal{X}_\eta$ if, for every $x, y \in \mathcal{X}_\eta$ and every $(a,\rho,z) \in \mathfrak{K}_\eta(x)$, there exists $(a',\rho',z') \in \mathfrak{K}_\eta(y)$ such that

$$\rho - \rho' \le L_r\|x - y\|_2 + \varepsilon^\eta_{\mathrm{rec},R} \quad \text{and} \quad \|z - z'\|_2 \le \alpha\|x - y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}.$$

Let $\mathfrak{R}_{j,\eta}$ be the collection of all additively recoverable finite-grid correspondences $\mathfrak{K}_\eta$ with $\mathfrak{K}_\eta(x) \subseteq \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$. The finite-grid recoverable core is

$$\mathfrak{K}^{\max}_{j,\eta}(x) := \bigcup_{\mathfrak{K}_\eta \in \mathfrak{R}_{j,\eta}} \mathfrak{K}_\eta(x), \qquad x \in \mathcal{X}_\eta. \tag{18}$$

Once the grid state space and the finite raw extended correspondences are fixed, the recoverable core is the greatest fixed point of a monotone pruning operator.

Algorithm 2 Finite-Grid Recoverable-Core Pruning

Require: Finite grid $\mathcal{X}_\eta$, finite raw extended correspondences $\mathfrak{E}_{j,\eta}(x)$, recovery constants L_r, α, additive slacks $\varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T}$.

Initialize $\mathfrak{K}^{(0)}_\eta(x) = \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$.
Set $r = 0$.
repeat
  For every $x \in \mathcal{X}_\eta$, initialize $\mathfrak{K}^{(r+1)}_\eta(x) = \emptyset$.
  for each $x \in \mathcal{X}_\eta$ **do**
    for each $e = (a, \rho, z) \in \mathfrak{K}^{(r)}_\eta(x)$ **do**
      Keep e if for every $y \in \mathcal{X}_\eta$, there exists $e' = (a', \rho', z') \in \mathfrak{K}^{(r)}_\eta(y)$ such that

$$\rho - \rho' \le L_r \|x - y\|_2 + \varepsilon^\eta_{\mathrm{rec},R}, \qquad \|z - z'\|_2 \le \alpha \|x - y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}.$$

      If e is kept, add it to $\mathfrak{K}^{(r+1)}_\eta(x)$.
    end for
  end for
  $r \leftarrow r + 1$.
until $\mathfrak{K}^{(r)}_\eta = \mathfrak{K}^{(r-1)}_\eta$
return $\mathfrak{K}^{(r)}_\eta$.

PROPOSITION 6 **(Correctness of finite-grid pruning)**. *Algorithm 2 terminates in finitely many iterations. Its output is the largest additively* $(L_r, \alpha, \varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T})$*-recoverable subcorrespondence of* $\mathfrak{E}_{j,\eta}$ *on the finite grid.*

Proof. The proof has two steps. First, we show that the pruning procedure terminates and that its terminal correspondence is additively recoverable. Second, we show that no additively recoverable subcorrespondence is ever deleted. This proves that the terminal correspondence is the largest one.

Let $N_\eta := \sum_{x \in \mathcal{X}_\eta} |\mathfrak{E}_{j,\eta}(x)|$ be the total number of finite-grid extended actions. The iterates satisfy $\mathfrak{K}^{(0)}_\eta \supseteq \mathfrak{K}^{(1)}_\eta \supseteq \mathfrak{K}^{(2)}_\eta \supseteq \cdots$, because the algorithm only deletes extended actions. Since $N_\eta < \infty$, after at most N_η strict deletions no further deletion is possible. Hence the algorithm terminates in finitely many iterations.

Let r_{term} be the terminal iteration and write $\mathfrak{K}^{(\infty)}_\eta := \mathfrak{K}^{(r_{\mathrm{term}})}_\eta$. By the stopping rule, $\mathfrak{K}^{(r_{\mathrm{term}})}_\eta = \mathfrak{K}^{(r_{\mathrm{term}}-1)}_\eta$. Therefore, every remaining extended action $e = (a, \rho, z) \in \mathfrak{K}^{(\infty)}_\eta(x)$ was kept in the final pruning pass. Hence, for every grid state $y \in \mathcal{X}_\eta$, it has a witness $e' = (a', \rho', z') \in \mathfrak{K}^{(r_{\mathrm{term}}-1)}_\eta(y) = \mathfrak{K}^{(\infty)}_\eta(y)$ satisfying $\rho - \rho' \le L_r \|x - y\|_2 + \varepsilon^\eta_{\mathrm{rec},R}$ and $\|z - z'\|_2 \le \alpha \|x - y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}$. Thus $\mathfrak{K}^{(\infty)}_\eta$ is additively $(L_r, \alpha, \varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T})$-recoverable.

It remains to prove maximality. Let $\mathfrak{K}_\eta$ be any additively recoverable subcorrespondence of $\mathfrak{E}_{j,\eta}$. We show by induction on r that $\mathfrak{K}_\eta(x) \subseteq \mathfrak{K}_\eta^{(r)}(x)$ for all $x \in \mathcal{X}_\eta$. The claim holds at $r=0$, because $\mathfrak{K}_\eta(x) \subseteq \mathfrak{E}_{j,\eta}(x) = \mathfrak{K}_\eta^{(0)}(x)$. Suppose the claim holds at iteration r. Fix any grid state $x \in \mathcal{X}_\eta$ and any extended action $e = (a, \rho, z) \in \mathfrak{K}_\eta(x)$. Since $\mathfrak{K}_\eta$ is additively recoverable, for every $y \in \mathcal{X}_\eta$ there exists $e_y = (a_y, \rho_y, z_y) \in \mathfrak{K}_\eta(y)$ such that $\rho - \rho_y \le L_r\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},R}$ and $\|z - z_y\|_2 \le \alpha\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}$. By the induction hypothesis, $e_y \in \mathfrak{K}_\eta^{(r)}(y)$ for every $y \in \mathcal{X}_\eta$. Therefore e has valid witnesses in the current iterate $\mathfrak{K}_\eta^{(r)}$ at every grid state. Hence the pruning step does not delete e, and $e \in \mathfrak{K}_\eta^{(r+1)}(x)$. This proves the induction.

Taking r to be the terminal iteration gives $\mathfrak{K}_\eta(x) \subseteq \mathfrak{K}_\eta^{(\infty)}(x)$ for all $x \in \mathcal{X}_\eta$. Thus every additively recoverable subcorrespondence of $\mathfrak{E}_{j,\eta}$ is contained in the terminal correspondence. Since the terminal correspondence is itself additively recoverable by the fixed-point argument, it is the largest additively recoverable subcorrespondence. □

The next result gives a primitive Lipschitz condition under which the finite grid contains an additively recoverable approximation to the true graph.

PROPOSITION 7 **(A sufficient condition for additive grid approximation)**. *Suppose Assumptions 1, 2, and Assumption 6 hold. Suppose also that the true reward and transition are Lipschitz in state and action on $\mathcal{X} \times \mathcal{A}$: there exist constants $L_{r,x}, L_{r,a}, L_{T,x}, L_{T,a}$ such that*

$$|r_\star(x,a) - r_\star(x',a')| \le L_{r,x}\|x-x'\|_2 + L_{r,a}\|a-a'\|_2,$$
$$\|T_\star(x,a) - T_\star(x',a')\|_2 \le L_{T,x}\|x-x'\|_2 + L_{T,a}\|a-a'\|_2.$$

Then Assumption 8 holds with additive slacks of order Δ_η. In particular, one may take

$$\varepsilon^\eta_{\mathrm{rec},R} = 2\eta_\rho + L_{r,a}\eta_a, \qquad \varepsilon^\eta_{\mathrm{rec},T} = 2\eta_x + L_{T,a}\eta_a,$$
$$C^\eta_{\mathrm{app}} \ge \max\{L_{r,x} + L_{r,a} + 1,\ 2 + L_{T,x} + L_{T,a}\}.$$

Proof of Proposition 7. For each grid state $x \in \mathcal{X}_\eta$ and grid action $a \in \mathcal{A}_\eta$, define

$$\rho_\eta(x,a) := \Pi_R r_\star(x,a), \qquad z_\eta(x,a) := \Pi_X T_\star(x,a).$$

These projections are well defined: Assumption 2 gives $r_\star(x,a) \in \mathcal{I}_R$, while Assumption 1 gives $T_\star(x,a) \in \mathcal{X}$. Moreover, the grid-covering properties imply

$$|\rho_\eta(x,a) - r_\star(x,a)| \le \eta_\rho, \qquad \|z_\eta(x,a) - T_\star(x,a)\|_2 \le \eta_x.$$

Let

$$\mathfrak{K}_{\star,j,\eta}(x) := \{(a, \rho_\eta(x,a), z_\eta(x,a)) : a \in \mathcal{A}_\eta\}.$$

We first check that this correspondence is contained in the finite raw correspondence. On the good event, for every grid (x,a), $|r_\star(x,a) - \hat{r}_j(x,a)| \le b_j^R(x,a)$ and $\|T_\star(x,a) - \hat{T}_j(x,a)\|_2 \le b_j^T(x,a)$. Therefore

$$|\rho_\eta(x,a) - \hat{r}_j(x,a)| \le |\rho_\eta(x,a) - r_\star(x,a)| + |r_\star(x,a) - \hat{r}_j(x,a)| \le \eta_\rho + b_j^R(x,a) \le b_j^R(x,a) + \varepsilon_R^\eta,$$

and

$$\|z_\eta(x,a)-\hat{T}_j(x,a)\|_2 \le \|z_\eta(x,a)-T_\star(x,a)\|_2 + \|T_\star(x,a)-\hat{T}_j(x,a)\|_2 \le \eta_x + b_j^T(x,a) \le b_j^T(x,a)+\varepsilon_T^\eta.$$

Thus $\mathfrak{K}_{\star,j,\eta}(x) \subseteq \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$.

We next prove additive grid recovery. Fix grid states $x, y \in \mathcal{X}_\eta$ and a triple $(a,\rho,z) = (a,\rho_\eta(x,a),z_\eta(x,a)) \in \mathfrak{K}_{\star,j,\eta}(x)$. By true latent recovery, there exists a continuous action $a' \in \mathcal{A}$ such that $|r_\star(x,a)-r_\star(y,a')| \le L_r\|x-y\|_2$ and $\|T_\star(x,a)-T_\star(y,a')\|_2 \le \alpha\|x-y\|_2$. Let $a'_\eta := \Pi_A a'$ and define the grid witness $(a'_\eta,\rho',z') := (a'_\eta,\rho_\eta(y,a'_\eta),z_\eta(y,a'_\eta)) \in \mathfrak{K}_{\star,j,\eta}(y)$. For the reward component, using the reward projection error and action Lipschitzness,

$$\begin{aligned}\rho-\rho' &\le r_\star(x,a)-r_\star(y,a'_\eta)+2\eta_\rho\\ &\le r_\star(x,a)-r_\star(y,a')+|r_\star(y,a')-r_\star(y,a'_\eta)|+2\eta_\rho\\ &\le L_r\|x-y\|_2+L_{r,a}\eta_a+2\eta_\rho.\end{aligned}$$

For the transition component, using the state projection error and action Lipschitzness,

$$\begin{aligned}\|z-z'\|_2 &\le \|T_\star(x,a)-T_\star(y,a'_\eta)\|_2+2\eta_x\\ &\le \|T_\star(x,a)-T_\star(y,a')\|_2+\|T_\star(y,a')-T_\star(y,a'_\eta)\|_2+2\eta_x\\ &\le \alpha\|x-y\|_2+L_{T,a}\eta_a+2\eta_x.\end{aligned}$$

Thus $\mathfrak{K}_{\star,j,\eta}$ is additively recoverable with the stated slacks.

Finally, fix a continuous $(x,a) \in \mathcal{X}\times\mathcal{A}$, set $\bar{x} = \Pi_X x$, $\bar{a} := \Pi_A a$, $\bar{\rho} := \rho_\eta(\bar{x},\bar{a})$, and $\bar{z} := z_\eta(\bar{x},\bar{a})$. Then $(\bar{a},\bar{\rho},\bar{z}) \in \mathfrak{K}_{\star,j,\eta}(\bar{x})$. The reward approximation satisfies

$$\begin{aligned}\bar{\rho} &\ge r_\star(\bar{x},\bar{a})-\eta_\rho\\ &\ge r_\star(x,a)-L_{r,x}\|x-\bar{x}\|_2-L_{r,a}\|a-\bar{a}\|_2-\eta_\rho\\ &\ge r_\star(x,a)-(L_{r,x}\eta_x+L_{r,a}\eta_a+\eta_\rho).\end{aligned}$$

For the next state, since $\Pi_X T_\star(x,a) \in \mathcal{X}_\eta$,

$$\begin{aligned}\|\bar{z}-\Pi_X T_\star(x,a)\|_2 &\le \|\bar{z}-T_\star(\bar{x},\bar{a})\|_2+\|T_\star(\bar{x},\bar{a})-T_\star(x,a)\|_2+\|T_\star(x,a)-\Pi_X T_\star(x,a)\|_2\\ &\le 2\eta_x+L_{T,x}\eta_x+L_{T,a}\eta_a.\end{aligned}$$

Both bounds are at most $C_{\mathrm{app}}^\eta\Delta_\eta$ for the displayed choice of C_{app}^η. This proves the assumption. □

F.1. Lazy Finite-Grid Algorithm

Algorithm 3 Recoverable-Core Latent UCB: Lazy Finite-Grid Implementation

Require: Horizon H, regularizers λ_R, λ_T, confidence levels δ_x, δ_R, feature maps ψ, Ψ, Φ, grids $\mathcal{X}_\eta, \mathcal{A}_\eta, \mathcal{R}_\eta$, recovery constants L_r, α, additive recovery slacks $\varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T}$

Initialize $\Lambda_1^R = \lambda_R I_p$ and $\Lambda_1^T = \lambda_T I_q$.
Set epoch index $j = 1$ and epoch start $\tau_1 = 1$.
Initialize estimators $\hat{\theta}_1$ and $\hat{W}_1$.
Form the finite raw extended correspondence $\mathfrak{E}_{1,\eta}$ from (17).
Compute its finite-grid recoverable core $\mathfrak{K}^{\max}_{1,\eta}$ using Algorithm 2.
Set $\widetilde{V}^1_{H+1,\eta} \equiv 0$, then compute raw backups $Q^1_{h,\eta}$ and envelope values $\widetilde{V}^1_{h,\eta}$ for $h = H, H-1, \ldots, 1$ by (19)–(20).
for $t = 1, \ldots, H$ **do**
    Observe state-estimation batch $Z^x_{t,1}, \ldots, Z^x_{t,n} \overset{\text{i.i.d.}}{\sim} d_t$.
    Compute $\hat{x}_t = n^{-1} \sum_{i=1}^n \psi(Z^x_{t,i})$ and $\bar{x}_t = \Pi_X(\hat{x}_t)$.
    if $t \geq 2$ **then**
        Update the transition regression using the previous transition sample $\Phi(\bar{x}_{t-1}, a_{t-1})$ and response $\bar{x}_t$.
        Update Λ_t^T, $\hat{W}_t$, and $\mathcal{C}_t^T$.
    end if
    if $\det(\Lambda_t^R) \geq 2\det(\Lambda^R_{\tau_j})$ or $\det(\Lambda_t^T) \geq 2\det(\Lambda^T_{\tau_j})$ **then**
        Start a new epoch: $j \leftarrow j+1$, $\tau_j \leftarrow t$.
        Form $\mathfrak{E}_{j,\eta}$ using statistics frozen at τ_j.
        Compute the finite-grid recoverable core $\mathfrak{K}^{\max}_{j,\eta}$ using Algorithm 2.
        Set $\widetilde{V}^j_{H+1,\eta} \equiv 0$, then compute raw backups $Q^j_{h,\eta}$ and envelope values $\widetilde{V}^j_{h,\eta}$ for $h = H, H-1, \ldots, t$.
    end if
    Choose

$$(a_t, \rho_t, z_t) \in \arg \max_{(a,\rho,z) \in \mathfrak{K}^{\max}_{j,\eta}(\bar{x}_t)} \left[\rho + \widetilde{V}^j_{t+1,\eta}(z)\right].$$

    Play only the real action a_t.
    Observe independent reward batch $Z^R_{t,1}, \ldots, Z^R_{t,n} \overset{\text{i.i.d.}}{\sim} d_t$ and rewards $Y_{t,i} = r(Z^R_{t,i}, a_t) + \xi_{t,i}$.
    Compute $\bar{Y}_t = n^{-1} \sum_{i=1}^n Y_{t,i}$.
    Update the reward regression using $\Psi(\bar{x}_t, a_t)$ and response $\bar{Y}_t$.
    Update Λ^R_{t+1}, $\hat{\theta}_{t+1}$, and $\mathcal{C}^R_{t+1}$.
end for
Observe the terminal state-estimation batch from d_{H+1}, compute $\bar{x}_{H+1}$, and perform the final transition update using $\Phi(\bar{x}_H, a_H)$ and response $\bar{x}_{H+1}$.

The following proposition is the finite-grid counterpart of Proposition 3. The Lipschitz envelope preserves value regularity, while the grid approximation introduces only an additive $O(\Delta_\eta)$ optimism loss.

Proposition 8 **(Finite-grid core regularity and approximate optimism)**. *On the good event, suppose Assumption 8 holds. Set*

$$\varepsilon^\eta_{\mathrm{val}} := \varepsilon^\eta_{\mathrm{rec},R} + L_{\mathrm{lat}} \varepsilon^\eta_{\mathrm{rec},T}.$$

Let $\widetilde{V}^j_{H+1,\eta}(x) := 0$ *for* $x \in \mathcal{X}_\eta$. *For* $h = H, H-1, \ldots, \tau_j$, *define the raw finite-grid backup*

$$Q^j_{h,\eta}(x) := \max_{(a,\rho,z) \in \mathfrak{K}^{\max}_{j,\eta}(x)} \left[\rho + \widetilde{V}^j_{h+1,\eta}(z)\right], \qquad x \in \mathcal{X}_\eta, \tag{19}$$

and the Lipschitz-envelope value

$$\widetilde{V}^j_{h,\eta}(x) := \max_{y \in \mathcal{X}_\eta} \left[Q^j_{h,\eta}(y) - L_{\mathrm{lat}} \|x - y\|_2\right], \qquad x \in \mathcal{X}_\eta. \tag{20}$$

Then $\mathfrak{K}^{\max}_{j,\eta}$ *is additively recoverable, the values* $\widetilde{V}^j_{h,\eta}$ *are* L_{lat}*-Lipschitz on the grid, and* $\operatorname{span}(\widetilde{V}^j_{h,\eta}) \le B_{\mathrm{lat}}$*. Moreover, for every* h *and grid state* x*,*

$$Q^j_{h,\eta}(x) \le \widetilde{V}^j_{h,\eta}(x) \le Q^j_{h,\eta}(x) + \varepsilon^\eta_{\mathrm{val}}. \tag{21}$$

Finally, for every $x \in \mathcal{X}$*,* $a \in \mathcal{A}$*, stage* h*, and grid epoch* j*,*

$$r_\star(x,a) \le \widetilde{V}^j_{h,\eta}(\Pi_X x) - \widetilde{V}^j_{h+1,\eta}(\Pi_X T_\star(x,a)) + C^\eta_{\mathrm{app}}(1+L_{\mathrm{lat}})\Delta_\eta.$$

Proof of Proposition 8. The proof has four parts. First, we show that the finite-grid core is nonempty, additively recoverable, and contains the finite-grid approximation of the true graph. Second, we show that the raw Bellman backup is approximately L_{lat}-Lipschitz whenever the continuation value is exactly L_{lat}-Lipschitz. Third, we use the Lipschitz envelope to convert this approximate regularity into an exactly Lipschitz continuation value, while losing only $\varepsilon^\eta_{\mathrm{val}}$ in the Bellman identity. Fourth, we prove the approximate supersolution inequality used in the regret proof.

Finite-grid core. By Assumption 8, on the good event there exists a finite-grid correspondence $\mathfrak{K}_{\star,j,\eta}$ such that $\mathfrak{K}_{\star,j,\eta}(x) \subseteq \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$, and $\mathfrak{K}_{\star,j,\eta}$ is additively $(L_r, \alpha, \varepsilon^\eta_{\mathrm{rec},R}, \varepsilon^\eta_{\mathrm{rec},T})$-recoverable. Therefore $\mathfrak{R}_{j,\eta}$ is nonempty.

The finite-grid core is defined by $\mathfrak{K}^{\max}_{j,\eta}(x) = \bigcup_{\mathfrak{K}_\eta \in \mathfrak{R}_{j,\eta}} \mathfrak{K}_\eta(x)$. Since every $\mathfrak{K}_\eta \in \mathfrak{R}_{j,\eta}$ is a subcorrespondence of $\mathfrak{E}_{j,\eta}$, the union also satisfies $\mathfrak{K}^{\max}_{j,\eta}(x) \subseteq \mathfrak{E}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$. Moreover, because $\mathfrak{K}_{\star,j,\eta} \in \mathfrak{R}_{j,\eta}$, $\mathfrak{K}_{\star,j,\eta}(x) \subseteq \mathfrak{K}^{\max}_{j,\eta}(x)$ for all $x \in \mathcal{X}_\eta$.

We also need that the union remains additively recoverable. Fix $x, y \in \mathcal{X}_\eta$ and $(a,\rho,z) \in \mathfrak{K}^{\max}_{j,\eta}(x)$. By definition of the union, there exists some $\mathfrak{K}_\eta \in \mathfrak{R}_{j,\eta}$ such that $(a,\rho,z) \in \mathfrak{K}_\eta(x)$. Since $\mathfrak{K}_\eta$ is additively recoverable, there exists $(a',\rho',z') \in \mathfrak{K}_\eta(y)$ satisfying $\rho - \rho' \le L_r\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},R}$ and $\|z-z'\|_2 \le \alpha\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}$. Since $\mathfrak{K}_\eta(y) \subseteq \mathfrak{K}^{\max}_{j,\eta}(y)$, the same triple is a valid recovery witness in the finite-grid core. Thus $\mathfrak{K}^{\max}_{j,\eta}$ is additively recoverable.

Approximate Lipschitzness of the raw backup. Recall that $\varepsilon^\eta_{\mathrm{val}} := \varepsilon^\eta_{\mathrm{rec},R} + L_{\mathrm{lat}}\varepsilon^\eta_{\mathrm{rec},T}$. At stage $H+1$, the continuation value $\widetilde{V}^j_{H+1,\eta} \equiv 0$ is exactly L_{lat}-Lipschitz. Suppose $\widetilde{V}^j_{h+1,\eta}$ is L_{lat}-Lipschitz on the grid.

Fix $x, y \in \mathcal{X}_\eta$, and let $(a,\rho,z) \in \mathfrak{K}^{\max}_{j,\eta}(x)$ attain the maximum in the raw backup at x, so that $Q^j_{h,\eta}(x) = \rho + \widetilde{V}^j_{h+1,\eta}(z)$. Since the core is additively recoverable, there exists $(a',\rho',z') \in \mathfrak{K}^{\max}_{j,\eta}(y)$ such that $\rho - \rho' \le L_r\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},R}$ and $\|z-z'\|_2 \le \alpha\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}$. Because (a',ρ',z') is feasible in the raw backup at y, $Q^j_{h,\eta}(y) \ge \rho' + \widetilde{V}^j_{h+1,\eta}(z')$. Therefore

$$\begin{aligned}
Q^j_{h,\eta}(x) - Q^j_{h,\eta}(y) &\le \rho - \rho' + \widetilde{V}^j_{h+1,\eta}(z) - \widetilde{V}^j_{h+1,\eta}(z') \\
&\le L_r\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},R} + L_{\mathrm{lat}}\|z-z'\|_2 \\
&\le (L_r + \alpha L_{\mathrm{lat}})\,\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},R} + L_{\mathrm{lat}}\varepsilon^\eta_{\mathrm{rec},T} \\
&= L_{\mathrm{lat}}\|x-y\|_2 + \varepsilon^\eta_{\mathrm{val}}.
\end{aligned}$$

In the last line we used $L_r + \alpha L_{\mathrm{lat}} = L_{\mathrm{lat}}$. Repeating the same argument with x and y interchanged gives

$$|Q^j_{h,\eta}(x) - Q^j_{h,\eta}(y)| \le L_{\mathrm{lat}}\|x-y\|_2 + \varepsilon^\eta_{\mathrm{val}} \qquad \text{for all } x, y \in \mathcal{X}_\eta.$$

Lipschitz envelope and sandwich bound. For any function Q on $\mathcal{X}_\eta$, define its L_{lat}-Lipschitz majorant by $\mathcal{M}Q(x) := \max_{y\in\mathcal{X}_\eta}[Q(y) - L_{\mathrm{lat}}\|x-y\|_2]$. This function is L_{lat}-Lipschitz, because it is the pointwise maximum of L_{lat}-Lipschitz functions of x. Applying this operator to $Q^j_{h,\eta}$ gives $\widetilde{V}^j_{h,\eta} = \mathcal{M}Q^j_{h,\eta}$, and hence $\widetilde{V}^j_{h,\eta}$ is exactly L_{lat}-Lipschitz on $\mathcal{X}_\eta$.

The lower sandwich inequality follows by choosing $y = x$ in the maximum: $\widetilde{V}^j_{h,\eta}(x) \ge Q^j_{h,\eta}(x)$. For the upper sandwich inequality, the approximate Lipschitz bound for $Q^j_{h,\eta}$ implies that, for every $y \in \mathcal{X}_\eta$, $Q^j_{h,\eta}(y) - L_{\mathrm{lat}}\|x-y\|_2 \le Q^j_{h,\eta}(x) + \varepsilon^\eta_{\mathrm{val}}$. Taking the maximum over y gives $\widetilde{V}^j_{h,\eta}(x) \le Q^j_{h,\eta}(x) + \varepsilon^\eta_{\mathrm{val}}$, which proves (21). Since the envelope step produces an exactly L_{lat}-Lipschitz value function, the backward induction on Lipschitz regularity closes.

The span bound follows immediately: $\mathrm{span}(\widetilde{V}^j_{h,\eta}) \le L_{\mathrm{lat}}\,\mathrm{diam}(\mathcal{X}) = B_{\mathrm{lat}}$.

Approximate supersolution inequality. Fix a continuous state $x \in \mathcal{X}$, action $a \in \mathcal{A}$, and stage h. Write $\bar{x} := \Pi_X x$ and $\bar{x}^+ := \Pi_X T_\star(x,a)$. By Assumption 8, there exists $(\bar{a}, \bar{\rho}, \bar{z}) \in \mathfrak{K}_{\star,j,\eta}(\bar{x}) \subseteq \mathfrak{K}^{\max}_{j,\eta}(\bar{x})$ such that $\bar{\rho} \ge r_\star(x,a) - C^\eta_{\mathrm{app}}\Delta_\eta$ and $\|\bar{z} - \bar{x}^+\|_2 \le C^\eta_{\mathrm{app}}\Delta_\eta$. Using this triple in the raw backup at $\bar{x}$ yields $Q^j_{h,\eta}(\bar{x}) \ge \bar{\rho} + \widetilde{V}^j_{h+1,\eta}(\bar{z})$. Since $\widetilde{V}^j_{h,\eta}(\bar{x}) \ge Q^j_{h,\eta}(\bar{x})$, we have $\widetilde{V}^j_{h,\eta}(\bar{x}) \ge \bar{\rho} + \widetilde{V}^j_{h+1,\eta}(\bar{z})$. Rearranging and using the lower bound on $\bar{\rho}$ gives $r_\star(x,a) \le \bar{\rho} + C^\eta_{\mathrm{app}}\Delta_\eta \le \widetilde{V}^j_{h,\eta}(\bar{x}) - \widetilde{V}^j_{h+1,\eta}(\bar{z}) + C^\eta_{\mathrm{app}}\Delta_\eta$. Finally, exact L_{lat}-Lipschitzness of $\widetilde{V}^j_{h+1,\eta}$ gives $-\widetilde{V}^j_{h+1,\eta}(\bar{z}) \le -\widetilde{V}^j_{h+1,\eta}(\bar{x}^+) + L_{\mathrm{lat}}\|\bar{z} - \bar{x}^+\|_2$. Therefore

$$\begin{aligned} r_\star(x,a) &\le \widetilde{V}^j_{h,\eta}(\bar{x}) - \widetilde{V}^j_{h+1,\eta}(\bar{x}^+) + C^\eta_{\mathrm{app}}\Delta_\eta + L_{\mathrm{lat}}C^\eta_{\mathrm{app}}\Delta_\eta \\ &= \widetilde{V}^j_{h,\eta}(\Pi_X x) - \widetilde{V}^j_{h+1,\eta}(\Pi_X T_\star(x,a)) + C^\eta_{\mathrm{app}}(1 + L_{\mathrm{lat}})\Delta_\eta. \end{aligned}$$

This is the desired approximate supersolution inequality. □

Remark 4 (Computational cost). The nonstandard computational step is the pruning step that constructs the recoverable core. Fix a grid state x and a candidate extended action $e = (a, \rho, z) \in \mathfrak{K}_\eta(x)$. The question in one pruning pass is whether this candidate can be matched from every other grid state. Thus, for each $y \in \mathcal{X}_\eta$, the algorithm looks for a recovery witness, namely a triple $(a', \rho', z') \in \mathfrak{K}_\eta(y)$ with $\rho' \ge \rho - L_r\|x-y\|_2 - \varepsilon^\eta_{\mathrm{rec},R}$ and $\|z - z'\|_2 \le \alpha\|x-y\|_2 + \varepsilon^\eta_{\mathrm{rec},T}$. If such a witness is missing for some y, then e is removed from $\mathfrak{K}_\eta(x)$. The pruning procedure repeats this test because removing one candidate may invalidate another candidate's witness.

A direct implementation is easy but conservative. Let $N_X = |\mathcal{X}_\eta|$ and let $N_E = \max_x |\mathfrak{E}_{j,\eta}(x)|$ at a fixed epoch. One pruning pass checks at most $N_X N_E$ candidate state-action triples, compares each of them against N_X possible recovery states, and may scan up to N_E candidates at each recovery state. Thus a naive pass costs on the order of $N_X^2 N_E^2$ comparisons. Since each nonterminal pass removes at least one candidate, the number of passes is at most $N_X N_E + 1$. This gives a crude

polynomial bound for one epoch; it is intended only to show implementability, not to describe the fastest possible data structure.

The witness search has useful structure. For fixed y, the next-state condition asks whether there is a candidate z' in a ball centered at z, and the reward condition asks whether at least one such candidate has reward above a threshold. Thus the inner scan can be viewed as a range-search query: among candidate next states near z, find the maximum available reward. Spatial indexing can reduce this inner cost, especially when the state grid has low dimension.

The core is recomputed only at determinant-doubling epochs, so the number of pruning calls over the horizon is $O(\gamma_R(H)+\gamma_T(H))$. If the grids are refined logarithmically in H and the covering dimensions are fixed, the total pruning cost is polynomial in $\log H$, with an exponent that depends on the grid dimensions. After pruning, the remaining computation is standard finite-horizon value iteration on the surviving extended actions. The Lipschitz envelope in (20) adds one deterministic maximization over the grid after each Bellman backup; on a one-dimensional uniform grid this envelope can be computed in linear time per layer by two running maxima, while on a general finite grid it is computed by the displayed finite maximum.

Appendix G: Regret Proofs

This appendix proves the regret accounting inequalities used in the main text. The first lemma is a warm-up for the full-model optimistic planner under robust recovery. In this setting, each Bellman backup optimizes directly over the parameter confidence class, and the value functions are continuous-state Bellman values. The proof isolates the basic telescoping mechanism: center the optimistic values within each lazy epoch, upper-bound the benchmark trajectory by optimism, lower-bound the realized trajectory by the optimistic Bellman backup, and pay only for statistical widths, finite-batch state error, and epoch changes.

LEMMA 3 **(Robust-recovery regret decomposition)**. *Suppose Assumption 5 holds. On $\mathcal{E}$, the lazy model-optimistic planner satisfies*

$$\mathrm{Reg}(H) \le 2B_{\mathrm{lat}}N_H + 2\sum_{t=1}^{H} b^R_{j(t)}(\hat{x}_t, a_t) + 2L_{\mathrm{lat}}\sum_{t=1}^{H} b^T_{j(t)}(\hat{x}_t, a_t) + C_x H\beta^x,$$

where $C_x = S_\theta L_\Psi + L_{\mathrm{lat}}(S_W L_\Phi + 1)$.

Proof. The proof is the continuous, full-model counterpart of the finite-grid decomposition below. We first define centered value functions, then derive a benchmark upper bound and a realized-trajectory lower bound, and finally subtract the two inequalities. For each epoch j and stage h, define

$$m^j_h := \min_{x\in\mathcal{X}} \bar{V}^j_h(x), \qquad \varphi^j_h(x) := \bar{V}^j_h(x) - m^j_h, \qquad g^j_h := m^j_h - m^j_{h+1}.$$

By Proposition 1, on $\mathcal{E}$, $0 \le \varphi^j_h(x) \le B_{\mathrm{lat}}$ for all $x \in \mathcal{X}$, and φ^j_h is L_{lat}-Lipschitz.

Benchmark upper bound. Because the true model belongs to $\mathcal{M}_j$ on $\mathcal{E}$, the optimistic Bellman recursion implies the centered supersolution inequality

$$r_\star(x,a) \le g_h^j + \varphi_h^j(x) - \varphi_{h+1}^j(T_\star(x,a)) \tag{22}$$

for every $x \in \mathcal{X}$, $a \in \mathcal{A}$, epoch j, and stage h. Indeed, using the true model and action a as a feasible choice in the Bellman backup gives $\bar{V}_h^j(x) \ge r_\star(x,a) + \bar{V}_{h+1}^j(T_\star(x,a))$. Subtracting m_{h+1}^j from both sides and using $\bar{V}_h^j(x) - m_{h+1}^j = g_h^j + \varphi_h^j(x)$ and $\bar{V}_{h+1}^j(T_\star(x,a)) - m_{h+1}^j = \varphi_{h+1}^j(T_\star(x,a))$ yields (22).

Let $(x_t^\star, a_t^\star)_{t=1}^H$ be an optimal true-model trajectory from x_1. Applying the centered supersolution inequality with $h = t$, epoch $j(t)$, state $x_t^\star$, and action $a_t^\star$, and using $x_{t+1}^\star = T_\star(x_t^\star, a_t^\star)$, gives

$$r_\star(x_t^\star, a_t^\star) \le g_t^{j(t)} + \varphi_t^{j(t)}(x_t^\star) - \varphi_{t+1}^{j(t)}(x_{t+1}^\star).$$

Summing over $t = 1, \dots, H$, the potential terms telescope within each lazy epoch. More explicitly, on an epoch interval $I_j = \{\tau_j, \dots, \tau_{j+1} - 1\}$,

$$\sum_{t \in I_j} \left[\varphi_t^j(x_t^\star) - \varphi_{t+1}^j(x_{t+1}^\star)\right] = \varphi_{\tau_j}^j(x_{\tau_j}^\star) - \varphi_{\tau_{j+1}}^j(x_{\tau_{j+1}}^\star) \le B_{\mathrm{lat}}.$$

For the final epoch $I_j = \{\tau_j, \dots, H\}$, the same bound holds with τ_{j+1} replaced by $H+1$. Therefore

$$V_1^\star(x_1) = \sum_{t=1}^H r_\star(x_t^\star, a_t^\star) \le \sum_{t=1}^H g_t^{j(t)} + B_{\mathrm{lat}} N_H. \tag{23}$$

Trajectory lower bound. Next consider the realized trajectory. At time t, let $j = j(t)$. The decision maker chooses a_t by evaluating the epoch-j Bellman backup at $\hat{x}_t$. Let $(\theta_t, W_t) \in \mathcal{M}_j$ be an optimistic model attaining this backup, and set $\rho_t := \theta_t^\top \Psi(\hat{x}_t, a_t)$ and $z_t := W_t \Phi(\hat{x}_t, a_t)$. If the supremum is not attained, the same argument applies to an ϵ_t-optimal triple and then lets $\sum_t \epsilon_t \downarrow 0$. For an exact maximizer, the Bellman backup gives $\bar{V}_t^j(\hat{x}_t) = \rho_t + \bar{V}_{t+1}^j(z_t)$. Subtracting m_{t+1}^j from both sides yields $g_t^j + \varphi_t^j(\hat{x}_t) = \rho_t + \varphi_{t+1}^j(z_t)$. Equivalently, and in the form used below,

$$g_t^j + \varphi_t^j(\hat{x}_t) \le \rho_t + \varphi_{t+1}^j(z_t). \tag{24}$$

We compare (ρ_t, z_t) with the true one-step outcome. Since both θ_t and $\theta_\star$ belong to $\mathcal{C}_{\tau_j}^R$,

$$\rho_t - r_\star(\hat{x}_t, a_t) \le 2 b_j^R(\hat{x}_t, a_t). \tag{25}$$

Moreover,

$$|r_\star(\hat{x}_t, a_t) - r_\star(x_t, a_t)| \le S_\theta L_\Psi \beta^x. \tag{26}$$

Similarly, since W_t and $W_\star$ belong to $\mathcal{C}_{\tau_j}^T$, $\|z_t - T_\star(\hat{x}_t, a_t)\|_2 \le 2 b_j^T(\hat{x}_t, a_t)$. Using $T_\star(x,a) = W_\star \Phi(x,a)$, the Lipschitzness of Φ, and the state-confidence event, we also have $\|T_\star(\hat{x}_t, a_t) - x_{t+1}\|_2 \le S_W L_\Phi \beta^x$

and $\|x_{t+1} - \hat{x}_{t+1}\|_2 \le \beta^x$. Therefore, $\|z_t - \hat{x}_{t+1}\|_2 \le 2b_j^T(\hat{x}_t, a_t) + (S_W L_\Phi + 1)\beta^x$. The Lipschitzness of φ_{t+1}^j then gives

$$\varphi_{t+1}^j(z_t) - \varphi_{t+1}^j(\hat{x}_{t+1}) \le 2L_{\text{lat}} b_j^T(\hat{x}_t, a_t) + L_{\text{lat}}(S_W L_\Phi + 1)\beta^x. \tag{27}$$

Combining (24), (25), (26), and (27) yields the one-step lower bound

$$r_\star(x_t, a_t) \ge g_t^j + \varphi_t^j(\hat{x}_t) - \varphi_{t+1}^j(\hat{x}_{t+1}) - 2b_j^R(\hat{x}_t, a_t) - 2L_{\text{lat}} b_j^T(\hat{x}_t, a_t) - C_x\beta^x.$$

Summing over $t = 1, \ldots, H$, the realized-trajectory potentials telescope within each lazy epoch. On an epoch interval I_j,

$$\sum_{t \in I_j} \left[\varphi_t^j(\hat{x}_t) - \varphi_{t+1}^j(\hat{x}_{t+1})\right] = \varphi_{\tau_j}^j(\hat{x}_{\tau_j}) - \varphi_{\tau_{j+1}}^j(\hat{x}_{\tau_{j+1}}) \ge -B_{\text{lat}}.$$

Again, the same bound applies to the final epoch with τ_{j+1} replaced by $H+1$. Thus

$$\sum_{t=1}^H r_\star(x_t, a_t) \ge \sum_{t=1}^H g_t^{j(t)} - B_{\text{lat}} N_H - 2\sum_{t=1}^H b_{j(t)}^R(\hat{x}_t, a_t) - 2L_{\text{lat}} \sum_{t=1}^H b_{j(t)}^T(\hat{x}_t, a_t) - C_x H\beta^x.$$

Subtracting this lower bound from (23) proves the claim. □

The finite-grid recoverable-core proof follows the same benchmark-upper-bound minus realized-trajectory-lower-bound template, but three new issues must be handled. First, the planner works with local extended triples (a, ρ, z) rather than parameter pairs from the confidence class. Second, those triples are restricted to a recoverable core, so the realized-trajectory lower bound must use the raw finite-grid backup associated with an actually executable core action. Third, projection to a finite grid introduces $O(\Delta_\eta)$ errors. The Lipschitz envelope supplies exact value regularity on the grid, and the approximate supersolution inequality absorbs the resulting discretization terms. The next lemma records the corresponding accounting bound.

LEMMA 4 **(Discretized regret decomposition)**. *On $\mathcal{E}$, the lazy finite-grid implementation of Recoverable-Core Latent UCB satisfies*

$$\text{Reg}(H) \le 2B_{\text{lat}} N_H + 2\sum_{t=1}^H \left[b_{j(t)}^R(\bar{x}_t, a_t) + L_{\text{lat}} b_{j(t)}^T(\bar{x}_t, a_t)\right] + C_x H\beta^x + C_{\text{grid}}^\eta H\Delta_\eta,$$

where C_{grid}^η is independent of H.

Proof. The proof converts the finite-grid optimistic values into centered potentials that telescope within each lazy epoch. The benchmark trajectory is upper bounded by these potentials using the approximate supersolution inequality, while the decision maker's realized reward is lower bounded using the raw Bellman backup, local confidence bounds, and Lipschitz regularity of the continuation potential. The only non-telescoping terms occur at epoch boundaries.

For each epoch j and stage h, define

$$m^j_{h,\eta} := \min_{y\in\mathcal{X}_\eta} \widetilde{V}^j_{h,\eta}(y), \qquad \varphi^j_{h,\eta}(x) := \widetilde{V}^j_{h,\eta}(x) - m^j_{h,\eta},$$

and $g^j_{h,\eta} := m^j_{h,\eta} - m^j_{h+1,\eta}$. By Proposition 8, on the good event, $0 \le \varphi^j_{h,\eta}(x) \le B_{\mathrm{lat}}$ for all $x \in \mathcal{X}_\eta$, and $\varphi^j_{h,\eta}$ is L_{lat}-Lipschitz on $\mathcal{X}_\eta$. Moreover, the approximate supersolution inequality in Proposition 8 can be written in centered form as

$$r_\star(x,a) \le g^j_{h,\eta} + \varphi^j_{h,\eta}(\Pi_X x) - \varphi^j_{h+1,\eta}(\Pi_X T_\star(x,a)) + C^\eta_{\mathrm{app}}(1+L_{\mathrm{lat}})\Delta_\eta. \tag{28}$$

Also, if (a_x, ρ_x, z_x) attains the raw backup $Q^j_{h,\eta}(x)$, then the envelope sandwich (21) gives

$$g^j_{h,\eta} + \varphi^j_{h,\eta}(x) \le \rho_x + \varphi^j_{h+1,\eta}(z_x) + \varepsilon^\eta_{\mathrm{val}}. \tag{29}$$

Indeed, this follows from $\widetilde{V}^j_{h,\eta}(x) \le Q^j_{h,\eta}(x) + \varepsilon^\eta_{\mathrm{val}} = \rho_x + \widetilde{V}^j_{h+1,\eta}(z_x) + \varepsilon^\eta_{\mathrm{val}}$ after subtracting $m^j_{h+1,\eta}$ from both sides.

Benchmark upper bound. Let $(x^\star_t, a^\star_t)_{t=1}^H$ be an optimal true-model trajectory from x_1. Applying (28) with $h = t$, epoch $j(t)$, state $x^\star_t$, and action $a^\star_t$, and using $x^\star_{t+1} = T_\star(x^\star_t, a^\star_t)$, gives

$$r_\star(x^\star_t, a^\star_t) \le g^{j(t)}_{t,\eta} + \varphi^{j(t)}_{t,\eta}(\Pi_X x^\star_t) - \varphi^{j(t)}_{t+1,\eta}(\Pi_X x^\star_{t+1}) + C^\eta_{\mathrm{app}}(1+L_{\mathrm{lat}})\Delta_\eta.$$

Summing over $t = 1, \dots, H$, the potential terms telescope within each lazy epoch. More explicitly, on an epoch interval $I_j = \{\tau_j, \dots, \tau_{j+1} - 1\}$,

$$\sum_{t\in I_j} \left[\varphi^j_{t,\eta}(\Pi_X x^\star_t) - \varphi^j_{t+1,\eta}(\Pi_X x^\star_{t+1})\right] = \varphi^j_{\tau_j,\eta}(\Pi_X x^\star_{\tau_j}) - \varphi^j_{\tau_{j+1},\eta}(\Pi_X x^\star_{\tau_{j+1}}) \le B_{\mathrm{lat}}.$$

For the final epoch, the same bound holds with τ_{j+1} replaced by $H+1$. Therefore

$$V^\star_1(x_1) = \sum_{t=1}^H r_\star(x^\star_t, a^\star_t) \le \sum_{t=1}^H g^{j(t)}_{t,\eta} + B_{\mathrm{lat}} N_H + C^\eta_{\mathrm{app}}(1+L_{\mathrm{lat}}) H \Delta_\eta. \tag{30}$$

Trajectory lower bound. At time t, let $j = j(t)$. The algorithm chooses $(a_t, \rho_t, z_t) \in \mathfrak{K}^{\max}_{j,\eta}(\bar{x}_t)$ to attain the raw backup at $\bar{x}_t$. Applying (29) at $x = \bar{x}_t$ gives

$$g^j_{t,\eta} + \varphi^j_{t,\eta}(\bar{x}_t) \le \rho_t + \varphi^j_{t+1,\eta}(z_t) + \varepsilon^\eta_{\mathrm{val}}. \tag{31}$$

We next relate the optimistic one-step quantities (ρ_t, z_t) to the true one-step reward and next latent state. Since $\mathfrak{K}^{\max}_{j,\eta}(\bar{x}_t) \subseteq \mathfrak{E}_{j,\eta}(\bar{x}_t)$, the definition of the inflated finite raw correspondence gives $|\rho_t - \hat{r}_j(\bar{x}_t, a_t)| \le b^R_j(\bar{x}_t, a_t) + \varepsilon^\eta_R$ and $\|z_t - \hat{T}_j(\bar{x}_t, a_t)\|_2 \le b^T_j(\bar{x}_t, a_t) + \varepsilon^\eta_T$. On the good event, $|r_\star(\bar{x}_t, a_t) - \hat{r}_j(\bar{x}_t, a_t)| \le b^R_j(\bar{x}_t, a_t)$, and hence $\rho_t - r_\star(\bar{x}_t, a_t) \le 2b^R_j(\bar{x}_t, a_t) + \varepsilon^\eta_R$. Using the Lipschitzness of Ψ, the bound $\|\theta_\star\|_2 \le S_\theta$, and $\|x_t - \bar{x}_t\|_2 \le \beta^x + \eta_x$, we obtain $|r_\star(\bar{x}_t, a_t) - r_\star(x_t, a_t)| \le S_\theta L_\Psi(\beta^x + \eta_x)$. Therefore

$$\rho_t - r_\star(x_t, a_t) \le 2b^R_j(\bar{x}_t, a_t) + S_\theta L_\Psi \beta^x + (S_\theta L_\Psi \eta_x + \varepsilon^\eta_R). \tag{32}$$

For the transition displacement, the good event gives

$$\|T_\star(\bar{x}_t, a_t) - \hat{T}_j(\bar{x}_t, a_t)\|_2 \le b_j^T(\bar{x}_t, a_t). \tag{33}$$

Feasibility of z_t in the finite raw correspondence gives $\|z_t - \hat{T}_j(\bar{x}_t, a_t)\|_2 \le b_j^T(\bar{x}_t, a_t) + \varepsilon_T^\eta$. Combining this feasibility bound with (33) gives $\|z_t - T_\star(\bar{x}_t, a_t)\|_2 \le 2b_j^T(\bar{x}_t, a_t) + \varepsilon_T^\eta$. Moreover, $T_\star(\bar{x}_t, a_t) = W_\star \Phi(\bar{x}_t, a_t)$ and $x_{t+1} = W_\star \Phi(x_t, a_t)$, so $\|T_\star(\bar{x}_t, a_t) - x_{t+1}\|_2 \le S_W L_\Phi \|\bar{x}_t - x_t\|_2 \le S_W L_\Phi(\beta^x + \eta_x)$. Since $\|x_{t+1} - \bar{x}_{t+1}\|_2 \le \beta^x + \eta_x$, we get

$$\|z_t - \bar{x}_{t+1}\|_2 \le 2b_j^T(\bar{x}_t, a_t) + (S_W L_\Phi + 1)\beta^x + ((S_W L_\Phi + 1)\eta_x + \varepsilon_T^\eta). \tag{34}$$

Because $\varphi_{t+1,\eta}^j$ is L_{lat}-Lipschitz on the grid, (34) implies

$$\varphi_{t+1,\eta}^j(z_t) - \varphi_{t+1,\eta}^j(\bar{x}_{t+1}) \le L_{\text{lat}}\left[2b_j^T(\bar{x}_t, a_t) + (S_W L_\Phi + 1)(\beta^x + \eta_x) + \varepsilon_T^\eta\right]. \tag{35}$$

Combining (31), (32), and (35), and using $\varepsilon_{\text{val}}^\eta = \varepsilon_{\text{rec},R}^\eta + L_{\text{lat}}\varepsilon_{\text{rec},T}^\eta$, gives the one-step trajectory lower bound

$$r_\star(x_t, a_t) \ge g_{t,\eta}^j + \varphi_{t,\eta}^j(\bar{x}_t) - \varphi_{t+1,\eta}^j(\bar{x}_{t+1}) - 2b_j^R(\bar{x}_t, a_t) - 2L_{\text{lat}} b_j^T(\bar{x}_t, a_t) - C_x\beta^x - C_{\text{loc}}^\eta \Delta_\eta.$$

Here C_{loc}^η is a finite constant, independent of H, that absorbs $S_\theta L_\Psi \eta_x + \varepsilon_R^\eta$, $L_{\text{lat}}((S_W L_\Phi + 1)\eta_x + \varepsilon_T^\eta)$, and $\varepsilon_{\text{val}}^\eta$. Since $\eta_x \le \Delta_\eta$, $\varepsilon_R^\eta = O(\Delta_\eta)$, $\varepsilon_T^\eta = O(\Delta_\eta)$, and $\varepsilon_{\text{val}}^\eta = O(\Delta_\eta)$, this constant does not depend on H.

Summing over $t = 1, \ldots, H$, the decision maker's potential terms telescope within each epoch. On an epoch interval $I_j = \{\tau_j, \ldots, \tau_{j+1} - 1\}$,

$$\sum_{t \in I_j} \left[\varphi_{t,\eta}^j(\bar{x}_t) - \varphi_{t+1,\eta}^j(\bar{x}_{t+1})\right] = \varphi_{\tau_j,\eta}^j(\bar{x}_{\tau_j}) - \varphi_{\tau_{j+1},\eta}^j(\bar{x}_{\tau_{j+1}}) \ge -B_{\text{lat}}.$$

Thus

$$\sum_{t=1}^H r_\star(x_t, a_t) \ge \sum_{t=1}^H g_{t,\eta}^{j(t)} - B_{\text{lat}} N_H - 2\sum_{t=1}^H \left[b_{j(t)}^R(\bar{x}_t, a_t) + L_{\text{lat}} b_{j(t)}^T(\bar{x}_t, a_t)\right] - C_x H\beta^x - C_{\text{loc}}^\eta H\Delta_\eta. \tag{36}$$

Combining the two bounds. Define $C_{\text{grid}}^\eta := C_{\text{loc}}^\eta + C_{\text{app}}^\eta(1 + L_{\text{lat}})$. Subtracting (36) from (30) yields the bound stated in Lemma 4. □

Proof of Theorem 1. We work on the good event $\mathcal{E}$, which has probability at least $1 - \delta$. By Lemma 4, it remains to bound the number of epochs and the two cumulative bonus terms.

Number of epochs. A new epoch starts only when the determinant of the reward design matrix or the transition design matrix doubles. Therefore $N_H \le 1 + \gamma_R(H)/\log 2 + \gamma_T(H)/\log 2$.

Reward bonus sum. Fix t, and let $j = j(t)$. By construction of the lazy epochs, $\Lambda^R_{\tau_j} \preceq \Lambda^R_t$ and $\det(\Lambda^R_t) < 2\det(\Lambda^R_{\tau_j})$. Lemma 8 therefore gives $\|\Psi(\bar{x}_t, a_t)\|_{(\Lambda^R_{\tau_j})^{-1}} \le \sqrt{2}\|\Psi(\bar{x}_t, a_t)\|_{(\Lambda^R_t)^{-1}}$. Since $\beta^R_{\tau_j} \le \bar{\beta}^R_H$, we have

$$\begin{aligned}\sum_{t=1}^{H} b^R_{j(t)}(\bar{x}_t, a_t) &= \sum_{t=1}^{H} \beta^R_{\tau_{j(t)}} \|\Psi(\bar{x}_t, a_t)\|_{(\Lambda^R_{\tau_{j(t)}})^{-1}} \\ &\le \sqrt{2}\,\bar{\beta}^R_H \sum_{t=1}^{H} \|\Psi(\bar{x}_t, a_t)\|_{(\Lambda^R_t)^{-1}} \\ &\le \sqrt{2}\,\bar{\beta}^R_H \sqrt{H \sum_{t=1}^{H} \|\Psi(\bar{x}_t, a_t)\|^2_{(\Lambda^R_t)^{-1}}}.\end{aligned}$$

By Lemma 7, $\sum_{t=1}^{H} \|\Psi(\bar{x}_t, a_t)\|^2_{(\Lambda^R_t)^{-1}} \le \kappa_R \gamma_R(H)$. Hence $\sum_{t=1}^{H} b^R_{j(t)}(\bar{x}_t, a_t) \le \bar{\beta}^R_H \sqrt{2H\kappa_R \gamma_R(H)}$.

Transition bonus sum. The same argument with transition features gives $\sum_{t=1}^{H} b^T_{j(t)}(\bar{x}_t, a_t) \le \bar{\beta}^T_H \sqrt{2H\kappa_T \gamma_T(H)}$.

Putting the bounds together. Substituting the epoch-count bound and the two cumulative bonus bounds into Lemma 4 yields

$$\begin{aligned}\mathrm{Reg}(H) \le\; & 2B_{\mathrm{lat}} \left(1 + \frac{\gamma_R(H)}{\log 2} + \frac{\gamma_T(H)}{\log 2}\right) \\ & + 2\bar{\beta}^R_H \sqrt{2H\kappa_R \gamma_R(H)} + 2L_{\mathrm{lat}} \bar{\beta}^T_H \sqrt{2H\kappa_T \gamma_T(H)} \\ & + C_x H \beta^x + C^{\eta}_{\mathrm{grid}} H \Delta_\eta.\end{aligned}$$

Finally, the determinant–trace bounds

$$\gamma_R(H) \le p \log\left(1 + \frac{HB^2_\Psi}{\lambda_R p}\right), \qquad \gamma_T(H) \le q \log\left(1 + \frac{HB^2_\Phi}{\lambda_T q}\right)$$

follow from Lemma 6. Since $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, the theorem holds with probability at least $1 - \delta$. □

Appendix H: Lower Bound Proofs

PROPOSITION 9 **(Fresh-state finite-batch observation lower bound)**. *There is a universal constant $c > 0$ such that the following holds. For every H and n, there exists a sequence of independent one-step latent population decision problems with scalar latent state $m = 1$ and known reward function such that any decision maker that observes only n samples from the current population before choosing its action suffers expected regret at least $cH/\sqrt{n}$ relative to a benchmark that observes the latent state x_t exactly before acting.*

The proposition isolates the statistical cost of observing a fresh latent population state only through a finite batch. It removes dynamic coupling across periods, so the loss is caused by repeated state estimation rather than transition learning. The proof is a Le Cam two-point argument (Tsybakov 2009, Chapter 2).

Proof of Proposition 9. Let $\mathcal{Z} = \{-1, +1\}$, let $\psi(z) = z$, and let the action space be $\mathcal{A} = [-1, 1]$. For a parameter $\sigma \in \{-1, +1\}$, define d_σ by $\mathbb{P}_{d_\sigma}(Z = 1) = (1 + \sigma\Delta)/2$ and $\mathbb{P}_{d_\sigma}(Z = -1) = (1 - \sigma\Delta)/2$.

Then $x(d_\sigma) = \mathbb{E}_{Z\sim d_\sigma}[Z] = \sigma\Delta$. At each round t, independently draw $\sigma_t \sim \mathrm{Unif}\{-1,+1\}$, set $d_t = d_{\sigma_t}$, and reveal to the decision maker only an independent batch of n samples from d_t. The reward is known and given by $r_\star(x,a) = ax$. There is no controlled transition map in this construction; the period-t distribution is redrawn independently in each one-step problem. A benchmark that observes $x_t = \sigma_t\Delta$ exactly chooses $a_t^\star = \sigma_t$ and receives reward Δ.

Any decision maker must effectively test whether the current distribution is d_+ or d_-. Let P_+^n and P_-^n denote the laws of the n-sample batch, and let $\widehat{\sigma} \in \{-1,+1\}$ denote an arbitrary, possibly randomized, estimator of the sign σ based on this batch. Let Prob denote probability under a uniformly random sign $\sigma \in \{-1,+1\}$ and an n-sample batch drawn from d_σ. The standard total-variation testing bound gives

$$\inf_{\widehat{\sigma}} \mathrm{Prob}(\widehat{\sigma} \neq \sigma) = \inf_{\widehat{\sigma}} \frac{1}{2}P_+^n(\widehat{\sigma} = -1) + \frac{1}{2}P_-^n(\widehat{\sigma} = +1) \geq \frac{1 - \|P_+^n - P_-^n\|_{\mathrm{TV}}}{2}.$$

Moreover,

$$\mathrm{KL}(P_+^n \,\|\, P_-^n) = n\,\mathrm{KL}(d_+ \,\|\, d_-) \leq 4n\Delta^2$$

for $\Delta \leq 1/2$. Hence Pinsker's inequality gives

$$\|P_+^n - P_-^n\|_{\mathrm{TV}} \leq \sqrt{2n\Delta^2}.$$

Choose $\Delta = 1/(4\sqrt{2n})$. Then every sign estimator has error probability at least $3/8$. For the decision maker's action a_t, define the induced sign estimate $\widehat{\sigma}_t = +1$ if $a_t \geq 0$ and $\widehat{\sigma}_t = -1$ otherwise. Since the testing bound applies to every sign estimator, this induced estimator has error probability at least $3/8$. If $\widehat{\sigma}_t \neq \sigma_t$, then $\sigma_t a_t \leq 0$, so the decision maker's reward is at most zero while the latent-state oracle obtains Δ. Hence the expected regret per round is $\Omega(1/\sqrt{n})$, and the cumulative regret is $\Omega(H/\sqrt{n})$. $\square$

Proof of Theorem 2. Let the latent state be $x_t = (u_t, \iota_t) \in [-1,1] \times \{0,1\}$, with initial state $x_1 = (0,0)$. The action space is $\mathcal{A} = [-1,1]$. There are two possible models, indexed by $\sigma \in \{+1,-1\}$. The transition is

$$u_{t+1} = \iota_t u_t + (1-\iota_t)\sigma a_t, \qquad \iota_{t+1} = 1.$$

Thus the first action determines the absorbing value of u. After the first round, $\iota_t = 1$ and $u_{t+1} = u_t$. These instances are not recoverable in the sense used in the positive results: for any two states $x = (u,1)$ and $y = (v,1)$, all actions a, a' satisfy $\|T_\sigma(x,a) - T_\sigma(y,a')\|_2 = |u-v| = \|x-y\|_2$, so no contraction with $\alpha < 1$ is possible when $u \neq v$.

This transition is linear in the known feature map and model-dependent matrix

$$\Phi((u,\iota),a) = \begin{pmatrix} \iota u \\ (1-\iota)a \\ 1 \end{pmatrix}, \qquad W_\sigma = \begin{pmatrix} 1 & \sigma & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The reward is known and given by $r(x,a)=(1+u)/2$. The full-information benchmark knows σ. It chooses $a_1=\sigma$, reaches $u_2=1$, and receives reward 1 in every round $t=2,\dots,H$.

Any algorithm chooses some (possibly randomized) first action $a_1\in[-1,1]$ before distinguishing $\sigma=+1$ from $\sigma=-1$. Under model σ, $u_2=\sigma a_1$. Hence the decision maker's reward in each round $t=2,\dots,H$ is $(1+\sigma a_1)/2$, whereas the benchmark receives 1. Therefore the regret under model σ is at least $(H-1)(1-\sigma a_1)/2$. For any possibly randomized a_1,

$$\sup_{\sigma\in\{+1,-1\}}\mathbb{E}_\sigma[\mathrm{Reg}(H)]\geq\sup_{\sigma\in\{+1,-1\}}\mathbb{E}\left[\frac{H-1}{2}(1-\sigma a_1)\right]\geq\frac{H-1}{2}.$$

The expectation in the middle term is over the algorithm's internal randomization. □

Appendix I: Linear Algebra Tools

We collect the standard determinant and elliptical-potential facts used in the regret analysis.

LEMMA 5 **(Matrix determinant lemma; Horn and Johnson (2012))**. *If $A\in\mathbb{R}^{d\times d}$ is invertible and $u,v\in\mathbb{R}^d$, then* $\det(A+uv^\top)=\det(A)(1+v^\top A^{-1}u)$.

LEMMA 6 **(Determinant–trace bound; Lemma 5.12 of Agarwal et al. (2019))**. *Let $\Lambda_{H+1}=\lambda I_d+\sum_{t=1}^H z_t z_t^\top$, where $\lambda>0$ and $\|z_t\|_2\leq B$. Then*

$$\log\frac{\det(\Lambda_{H+1})}{\det(\lambda I_d)}\leq d\log\left(1+\frac{HB^2}{\lambda d}\right).$$

LEMMA 7 **(Elliptical potential)**. *Adapted from Lemmas 5.11–5.12 of Agarwal et al. (2019), let $\Lambda_t=\lambda I_d+\sum_{s<t}z_s z_s^\top$, where $\lambda>0$ and $\|z_t\|_2\leq B$. Define*

$$\gamma_H:=\log\frac{\det(\Lambda_{H+1})}{\det(\lambda I_d)},\qquad\kappa(B,\lambda):=\frac{B^2/\lambda}{\log(1+B^2/\lambda)}.$$

Then

$$\sum_{t=1}^H\|z_t\|_{\Lambda_t^{-1}}\leq\sqrt{H\kappa(B,\lambda)\gamma_H}.$$

Proof. Set $u_t:=\|z_t\|^2_{\Lambda_t^{-1}}$. The matrix determinant lemma gives $\det(\Lambda_{t+1})=\det(\Lambda_t)(1+u_t)$, hence $\gamma_H=\sum_{t=1}^H\log(1+u_t)$. Also, $u_t\leq B^2/\lambda=:L$. Since concavity gives $\log(1+u)\geq u\log(1+L)/L$ for $0\leq u\leq L$,

$$\sum_{t=1}^H u_t\leq\kappa(B,\lambda)\gamma_H,\qquad\sum_{t=1}^H\|z_t\|_{\Lambda_t^{-1}}=\sum_{t=1}^H\sqrt{u_t}\leq\sqrt{H\sum_{t=1}^H u_t},$$

and the claim follows. □

LEMMA 8 **(Frozen norm comparison)**. *Let $A\preceq B$ be positive definite and* $\det(B)\leq c\det(A)$ *for some $c\geq 1$. Then*

$$\|z\|_{A^{-1}}\leq\sqrt{c}\,\|z\|_{B^{-1}}.$$

Proof. For $M=A^{-1/2}BA^{-1/2}$, we have $M\succeq I$ and $\det(M)\leq c$. Thus every eigenvalue lies in $[1,c]$, so $M\preceq cI$ and $B^{-1}=A^{-1/2}M^{-1}A^{-1/2}\succeq c^{-1}A^{-1}$. Hence $z^\top A^{-1}z\leq cz^\top B^{-1}z$, proving the claim. □