MINIMAX-OPTIMAL ROBUST REINFORCEMENT LEARNING WITH χ^2-DIVERGENCE UNCERTAINTY SET UNDER $\mathcal{SA}$- AND $\mathcal{S}$-RECTANGULARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the sample complexity of model-based distributionally robust reinforcement learning (DR-RL) with χ^2-divergence defined uncertainty set under balanced nominal generative-model sampling. For $\mathcal{SA}$-rectangular uncertainty set, existing upper and lower bounds do not match in the uncertainty level ρ. We close this gap by establishing a tighter upper bound of $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\left(1+\frac{\rho}{1-\gamma}\right)\right)$, which matches with the minimax lower bound and thus we establish the minimax optimality. Here, γ is the discount factor, $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, and ε is the optimality gap. Our proof develops a novel variance-sensitive framework that exploits the second-moment structure of adversarial density ratios to control robust Bellman estimation and to propagate statistical errors while preserving dependence linear in the divergence. For $\mathcal{S}$-rectangular uncertainty set with shared state-wise budget of $|\mathcal{A}|\rho$, we further establish an upper bound of $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\left(1+\frac{|\mathcal{A}|\rho}{1-\gamma}\right)\right)$ and also a matching minimax lower bound, and therefore we establish the minimax optimality also for $\mathcal{S}$-rectangular uncertainty set. Our results show that the sample complexity shall not depend on the minimum of non-zero transition probability of the nominal transition kernel, and provide the correct scaling with the uncertainty level $|\mathcal{A}|\rho$.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 2018) has achieved notable successes in game playing (Mnih et al., 2013), robotic control (Kober et al., 2013), and aligning language models with human preferences (Ouyang et al., 2022). A central challenge in deploying learned policies is that the transition dynamics encountered in practice may differ from those used for training. In robotic control, for example, discrepancies in physical parameters and contact dynamics can undermine the performance of policies transferred from simulation to real systems (Peng et al., 2018). Policies optimized solely for a nominal transition model can be sensitive to such discrepancies, even when they perform well in the training environment (Ou & Bi, 2025; Suilen et al., 2024).

Distributionally robust reinforcement learning (DR-RL) addresses this problem by optimizing over an uncertainty set of transition kernels centered on a nominal kernel P_0. The agent seeks a policy that maximizes the worst-case expected discounted return over this set, whose geometry and size specify the allowed model variations (Ou & Bi, 2025). Statistical guarantees for DR-RL have been developed in offline settings e.g., (Zhou et al., 2021; Blanchet et al., 2023; Wang et al., 2024b), and under generative-model access to the nominal dynamics, e.g., (Panaganti & Kalathil, 2022; Yang et al., 2022; Shi et al., 2023; Wang et al., 2024a). A central statistical question is how many nominal samples are required to learn an ε-optimal robust policy with high probability, and how this sample complexity depends on the size and structure of the uncertainty set.

The rectangularity structure of the uncertainty set specifies how transition perturbations can be coupled across states and actions. Under $\mathcal{SA}$-rectangular uncertainty, the set factors across state–action pairs, so each conditional next-state distribution $P(\cdot|s,a)$ can be perturbed independently within its corresponding uncertainty set. Under $\mathcal{S}$-rectangular uncertainty, perturbations remain independent across states but may be coupled across actions within each state. For example, a shared

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

state-wise divergence budget requires the adversary to jointly choose the action-specific perturbations, with an optimal budget allocation that can depend on the policy's action probabilities. This distinction affects the structure of optimal policies: whereas $\mathcal{SA}$-rectangular uncertainty admits a stationary deterministic optimal policy, $\mathcal{S}$-rectangular uncertainty may require randomization to achieve robust optimality (Yang et al., 2022; Li et al., 2025).

We focus on the uncertainty set defined by χ^2-divergence. It directly controls the second moment of the adversarial-to-nominal density ratio, providing a natural basis for variance-sensitive statistical analysis. However, existing sample-complexity bounds leave unresolved the precise dependence on the uncertainty level ρ. For $\mathcal{SA}$-rectangular uncertainty, Shi et al. (2023) establish the following upper and lower bounds:

$$\text{Upper bound:} \quad \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\left(1+\frac{\max\{\sqrt{\rho},\rho\}}{1-\gamma}\right)\right), \tag{1}$$

$$\text{Lower bound:} \quad \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\left(1+\frac{\rho}{1-\gamma}\right)\right). \tag{2}$$

Here, γ is the discount factor, $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, and ε is the optimality gap. These bounds match up to logarithmic factors when $\rho \lesssim (1-\gamma)^2$ or $\rho \gtrsim 1$, but leave a gap in the intermediate regime $(1-\gamma)^2 \ll \rho \ll 1$. This gap leaves open whether the square-root dependence in the upper bound is statistically unavoidable or an artifact of the analysis.

For $\mathcal{S}$-rectangular uncertainty set, Li et al. (2025) provide finite-sample guarantees for divergence-based models, but their χ^2 bound retains an inverse dependence on the smallest positive nominal transition probability. A sharp characterization of the joint dependence on the uncertainty budget and discount factor, without this additional dependence, therefore remains unresolved. These gaps motivate the following questions:

> *Can a χ^2-specific analysis close the uncertainty-level gap under $\mathcal{SA}$-rectangularity? What is the minimax sample complexity under $\mathcal{S}$-rectangular uncertainty with a shared state-wise divergence budget?*

We address these questions through a variance-sensitive analysis of empirical robust planning. The analysis preserves the quadratic structure of the χ^2-defined constraint when comparing nominal and adversarial variances, and controls how statistical errors propagate through the robust Bellman equations. This yields sample-complexity upper bounds with explicit dependence on the uncertainty budget and discount factor for both rectangularity structures. Together with matching lower bounds, these results establish minimax-optimal sample complexities up to logarithmic factors.

Our main contributions are summarized as follows.

- **Minimax optimality for $\mathcal{SA}$-rectangularity.** For $\mathcal{SA}$-rectangular χ^2-divergence defined uncertainty set, we first develop a sample complexity upper bound: $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3}+\frac{\rho}{(1-\gamma)^4}\right)\right)$. This matches the lower bound in eq. (2) established by Shi et al. (2023), up to logarithmic factors, throughout the uncertainty-level range covered by our theorems, establishing minimax optimality. For $0<\rho<1$, our bound replaces the $\sqrt{\rho}$ term in eq. (1) with ρ, closing the gap in the intermediate regime. It also shows that the standard RL sample-complexity is retained whenever $\rho \lesssim 1-\gamma$.
- **A new variance-sensitive framework for DR-RL with χ^2-divergence defined uncertainty set.** We develop a new analytical framework that connects the statistical sensitivity of robust Bellman estimation to the second-moment structure of adversarial density ratios. The key novelty in the proof is a variance decomposition technique that controls normalized dual fluctuations, including their higher-order terms, by variance under the worst-case transition kernel and a remainder linear in ρ. We further establish a nominal-to-adversarial variance comparison valid for every feasible transition kernel. This enables variance propagation through the robust Bellman equations even when the minimizing kernels appearing in the error comparison differ. To obtain finite-sample guarantees, we combine these structural bounds with concentration uniform over clipping thresholds and uncertainty allocations, and use an absorbing-state construction to handle sample-dependent value functions.
  Beyond establishing minimax-optimal sample complexity results, this framework reveals how χ^2 uncertainty determines statistical difficulty: the sample complexity separates into the ordinary transition-estimation cost and an additional cost linear in the uncertainty radius. In particular, it

explains why $\rho \lesssim 1-\gamma$ preserves the ordinary-RL sample-complexity order, up to logarithmic factors, and identifies the regime in which uncertainty becomes the dominant statistical cost.

- **Minimax optimality bounds for shared-budget $\mathcal{S}$-rectangularity.** For the statewise uncertainty constraint: $\sum_{a\in\mathcal{A}} d_{\chi^2}(P_{s,a}, P_{0,s,a}) \le |\mathcal{A}|\rho$, we establish an upper bound of sample complexity to guarantee an ε-optimality: $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3}+\frac{|\mathcal{A}|\rho}{(1-\gamma)^4}\right)\right)$. The proof accounts for randomized policies and the joint allocation of uncertainty across actions, rather than replacing the shared constraint by independent row-wise balls. We also develop a matching minimax lower bound by separating deterministic branch selection from stochastic evaluation and reducing robust policy learning to binary hypothesis testing. Together, we demonstrate that the minimax optimal sample complexity shall not depend on the inverse of the minimum non-zero transition probability of the nominal transition kernel in contrast to the upper bound results in (Li et al., 2025).

2 Related Work

Upper Bound: $\mathcal{SA}$-Rectangular Uncertainty Set. Early finite-sample analyses established the statistical feasibility of model-based robust planning with an unknown nominal kernel. Panaganti & Kalathil (2022); Yang et al. (2022); Xu et al. (2023); Zhou et al. (2021); Blanchet et al. (2023); Clavier et al. (2023); Duchi & Namkoong (2018); Roch et al. (2025b) studied the $\mathcal{S}$-rectangular DR-RL problem from different uncertainty sets. For the model-free setting, Wang & Zou (2021); Wang et al. (2023; 2024a); Liu et al. (2022); Roch et al. (2025a); Zhang et al. (2025) also provided related algorithm and sample size results. Specifically, for χ^2-divergence defined DR-RL, Shi et al. (2023, Theorem 3) established the upper bound $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}(1+\frac{\max\{\sqrt{\rho},\rho\}}{1-\gamma})\right)$. Their upper and lower bounds match when $\rho \lesssim (1-\gamma)^2$ or $\rho \gtrsim 1$, but differ in the intermediate regime $(1-\gamma)^2 \ll \rho \ll 1$. Our analysis gives $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}(1+\frac{\rho}{1-\gamma})\right)$, replacing the square-root radius dependence by a linear term in ρ. This closes the intermediate-radius gap and matches the lower bound up to logarithmic factors over the common parameter range of our theorems. It also identifies $\rho \asymp 1-\gamma$ as the transition between the standard RL statistical sample complexity and the uncertainty-dominated sample complexity.

Upper Bound: $\mathcal{S}$-Rectangular Uncertainty Set. For $\mathcal{S}$-rectangular uncertainty sets, the coupling of transition perturbations across actions makes statistical analysis more complicated Kumar et al. (2024). Yang et al. (2022) provided finite-sample guarantees covering randomized policies; their χ^2-divergence upper bound has quadratic state space size dependence and cubic action space size dependence. For generalized L_p-norm uncertainty, Clavier et al. (2024) subsequently obtained improved bounds under both rectangularity structures. These norm-based results concern a different uncertainty geometry from the nominal-probability-weighted χ^2 divergence uncertainty set studied here. Ho et al. (2026) develops projection-based planning algorithms for discounted robust Markov decision processes (MDPs) with $\mathcal{S}$-rectangular ambiguity sets, including χ^2 uncertainty; their computational-complexity guarantees are not directly comparable to our statistical sample-complexity results.

The closest related divergence-based result is Li et al. (2025). They analyze the Cressie–Read family ϕ-divergence defined robust RL, which includes χ^2-divergence as a special case. This work provides a uniform upper bound for the entire Cressie–Read family, but their results depend on the inverse of the minimum positive probability of the nominal transition kernel. In contrast, our work develops a novel variance-sensitive analysis. For the shared budget $\sum_{a\in\mathcal{A}} d_{\chi^2}(P_{s,a}|P_{0,s,a}) \le |\mathcal{A}|\rho$, we obtain a sample complexity upper bound of $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3}+\frac{|\mathcal{A}|\rho}{(1-\gamma)^4}\right)\right)$, which does not depend on the minimum non-zero transition probability of the nominal transition kernel, and reveals the transition between small and large uncertainty at $|\mathcal{A}|\rho \asymp 1-\gamma$. It characterizes the cost of the actual coupled budget rather than replacing that budget with independent action-wise constraints.

Minimax Lower Bound: $\mathcal{SA}$- and $\mathcal{S}$-Rectangular Uncertainty Sets. Yang et al. (2022) established early information-theoretic minimax lower bounds using perturbed Bernoulli hard instances. For the $\mathcal{SA}$-rectangular policy-learning problem, Shi et al. (2023, Theorem 4) strengthened the lower bound to $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3}+\frac{\rho}{(1-\gamma)^4}\right)\right)$. Lower bounds for $\mathcal{S}$-rectangular models are not entirely absent

from the literature. The comparison in Li et al. (2025, Table 1) includes the χ^2 lower bound of Yang et al. (2022) for this setting, and Clavier et al. (2024) establish an $\mathcal{S}$-rectangular lower bound for L_∞ uncertainty. The issue addressed here is instead a matching characterization of the radius-dependent complexity for the shared-budget χ^2 model.

We first construct a class of hard MDP instances and establish the lower bound of $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3}+\frac{|\mathcal{A}|\rho}{(1-\gamma)^4}\right)\right)$. To the best of our knowledge, this is the first such lower bound for the $\mathcal{S}$-rectangular setting. It provides a complete understanding of the intrinsic difficulty of this setting and enables a sharper comparison of the sample complexities under $\mathcal{S}$-rectangular and $\mathcal{S},\mathcal{A}$-rectangular uncertainty sets.

3 Preliminaries and Problem Formulation

Notations: Denote by $\Delta(\mathcal{S})$ (resp., $\Delta(\mathcal{A})$) the probability simplex over $\mathcal{S}$ (resp., $\mathcal{A}$). We write $P(\cdot|s,a)=P_{s,a}(\cdot)$. Let $P_s=\otimes_{a\in\mathcal{A}}P_{s,a}$. For two distributions P_1 and P_2 over the same probability space X, where P_1 is absolutely continuous with respect to P_2, the χ^2-divergence $d(P_1,P_2)$ is defined as $d(P_1,P_2)=\sum_X\frac{(P_1-P_2)^2}{P_2}$. Here, $P_{s,a}\ll P_{0,s,a}$ denotes absolute continuity of $P_{s,a}$ with respect to $P_{0,s,a}$, meaning that $P_{s,a}(s')=0$ whenever $P_{0,s,a}(s')=0$. We use $\otimes$ to denote a Cartesian product of uncertainty sets.

3.1 Markov Decision Processes

In this paper, we consider the discounted infinite horizon MDPs, which can be defined by a tuple $(\mathcal{S},\mathcal{A},R,P,\gamma)$. Here $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is a finite action space, $R:\mathcal{S}\times\mathcal{A}\to[0,1]$ is reward function, $P=\{P_{s,a}\}_{s\in\mathcal{S},a\in\mathcal{A}}$ is the transition kernel, where $P_{s,a}\in\Delta(\mathcal{S})$ and γ is the discounted factor. Given an MDP, the policy $\pi:\mathcal{S}\to\Delta(\mathcal{A})$ specifies how to choose an action based on the current state. Given the policy π and transition P, the value function is defined as $V(s)=\mathbb{E}\left[\sum_t\gamma^tR(s_t,a_t)|s_0=s,P,\pi\right]$. The action value function is defined as $Q(s,a)=\mathbb{E}\left[\sum_t\gamma^tR(s_t,a_t)|s_0=s,a_0=a,P,\pi\right]$. The optimal policy in a standard discounted infinite-horizon MDP is Markovian, i.e., the agent takes actions based solely on the current state.

3.2 Distributionally Robust MDPs

A distributionally robust MDP accounts for uncertainty in the transition kernel. Rather than optimizing performance under the nominal kernel $P_0=\{P_{0,s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ alone, the agent optimizes its worst-case performance over an uncertainty set $\mathcal{P}$ centered at P_0. Let $\rho\geq 0$ be the uncertainty level. For a policy π, with $\pi(\cdot|s)\in\Delta(\mathcal{A})$, define the robust value function for a given policy π and the optimal robust value function as follows:

$$V_\pi(s)=\inf_{P\in\mathcal{P}}\mathbb{E}^{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^tR(S_t,A_t)\,|\,S_0=s\right],\qquad V^*(s)=\sup_\pi V_\pi(s).$$

$\mathcal{SA}$-Rectangular Uncertainty Set. For $\mathcal{SA}$-rectangular uncertainty sets, the transition kernel can be perturbed independently across state–action pairs:

$$\mathcal{P}=\otimes_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathcal{P}_{s,a},\quad \mathcal{P}_{s,a}=\{P_{s,a}\in\Delta(\mathcal{S}):d\left(P_{s,a},P_{0,s,a}\right)\leq\rho\}.$$

For a fixed policy π, define its robust Bellman operator by

$$(\mathcal{T}_\pi V)(s)=\sum_{a\in\mathcal{A}}\pi(a|s)\left(R(s,a)+\gamma\inf_{P_{s,a}\in\mathcal{P}_{s,a}}\sum_{s'\in\mathcal{S}}P_{s,a}(s')V(s')\right).$$

The robust value satisfies $V_\pi=\mathcal{T}_\pi V_\pi$. The Bellman optimality operator can also be written as $(\mathcal{T}V)(s)=\max_{a\in\mathcal{A}}R(s,a)+\gamma\inf_{P_{s,a}\in\mathcal{P}_{s,a}}\sum_{s'\in\mathcal{S}}P_{s,a}(s')V(s')$. Its fixed point is V^*, and choosing a maximizing action at every state yields a stationary deterministic optimal policy (Nilim & El Ghaoui, 2004; Iyengar, 2005). We then present the following duality results, which will be useful for our analysis:

Lemma 1. *Given the χ^2-divergence defined $\mathcal{SA}$-rectangular uncertainty set, the primal problem of finding the worst-case performance is equivalent to the following dual problem:*

$$\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[V(s')] = \sup_{\eta\in\mathbb{R}} \left\{ -\sqrt{2\rho+1}\sqrt{\mathbb{E}_{s'\sim P_{0,s,a},}[(\eta - V(s'))_+^2]} + \eta \right\},$$

where $(\cdot)_+ = \max\{\cdot, 0\}$.

We note that the duality in Lemma 1 is slightly different from the existing one in e.g., (Iyengar, 2005). We provide the proof in the Appendix Section F for completeness.

$\mathcal{S}$-Rectangular Uncertainty Set. For $\mathcal{S}$-rectangular uncertainty sets, perturbations are independent across states but coupled across actions. Let $P_s = \{P_{s,a}\}_{a\in\mathcal{A}}$ and $P_{0,s} = \{P_{0,s,a}\}_{a\in\mathcal{A}}$. We then define the shared-budget uncertainty set:

$$\mathcal{P} = \otimes_{s\in\mathcal{S}}\mathcal{P}_s, \quad \mathcal{P}_s = \left\{ P_s \in \Delta(\mathcal{S})^{\mathcal{A}} : \sum_{a\in\mathcal{A}} d_{P_{s,a},P_{0,s,a}} \le |\mathcal{A}|\rho \right\}.$$

Here, $|\mathcal{A}|\rho$ is the total budget per state. This constraint is not weighted by the policy and does not impose an uncertainty budget ρ on each row (action) separately. The set $\mathcal{P}_s$ is determined by $P_{0,s}$ and ρ. For a fixed policy π and value function V, the robust Bellman operator is defined as

$$(\mathcal{T}_\pi V)(s) = \sum_{a\in\mathcal{A}} \pi(a|s)R(s,a) + \gamma \inf_{P_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a)V(s').$$

Again, we have $V_\pi = \mathcal{T}_\pi V_\pi$. The robust Bellman optimality operator is

$$(\mathcal{T}V)(s) = \max_{\pi(\cdot|s)\in\Delta(\mathcal{A})} (\mathcal{T}_\pi V)(s),$$

and V^* is its fixed point $V^* = \mathcal{T}V^*$. The minimization must act jointly on the action-indexed tuple P_s. Unlike the $\mathcal{SA}$-rectangular case, the expression being maximized need not be linear in $\pi(\cdot|s)$ after minimization, and an optimal policy may require randomization (Li et al., 2025).

We then present the following duality results, which will be useful for our analysis:

Lemma 2. *Given the χ^2-divergence defined $\mathcal{S}$-rectangular uncertainty set, the primal problem of finding the worst-case performance is equivalent to the following dual problem:*

$$\begin{aligned}
&\inf_{P_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} \pi(a|s)\, \mathbb{E}_{s'\sim P_{0,s,a}}[V(s')] \\
&= \sup_{\eta\in\mathbb{R}^{|\mathcal{A}|}} \left\{ -\sqrt{|\mathcal{A}|(2\rho+1)}\sqrt{\sum_{a\in\mathcal{A}} \mathbb{E}_{s'\sim P_{0,s,a}}[(\eta_a - \pi(a|s)V(s'))_+^2]} + \sum_{a\in\mathcal{A}} \eta_a \right\}. \qquad (3)
\end{aligned}$$

The proof can be found in Appendix Section F.

Generative Model. We assume access to a generative model of the nominal transition kernel (Kearns, 1989): for any queried $(s,a) \in \mathcal{S}\times\mathcal{A}$, the model returns an independent next state distributed according to $P_{0,s,a}$. We collect N independent samples for every state–action pair, $s_i^{s,a} \sim P_{0,s,a}$ for $i = 1,\dots,N$, and construct the empirical estimate of P_0:

$$\widehat{P}_{0,s,a}(s') = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{s_i^{s,a} = s'\}, \qquad \widehat{P}_0 = \{\widehat{P}_{0,s,a}\}_{s,a}. \qquad (4)$$

The total number of nominal samples is $N|\mathcal{S}||\mathcal{A}|$.

4 Algorithms for $\mathcal{S}$- and $\mathcal{SA}$-Rectangular Uncertainty Sets

In this section, we introduce the empirical robust value iteration algorithm for both $\mathcal{S}$- and $\mathcal{SA}$-rectangular uncertainty sets using the empirical estimate of P_0 in eq. (4).

4.1 $\mathcal{SA}$-RECTANGULAR UNCERTAINTY SET

The algorithm is essentially robust value iteration over an empirical uncertainty set centered at $\widehat{P}_0$ from equation 4. This set is defined by

$$\widehat{\mathcal{P}} = \otimes_{(s,a)\in\mathcal{S}\times\mathcal{A}}\widehat{\mathcal{P}}_{s,a}, \quad \widehat{\mathcal{P}}_{s,a} = \left\{P_{s,a} \in \Delta(\mathcal{S}) : d\left(P_{s,a}, \widehat{P}_{0,s,a}\right) \le \rho\right\}.$$

Set $\widehat{V}_0 = 0$ be the initialization. For each $k = 0, \ldots, K-1$ and every state $s \in \mathcal{S}$, perform the update

$$\widehat{V}_{k+1}(s) \leftarrow (\widehat{\mathcal{T}}\widehat{V}_k)(s) = \max_{a\in\mathcal{A}}\left\{R(s,a) + \gamma \inf_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \sum_{s'\in\mathcal{S}} P(s'|s,a)\widehat{V}_k(s')\right\}. \tag{5}$$

For a given value vector $\widehat{V}_K$, each inner minimization problem can be solved using Lemma 1. After K iterations, the output policy is chosen in the deterministic way: for each state,

$$\widehat{\pi}_K(a|s) \in \arg\max_{a\in\mathcal{A}}\left\{R(s,a) + \gamma \inf_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \sum_{s'\in\mathcal{S}} P(s'|s,a)\widehat{V}_K(s')\right\}.$$

and return the deterministic policy $\widehat{\pi}_K(a|s)$.

4.2 $\mathcal{S}$-RECTANGULAR UNCERTAINTY SET

Using the empirical distribution $\widehat{P}_0$ defined in equation 4, we construct the empirical uncertainty set:

$$\widehat{\mathcal{P}} = \otimes_{s\in\mathcal{S}}\widehat{\mathcal{P}}_s, \quad \widehat{\mathcal{P}}_s = \left\{P_s \in \Delta(\mathcal{S})^{\mathcal{A}} : \sum_{a\in\mathcal{A}} d_{\chi^2}\left(P_{s,a}, \widehat{P}_{0,s,a}\right) \le |\mathcal{A}|\rho\right\}.$$

Let $\widehat{\mathcal{T}}_\pi$ and $\widehat{\mathcal{T}}$ be the empirical counterparts of the robust Bellman operator for policy π and the robust Bellman optimality operator under $\mathcal{S}$-rectangular uncertainty sets. Initialize $\widehat{V}_0 = 0$ and perform the following update at each iteration: $k = 0, \ldots, K-1$, do update for every state $s \in \mathcal{S}$,

$$\widehat{V}_{k+1}(s) = (\widehat{\mathcal{T}}\widehat{V}_k)(s) = \max_{\pi(\cdot|s)\in\Delta(\mathcal{A})}\left\{\sum_{a\in\mathcal{A}}\pi(a|s)R(s,a) + \gamma \inf_{P_s\in\widehat{\mathcal{P}}_s} \mathbb{E}_{a\sim\pi(\cdot|s),s'\sim P_{s,a}}\widehat{V}_k(s')\right\}.$$

The maximization is over a distribution $\pi(\cdot|s)$. Due to the coupling over the actions, the inner infimum cannot be equivalently achieved by the sum of the infimum, which is the key challange of $\mathcal{S}$-rectangular uncertainty set. By Lemma 2, each Bellman update can be computed by jointly optimizing the statewise action probabilities and the dual vector, with the reward term added as in the update. The common square-root term encodes the coupling across actions.

After K iterations, the output of the algorithm is a stationary randomized policy:

$$\widehat{\pi}_K(\cdot|s) \in \arg\max_{\pi(\cdot|s)\in\Delta(\mathcal{A})} (\widehat{\mathcal{T}}_\pi\widehat{V}_K)(s), \quad s \in \mathcal{S}.$$

Unlike the $\mathcal{SA}$-rectangular algorithm, this step retains the maximizing action distribution rather than selecting its largest-probability action. This is the empirical robust planning formulation for the coupled model (Li et al., 2025).

4.3 CONTRACTION AND CONVERGENCE

For both uncertainty structures, the robust Bellman operator and the robust Bellman optimality operator are γ-contractions in the supremum norm. Let $\widehat{V}^*$ be the optimal robust value for the empirical uncertainty set, and $\widehat{V}_{\widehat{\pi}}$ the value of the returned policy under $\widehat{\mathcal{P}}$. Since $\widehat{V}_0 = 0$ and rewards lie in $[0,1]$, we can easily obtain that

$$\|\widehat{V}_K - \widehat{V}^*\|_\infty \le \frac{\gamma^K}{1-\gamma}, \qquad \|\widehat{V}^* - \widehat{V}_{\widehat{\pi}}\|_\infty \le \frac{2\gamma^{K+1}}{(1-\gamma)^2}.$$

From the above two inequalities, together with $\widehat{\mathcal{T}}\widehat{V}_K = \widehat{\mathcal{T}}_{\widehat{\pi}}\widehat{V}_K$, we can easily obtain that

$$(1-\gamma)\|\widehat{V}^* - \widehat{V}_{\widehat{\pi}}\|_\infty \le 2\gamma\|\widehat{V}^* - \widehat{V}_K\|_\infty.$$

This bound quantifies planning error using the empirical uncertainty set. The subsequent sample-complexity analysis controls the additional error from replacing P_0 by $\widehat{P}$.

5 MINIMAX OPTIMALITY

In this section, the theoretical results will be proposed.

5.1 $\mathcal{SA}$-RECTANGULAR UNCERTAINTY SET

Theorem 1 (Upper bound under $\mathcal{SA}$-rectangular uncertainty)**.** *Consider $\mathcal{P} = \otimes_{s,a}\mathcal{P}_{s,a}$, where $\mathcal{P}_{s,a} = \{P_{s,a} \in \Delta(\mathcal{S}) : d_{P_{s,a},P_{0,s,a}} \le \rho\}$. Suppose $N \ge 2$ and $N \ge 16\log\big(12|\mathcal{S}||\mathcal{A}|/\delta\big)/(1-\gamma)^2$. For every integer $K \ge 0$, the policy returned by Algorithm in Section 4 satisfies, with probability at least $1-\delta$,*

$$\|V^* - V_{\widehat{\pi}}\|_\infty \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{1}{N(1-\gamma)^3}} + \sqrt{\frac{\rho}{N(1-\gamma)^4}} + \frac{1}{N(1-\gamma)^2} + \frac{\gamma^{K+1}}{(1-\gamma)^2}\right). \tag{6}$$

For an exact empirical robust optimal policy, the last term is omitted.

The first term is the ordinary transition-estimation contribution, whereas the second quantifies the additional effect of χ^2 uncertainty. The remaining terms are a lower-order statistical remainder and the finite-iteration planning error. The two leading statistical terms have the same order when $\rho \asymp 1-\gamma$.

5.2 $\mathcal{S}$-RECTANGULAR UNCERTAINTY SET

Theorem 2 (Upper bound under $\mathcal{S}$-rectangular uncertainty)**.** *Consider $\mathcal{P} = \otimes_s\mathcal{P}_s$, where $\mathcal{P}_s = \{P_s \in \Delta(\mathcal{S})^{|\mathcal{A}|} : \sum_{a\in\mathcal{A}} d_{P_{s,a},P_{0,s,a}} \le |\mathcal{A}|\rho\}$. Suppose $N \ge 2$ and $N \ge 16\log\big(12|\mathcal{S}||\mathcal{A}|/\delta\big)/(1-\gamma)^2$. For every integer $K \ge 0$, the stationary randomized policy returned by Algorithm in Section 4 satisfies, with probability at least $1-\delta$,*

$$\|V^* - V_{\widehat{\pi}}\|_\infty \le \widetilde{\mathcal{O}}\left(\sqrt{\frac{1}{N(1-\gamma)^3}} + \sqrt{\frac{\rho|\mathcal{A}|}{N(1-\gamma)^4}} + \frac{1}{N(1-\gamma)^2} + \frac{\gamma^{K+1}}{(1-\gamma)^2}\right). \tag{7}$$

The guarantee allows the minimizing transition tuple to depend on the policy and does not require the optimal policy to be deterministic.

Here $|\mathcal{A}|\rho$ is the total statewise uncertainty budget. The bound applies to the coupled set $\mathcal{P}_s$, not to a product of independent row-wise balls. The leading statistical terms balance when $|\mathcal{A}|\rho \asymp 1-\gamma$.

5.3 SAMPLE-COMPLEXITY CONSEQUENCES

Corollary 1 (Sufficient samples for robust policy learning)**.** *Let $\varepsilon \in (0,1]$ and $\delta \in (0,1)$. Choose $K \ge \left\lceil \frac{\log\left(8/[\varepsilon(1-\gamma)^2]\right)}{-\log\gamma} \right\rceil$. For $\mathcal{SA}$-rectangular uncertainty, it is sufficient to take $N \ge \left(\frac{1}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3} + \frac{\rho}{(1-\gamma)^4}\right)\right)$. For $\mathcal{S}$-rectangular uncertainty, it is sufficient to take $N \ge \widetilde{O}\left(\frac{1}{\varepsilon^2}\left(\frac{1}{(1-\gamma)^3} + \frac{|\mathcal{A}|\rho}{(1-\gamma)^4}\right)\right)$.*

Under the respective sample condition, the corresponding algorithm returns a policy satisfying $\|V^* - V_{\widehat{\pi}}\|_\infty \le \varepsilon$ with probability at least $1-\delta$.

Consequently, the sufficient total sample complexities $|\mathcal{S}||\mathcal{A}|N$ are

$$\begin{aligned} \mathcal{SA}\text{-rectangular}: &\quad \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\left(1 + \frac{\rho}{1-\gamma}\right)\right), \\ \mathcal{S}\text{-rectangular}: &\quad \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\left(1 + \frac{|\mathcal{A}|\rho}{1-\gamma}\right)\right). \end{aligned} \tag{8}$$

Remark 1 (Discussion)**.** *For the $\mathcal{SA}$-rectangular problem, our sample complexity matches the lower bound of Shi et al. (2023) up to logarithmic factors under balanced nominal sampling and within the*

common parameter range of the theorems. Compared with their upper bound, our result replaces $\max\{\sqrt{\rho}, \rho\}$ *by* ρ*, closing the intermediate-radius gap. Beyond improving the rate, this identifies* $\rho \asymp 1-\gamma$ *as the statistical transition: the ordinary-RL sample-complexity order persists when* $\rho \lesssim 1-\gamma$*, whereas larger radii introduce an additional cost linear in* ρ*.*

For the $\mathcal{S}$*-rectangular problem, we characterize the sharp joint dependence on the uncertainty budget and discount factor. The guarantee of Li et al. (2025) applies to every positive radius, but its final* χ^2 *bound retains an inverse minimum-positive-transition-probability factor and does not distinguish the small- and large-uncertainty regimes. Our bound removes this instance-dependent factor and separates the ordinary transition-estimation cost from the additional cost of the shared budget. Under* $\sum_{a\in\mathcal{A}} d_{P_{s,a},P_{0,s,a}} \le |\mathcal{A}|\rho$*, the transition occurs at* $|\mathcal{A}|\rho \asymp 1-\gamma$*. In particular, the ordinary-RL order is recovered when* $|\mathcal{A}|\rho \lesssim 1-\gamma$*. Together with our matching lower bound, this yields minimax-optimal dependence up to logarithmic factors under the stated sampling model.*

5.4 Lower Bound

Theorem 3 (Lower bound under balanced nominal sampling)**.** *Let* $|\mathcal{S}| \ge 4$*,* $|\mathcal{A}| \ge 2$*,* $\gamma \in [3/4, 1)$*,* $0 \le \rho < \infty$*, constant* C*, and* $0 < \varepsilon \le 1/(128(1-\gamma))$*. There exist two nominal MDPs* M_0, M_1 *with the same rewards and a common distinguished state* d*, such that every estimator* $\widehat{\pi}$ *based on the balanced experiment satisfies*

$$\max_{\theta\in\{0,1\}} \mathbb{P}_\theta \left\{ V_\theta^*(d) - V_\theta^{\widehat{\pi}}(d) > \varepsilon \right\} \ge \frac{1}{4}, \qquad \textit{whenever } n \le \frac{1-\gamma+2\rho}{C\varepsilon^2(1-\gamma)^4}. \tag{9}$$

Here $V_\theta^\pi = V_{P_\theta}^\pi$*,* $V_\theta^* = V_{P_\theta}^*$*, and* $\mathbb{P}_\theta$ *includes both the nominal-data randomness and the estimator's internal randomness. Consequently, for any fixed failure probability* $\delta < 1/4$*, the necessary balanced total sample size is*

$$N = \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3} \left(1 + \frac{2|\mathcal{A}|\rho}{1-\gamma} \right) \right). \tag{10}$$

The same lower bound holds for the uniform-in-state requirement $\|V^* - V^{\widehat{\pi}}\|_\infty \le \varepsilon$*.*

Minimax optimality under $\mathcal{S}$*-rectangularity.* For the $\mathcal{S}$-rectangular setting, the upper bound in equation 8 matches the lower bound in Theorem 3 up to logarithmic factors. Thus, under the stated sampling model, the minimax sample complexity has order $\widetilde{\Theta}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\left(1+\frac{|\mathcal{A}|\rho}{1-\gamma}\right)\Big)$. Our bounds therefore characterize the joint dependence on the uncertainty budget and the discount factor.

Remark 2. *The lower bound in eq.* (10) *also contains the weaker numerical rate* $N|\mathcal{S}||\mathcal{A}| = \Omega\Big(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\left(1+\frac{\rho}{1-\gamma}\right)\Big)$*, which has the same order as the* $\mathcal{SA}$*-rectangular lower bound of (Shi et al., 2023, Theorem 4) under balanced nominal sampling. This comparison of rates does not make the proofs interchangeable. Under* $\mathcal{SA}$*-rectangularity, each state–action transition has a separate uncertainty set. Under* $\mathcal{S}$*-rectangularity, the actionwise transitions share a statewise uncertainty budget. The adversary may allocate this budget unevenly across actions, and the allocation depends on the policy and the transition values. This coupling requires a separate analysis of policy error.*

6 Proof Sketch and Technique Novelty

Upper Bound Technique Novelty. For the $\mathcal{SA}$-rectangular problem, the KKT conditions give $Z = \mathbb{E}_{P_{0,s,a}}[(\eta^* - V)_+]$, $\underline{P}_{s,a}^V = P_{0,s,a} \circ w$ with $w = (\eta^* - V)_+/Z$, and $\eta^* - \bar{V}_{s,a} = (1+2\rho)Z$ whenever $Z > 0$. We first prove the kernel formula and treat saturation and zero radius separately. We then finish the error decomposition for $g(t, \eta^*(t))$ and bound the variance of $h(V) = (\eta^* - V)_+^2/Z$. For the $\mathcal{S}$-rectangular problem, the same KKT argument gives one common Z across actions and allocated divergences ρ_a^* with $\sum_a \rho_a^* \le |\mathcal{A}|\rho$.

The derivative formula alone does not justify concentration of a sample-dependent ratio. To complete that step, we prove an equivalent bounded expression for the same robust expectation, using $V \wedge \eta$ rather than dividing by Z. Concentration of its mean and standard deviation is simultaneous in η and in the row radius. The joint statewise minimization is exactly a minimization over allocations $\rho_a \ge 0$,

$\sum_a \rho_a \le |\mathcal{A}|\rho$, so this same event covers randomized policies and sample-dependent allocations. For any feasible row, the density-ratio inequality $1 \le 2w + (w-1)^2$ gives

$$\mathrm{Var}_{P_{0,s,a}}(V) \le 2\mathrm{Var}_{P_{s,a}}(V) + \frac{2d_{\chi^2}(P_{s,a}|P_{0,s,a})}{(1-\gamma)^2}.$$

An absorbing-state construction removes all samples at one state; a scalar grid then establishes concentration at $\widehat{V}^*$ and $\widehat{V}_{\pi^*}$. Finally, for every feasible mixed kernel P^π, the robust Bellman inequality $V_\pi \le r_\pi + \gamma P^\pi V_\pi$ controls variance propagation. A two-sided resolvent comparison and absorption is proved.

Lower Bound Design and Proof. Under hypothesis $\theta \in \{0,1\}$, action 0 at d selects u_0 deterministically; other actions select u_1. At u_j, action 0 exits to an absorbing state g with probability t_+ if $j = \theta$ and t_- otherwise; other actions exit with probability $(1-\gamma)/2$. Failed exits self-loop, only g earns reward 1, and padding states are zero-reward self-loops. Forward Pearson divergence fixes deterministic rows, while the shared budget gives the informative row minimum exit probability $\alpha_\beta(t) = [t - \sqrt{\beta t(1-t)}]_+$, where $\beta = 2A\rho$. Choose the robust exit levels

$$a_0 = 1-\gamma, \qquad a_1 = a_0 + \eta, \qquad \eta = 32\varepsilon(1-\gamma)^2 \le \frac{1-\gamma}{4}. \tag{11}$$

Lemma 3 (Inverse parametrization and sensitivity)**.** *For $\beta \ge 0$ and $a \in (0,1)$, define*

$$p_\beta(a) = \frac{\beta + 2a + \sqrt{\beta^2 + 4\beta a(1-a)}}{2(1+\beta)}. \tag{12}$$

Then $p_\beta(a) \in (0,1)$, $p'_\beta(a) > 0$, and $\alpha_\beta(p_\beta(a)) = a$, $\quad \frac{[p'_\beta(a)]^2}{p_\beta(a)(1-p_\beta(a))} = \frac{4}{\beta + 4a(1-a)}$.

For $\gamma \in [3/4, 1)$ and the levels in eq. (11)*, $t_- = p_\beta(a_0)$ and $t_+ = p_\beta(a_1)$ satisfy*

$$\mathrm{Hel}^2\ (\mathrm{Ber}(t_-), \mathrm{Ber}(t_+)) \le \frac{\eta^2}{1-\gamma+\beta}, \tag{13}$$

where $\mathrm{Hel}^2(P,Q) = \sum_x (\sqrt{P(x)} - \sqrt{Q(x)})^2$.

Policy gap and testing. Let w_θ be the probability of selecting the correct branch and $G(z) = \gamma z/[(1-\gamma)(1-\gamma+\gamma z)]$. A feasible adversary sets informative rows to their robust endpoints and leaves other rows nominal.

$$V_\theta^*(d) - V_\theta^\pi(d) \ge \gamma(1-w_\theta)[G(a_1) - G(a_0)] \ge 4\varepsilon(1-w_\theta). \tag{14}$$

Thus a majority-mass decoding error implies loss at least 2ε. Only two nominal rows differ, so n independent samples per pair give full-data squared Hellinger distance at most $2n\eta^2/(1-\gamma+\beta)$. At the theorem's sample threshold, total variation is at most $1/2$, forcing maximum testing error at least $1/4$ and hence the claimed policy lower bound. Appendix D provides the complete proof.

7 Conclusion

We established sample-complexity bounds for χ^2 DR-RL under $\mathcal{SA}$-rectangular and $\mathcal{S}$-rectangular uncertainty. Under balanced nominal generative-model sampling, our upper and lower bounds match up to logarithmic factors within the stated parameter ranges. This closes the intermediate-radius gap for $\mathcal{SA}$-rectangular uncertainty and characterizes the cost of shared statewise budgets while accounting for randomized policies and policy-dependent adversarial allocations.

Our variance-sensitive framework connects robust Bellman estimation to the second-moment structure of adversarial density ratios. By controlling normalized dual fluctuations and comparing nominal and adversarial variances, it preserves dependence linear in the allocated divergence. Combined with uniform concentration and robust variance propagation, this yields guarantees without inverse dependence on the minimum positive nominal transition probability.

The results separate ordinary transition-estimation cost from the additional cost of uncertainty. The ordinary-RL order persists when $\rho \lesssim 1-\gamma$ for $\mathcal{SA}$-rectangular uncertainty, or $|\mathcal{A}|\rho \lesssim 1-\gamma$ for the shared-budget $\mathcal{S}$-rectangular model; larger budgets lead to uncertainty-dominated complexity. Extending these characterizations to unrestricted adaptive sampling and offline learning with partial coverage remains a direction for future research.

REFERENCES

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36:66845–66859, 2023.

Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

Pierre Clavier, Laixi Shi, Erwan Le Pennec, Eric Mazumdar, Adam Wierman, and Matthieu Geist. Near-optimal distributionally robust reinforcement learning with general l_p norms. *Advances in Neural Information Processing Systems*, 37:1750–1810, 2024.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Efficient algorithms for robust markov decision processes with s-rectangular ambiguity sets. *arXiv preprint arXiv:2602.05591*, 2026.

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

M. J. Kearns. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Navdeep Kumar, Kaixin Wang, Kfir Yehuda Levy, and Shie Mannor. Efficient value iteration for s-rectangular robust markov decision processes. In *ICML*, pp. 25682–25725, 2024.

Zhenghao Li, Shengbo Wang, and Nian Si. Near-optimal sample complexities of divergence-based s-rectangular distributionally robust reinforcement learning. *arXiv preprint arXiv:2505.12202*, 2025.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *International Conference on Machine Learning*, pp. 13623–13643. PMLR, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *NIPS Deep Learning Workshop*, 2013.

Arnab Nilim and Laurent El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 839–846, 2004.

Wenfan Ou and Sheng Bi. Sequential decision-making under uncertainty: a robust mdps review. *Annals of Operations Research*, 353(3):1239–1285, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

Zachary Roch, George Atia, and Yue Wang. Model-free robust average-reward reinforcement learning with sample complexity analysis. *arXiv preprint arXiv:2505.12462*, 2025a.

Zachary Andrew Roch, George K. Atia, and Yue Wang. A reduction framework for distributionally robust reinforcement learning under average reward. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=B5p7bjTI8v.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Marnix Suilen, Thom Badings, Eline M Bovy, David Parker, and Nils Jansen. Robust markov decision processes: A place where ai and formal methods meet. In *Principles of Verification: Cycling the Probabilistic Landscape: Essays Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday, Part III*, pp. 126–154. Springer, 2024.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398. PMLR, 2023.

Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Uncertainty in Artificial Intelligence*, pp. 3470–3513. PMLR, 2024a.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Yue Wang, Zhongchang Sun, and Shaofeng Zou. A unified principle of pessimism for offline reinforcement learning under model mismatch. *Advances in Neural Information Processing Systems*, 37:9281–9328, 2024b.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022.

Chi Zhang, Zain Ulabedeen Farhat, George Atia, and Yue Wang. Model-free offline reinforcement learning with enhanced robustness. In *International Conference on Learning Representations*, volume 2025, pp. 61209–61253, 2025.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, grammatical review, and the typesetting of mathematical equations in LaTeX.

A NOTATIONS

Firstly, we introduce the following notation for the value functions of a policy π under the transition P and $\widehat{P}$:

Value function (for nominal transition P and policy π):

$$V_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}[V_\pi(s')]\right] = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim\underline{P}(\cdot|s,a)}[V_\pi(s')]\right].$$

Empirical value function (for empirical transition $\widehat{P}$ and policy π):

$$\widehat{V}_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma \inf_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \mathbb{E}_{s'\sim q}[\widehat{V}_\pi(s')]\right] = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim\underline{\widehat{P}}(\cdot|s,a)}[\widehat{V}_\pi(s')]\right].$$

Q-value recursion. The optimal value function satisfies

$$V^*_{\pi^*}(s) = \sum_{a\in\mathcal{A}} \pi^*(a|s) Q^*_{\pi^*}(s,a),$$

where the optimal Q-function obeys the robust Bellman recursion

$$Q^*_{\pi^*}(s,a) = r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim P_{s,a}}\left[V^*_{\pi^*}(s')\right].$$

Similarly, under the empirical transition model $\widehat{P}$, we have

$$V^*_{\widehat{\pi}^*}(s) = \sum_{a\in\mathcal{A}} \widehat{\pi}^*(a|s) Q^*_{\widehat{\pi}^*}(s,a),$$

with

$$Q^*_{\widehat{\pi}^*}(s,a) = r(s,a) + \gamma \inf_{P_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \mathbb{E}_{s'\sim q}\left[V^*_{\widehat{\pi}^*}(s')\right].$$

We summarize the notations as follows.

- P: true transition kernel matrix.
- $\widehat{P}$: empirical transition kernel,i.e. $\widehat{P}_{s,a}(s') = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}_{s'_n = s'}$. .
- π^*: optimal policy under P.
- $\widehat{\pi}$: output policy under $\widehat{P}$.
- $\widehat{\pi}^*$: optimal empirical policy under $\widehat{P}$.

We decompose the difference between the value function of the optimal policy and the empirical/output policy:

$$V_{\pi^*} - V_{\widehat{\pi}} = \underbrace{\left(V_{\pi^*} - \widehat{V}_{\pi^*}\right)}_{\text{Term I: estimation error}} + \underbrace{\left(\widehat{V}_{\pi^*} - \widehat{V}_{\widehat{\pi}^*}\right)}_{\text{Term II: policy optimality gap}} + \underbrace{\left(\widehat{V}_{\widehat{\pi}^*} - \widehat{V}_{\widehat{\pi}}\right)}_{\text{Term III: sub-optimality of } \widehat{\pi}} + \underbrace{\left(\widehat{V}_{\widehat{\pi}} - V_{\widehat{\pi}}\right)}_{\text{Term IV:estimation error (empirical optimal)}}.$$

Note that the term III is negative, and the term II can be bounded by the γ-contraction property. Therefore, the key challenge lies in the bounds of term I and term IV, which are due to the estimation error given the same policy.

Here, we consider a general bound of $||V_\pi - \widehat{V}_\pi||_\infty$ given a policy π.

Lemma 4 (Shi et al. (2023))**.** *The estimation error can be bounded by as follows:*

$$\left\|V_\pi - \widehat{V}_\pi\right\|_\infty \leq \gamma \max\left\{ \left\|(I - \gamma \widehat{\underline{P}}^{V_\pi,\pi})^{-1}\left(\widehat{\underline{P}}^{V_\pi,\pi}V_\pi - \underline{P}^{V_\pi,\pi}V_\pi\right)\right\|_\infty, \right.$$
$$\left. \left\|(I - \gamma \widehat{\underline{P}}^{\widehat{V}_\pi,\pi})^{-1}\left(\widehat{\underline{P}}^{V_\pi,\pi}V_\pi - \underline{P}^{V_\pi,\pi}V_\pi\right)\right\|_\infty \right\}.$$

Therefore, the key challenge lies in the bound of $\left\|\left(\widehat{\underline{P}}^{V_\pi,\pi}V_\pi - \underline{P}^{V_\pi,\pi}V_\pi\right)\right\|_\infty$. It holds that

$$\left\|\left(\widehat{\underline{P}}^{V_\pi,\pi}V_\pi - \underline{P}^{V_\pi,\pi}V_\pi\right)\right\|_\infty = \max_{s,a}\left\{\left|\mathbb{E}_{\underline{P}_{s,a}}[V(s')] - \mathbb{E}_{\widehat{\underline{P}}_{s,a}}[V(s')]\right|\right\}. \tag{15}$$

Besides, it also holds that

$$\left\|\left(\widehat{\underline{P}}^{V_\pi,\pi}V_\pi - \underline{P}^{V_\pi,\pi}V_\pi\right)\right\|_\infty = \max_{s}\left\{\left|\mathbb{E}_{\underline{P}_s}[V(s')] - \mathbb{E}_{\widehat{\underline{P}}_s}[V(s')]\right|\right\}. \tag{16}$$

From this point onward, the proofs for the $\mathcal{SA}$-rectangular and $\mathcal{A}$-rectangular uncertainty sets differ. We therefore treat them separately.

B $\mathcal{SA}$-RECTANGULAR UNCERTAINTY SET

In this section, we provide the proof of Theorem 1.

Notations in $\mathcal{SA}$-Rectangular Section Firstly, denote the Lagrange multiplier by

$$\eta^* = \max\left\{\arg\max_{\eta\in\mathbb{R}}\left\{-\sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta - V(s'))_+^2]} + \eta\right\}\right\}.$$

Remark 3. *We now explain why the minimum operator is included in the definition of η^*. For a DR-RL problem with a χ^2 uncertainty set, the maximizer of the dual objective need not be unique; several choices of η^* may yield the same optimal value. As a simple example, consider $P(s_b|a=0,s_0) = \frac{1}{1+2\rho}$, $P(s_g|a=0,s_0) = \frac{2\rho}{1+2\rho}$, $V(s_g) = \frac{1}{1-\gamma}$, and $V(s_b) = 0$. Under this setting, any value of η^* between 0 and $\frac{1}{1-\gamma}$ is optimal. To make the definition unambiguous and avoid confusion, we select the largest maximizer.*

Therefore, we can rewrite the $\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{P(s'|s,a)}[V(s')]$ as following,

$$\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{P(s'|s,a)}[V(s')] = \eta^* - \sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2]}.$$

Here, we define the robust value function expectation $\bar{V}_{s,a} = \inf_{P_{s,a}\in\mathcal{P}_{s,a}} \sum_{a\in\mathcal{A}} \mathbb{E}_{P(s'|s,a)}[V(s')]$.

B.1 VARIANCE-SENSITIVE ANALYSIS METHOD

To obtain a sharper sample complexity bound, we develop a variance-sensitive method. We begin by presenting a closed-form characterization of the worst-case transition distributions.

Define the $\underline{P}^V_{s,a}$ the worst-case transitions given the value function V, i.e.

$$\underline{P}^V_{s,a} = \arg\min_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{P_{s,a}}[V].$$

Proposition 1. *Under the χ^2 $\mathcal{SA}$-rectangular uncertainty set, if η^* is the Lagrangian parameter, the worst-case distribution within the uncertainty set can be written as*

$$\underline{P}^V_{s,a} = P_{0,s,a} \circ \frac{(\eta^* - V(s'))_+}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]},$$

which is equivalent to

$$\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{P(s'|s,a)}[V(s')] = \frac{\mathbb{E}_{P_0(\cdot|s,a)}[V(s')(\eta^* - V(s'))_+]}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]}.$$

Here, we provide the proof of Proposition 1. We prove the proposition by considering two cases: the regular case and the degenerate case.

Regular Case.

In the regular case, the Karush–Kuhn–Tucker (KKT) conditions apply. Then, by KKT conditions, we can get that

$$\begin{aligned}
&-\sqrt{2\rho+1}\frac{\mathbb{E}_{P_0(\cdot|s,a)}[2(\eta^* - V(s'))_+]}{2\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2]}} + 1 = 0;\\
\rightarrow &-\sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2]}\frac{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2]} + 1 = 0. \qquad (17)
\end{aligned}$$

Therefore, we have the following results:

$$(2\rho+1)\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]^2 = \mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2].$$

Then, we can rewrite the robust value function expectation as follows:

$$\begin{aligned}\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \mathbb{E}_{P(s'|s,a)}[V(s')] &= \eta^* - \sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2]} \\ &= \eta^* - \frac{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+^2]}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]} \\ &= \frac{\mathbb{E}_{P_0(\cdot|s,a)}[\eta^*(\eta^* - V(s'))_+ - (\eta^* - V(s'))_+^2]}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]} \\ &\overset{(a)}{=} \frac{\mathbb{E}_{P_0(\cdot|s,a)}[V(s')(\eta^* - V(s'))_+]}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^* - V(s'))_+]},\end{aligned}$$

where (a) follows from the fact when $\eta^* < V(s')$, $V(s')(\eta^* - V(s'))_+ = 0$.

Degenerate Case.

In this case, the uncertainty set is degenerate, and the worst-case transition kernel has a $0-1$ form: it assigns probability 1 to the worst state and probability 0 to every other state.

It remains to verify that, under this condition, the worst-case transition kernel satisfies the closed-form solution. Specifically, η^* is equal to the value function at the state with the second-smallest value. Consequently, the worst-case transition kernel assigns probability 1 to the worst state and probability 0 to all other states, confirming the validity of the closed-form solution.

Therefore, the proof is complete.

We next present the variance-sensitive transfer procedure, beginning with the lemma needed for its analysis.

Lemma 5. *Let X be a real-valued random variable satisfying $\mathbb{E}[X^2] < \infty$. Then, for any $m \in \mathbb{R}$,*

$$\mathbb{E}\left[(X - m\mathbb{E}[X])^2\right] = \mathrm{Var}(X) + (m-1)^2\left(\mathbb{E}[X]\right)^2.$$

We next define a weighted distribution as a linear combination of the nominal and empirical distributions: $P_{s,a}(t) = tP_{s,a} + (1-t)\hat{P}_{s,a}$; $\quad \Delta P_{s,a} = P_{s,a} - \hat{P}_{s,a}$.

Define $g(t,\eta(t)) = -\sqrt{2\rho+1}\left(\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta(t) - V(s))_+^2\right]\right)^{1/2} + \eta(t)$. Next, define the $\eta^*(t) = \arg\max_{\eta\in\mathbb{R}} g(t,\eta)$. Combined with equation 17, we can conclude that

$$\begin{aligned}\partial_t g(t,\eta^*(t)) &= -\sqrt{2\rho+1}\frac{\mathbb{E}_{\Delta P_{0,s,a}}\left[(\eta^*(t) - V(s))_+^2\right]}{2\sqrt{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t) - V(s))_+^2\right]}} \\ &= -\frac{1}{2}\frac{\mathbb{E}_{\Delta P_{0,s,a}}\left[(\eta^*(t) - V(s))_+^2\right]}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t) - V(s))_+\right]}. \end{aligned} \tag{18}$$

Moreover, we can rewrite the derivative function as:

$$\partial_t g(t,\eta^*(t)) = -\frac{1}{2}\frac{\mathbb{E}_{\Delta P_{0,s,a}}\left[\eta^*(t)(\eta^*(t) - V(s))_+\right]}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t) - V(s))_+\right]} + \frac{1}{2}\frac{\mathbb{E}_{\Delta P_{0,s,a}}\left[V(s)(\eta^*(t) - V(s))_+\right]}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t) - V(s))_+\right]}. \tag{19}$$

Since $\mathbb{E}\left[\Delta P_{s,a}\right] = 0$. Then, the expectation of above equitation at $t = 0, 1$ satisfies that

$$\mathbb{E}\left[\partial_t g(t=0,\eta^*(t=0))\right] = 0; \mathbb{E}\left[\partial_t g(t=1,\eta^*(t=1))\right] = 0.$$

Furthermore, we establish the convexity of $g(t,\eta^*(t))$ as a function of t.

Lemma 6. *Define $F(t) = g(t,\eta^*(t))$. Then F is convex in t.*

We next apply the variance-sensitive transfer argument, bounding the relevant term separately under two conditions.

B.2 CASE A: $g(0,\eta^*(0)) \geq g(1,\eta^*(1))$.

Combining with the convexity of $g(t,\eta^*(t))$ in t, established in Lemma 6, we obtain

$$|\partial_t g(0,\eta^*(0))| \geq |\partial_t g(1,\eta^*(1))|.$$

We then obtain the bound using the partial derivative $\partial_t g(0,\eta^*(0))$:

$$\begin{aligned}&\left|\mathbb{E}_{P_{s,a}}[h(V(s'),P_{0,s,a})V(s')] - \mathbb{E}_{\widehat{P}_{s,a}}[h(V(s'),\widehat{P}_{0,s,a})V(s')\right| = g(0,\eta^*(0)) - g(1,\eta^*(1))\\ &= |\partial_t g(t,\eta^*(t))| \leq \max\left\{|\partial_t g(0,\eta^*(0))|, |\partial_t g(1,\eta^*(1))|\right\} \leq |\partial_t g(0,\eta^*(0))|,\end{aligned}$$

where the first equality follows from the fact that

$$\partial_{\eta^*(t)} g(t,\eta^*(t)) = 0.$$

On the other side, it also holds that

$$\begin{aligned}&\left|\mathbb{E}_{P_{s,a}}[h(V(s'),P_{0,s,a})V(s')] - \mathbb{E}_{\widehat{P}_{s,a}}[h(V(s'),\widehat{P}_{0,s,a})V(s')\right|\\ &= g(0,\eta^*(0)) - g(1,\eta^*(1))\\ &\leq g(0,\eta^*(0)) - g(1,\eta^*(0))\\ &\leq \sqrt{2\rho+1}\left|\sqrt{\mathbb{E}_{\widehat{P}_0(\cdot|s,a)}[(\eta^*(0)-V(s'))_+^2]} - \sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^*(0)-V(s'))_+^2]}\right|.\end{aligned}$$

where the first inequality follows from

$$\begin{aligned}g(1,\eta^*(1)) &= \sup_{\eta\in\mathbb{R}}\left\{-\sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta-V(s'))_+^2]} + \eta\right\}\\ &\geq \left\{-\sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^*(0)-V(s'))_+^2]} + \eta^*(0)\right\} = g(1,\eta^*(0)).\end{aligned}$$

Combining the preceding results, we obtain the following bound in Case A:

$$\begin{aligned}\mathbb{E}_{P_{s,a}}[h(V(s'),P_{0,s,a})V(s')] - \mathbb{E}_{\widehat{P}_{s,a}}[h(V(s'),\widehat{P}_{0,s,a})V(s') \leq \min\Bigg\{&|\partial_t g(0,\eta^*(0))|,\\ &\sqrt{2\rho+1}\{\sqrt{\mathbb{E}_{\widehat{P}_0(\cdot|s,a)}[(\eta^*(0)-V(s'))_+^2]} - \sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta^*(0)-V(s'))_+^2]}\Bigg\}\end{aligned}$$

To obtain a sharper bound, we divide the analysis into two subcases according to whether $\mathbb{E}_{\widehat{P}_{0,s,a}(t)}[(\eta^*(0)-V(0))_+] \geq 1$.

Case A.1: $\mathbb{E}_{\widehat{P}_{0,s,a}(t)}[(\eta^*(0)-V(0))_+] \geq 1$**.**

Applying the Bernstein inequality, with probability at least $1-\frac{\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds

$$\begin{aligned}\left|\frac{\mathbb{E}_{\Delta P_{0,s,a}}\left[(\eta^*(t)-V(s))_+^2\right]}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right| \leq &\sqrt{\frac{2\mathrm{Var}_{P_{0,s,a}}\left(\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta^*(t)-V(s))_+]}\right)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}}\\ &+\frac{2Span(\frac{(\eta^*-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta^*(t)-V(s))_+]})\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3N}.\end{aligned}$$

Then, we analyze the upper bound of term $Span(\frac{(\eta^*-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta^*(t)-V(s))_+]})$.

If $\eta^*(0) \geq \frac{2}{1-\gamma} \geq \max_s V(s)$, then $\eta^*(0) - V(s) \geq \frac{\eta^*(0)}{2}$ for all $s \in \mathcal{S}$. It holds that

$$\mathbb{E}_{\widehat{P}_{0,s,a}}[(\eta^*(0)-V(s))_+] \geq \mathbb{E}_{\widehat{P}_{0,s,a}}\left[\frac{\eta^*(0)}{2}\right] \geq \frac{\eta^*(0)}{2}.$$

The term can be bounded as:

$$Span(\frac{(\eta^*(0)-V(s))_+^2}{\mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta^*(0)-V(s))_+\right]}) \leq \frac{\eta^*(0)\frac{1}{1-\gamma}}{\frac{\eta^*(0)}{2}} \leq \frac{2}{1-\gamma}.$$

Otherwise, $\eta^*(0) \leq \frac{2}{1-\gamma}$. It holds that

$$Span(\frac{(\eta^*(0)-V(s))_+^2}{\mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta^*(0)-V(s))_+\right]}) \leq Span((\eta^*(0)-V(s))_+^2) \leq \frac{\eta^*(0)}{1-\gamma} \leq \frac{1}{(1-\gamma)^2}.$$

Above all, the term $Span(\frac{(\eta^*(0)-V(s))_+^2}{\mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta^*(0)-V(s))_+\right]})$ can be bounded by $\frac{2}{(1-\gamma)^2}$ when Case A.1.

We next derive a sharper bound for the variance term $\mathrm{Var}_{P_{0,s,a}(t)}\left(\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right)$, which is key to the final result order.

Firstly, we present the expectation of the term $\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}$,

$$\begin{aligned}
&\mathbb{E}_{P_{0,s,a}(t)}\left[\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right] \\
&= \mathbb{E}_{P_{0,s,a}(t)}\left[\frac{\eta^*(t)(\eta^*(t)-V(s))_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right] - \mathbb{E}_{P_{0,s,a}(t)}\left[\frac{V(s)(\eta^*(t)-V(s))_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right] \\
&= \eta^*(t) - \bar{V}.
\end{aligned}$$

Define the function $h_t(v) = \frac{(\eta^*(t)-v)_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}$. According to the equation 18, it holds that $\eta^*(t) - \bar{V} = (2\rho+1)\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]$. Therefore, $h_t(\bar{V}) = (2\rho+1)(\eta^*(t)-\bar{V})$. Besides, by the mean value theorem, we can rewrite the bound as

$$\left|h_t(v) - h_t(\bar{V})\right| = |\frac{\partial_v h_t(\tilde{v})}{\partial v}||\bar{V}-v| = \left|\frac{(\eta^*(t)-\tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right||v-\bar{V}|,$$

where $\tilde{v} \in [\min\{v,\bar{V}\}, \max\{v,\bar{V}\}]$.

If $v \leq \bar{V}$, $\left|\frac{(\eta^*(t)-\tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right| \leq \left|\frac{(\eta^*(t)-v)_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right|$.

Otherwise, $\eta^*(t) > v > \bar{V}$, $\left|\frac{(\eta^*(t)-\tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right| \leq \left|\frac{(\eta^*(t)-\bar{V})_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right| = 2\rho+1$. Besides, $\left(h_t(v) - h_t(\bar{V})\right)^2 \leq h_t^2(\bar{V})$.

Else, $v \geq \eta^*(t)$, $\left(h_t(v) - h_t(\bar{V})\right)^2 = h_t(\bar{V})^2$.

Lemma 7. *If $\eta^*(t) \geq \frac{2}{1-\gamma}$, it holds that $\rho \leq 1$.*

Next, combined with Lemma 5 and 7, if $\rho \geq 1$, $\eta^*(t) \leq \frac{2}{1-\gamma}$. Therefore, we do the error decomposition as:

$$
\begin{aligned}
&\mathrm{Var}_{P_{0,s,a}(t)}\left(\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right) \\
&= \mathbb{E}_{P_{0,s,a}(t)}\left[\left(h_t(v)-h_t(\bar{V})\right)^2\right] - (2\rho)^2\mathbb{E}_{P_{0,s,a}(t)}\left[\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right]^2 \\
&= \mathbb{E}_{P_{0,s,a}(t)}\left[\left(\frac{(\eta^*(t)-\tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right)^2(V-\bar{V})^2\right] - 4\rho^2\mathbb{E}_{P_{0,s,a}(t)}\left[\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right]^2 \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] - 4\rho^2\mathbb{E}_{P_{0,s,a}(t)}\left[\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right]^2 \\
&\quad + \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-\bar{V})_+^4}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\mathbf{1}(V(s)>\bar{V})\right] \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + (2\rho+1)^2(\eta^*-\bar{V})^2 - 4\rho^2(\eta^*-\bar{V})^2 \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + (4\rho+1)(\eta^*-\bar{V})^2 \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + \frac{20\rho}{(1-\gamma)^2},
\end{aligned}
$$

where the last inequality follows from Lemma 7.

Otherwise, if $\rho < 1$, we do the error decomposition as follows:

$$
\begin{aligned}
&\mathrm{Var}_{P_{0,s,a}(t)}\left(\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right) \\
&= \mathbb{E}_{P_{0,s,a}(t)}\left[\left(h_t(v)-h_t(\bar{V})\right)^2\right] - (2\rho)^2\mathbb{E}_{P_{0,s,a}(t)}\left[\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right]^2 \\
&= \mathbb{E}_{P_{0,s,a}(t)}\left[\left(\frac{(\eta^*(t)-\tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right)^2(V-\bar{V})^2\right] - 4\rho^2\mathbb{E}_{P_{0,s,a}(t)}\left[\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right]^2 \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-\bar{V})_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)>\bar{V})\right] \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + (2\rho+1)^2\mathbb{E}_{P_{0,s,a}}\left[\|V-\bar{V}\|^2\mathbf{1}(V(s)>\bar{V})\right] \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + 9\mathbb{E}_{P_{0,s,a}}\left[\|V-\bar{V}\|^2\mathbf{1}(V(s)>\bar{V})\right].
\end{aligned}
$$

Therefore, above all, we get the bound of variance term as follows

$$
\begin{aligned}
&\mathrm{Var}_{P_{0,s,a}(t)}\left(\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t)-V(s))_+\right]}\right) \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + 9\mathbb{E}_{P_{0,s,a}}\left[\|V-\bar{V}\|^2\mathbf{1}(V(s)>\bar{V})\right] + \frac{20\rho}{(1-\gamma)^2}.
\end{aligned}
$$

Here, we define term I$= \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s) \leq \bar{V})\right]$. Term II $=$ $\mathbb{E}_{P_{0,s,a}}\left[\|V-\bar{V}\|^2\mathbf{1}(V(s)>\bar{V})\right]$.

Firstly, define $w(s) = \frac{(\eta^*-V(s))_+}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]}$. Then, according to the constrict, it holds that

$$\underline{P}_{s,a}(s') = w(s') \circ P_{0,s,a}(s'), \qquad \mathbb{E}_{P_{0,s,a}}[w(s')] = 1, \qquad \mathbb{E}_{P_{0,s,a}}\left[\frac{(w-1)^2}{2}\right] = \rho.$$

Besides, we could also get the result,

$$\mathbb{E}_{P_{0,s,a}}[w^2 - 2w + 1] = 2\rho \to \mathbb{E}_{P_{0,s,a}}[w^2] = 2\rho - 1 + 2\mathbb{E}_{P_{0,s,a}}[w] = 2\rho + 1.$$

Moreover, according to the equation 15, we can conclude that

$$(2\rho+1)\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+] = (\eta^*-\bar{V}).$$

Bounding term (1). We start from the decomposition derived previously and bound the expectation by separating the centered term and using $\bar{V} = \mathbb{E}_{\underline{P}_{s,a}}[V]$. Given the fact when $V(s) \leq \bar{V}$,

$$\bar{V} - V = (\eta^* - V) - (\eta^* - \bar{V}) = (\eta^* - V)_+ - (\eta^* - \bar{V})_+,$$

we can bound the term I as follows

$$\begin{aligned}
&\mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] \\
&= \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+\Big((\eta^*-V)_+ - (\eta^*-\bar{V})\Big)}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] \\
&\quad + \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+(\eta^*-\bar{V})}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+\Big((\eta^*-V)_+ - (\eta^*-\bar{V})\Big)}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + (2\rho+1)\operatorname{Var}_{\underline{P}_{s,a}}(V) \\
&= \mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+(\bar{V}-V)}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + (2\rho+1)\operatorname{Var}_{\underline{P}_{s,a}}(V) \\
&\leq \frac{1}{2}\mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2 + (\bar{V}-V)^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + (2\rho+1)\operatorname{Var}_{\underline{P}_{s,a}}(V),
\end{aligned}$$

where the first equality follows from $(\eta^* - V)_+ - (\eta^* - \bar{V}) = (\bar{V} - V)$ and the last inequality follows from the fact that $(\bar{V} - V)\mathbf{1}(V(s) \leq \bar{V}) \geq 0$.

Then, we can conclude that

$$\begin{aligned}
&\mathbb{E}_{P_{0,s,a}}\left[\frac{(\eta^*-V)_+^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{(\bar{V}-V)^2}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\|V-\bar{V}\|^2\mathbf{1}(V(s)\leq\bar{V})\right] + 2\,(2\rho+1)\operatorname{Var}_{\underline{P}_{s,a}}(V) \\
&\leq \mathbb{E}_{P_{0,s,a}}\left[\frac{\|V-\bar{V}\|^4}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\mathbf{1}(V(s)\leq\bar{V})\right] + 2\operatorname{Var}_{\underline{P}_{s,a}}(V) + \frac{4\rho}{(1-\gamma)^2}.
\end{aligned}$$

Above all, we can conclude that

Bounding the higher–order term. Consider the term

$$\mathbb{E}_{P_{0,s,a}}\left[\frac{\|V-\bar{V}\|^4\mathbf{1}(V(s)\le\bar{V})}{\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]^2}\right].$$

Define

$$Z:=\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+],\qquad w(s):=\frac{(\eta^*-V(s))_+}{Z}.$$

According to the equality,

$$\begin{aligned}\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+]&=(1+2\rho)^{-1}(\eta^*-\bar{V})\\&=(1+2\rho)^{-1}\mathbb{E}_{P_{0,s,a}}[(\eta^*-\bar{V})_+]\\&=(1+2\rho)^{-1}\mathbb{E}_{P_{0,s,a}}[w(\eta^*-\bar{V})_+],\end{aligned}$$

we can express the deviation from the mean through the weight w.

When $V(s)\le\bar{V}$, we can get that

$$\frac{\bar{V}-V(s)}{Z}=\frac{\eta^*-V}{Z}-\frac{\eta^*-\bar{V}}{Z}=w-(1+2\rho).$$

Therefore, we can conclude that

$$\frac{(V-\bar{V})^2}{Z^2}=\big(w-(1+2\rho)\big)^2\qquad\text{if } V(s)\le\bar{V}.$$

Besides, when $V(s)\le\bar{V}$, $w=\frac{(\eta^*-V(s))_+}{Z}\ge\frac{(\eta^*-\bar{V})_+}{Z}=1+2\rho$. Therefore, when $V(s)\le\bar{V}$, $w-1\ge2\rho$, it holds

$$(w-(1+2\rho))^2=(w-1-2\rho)^2\le(w-1)^2,$$

we obtain the bound

$$\begin{aligned}\mathbb{E}_{P_{0,s,a}}\left[\frac{\|V-\bar{V}\|^4}{Z^2}\mathbf{1}(V(s)\le\bar{V})\right]&\le\frac{1}{(1-\gamma)^2}\mathbb{E}_{P_{0,s,a}}[(w-1)^2]\\&\le\frac{2\rho}{(1-\gamma)^2}.\end{aligned}$$

Final bound for term (1). Combining the previous estimates gives

$$\text{Term (1)}\ \le\ \frac{2\rho}{(1-\gamma)^2}+2\mathrm{Var}_{\underline{P}_{s,a}}(V)+\frac{4\rho}{(1-\gamma)^2}.$$

Bounding term (2). Next consider term (2):

$$9\mathbb{E}_{P_{0,s,a}}\left[\|V-\bar{V}\|^2\mathbf{1}(V(s)<\bar{V})\right].$$

Let

$$Y=\|V(s)-\bar{V}\|^2\mathbf{1}(V(s)<\bar{V}),\qquad Y\in\Big[0,\frac{1}{(1-\gamma)^2}\Big].$$

Define the weight

$$w=\frac{(\eta^*-V)_+}{Z},\qquad Z=\mathbb{E}_{P_{0,s,a}}[(\eta^*-V)_+].$$

We use the inequality valid for any $w\ge0$:

$$1\le4w+4(w-1)^2.$$

Indeed,

$$4w + 4(w-1)^2 = 4\Big(w - \tfrac{1}{2}\Big)^2 + 3 \geq 3.$$

Multiplying both sides by Y and taking expectations yields

$$\begin{aligned}\mathbb{E}[Y] &\leq \mathbb{E}[4wY + 4(w-1)^2 Y] \\ &= 4\mathbb{E}[wY] + 4\mathbb{E}[(w-1)^2 Y].\end{aligned}$$

Using the bound $Y \leq (1-\gamma)^{-2}$ we obtain

$$\mathbb{E}[Y] \leq 4\mathbb{E}[wY] + \frac{4}{(1-\gamma)^2}\mathbb{E}[(w-1)^2].$$

Since $\mathbb{E}[(w-1)^2] = 2\rho$, this gives

$$\mathbb{E}[Y] \leq 4\operatorname{Var}_{\underline{P}_{s,a}}(V) + \frac{8\rho}{(1-\gamma)^2}.$$

Final variance bound. Combining the bounds for term (1) and term (2) yields

$$\begin{aligned}&\operatorname{Var}_{P_{0,s,a}(t)}\left(\frac{(\eta^*(t) - V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t) - V(s))_+\right]}\right) \\ &\leq \frac{2\rho}{(1-\gamma)^2} + 2\operatorname{Var}_{\underline{P}_{s,a}}(V) \quad + \frac{4\rho}{(1-\gamma)^2} + 36\operatorname{Var}_{\underline{P}_{s,a}}(V) + \frac{72\rho}{(1-\gamma)^2} + \frac{20\rho}{(1-\gamma)^2} \\ &\leq 38\operatorname{Var}_{\underline{P}_{s,a}}(V) + \frac{98\rho}{(1-\gamma)^2}.\end{aligned}$$

Then, we can get

$$\begin{aligned}&\mathbb{E}_{P_{s,a}}[h(V(s'), P_{0,s,a})V(s')] - \mathbb{E}_{\widehat{P}_{s,a}}[h(V(s'), \widehat{P}_{0,s,a})V(s') \\ &\leq \left|\frac{\mathbb{E}_{\Delta P_{0,s,a}}\left[(\eta^*(t) - V(s))_+^2\right]}{\mathbb{E}_{P_{0,s,a}(t)}\left[(\eta^*(t) - V(s))_+\right]}\right| \\ &\leq \sqrt{\frac{2\mathrm{Var}_{P_{0,s,a}}\left(\frac{(\eta^*(t)-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta^*(t)-V(s))_+]}\right)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}} \\ &\qquad + \frac{2Span(\frac{(\eta^*-V(s))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta^*(t)-V(s))_+]})\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3N} \\ &\leq \sqrt{\frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(38\operatorname{Var}_{\underline{P}_{s,a}}(V) + \frac{98\rho}{(1-\gamma)^2}\right)} + \frac{4\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3N(1-\gamma)^2}.\end{aligned}$$

Case A.2: $\mathbb{E}_{\widehat{P}_{0,s,a}(t)}\left[(\eta^*(0) - V(0))_+\right] < 1$**.**

Under these condition, we could get the following equation,

$$\mathbb{E}_{\widehat{P}_{0,s,a}(0)}\left[(\eta^*(0) - V(s))_+^2\right] = (2\rho+1)\mathbb{E}_{\widehat{P}_{0,s,a}(t)}\left[(\eta^*(0) - V(s))_+\right]^2.$$

Furthermore, we can get a upper bound of variance. It holds that

$$\begin{aligned}\operatorname{Var}_{\widehat{P}_{0,s,a}(0)}\left((\eta^*(0) - V(s))_+\right) &= \mathbb{E}_{\widehat{P}_{0,s,a}(0)}\left[(\eta^*(0) - V(s))_+^2\right] - \mathbb{E}_{\widehat{P}_{0,s,a}(t)}\left[(\eta^*(0) - V(s))_+\right]^2 \\ &= 2\rho\mathbb{E}_{\widehat{P}_{0,s,a}(t)}\left[(\eta^*(0) - V(s))_+\right]^2 \leq 2\rho.\end{aligned}$$

Next, we introduce an important lemma.

Lemma 8. *Given the χ^2 based $\mathcal{S}$-rectangular uncertainty set, the primal problem is equivalent to the following dual function:*

$$\begin{aligned}\inf_{P_{s,a}\in\mathcal{P}_{s,a}} \sum_{a\in\mathcal{A}} \mathbb{E}_{P_{s,a}}[V(s')] &= \sup_{\eta\in\mathbb{R}} \left\{ -\sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta - V(s'))_+^2]} + \eta \right\} \\ &= \sup_{\eta\in\mathbb{R}} \left\{ \eta - \mathbb{E}_{P_0(\cdot|s,a)}[(\eta - V)_+] - \sqrt{2\rho}\sqrt{\mathrm{Var}_{P_0(\cdot|s,a)}((\eta - V(s'))_+)} \right\}. \end{aligned} \tag{20}$$

Here, the optimizer, η^* of these two different formulation are same.

Then, under the case A.2, $g(0,\eta^*(0)) \geq g(1,\eta^*(1))$, we can conclude that

$$\begin{aligned}&\mathbb{E}_{P_{s,a}}[h(V(s'),P_{0,s,a})V(s')] - \mathbb{E}_{\widehat{P}_{s,a}}[h(V(s'),\widehat{P}_{0,s,a})V(s') \\ &= g(0,\eta^*(0)) - g(1,\eta^*(1)) \\ &\leq g(0,\eta^*(0)) - g(1,\eta^*(0)) \\ &\leq \left| \mathbb{E}_{P_0(\cdot|s,a)}[(\eta^*(0) - V)_+] - \mathbb{E}_{\widehat{P}_0(\cdot|s,a)}[(\eta^*(0) - V)_+] \right| \\ &\quad + \sqrt{2\rho} \left| \sqrt{\mathrm{Var}_{P_0(\cdot|s,a)}((\eta^*(0) - V(s'))_+)} - \sqrt{\mathrm{Var}_{\widehat{P}_0(\cdot|s,a)}((\eta^*(0) - V(s'))_+)} \right|. \end{aligned}$$

Here, under the case A.2, $\mathrm{Var}_{\widehat{P}_{0,s,a}(0)}\left((\eta^*(0) - V(s))_+\right) \leq 2\rho$. It holds that with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|}$,

$$\begin{aligned}&\left| \mathbb{E}_{P_0(\cdot|s,a)}[(\eta^*(0) - V)_+] - \mathbb{E}_{\widehat{P}_0(\cdot|s,a)}[(\eta^*(0) - V)_+] \right| \\ &\leq \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_{0,s,a}}((\eta^*(0) - V)_+)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}} + \frac{2Span((\eta^*(0) - V)_+)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3N} \\ &\leq \sqrt{\frac{4\rho\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}} + \frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3(1-\gamma)N}. \end{aligned}$$

Lemma 9 (Lemma 6 Panaganti & Kalathil (2022))**.** *Given P and V, with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds that*

$$\left| \sqrt{\mathrm{Var}_P(V)} - \sqrt{\mathrm{Var}_{\widehat{P}}(V)} \right| \leq \sqrt{\frac{2Span(V)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}}.$$

Then, applying the Lemma 6 in Panaganti & Kalathil (2022), we can conclude that with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds that

$$\left| \sqrt{\mathrm{Var}_{P_0(\cdot|s,a)}((\eta^*(0) - V(s'))_+)} - \sqrt{\mathrm{Var}_{\widehat{P}_0(\cdot|s,a)}((\eta^*(0) - V(s'))_+)} \right| \leq \sqrt{\frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}}.$$

Above all, with probability at least $1 - 2\frac{\delta}{|\mathcal{S}||\mathcal{A}|}$, wit holds that

$$g(0,\eta^*(0)) - g(1,\eta^*(1)) \leq \sqrt{\frac{4\rho\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}} + \frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3(1-\gamma)N} + \sqrt{\frac{4\rho\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}}.$$

Above the Case A.1 and A.1, with probability at least $1 - \frac{3\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds that

$$\begin{aligned}
&\left|\mathbb{E}_{P_{s,a}}[h(V(s'), P_{0,s,a})V(s')] - \mathbb{E}_{\widehat{P}_{s,a}}[h(V(s'), \widehat{P}_{0,s,a})V(s')\right| \\
&\leq \frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3(1-\gamma)N} + 4\sqrt{\frac{\rho\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}} + \sqrt{\frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(38\,\mathrm{Var}_{\underline{P}_{s,a}}(V) + \frac{98\rho}{(1-\gamma)^2}\right)} \\
&\quad + \frac{4\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3N(1-\gamma)^2} \\
&\leq \frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)N} + 34\sqrt{\frac{\rho\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}} + 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\mathrm{Var}_{\underline{P}_{s,a}}(V)\right)}
\end{aligned}$$

Otherwise, if $g(0, \eta^*(0)) < g(1, \eta^*(1))$, the same conclusions follow by an argument analogous to the one above.

C $\mathcal{S}$-RECTANGULAR UNCERTAINTY SET

In this section, we present the analysis of the $\mathcal{S}$-rectangular uncertainty-set DR-RL problem.

Similar, we define $g(\eta) = -\sqrt{|\mathcal{A}|(2\rho+1)}\left(\sum_{a\in\mathcal{A}}\mathbb{E}_{P_{0,s,a}}\left[(\eta_a - \pi(a|s)V(s))_+^2\right]\right)^{1/2} + \sum_{a\in\mathcal{A}}\eta_a$. Then, the $\eta^* = \arg\max_{\eta\in\mathbb{R}^{|\mathcal{A}|}} g(\eta)$.

Proposition 2. *Under the χ^2 $\mathcal{S}$-rectangular uncertainty set, if $\eta^* = \{\eta_a^*\}_{a\in\mathcal{A}}$ is the Lagrange parameters vector, given the policy π, the worst-case distribution within the uncertainty set can be written as*

$$\underline{P}^V_{s,a} = P_{0,s,a} \circ \frac{(\eta_a^* - \pi(a|s)V(s'))_+}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta_a^* - \pi(a|s)V(s'))_+]},$$

which is equivalent to

$$\inf_{P_s\in\mathcal{P}_s}\sum_{a\in\mathcal{A}}\pi(a|s)\mathbb{E}_{P(s'|s,a)}[V(s')] = \sum_{a\in\mathcal{A}}\frac{\mathbb{E}_{P_0(\cdot|s,a)}[\pi(a|s)V(s')(\eta_a^* - \pi(a|s)V(s'))_+]}{\mathbb{E}_{P_0(\cdot|s,a)}[(\eta_a^* - \pi(a|s)V(s'))_+]}.$$

Proof. Here, we will prove the proposition via two different conditions.

REGULAR CASE.

First, we prove the proposition in the regular case. For a given state, the worst-case transition kernels associated with the available actions do not all concentrate their probability mass on the worst state. Consequently, the uncertainty-set constraint is active.

Next, combined with KKT conditions, for every η_a, it holds that

$$\partial_{\eta_a} g(\eta^*) = -\sqrt{|\mathcal{A}|(2\rho+1)}\frac{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+\right]}{\sqrt{\left(\sum_{a'}\mathbb{E}_{P_{0,s,a'}}\left[(\eta_{a'}^* - \pi(a'|s)V(s'))_+^2\right]\right.}} + 1 = 0;$$

$$\to -\sqrt{|\mathcal{A}|(2\rho+1)}\sqrt{\left(\sum_{a'}\mathbb{E}_{P_{0,s,a'}}\left[(\eta_{a'}^* - \pi(a'|s)V(s'))_+^2\right]\right.} = -\frac{\left(\sum_{a'}\mathbb{E}_{P_{0,s,a'}}\left[(\eta_{a'}^* - \pi(a'|s)V(s'))_+^2\right]\right.}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+\right]};$$

$$\to \frac{\sqrt{\left(\sum_{a'}\mathbb{E}_{P_{0,s,a'}}\left[(\eta_{a'}^* - \pi(a'|s)V(s'))_+^2\right]\right.}}{\sqrt{|\mathcal{A}|(2\rho+1)}} = \mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+\right].$$

Then, we can conclude the following lemma.

Lemma 10. *Since these equations hold for all $a \in \mathcal{A}$, we can get*

$$\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+\right] = \mathbb{E}_{P_{0,s,a'}}\left[(\eta_{a'}^* - \pi(a'|s)V(s))_+\right],$$

for any $a, a' \in \mathcal{A}$. Moreover, it holds that

$$\sum_{a\in\mathcal{A}}(2\rho+1)\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+\right]^2 = \sum_{a\in\mathcal{A}}\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+^2\right].$$

Therefore, it holds that

$$\sum_{a\in\mathcal{A}}\mathrm{Var}_{P_{0,s,a}}((\eta^* - \pi(a|s)V(s'))_+) \le 2\rho\sum_{a\in\mathcal{A}}\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s))_+\right]^2.$$

Then, we can conclude that

$$\begin{aligned} g(\eta^*) &= -\frac{\sum_{a\in\mathcal{A}}\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s'))_+^2\right]}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s'))_+\right]} + \sum_a \eta_a^* \\ &= -\sum_{a\in\mathcal{A}}\frac{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s'))_+^2\right]}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s'))_+\right]} + \sum_a \eta_a^* \\ &= \sum_{a\in\mathcal{A}}\frac{\mathbb{E}_{P_{0,s,a}}\left[\pi(a|s)V(s')\cdot(\eta_a^* - \pi(a|s)V(s'))_+\right]}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^* - \pi(a|s)V(s'))_+\right]}. \end{aligned}$$

Therefore, for the $\mathcal{S}$-rectangular uncertainty set, the worst-case transitions can be written as

$$P_{s,a} = P^0_{s,a} \circ \frac{(\eta^*_a - \pi(a|s)V(s'))_+}{\mathbb{E}_{p_{s,a}}\left[(\eta^*_a - \pi(a|s)V(s'))_+\right]}. \tag{21}$$

DEGENERATE CASE.

In this case, the uncertainty set is degenerate, and the worst-case transition kernel has a $0-1$ form: it assigns probability 1 to the worst state and probability 0 to every other state.

It remains to verify that, under this condition, the worst-case transition kernel satisfies the closed-form solution. Specifically, η^* is equal to the value function at the state with the second-smallest value. Consequently, the worst-case transition kernel assigns probability 1 to the worst state and probability 0 to all other states, confirming the validity of the closed-form solution.

□

Then, consider variance. Next, we define $P_{s,a},\ \hat{P}_{s,a}, \quad P_s(t) = tP_s + (1-t)\hat{P}_s; \quad \Delta P_{s,a} = P_{s,a} - \hat{P}_{s,a}$.

Next, we can get the weighted function

$$g(t, \eta^*(t)) = -\sqrt{|\mathcal{A}|(2\rho+1)}\sqrt{\sum_a \mathbb{E}_{P_{s,a}(t)}\left[(\eta^*_a(t) - \pi(a|s)V(s'))^2_+\right]} + \sum_a \eta^*_a(t).$$

Next, combined with Mean Value Theorem, we can get

$$\begin{aligned}
&|g(0,\eta^*(0)) - g(1,\eta^*(1))| = |d_t g(t,\eta^*(t))| = |\partial_t g(t,\eta^*(t))| \\
&= \sqrt{|\mathcal{A}|(2\rho+1)} \cdot \frac{1}{2} \frac{\partial_t \sum_a \mathbb{E}_{P_{s,a}(t)}\left[(\eta^*_a(t) - \pi(a|s)V(s))^2_+\right]}{\sqrt{\sum_a \mathbb{E}_{P_{s,a}(t)}\left[(\eta^*_a(t) - \pi(a|s)V(s))^2_+\right]}} \\
&= \frac{\sqrt{|\mathcal{A}|(2\rho+1)}}{2} \sum_a \frac{\left\langle \Delta P_{s,a},\ (\eta^*_a(t) - \pi(a|s)V(s))^2_+ \right\rangle}{\sqrt{\mathbb{E}_{P_{s,a}(t)}\left[(\eta^*_a(t) - \pi(a|s)V(s))^2_+\right]}} \\
&= \frac{1}{2} \sum_a \frac{\left\langle \Delta P_{s,a},\ (\eta^*_a(t) - \pi(a|s)V(s))^2_+ \right\rangle}{\mathbb{E}_{P_{s,a}(t)}\left[(\eta^*_a(t) - \pi(a|s)V(s))_+\right]},
\end{aligned}$$

where $t \in (0,1)$.

According to the discussion in Section A, the key challenge lies in the bound of $\left\| \left(\underline{\widehat{P}}^{V_\pi,\pi} V_\pi - \underline{P}^{V_\pi,\pi} V_\pi \right) \right\|_\infty$.

Lemma 11. *Set $F(t) = g(t, \eta^*(t))$. The function $F(t)$ is convex to t.*

Lemma 12. *Given policy π and value function V, it holds that*

$$\begin{aligned}
&\inf_{P_s \in \mathcal{P}_s} \sum_{a\in\mathcal{A}} \pi(a|s)\, \mathbb{E}_{P_0(s'|s,a)}[V(s')] \\
&= \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} \left\{ -\sqrt{|\mathcal{A}|(2\rho+1)} \left(\sum_{a\in\mathcal{A}} \mathbb{E}_{P_{0,s,a}}\left[(\eta_a - \pi(a|s)V(s'))^2_+\right] \right)^{1/2} + \sum_{a\in\mathcal{A}} \eta_a \right\} \\
&= \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} \left\{ \sum_{a\in\mathcal{A}} \mathbb{E}_{P_{0,s,a}}\left[\eta_a - (\eta_a - \pi(a|s)V(s'))_+\right] - \sqrt{2|\mathcal{A}|\rho} \sqrt{\sum_a \mathrm{Var}_{P_{0,s,a}}((\eta_a - \pi(a|s)V(s'))_+)} \right\}.
\end{aligned}$$

Here, the multipliers are same for these two different formulations.

Next, we also divide the proof under two cases.

Case 1: $g(1,\eta^*(1)) \geq g(0,\eta^*(0))$. Then, it can get

$$\begin{aligned}|g(0,\eta^*(0)) - g(1,\eta^*(1))| &= g(1,\eta^*(1)) - g(0,\eta^*(0)) \leq |\partial_t g(1,\eta^*(1))| \\ &= \frac{1}{2}\left|\sum_a \frac{\left\langle \Delta P_{s,a},\ (\eta_a^*(1) - \pi(a|s)V(s))_+^2 \right\rangle}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s))_+\right]}\right| \\ &\leq \frac{1}{2}\sum_a \frac{\left|\left\langle \Delta P_{s,a},\ (\eta_a^*(1) - \pi(a|s)V(s))_+^2 \right\rangle\right|}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s))_+\right]}.\end{aligned}$$

Combined with Lemma 12, we also can get

$$\begin{aligned}&|g(0,\eta^*(0)) - g(1,\eta^*(1))| = g(1,\eta^*(1)) - g(0,\eta^*(0)) \leq g(1,\eta^*(1)) - g(0,\eta^*(1)) \\ &\quad\leq \sum_{a\in\mathcal{A}}\left|\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] - \mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]\right| \\ &+ \sqrt{2|\mathcal{A}|\rho}\left|\sqrt{\sum_a \mathrm{Var}_{P_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)} - \sqrt{\sum_a \mathrm{Var}_{\widehat{P}_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)}\right| \\ &\leq \sum_{a\in\mathcal{A}}\left|\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] - \mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]\right| \\ &+ \sqrt{2|\mathcal{A}|\rho}\sum_{a\in\mathcal{A}}\left|\sqrt{\mathrm{Var}_{P_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)} - \sqrt{\mathrm{Var}_{\widehat{P}_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)}\right|.\end{aligned}$$

Next, we consider two different cases.

Case A: $\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] > 1$. On this condition, with probability at least $1 - \frac{\delta}{|\mathcal{A}||\mathcal{S}|}$, it holds that

$$\begin{aligned}&\frac{\left|\left\langle \Delta P_{s,a},\ (\eta_a^*(1) - \pi(a|s)V(s'))_+^2 \right\rangle\right|}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]} \\ &\leq \sqrt{\frac{2\mathrm{Var}_{P_{0,s,a}}\left(\frac{(\eta_a^*(1) - \pi(a|s)V(s'))_+^2}{\mathbb{E}_{P_{0,s,a}(1)}[(\eta_a^*(t) - \pi(a|s)V(s'))_+]}\right)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}} \\ &\quad + \frac{2Span(\frac{(\eta_a^*(1) - \pi(a|s)V(s'))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta_a^*(1) - \pi(a|s)V(s'))_+]})\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{3N}.\end{aligned}$$

Here, the term $Span(\frac{(\eta_a^*(1) - \pi(a|s)V(s'))_+^2}{\mathbb{E}_{P_{0,s,a}(t)}[(\eta_a^*(1) - \pi(a|s)V(s'))_+]}) \leq \frac{2}{(1-\gamma)^2}$ following the discussion in Section B.

Similar, according to the discussion in Section B, we could get with probability at least $1 - \frac{\delta}{|\mathcal{A}||\mathcal{S}|}$, it holds that

$$\begin{aligned}&\frac{\left|\left\langle \Delta P_{s,a},\ (\eta_a^*(1) - \pi(a|s)V(s'))_+^2 \right\rangle\right|}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]} \\ &\leq \frac{2\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)N} + 34\sqrt{\frac{\rho_a\pi(a|s)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2N}} + 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\mathrm{Var}_{\underline{P}_{s,a}}(\pi(a|s)V(s'))\right)}\end{aligned}$$

Above all, we can conclude that

$$\begin{aligned}&\sum_{a\in\mathcal{A}}\frac{\left|\left\langle \Delta P_{s,a},\ (\eta_a^*(1) - \pi(a|s)V(s'))_+^2 \right\rangle\right|}{\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]} \\ &\leq \frac{2|\mathcal{A}|\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)N} + \sum_{a\in\mathcal{A}} 34\sqrt{\frac{\rho_a\pi(a|s)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2N}} + 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\mathrm{Var}_{\underline{P}_{s,a}}(\pi(a|s)V(s'))\right)}.\end{aligned}$$

Case A.2: $\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] \leq 1$. Then, we do error decomposition as follows: Combined with Lemma 12, we also can get

$$|g(0,\eta^*(0)) - g(1,\eta^*(1))| = g(1,\eta^*(1)) - g(0,\eta^*(0)) \leq g(1,\eta^*(1)) - g(0,\eta^*(1))$$
$$\leq \left|\sum_{a\in\mathcal{A}} \mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] - \sum_{a\in\mathcal{A}} \mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]\right|$$
$$+ \sqrt{2|\mathcal{A}|\rho}\left|\sqrt{\sum_a \mathrm{Var}_{P_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)} - \sqrt{\sum_a \mathrm{Var}_{\widehat{P}_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)}\right|.$$

To bound the first term, we introduce the following lemma.

Lemma 13 (Bernstein concentration for a sum of bounded random variables)**.** *Let $M, T \geq 1$ and $a > 0$. Suppose that $X_1, \ldots, X_M$ satisfy*

$$0 \leq X_m \leq a, \qquad \mathbb{E}[X_m] = u_m, \qquad \mathrm{Var}(X_m) = \sigma_m, \qquad m = 1, \ldots, M.$$

Here σ_n denotes the variance, rather than the standard deviation. Let $(X_{1,t}, \ldots, X_{M,t})$, $t = 1, \ldots, T$, be independent copies of $(X_1, \ldots, X_M)$, and define

$$H = \sum_{m=1}^{M} X_m, \qquad \widehat{H}_T = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} X_{m,t}.$$

If $X_1, \ldots, X_M$ are mutually independent, then, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\widehat{H}_T - \sum_{m=1}^{M} u_m\right| < \varepsilon\right) \geq 1 - 2\exp\left(-\frac{T\varepsilon^2}{2\sum_{m=1}^{M}\sigma_m + \frac{2}{3}a\varepsilon}\right). \tag{22}$$

Consequently, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\widehat{H}_T - \sum_{m=1}^{M} u_m\right| \leq \sqrt{\frac{2\log(2/\delta)}{T}\sum_{m=1}^{M}\sigma_m} + \frac{2a\log(2/\delta)}{3T}. \tag{23}$$

In particular,

$$T \geq \left(\frac{2\sum_{m=1}^{M}\sigma_m}{\varepsilon^2} + \frac{2a}{3\varepsilon}\right)\log\frac{2}{\delta} \tag{24}$$

is sufficient for the probability in equation 22 to be at least $1-\delta$.

Without independence among $X_1, \ldots, X_N$, but retaining independence across the T copies, the following bound holds:

$$\mathbb{P}\left(\left|\widehat{H}_T - \sum_{n=1}^{N} u_n\right| < \varepsilon\right) \geq 1 - 2\exp\left(-\frac{T\varepsilon^2}{2\left(\sum_{n=1}^{N}\sqrt{\sigma_n}\right)^2 + \frac{2}{3}Na\varepsilon}\right). \tag{25}$$

Then, with probability at least $1 - \frac{\delta}{|\mathcal{A}||\mathcal{S}|}$, it holds that

$$\left|\sum_{a\in\mathcal{A}} \mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] - \sum_{a\in\mathcal{A}} \mathbb{E}_{\widehat{P}_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right]\right|$$
$$\leq \sqrt{\frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{N}\sum_a \mathrm{Var}_{P_{0,s,a}}\left((\eta_a^*(1) - \pi(a|s)V(s'))_+\right)} + \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2}$$
$$\leq \sqrt{\frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{N}\sum_a \mathrm{Var}_{P_{0,s,a}}\left((\eta_a^*(1) - \pi(a|s)V(s'))_+\right)} + \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3(1-\gamma)^2 N}$$
$$\leq \sqrt{\frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{N}|\mathcal{A}|\rho} + \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2}.$$

Next, we try to bound the variance term. We firstly introduce the Lemma.

Lemma 14 (Range-based concentration for the square root of a sum of variances)**.** *Let $N \geq 1$, $T \geq 2$, and $a > 0$. Suppose that $\{X_{n,t} : 1 \leq n \leq N,\ 1 \leq t \leq T\}$ are mutually independent. For each n, let $X_{n,1}, \ldots, X_{n,T}$ be identically distributed copies of X_n, where*

$$0 \leq X_n \leq a, \qquad \mathbb{E}[X_n] = u_n, \qquad \mathrm{Var}(X_n) = \sigma_n.$$

Define

$$V \triangleq \sum_{n=1}^{N} \sigma_n, \qquad \overline{X}_{n,T} \triangleq \frac{1}{T}\sum_{t=1}^{T} X_{n,t}, \qquad \widehat{V}_T \triangleq \sum_{n=1}^{N} \frac{1}{T-1}\sum_{t=1}^{T}(X_{n,t} - \overline{X}_{n,T})^2.$$

Thus $\widehat{V}_T$ is the sum of the unbiased sample variances. For every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\sqrt{\widehat{V}_T} - \sqrt{V}\right| < \varepsilon\right) \geq 1 - 2\exp\left(-\frac{2\lfloor T/2 \rfloor \varepsilon^2}{a^2}\right). \tag{26}$$

Consequently, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\sqrt{\widehat{V}_T} - \sqrt{V}\right| \leq a\sqrt{\frac{\log(2/\delta)}{2\lfloor T/2 \rfloor}} \leq a\sqrt{\frac{\log(2/\delta)}{T-1}}. \tag{27}$$

Then, according to Lemma 14, we can conclude that with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds that

$$\left|\sqrt{\sum_a \mathrm{Var}_{P_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)} - \sqrt{\sum_a \mathrm{Var}_{\widehat{P}_{0,s,a}}((\eta_a^*(1) - \pi(a|s)V(s'))_+)}\right| \leq \sqrt{\frac{\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N(1-\gamma)^2}}.$$

Therefore, when $\mathbb{E}_{P_{0,s,a}}\left[(\eta_a^*(1) - \pi(a|s)V(s'))_+\right] \leq 1$, with probability at least $1 - \frac{2\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds that

$$\begin{aligned}
&|g(0,\eta^*(0)) - g(1,\eta^*(1))| = g(1,\eta^*(1)) - g(0,\eta^*(0)) \leq g(1,\eta^*(1)) - g(0,\eta^*(1)) \\
&\leq \sqrt{\frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{N}|\mathcal{A}|\rho} + \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2} + \sqrt{2\rho|\mathcal{A}|}\sqrt{\frac{\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N(1-\gamma)^2}} \\
&\leq \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2} + 2\sqrt{\rho|\mathcal{A}|}\sqrt{\frac{\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N(1-\gamma)^2}}.
\end{aligned}$$

Above all, with probability at least $1 - \frac{2\delta}{|\mathcal{S}||\mathcal{A}|}$, it holds that

$$
\begin{aligned}
&|g(0,\eta^*(0)) - g(1,\eta^*(1))| = g(1,\eta^*(1)) - g(0,\eta^*(0)) \le g(1,\eta^*(1)) - g(0,\eta^*(1)) \\
&\le \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2} + 2\sqrt{\rho|\mathcal{A}|}\sqrt{\frac{\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N(1-\gamma)^2}} + \frac{2|\mathcal{A}|\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)N} \\
&\quad + \sum_{a\in\mathcal{A}} 34\sqrt{\frac{\rho_a \pi(a|s)\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}} + \sum_{a\in\mathcal{A}} 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\mathrm{Var}_{\underline{P}_{s,a}}(\pi(a|s)V(s'))\right)} \\
&\le \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2} + 2\sqrt{\rho|\mathcal{A}|}\sqrt{\frac{\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N(1-\gamma)^2}} + \frac{2|\mathcal{A}|\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)N} \\
&\quad + 34\sqrt{\frac{\sum_{a\in\mathcal{A}}\rho_a \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}} + 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\sum_{a\in\mathcal{A}}\pi(a|s)\,\mathrm{Var}_{\underline{P}_{s,a}}(V(s'))\right)} \\
&\le \frac{2\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2} + 2\sqrt{\rho|\mathcal{A}|}\sqrt{\frac{\log(\frac{|\mathcal{S}||\mathcal{A}|}{\delta})}{N(1-\gamma)^2}} + \frac{2|\mathcal{A}|\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)N} \\
&\quad + 34\sqrt{\frac{\rho|\mathcal{A}|\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}} + 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\mathrm{Var}_{\underline{P}_{s,\pi}}(V(s'))\right)} \\
&\le \frac{8\log\left(\frac{|\mathcal{A}||\mathcal{S}|}{\delta}\right)}{3N(1-\gamma)^2} + 36\sqrt{\frac{\rho|\mathcal{A}|\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{(1-\gamma)^2 N}} + 9\sqrt{\frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{N}\left(\mathrm{Var}_{\underline{P}_{s,\pi}}(V(s'))\right)}.
\end{aligned}
$$

Otherwise, if $g(0,\eta^*(0)) > g(1,\eta^*(1))$, the same conclusions follow by an argument analogous to the one above.

D LOWER BOUND FOR $\mathcal{S}$-RECTANGULAR DR-RL

D.1 MODEL, ASSUMPTIONS, AND BALANCED-SAMPLING STATEMENT

Let $S \geq 4$ and $A \geq 2$ be integers. Throughout the analysis, assume

$$\gamma \in [3/4, 1), \qquad \rho \geq 0, \qquad 0 < \varepsilon \leq \frac{1}{128(1-\gamma)}, \qquad \beta \triangleq 2A\rho. \tag{28}$$

The learner observes transitions from the nominal kernel. The ambiguity set is used to evaluate the returned policy; adversarial transitions are not observed.

For distributions P, Q on a finite set, use the normalization

$$d_{\chi^2}(Q|P) \triangleq \frac{1}{2} \sum_{x:P(x)>0} \frac{(Q(x)-P(x))^2}{P(x)}, \qquad Q \ll P, \tag{29}$$

and set the divergence to $+\infty$ when $Q \not\ll P$. Here $Q \ll P$ means that $Q(x) = 0$ whenever $P(x) = 0$. For a nominal transition kernel P, define

$$\mathcal{U}(P) \triangleq \left\{ Q : \begin{array}{l} Q(\cdot|s,\ell) \in \Delta(\mathcal{S}), \quad Q(\cdot|s,\ell) \ll P(\cdot|s,\ell) \quad \text{for all } s,\ell, \\ \sum_{\ell \in \mathcal{A}} d_{\chi^2}\big(Q(\cdot|s,\ell)|P(\cdot|s,\ell)\big) \leq A\rho \quad \text{for every } s \end{array} \right\}, \tag{30}$$

where $\Delta(\mathcal{S})$ is the probability simplex on $\mathcal{S}$. Thus the budget is independent across states and shared across actions at each state. Nominal kernels may have zero entries; in particular, absolute continuity fixes every deterministic row.

Let Π_{stat} be the class of stationary randomized Markov policies. For $\pi \in \Pi_{\mathrm{stat}}$, write

$$\begin{aligned} V^{\pi,Q}(s) &\triangleq \mathbb{E}_s^{\pi,Q}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t)\right], \\ V_P^{\pi}(s) &\triangleq \inf_{Q \in \mathcal{U}(P)} V^{\pi,Q}(s), \qquad V_P^*(s) \triangleq \sup_{\pi \in \Pi_{\mathrm{stat}}} V_P^{\pi}(s). \end{aligned} \tag{31}$$

Rewards lie in $[0, 1]$ and are received at the current state and action. All values are therefore finite. The constructions below use the same reward function and radius under every hypothesis, so these quantities may be known to the learner.

Theorem 4 (Balanced nominal sampling)**.** *Under equation 28, let $n \geq 0$ be an integer. Suppose a learner receives n independent nominal samples from every state–action pair and returns $\widehat{\pi} \in \Pi_{\mathrm{stat}}$, possibly using internal randomness. If*

$$n \leq \frac{1-\gamma+2|\mathcal{A}|\rho}{8192\varepsilon^2(1-\gamma)^4}, \tag{32}$$

then there exists an S-state, A-action MDP for which

$$\mathbb{P}\Big\{\|V_P^* - V_P^{\widehat{\pi}}\|_\infty > \varepsilon\Big\} \geq \frac{1}{4}. \tag{33}$$

Consequently, a uniform guarantee with failure probability strictly below $1/4$ requires a total balanced sample count $N = |\mathcal{S}||\mathcal{A}|n$ satisfying

$$N > \frac{|\mathcal{S}||\mathcal{A}|(1-\gamma+2A\rho)}{8192\varepsilon^2(1-\gamma)^4}. \tag{34}$$

In particular, the required sample count is

$$N = \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3} + \frac{|\mathcal{S}||\mathcal{A}|^2\rho}{\varepsilon^2(1-\gamma)^4}\right). \tag{35}$$

D.2 BERNOULLI CALIBRATION AND SENSITIVITY

The construction is parameterized by the lowest feasible exit probability of an informative action. For $t \in (0,1)$, define

$$\alpha_\beta(t) \triangleq \max\{0, t - \sqrt{\beta t(1-t)}\}. \tag{36}$$

For $a \in (0,1)$, its positive-level inverse is

$$p_\beta(a) \triangleq \frac{\beta + 2a + \sqrt{\beta^2 + 4\beta a(1-a)}}{2(1+\beta)}. \tag{37}$$

In particular, $p_0(a) = a$.

Set

$$\eta \triangleq 32\varepsilon(1-\gamma)^2, \qquad a_0 \triangleq 1-\gamma, \qquad a_1 \triangleq 1-\gamma+\eta. \tag{38}$$

The equation 28 gives

$$0 < \eta \le \frac{1-\gamma}{4}, \qquad 0 < a_0 < a_1 \le \frac{5(1-\gamma)}{4} \le \frac{5}{16} < 1. \tag{39}$$

For discrete probability laws, define Hellinger affinity and squared Hellinger distance by

$$\mathrm{Aff}(P,Q) \triangleq \sum_x \sqrt{P(x)Q(x)}, \qquad \mathrm{Hel}^2(P,Q) \triangleq \sum_x (\sqrt{P(x)} - \sqrt{Q(x)})^2 = 2[1 - \mathrm{Aff}(P,Q)]. \tag{40}$$

Lemma 15 (Inverse calibration and sensitivity)**.** *For every finite $\beta \ge 0$ and $a \in (0,1)$, $p_\beta(a) \in (0,1)$ and*

$$\alpha_\beta(p_\beta(a)) = a, \qquad p'_\beta(a) > 0, \qquad \frac{[p'_\beta(a)]^2}{p_\beta(a)(1-p_\beta(a))} = \frac{4}{\beta + 4a(1-a)}. \tag{41}$$

If $t_- \triangleq p_\beta(a_0)$ and $t_+ \triangleq p_\beta(a_1)$, then

$$\mathrm{Hel}^2\big(\mathrm{Ber}(t_-), \mathrm{Ber}(t_+)\big) \le \frac{\eta^2}{1-\gamma+\beta}. \tag{42}$$

Proof. We first verify that equation 37 calibrates the robust endpoint. When $\beta = 0$, the identities $p_0(a) = a$ and $p'_0(a) = 1$ give all statements in equation 41 directly. For $\beta > 0$, $p_\beta(a)$ is the larger root of

$$(p-a)^2 - \beta p(1-p) = 0. \tag{43}$$

The polynomial has positive leading coefficient $1+\beta$, is negative at $p = a$, and is positive at $p = 1$. Its larger root lies in $(a,1)$. Since $p_\beta(a) - a > 0$, taking the nonnegative square root in equation 43 gives

$$p_\beta(a) - \sqrt{\beta p_\beta(a)(1-p_\beta(a))} = a > 0. \tag{44}$$

This proves the endpoint identity.

Next, write $p = p_\beta(a)$ and differentiate equation 43:

$$2(p-a)(p'_\beta(a) - 1) - \beta(1-2p)p'_\beta(a) = 0.$$

Rearranging yields

$$[2(p-a) - \beta(1-2p)]p'_\beta(a) = 2(p-a). \tag{45}$$

By equation 37, the coefficient on the left is

$$2(p-a) - \beta(1-2p) = 2(1+\beta)p - (\beta + 2a) = \sqrt{\beta^2 + 4\beta a(1-a)} > 0.$$

The right side is positive as well, so $p'_\beta(a) > 0$. Squaring equation 45 and using $(p-a)^2 = \beta p(1-p)$ gives

$$\frac{[p'_\beta(a)]^2}{p(1-p)} = \frac{4(p-a)^2}{\big(\beta^2 + 4\beta a(1-a)\big)p(1-p)} = \frac{4\beta}{\beta^2 + 4\beta a(1-a)} = \frac{4}{\beta + 4a(1-a)}.$$

All denominators are positive.

To obtain the finite-perturbation bound, define $v(a) = (\sqrt{p_\beta(a)}, \sqrt{1-p_\beta(a)})$. Differentiation and equation 41 give

$$\|v'(a)\|_2^2 = \frac{[p'_\beta(a)]^2}{4p_\beta(a)(1-p_\beta(a))} = \frac{1}{\beta + 4a(1-a)}.$$

For $a \in [a_0, a_1]$, equation 39 implies $a \geq 1-\gamma$ and $1-a \geq 11/16$. Hence $4a(1-a) \geq 11(1-\gamma)/4 \geq 1-\gamma$. The fundamental theorem of calculus and Cauchy–Schwarz now yield

$$\begin{aligned}\mathrm{Hel}^2\big(\mathrm{Ber}(t_-), \mathrm{Ber}(t_+)\big) &= \left\| \int_{a_0}^{a_1} v'(a)\, da \right\|_2^2 \\ &\leq \eta \int_{a_0}^{a_1} \|v'(a)\|_2^2\, da \leq \frac{\eta^2}{1-\gamma+\beta}.\end{aligned}$$

The argument includes $\beta = 0$, for which the derivative identity was verified separately. □

The identity equation 41 controls statistical separation in terms of the robust endpoint gap η. This is the link between the policy-loss calculation and the testing bounds below.

D.3 Construction of the hard MDPs

States, actions, and rewards. Let $\Theta = \{0, 1\}$. For each $\theta \in \Theta$, define the nominal MDP $M_\theta = (\mathcal{S}, \mathcal{A}, P_\theta, r, \gamma)$ with

$$\mathcal{S} = \{d, u_0, u_1, g\} \cup \mathcal{V}, \qquad \mathcal{A} = \{0, 1, \ldots, A-1\}, \qquad \mathcal{V} = \{v_1, \ldots, v_{S-4}\}. \tag{46}$$

All listed states are distinct, and $\mathcal{V} = \varnothing$ when $S = 4$. State d is the decision state, u_0, u_1 are the evaluation states, and g is an absorbing good state. The padding states in $\mathcal{V}$ are absorbing and have zero reward. Every state has the same A actions, and

$$r(s, \ell) = \mathbf{1}\{s = g\}. \tag{47}$$

Thus a transition from an evaluation state to g earns its first positive reward at the next time step.

Nominal transition probabilities. Let $b \triangleq (1-\gamma)/2$. For $j \in \{0, 1\}$, define

$$a_j^{(\theta)} = \begin{cases} a_1, & j = \theta, \\ a_0, & j \neq \theta, \end{cases} \qquad t_j^{(\theta)} \triangleq p_\beta(a_j^{(\theta)}), \qquad \tau_{j,\ell}^{(\theta)} = \begin{cases} t_j^{(\theta)}, & \ell = 0, \\ b, & \ell \geq 1. \end{cases} \tag{48}$$

Here j labels an evaluation branch, ℓ labels an action, and θ identifies the high branch. The quantities a_0, a_1 are robust exit probabilities, whereas t_-, t_+ are nominal probabilities. Lemma 15 and equation 39 give $0 < b < a_0 < a_1 < 1$, $0 < t_- < t_+ < 1$, and $\alpha_\beta(t_j^{(\theta)}) = a_j^{(\theta)}$.

For all $s, s' \in \mathcal{S}$ and $\ell \in \mathcal{A}$, set

$$P_\theta(s'|s, \ell) = \begin{cases} \mathbf{1}\{s' = u_0\}, & s = d,\ \ell = 0, \\ \mathbf{1}\{s' = u_1\}, & s = d,\ \ell \geq 1, \\ \tau_{j,\ell}^{(\theta)} \mathbf{1}\{s' = g\} + (1 - \tau_{j,\ell}^{(\theta)}) \mathbf{1}\{s' = u_j\}, & s = u_j,\ j \in \{0, 1\}, \\ \mathbf{1}\{s' = g\}, & s = g, \\ \mathbf{1}\{s' = v_m\}, & s = v_m \in \mathcal{V}. \end{cases} \tag{49}$$

Every row is a probability distribution. A trajectory starting at d visits it only at time zero, then remains in its selected evaluation branch until reaching g. It cannot reach a padding state.

The decision-action classes are

$$\mathcal{A}_0 \triangleq \{0\}, \qquad \mathcal{A}_1 \triangleq \{1, \ldots, A-1\}. \tag{50}$$

An action in $\mathcal{A}_j$ selects branch u_j. The only nominal rows that differ between the two hypotheses are

$$\mathcal{I} \triangleq \{(u_0, 0), (u_1, 0)\}. \tag{51}$$

All other rows, rewards, and the radius are common to both models. Each model uses its own nominal center in equation 30.

The joint statewise ambiguity set. Absolute continuity fixes every deterministic row. At an evaluation state u_j, it restricts the support of each action row to $\{u_j, g\}$. Write $z_{j,\ell} \triangleq Q(g|u_j, \ell)$, so that $Q(u_j|u_j, \ell) = 1 - z_{j,\ell}$, and let $\boldsymbol{z}_j = (z_{j,0}, \ldots, z_{j,A-1})$. For a Bernoulli row,

$$d_{\chi^2}\big(\mathrm{Ber}(z)|\,\mathrm{Ber}(t)\big) = \frac{1}{2}\left[\frac{(z-t)^2}{t} + \frac{((1-z)-(1-t))^2}{1-t}\right] = \frac{(z-t)^2}{2t(1-t)}. \tag{52}$$

Therefore the complete ambiguity set at u_j is

$$\mathcal{Z}_{\theta,j} = \left\{\boldsymbol{z}_j \in [0,1]^A : \frac{(z_{j,0} - t_j^{(\theta)})^2}{t_j^{(\theta)}(1-t_j^{(\theta)})} + \sum_{\ell=1}^{A-1} \frac{(z_{j,\ell} - b)^2}{b(1-b)} \le \beta\right\}. \tag{53}$$

The constraints at u_0 and u_1 are independent, but action coordinates within either state share one budget.

Marginal range of the informative action. Since every summand in equation 53 is nonnegative, feasibility requires $(z_{j,0} - t_j^{(\theta)})^2 \le \beta t_j^{(\theta)}(1 - t_j^{(\theta)})$. Conversely, every $z_{j,0} \in [0,1]$ satisfying this inequality is jointly feasible by setting $z_{j,\ell} = b$ for $\ell \ge 1$. The projection onto the informative coordinate is thus exactly

$$\left[\max\left\{0, t_j^{(\theta)} - \sqrt{\beta t_j^{(\theta)}(1-t_j^{(\theta)})}\right\},\ \min\left\{1, t_j^{(\theta)} + \sqrt{\beta t_j^{(\theta)}(1-t_j^{(\theta)})}\right\}\right]. \tag{54}$$

Its lower endpoint is $a_j^{(\theta)}$. When $\rho > 0$, attaining this endpoint spends the entire statewise divergence budget $A\rho$, so all other rows at that state must remain nominal. Other action rows can be perturbed when the informative row does not exhaust the budget. When $\rho = 0$, the ambiguity set contains only the nominal kernel.

D.4 Robust policy-loss gap

For the constructed models, abbreviate $V_{P_\theta}^{\pi}$ and $V_{P_\theta}^{*}$ as V_θ^π and V_θ^*. The probability that π selects the correct branch is

$$w_\theta(\pi) \triangleq \sum_{\ell \in \mathcal{A}_\theta} \pi(\ell|d) = \begin{cases} \pi(0|d), & \theta = 0, \\ 1 - \pi(0|d), & \theta = 1. \end{cases} \tag{55}$$

Proposition 3 (Loss from selecting the low branch)**.** *Define*

$$G(z) \triangleq \frac{\gamma z}{(1-\gamma)(1-\gamma+\gamma z)}, \qquad z \in [0,1]. \tag{56}$$

Then $V_\theta^*(d) = \gamma G(a_1)$, and every $\pi \in \Pi_{\mathrm{stat}}$ satisfies

$$V_\theta^*(d) - V_\theta^\pi(d) \ge \frac{\gamma^2 \eta[1 - w_\theta(\pi)]}{(1-\gamma+\gamma a_0)(1-\gamma+\gamma a_1)} \tag{57}$$

$$\ge 4\varepsilon[1 - w_\theta(\pi)]. \tag{58}$$

Proof.

Evaluation under a fixed kernel. The value at g is $1/(1-\gamma)$. At a zero-reward evaluation state with fixed effective exit probability z, the Bellman equation is

$$V = \gamma\left[\frac{z}{1-\gamma} + (1-z)V\right].$$

Solving it gives equation 56, since $1 - \gamma(1-z) = 1 - \gamma + \gamma z > 0$. The quotient rule gives

$$G'(z) = \frac{\gamma(1-\gamma+\gamma z) - \gamma^2 z}{(1-\gamma)(1-\gamma+\gamma z)^2} = \frac{\gamma}{(1-\gamma+\gamma z)^2} > 0.$$

Thus the value increases with the exit probability.

A common feasible adversarial kernel. Define Q_θ^{w} by setting $Q_\theta^{\mathrm{w}}(g|u_j, 0) = a_j^{(\theta)}$ for both $j = 0, 1$, and leaving every other row nominal. Equations equation 44 and equation 52 give, for $i = 0, 1$,

$$\frac{(p_\beta(a_i) - a_i)^2}{2p_\beta(a_i)(1 - p_\beta(a_i))} = \frac{\beta}{2} = A\rho.$$

The equality also holds at $\beta = 0$. All other row divergences are zero. Because the budgets at the two evaluation states are independent, both changes are feasible in the same kernel: $Q_\theta^{\mathrm{w}} \in \mathcal{U}(P_\theta)$.

Let $q_j \triangleq \pi(0|u_j)$. Under Q_θ^{w}, the effective exit probability at u_j is $q_j a_j^{(\theta)} + (1 - q_j)b \le a_j^{(\theta)}$, since $b < a_j^{(\theta)}$ and $q_j \in [0, 1]$. Consequently,

$$V_\theta^\pi(u_j) \le V^{\pi, Q_\theta^{\mathrm{w}}}(u_j) = G\big(q_j a_j^{(\theta)} + (1 - q_j)b\big) \le G(a_j^{(\theta)}). \tag{59}$$

The first inequality follows from taking an infimum over a set containing Q_θ^{w}.

Optimal values and the decision-state gap. A policy that always selects action 0 at u_j has value $G(z)$ under any admissible kernel, with $z \ge a_j^{(\theta)}$ by equation 54. Monotonicity of G gives a lower bound of $G(a_j^{(\theta)})$, which the witness attains. Together with equation 59, this proves that the optimal robust evaluation value is $G(a_j^{(\theta)})$.

The decision rows are fixed. Applying the same witness to both branches yields

$$V_\theta^\pi(d) \le V^{\pi, Q_\theta^{\mathrm{w}}}(d) \le \gamma\big[w_\theta(\pi)G(a_1) + (1 - w_\theta(\pi))G(a_0)\big].$$

Selecting the high branch and then always using action 0 attains robust value $\gamma G(a_1)$, so this is $V_\theta^*(d)$. Subtracting gives

$$V_\theta^*(d) - V_\theta^\pi(d) \ge \gamma[1 - w_\theta(\pi)]\,[G(a_1) - G(a_0)]. \tag{60}$$

Using $a_1 - a_0 = \eta$ and a common denominator,

$$\begin{aligned}\gamma[G(a_1) - G(a_0)] &= \frac{\gamma^2}{1 - \gamma} \frac{a_1(1 - \gamma + \gamma a_0) - a_0(1 - \gamma + \gamma a_1)}{(1 - \gamma + \gamma a_0)(1 - \gamma + \gamma a_1)} \\ &= \frac{\gamma^2 \eta}{(1 - \gamma + \gamma a_0)(1 - \gamma + \gamma a_1)}.\end{aligned}$$

Here $\gamma^2 \ge 9/16$, $1 - \gamma + \gamma a_0 \le 2(1 - \gamma)$, and $1 - \gamma + \gamma a_1 \le 9(1 - \gamma)/4$. The denominators are positive, so

$$\gamma[G(a_1) - G(a_0)] \ge \frac{(9/16)\eta}{2(1 - \gamma)\, 9(1 - \gamma)/4} = \frac{\eta}{8(1 - \gamma)^2} = 4\varepsilon.$$

Substitution into equation 60 proves both inequalities. □

D.5 Testing reduction and proof of the balanced lower bound

Proof of Theorem 4.

The statistical experiment. For every (s, ℓ), let $Y_{s,\ell,m}$, $1 \le m \le n$, be independent samples from $P_\theta(\cdot|s, \ell)$, independent across all state–action pairs. Write

$$\mathcal{D} = (Y_{s,\ell,m})_{s,\ell,m}, \qquad \mathsf{P}_\theta = \otimes_{s \in \mathcal{S}} \otimes_{\ell \in \mathcal{A}} P_\theta(\cdot|s, \ell)^{\otimes n}.$$

Here P_θ is a transition kernel and P_θ is the law of the entire dataset. Include the learner's independent random seed U with common law ν, so that $\widehat{\pi} = \widehat{\pi}(\mathcal{D}, U)$ is measurable. Let $\mathbb{P}_\theta$ denote probability under $\mathsf{P}_\theta \otimes \nu$.

Decoding the high branch. Define

$$\widehat{\theta}(\widehat{\pi}) = \begin{cases} 0, & \widehat{\pi}(0|d) > 1/2, \\ 1, & \widehat{\pi}(0|d) \le 1/2. \end{cases} \tag{61}$$

If $\widehat{\theta} \neq \theta$, then $w_\theta(\widehat{\pi}) \leq 1/2$ under either hypothesis, including the tie case. Proposition 3 therefore gives

$$\{\widehat{\theta} \neq \theta\} \subseteq \{V_\theta^*(d) - V_\theta^{\widehat{\pi}}(d) \geq 2\varepsilon\} \subseteq \{V_\theta^*(d) - V_\theta^{\widehat{\pi}}(d) > \varepsilon\}. \tag{62}$$

This implication holds for every stationary randomized output policy, regardless of its actions at the evaluation states.

Hellinger distance of the nominal data laws. Let $c \triangleq \mathrm{Aff}(\mathrm{Ber}(t_-), \mathrm{Ber}(t_+)) \in [0,1]$. Only the two rows in equation 51 differ between hypotheses. Each contributes affinity c^n, so independence gives $\mathrm{Aff}(\mathsf{P}_0, \mathsf{P}_1) = c^{2n}$. For $n \geq 1$, $1 - c^{2n} = (1-c)\sum_{m=0}^{2n-1} c^m \leq 2n(1-c)$. It follows from Lemma 15 that

$$\begin{aligned} \mathrm{Hel}^2(\mathsf{P}_0, \mathsf{P}_1) &= 2(1 - c^{2n}) \\ &\leq 4n(1-c) = 2n\,\mathrm{Hel}^2(\mathrm{Ber}(t_-), \mathrm{Ber}(t_+)) \\ &\leq \frac{2n\eta^2}{1-\gamma+\beta}. \end{aligned} \tag{63}$$

For $n = 0$, the two data laws coincide and the same bound holds.

Define total variation by $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$. Cauchy–Schwarz yields

$$\begin{aligned} \mathrm{TV}(P,Q) &= \frac{1}{2}\sum_x |\sqrt{P(x)} - \sqrt{Q(x)}|(\sqrt{P(x)} + \sqrt{Q(x)}) \\ &\leq \frac{1}{2}\,\mathrm{Hel}(P,Q)\sqrt{2 + 2\,\mathrm{Aff}(P,Q)} \leq \mathrm{Hel}(P,Q). \end{aligned}$$

Taking a product with the same independent seed preserves both distances: integrating over ν multiplies their defining sums by $\int d\nu = 1$. Thus

$$n \leq \frac{1-\gamma+\beta}{8\eta^2} \quad \Longrightarrow \quad \mathrm{TV}(\mathsf{P}_0 \otimes \nu, \mathsf{P}_1 \otimes \nu) \leq \frac{1}{2}. \tag{64}$$

Testing risk and sample count. Let $E = \{\widehat{\theta} = 0\}$. The defining event bound for total variation gives

$$\begin{aligned} \mathbb{P}_0\{\widehat{\theta} \neq 0\} + \mathbb{P}_1\{\widehat{\theta} \neq 1\} &= \mathbb{P}_0(E^c) + \mathbb{P}_1(E) \\ &= 1 - [\mathbb{P}_0(E) - \mathbb{P}_1(E)] \\ &\geq 1 - \mathrm{TV}(\mathsf{P}_0 \otimes \nu, \mathsf{P}_1 \otimes \nu). \end{aligned}$$

Under equation 64, the sum is at least $1/2$, so one of the error probabilities is at least $1/4$. By equation 62,

$$\max_{\theta \in \{0,1\}} \mathbb{P}_\theta\{V_\theta^*(d) - V_\theta^{\widehat{\pi}}(d) > \varepsilon\} \geq \frac{1}{4}.$$

Since $\eta^2 = 1024\varepsilon^2(1-\gamma)^4$, condition equation 64 equals equation 32. Failure at d implies failure of the uniform-in-state guarantee. Finally, the balanced design prescribes n samples at each of the $|\mathcal{S}||\mathcal{A}|$ pairs, so $N = |\mathcal{S}||\mathcal{A}|n$ gives equation 34 and equation 35. □

E Extension to adaptive nominal queries

We consider the same nominal MDP class and ambiguity set, now allowing the learner to select each state–action query using its previous observations and internal randomness. Conditional on the selected pair and the query history, the response is a fresh draw from the corresponding nominal transition law. The learner returns a stationary randomized policy in Π_{stat} that must be accurate simultaneously at every state. Each state has its own ambiguity budget $A\rho$, shared across its A actions.

Throughout this section, write $h \triangleq 1 - \gamma$ and assume

$$S \geq 4, \qquad A \geq 2, \qquad \gamma \in [3/4, 1), \qquad \rho \geq 0, \qquad 0 < \varepsilon \leq \frac{1}{128h},$$

where S and A are integers. Recall the calibrated parameters

$$\beta = 2A\rho, \qquad \eta = 32\varepsilon h^2, \qquad a_0 = h, \qquad a_1 = h + \eta.$$

Thus $0 < \eta \le h/4$ and $0 < a_0 < a_1 \le 5h/4 \le 5/16$. For a nominal kernel P, write V_P^π and V_P^* for the corresponding robust policy value and optimal robust value. Define the binary relative entropy, using natural logarithms, by

$$\mathrm{kl}(u, v) \triangleq u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v}, \qquad u, v \in (0, 1),$$

with its usual extended-real boundary values.

Theorem 5 (Adaptive-query lower bound)**.** *Let $k \triangleq \lfloor (S-1)/3 \rfloor$ and $0 < \delta < 1/2$. Suppose an algorithm makes at most T nominal generative-model queries almost surely and, for every nominal MDP in the class, outputs $\widehat{\pi} \in \Pi_{\mathrm{stat}}$ satisfying*

$$\mathbb{P}_P\Big\{ \|V_P^* - V_P^{\widehat{\pi}}\|_\infty \le \varepsilon \Big\} \ge 1 - \delta.$$

The probability includes both the nominal observations and the algorithm's internal randomness. Then

$$T \ge \frac{k(1-\gamma+\beta)}{8192\varepsilon^2(1-\gamma)^4} \, \mathrm{kl}(1-\delta, \delta). \tag{65}$$

For each fixed $\delta \in (0, 1/2)$, this implies

$$T = \Omega\left(\frac{S}{\varepsilon^2(1-\gamma)^3} + \frac{|\mathcal{S}||\mathcal{A}|\rho}{\varepsilon^2(1-\gamma)^4} \right), \tag{66}$$

where the implicit constant may depend on δ.

E.1 A FINITE-PERTURBATION KL BOUND

Lemma 16 (Nominal Bernoulli KL control)**.** *Let $p_i \triangleq p_\beta(a_i)$ for $i \in \{0, 1\}$, with p_β defined in equation 12. Then*

$$\max\{\mathrm{kl}(p_0, p_1), \mathrm{kl}(p_1, p_0)\} \le \frac{8\eta^2}{h+\beta}. \tag{67}$$

Proof. We first compare the nominal Bernoulli variances along $[a_0, a_1]$. For $\beta > 0$, differentiating equation 12 gives

$$p'_\beta(a) = \frac{1 + \beta(1-2a)/\sqrt{\beta^2 + 4\beta a(1-a)}}{1+\beta}.$$

On $[a_0, a_1] \subset (0, 1/2)$, the fraction in the numerator lies in $[0, 1]$: its numerator is nonnegative and at most β, and its denominator is at least $\beta > 0$. Consequently, $0 < p'_\beta(a) \le 2/(1+\beta)$. For $\beta = 0$, $p_\beta(a) = a$ and $p'_\beta(a) = 1$, so the same bounds hold. Since $p_0 \ge a_0 = h$ and $\eta \le h/4$,

$$0 < p_1 - p_0 \le \frac{2\eta}{1+\beta} \le \frac{h}{2(1+\beta)} \le \frac{p_0}{2}. \tag{68}$$

In particular, $p_1 \le 3p_0/2$.

To control the complementary probabilities, observe that

$$(\beta + 2h)^2 - \big[\beta^2 + 4\beta h(1-h)\big] = 4h^2(1+\beta) > 0.$$

Both quantities being compared are nonnegative, so $\sqrt{\beta^2 + 4\beta h(1-h)} \le \beta + 2h$. Substituting this bound into equation 12 at $a_0 = h$ yields

$$\begin{aligned} 1 - p_0 &= \frac{2 + \beta - 2h - \sqrt{\beta^2 + 4\beta h(1-h)}}{2(1+\beta)} \\ &\ge \frac{1-2h}{1+\beta} \ge \frac{1}{2(1+\beta)}, \end{aligned}$$

where the last inequality uses $h \le 1/4$. Together with equation 68, this gives

$$\frac{p_1 - p_0}{1 - p_0} \le h \le \frac{1}{4}, \qquad 1 - p_1 \ge \frac{3}{4}(1 - p_0).$$

Set $v_0 \triangleq p_0(1-p_0) > 0$. Because $p_0 \le p_\beta(a) \le p_1$ for $a \in [a_0, a_1]$, we can bound the two variance factors separately:

$$p_\beta(a)(1 - p_\beta(a)) \le p_1(1 - p_0) \le \frac{3}{2}v_0, \qquad p_1(1 - p_1) \ge \frac{3}{4}v_0.$$

The sensitivity identity in Lemma 3 gives

$$[p'_\beta(a)]^2 = \frac{4p_\beta(a)(1 - p_\beta(a))}{\beta + 4a(1 - a)}.$$

Here $a \ge h$ and $1 - a \ge 11/16$, so $4a(1 - a) \ge 11h/4 \ge h$. Hence

$$[p'_\beta(a)]^2 \le \frac{6v_0}{h + \beta}.$$

The fundamental theorem of calculus and Cauchy–Schwarz imply

$$\begin{aligned}(p_1 - p_0)^2 &= \left(\int_{a_0}^{a_1} p'_\beta(a)\, da\right)^2 \\ &\le \eta \int_{a_0}^{a_1} [p'_\beta(a)]^2\, da \le \frac{6\eta^2 v_0}{h + \beta}.\end{aligned} \tag{69}$$

Finally, $\log x \le x - 1$ implies, for $u, v \in (0, 1)$,

$$\begin{aligned}\mathrm{kl}(u, v) &\le u\left(\frac{u}{v} - 1\right) + (1 - u)\left(\frac{1 - u}{1 - v} - 1\right) \\ &= (u - v)\left[\frac{u}{v} - \frac{1 - u}{1 - v}\right] = \frac{(u - v)^2}{v(1 - v)}.\end{aligned}$$

For $\mathrm{kl}(p_1, p_0)$, the denominator is v_0, so equation 69 gives the constant 6. For $\mathrm{kl}(p_0, p_1)$, the denominator is at least $3v_0/4$, giving the constant $6/(3/4) = 8$. This proves equation 67 in both directions. □

E.2 Proof of the adaptive-query lower bound

Proof of Theorem 5. **Step 1: replicated hard gadgets.** Let $\Theta_k \triangleq \{0, 1\}^k$, with hypotheses $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$. For each ϑ, construct a nominal MDP with state set

$$\mathcal{S}_k = \{g\} \cup \bigcup_{i=1}^{k} \{d_i, u_{i,0}, u_{i,1}\} \cup \mathcal{V}_k, \qquad |\mathcal{V}_k| = S - (3k + 1).$$

All listed states are distinct, and every state has action set $\mathcal{A} = \{0, 1, \ldots, A - 1\}$. At d_i, action 0 moves deterministically to $u_{i,0}$, and every other action moves deterministically to $u_{i,1}$. Write $t_- = p_\beta(a_0)$, $t_+ = p_\beta(a_1)$, and $b = h/2$. At $u_{i,j}$, action 0 exits to g with probability

$$t_{i,j}^{(\vartheta)} = \begin{cases} t_+, & j = \vartheta_i, \\ t_-, & j \ne \vartheta_i, \end{cases}$$

and every other action exits with probability b. The complementary probability is a self-loop. State g is absorbing with reward 1; all other rewards are zero, and each padding state is absorbing.

Use the ambiguity set centered at this nominal kernel, denoted P_ϑ. The budgets are separate across states and shared across actions within each state. Absolute continuity fixes all deterministic rows, including those at the common absorbing state g. Therefore, it will be imposed simultaneously at every evaluation state. Thus the proposition applies at each d_i: writing

$$w_{\vartheta,i}(\pi) \triangleq \begin{cases} \pi(0|d_i), & \vartheta_i = 0, \\ 1 - \pi(0|d_i), & \vartheta_i = 1, \end{cases}$$

we have, for every $\pi \in \Pi_{\mathrm{stat}}$,

$$V_\vartheta^*(d_i) - V_\vartheta^\pi(d_i) \geq 4\varepsilon\big[1 - w_{\vartheta,i}(\pi)\big]. \tag{70}$$

Here V_ϑ^π and V_ϑ^* abbreviate the robust values associated with P_ϑ.

Use $\vartheta = \mathbf{0}$ as the reference model. For each i, let e_i denote the vector whose ith coordinate is 1 and whose other coordinates are 0. The nominal kernels $P_{\mathbf{0}}$ and P_{e_i} differ only at $(u_{i,0}, 0)$ and $(u_{i,1}, 0)$.

Step 2: testing events implied by policy accuracy. Define $E_i \triangleq \{\widehat{\pi}(0|d_i) > 1/2\}$. Under the reference model, E_i^c implies $w_{\mathbf{0},i}(\widehat{\pi}) \leq 1/2$, and hence loss at least 2ε at d_i by equation 70. Under e_i, the event E_i implies $w_{e_i,i}(\widehat{\pi}) < 1/2$, and hence loss greater than 2ε there. The uniform accuracy guarantee consequently gives

$$\mathbb{P}_{\mathbf{0}}(E_i) \geq 1 - \delta, \qquad \mathbb{P}_{e_i}(E_i) \leq \delta. \tag{71}$$

Each inequality follows directly from the assumed guarantee; no union bound over gadgets is required.

Step 3: change of measure for adaptive queries. Let U contain all internal randomness of the algorithm, including any output randomization. The seed has the same distribution under every hypothesis and is independent of the oracle's sampling randomness. Conditional on U and the past observations, the query, stopping, and output rules are deterministic and identical under all hypotheses. If the algorithm stops before T, pad the transcript with queries to $(g, 0)$ and retain its original output. These added observations have the same deterministic response under every hypothesis.

Write (s_t, ℓ_t) for query t and X_t for its response. Let L_ϑ be the law of the augmented transcript containing U, all queries and responses, and $\widehat{\pi}$. Define

$$N_{i,j} \triangleq \sum_{t=1}^{T} \mathbf{1}\{(s_t, \ell_t) = (u_{i,j}, 0)\}, \qquad N_i \triangleq N_{i,0} + N_{i,1}.$$

For probability laws L, K, write $\mathrm{KL}(L|K) = \int \log(dL/dK)\, dL$ when $L \ll K$, and set it to $+\infty$ otherwise. All nominal kernels in the construction have the same rowwise supports, with $0 < t_- < t_+ < 1$. Therefore their augmented transcript laws are mutually absolutely continuous. The common seed law and deterministic query rules cancel from the likelihood ratio, giving

$$\log \frac{d\mathsf{L}_{\mathbf{0}}}{d\mathsf{L}_{e_i}} = \sum_{t=1}^{T} \log \frac{P_{\mathbf{0}}(X_t|s_t, \ell_t)}{P_{e_i}(X_t|s_t, \ell_t)} \qquad \mathsf{L}_{\mathbf{0}}\text{-almost surely.}$$

The output is a measurable function of the seed and the observations, so its inclusion introduces no additional term. Taking conditional expectations under the reference model and using the two changed rows yields

$$\begin{aligned} \mathrm{KL}(\mathsf{L}_{\mathbf{0}}|\mathsf{L}_{e_i}) &= \mathbb{E}_{\mathbf{0}}[N_{i,0}]\,\mathrm{kl}(t_+, t_-) + \mathbb{E}_{\mathbf{0}}[N_{i,1}]\,\mathrm{kl}(t_-, t_+) \\ &\leq \frac{8\eta^2}{h+\beta}\mathbb{E}_{\mathbf{0}}[N_i], \end{aligned} \tag{72}$$

where the inequality uses Lemma 16.

Applying the log-sum inequality on E_i and its complement gives

$$\mathrm{KL}(\mathsf{L}_{\mathbf{0}}|\mathsf{L}_{e_i}) \geq \mathrm{kl}\big(\mathbb{P}_{\mathbf{0}}(E_i), \mathbb{P}_{e_i}(E_i)\big).$$

For $0 < v < u < 1$,

$$\partial_u\, \mathrm{kl}(u, v) = \log \frac{u(1-v)}{v(1-u)} > 0, \qquad \partial_v\, \mathrm{kl}(u, v) = -\frac{u}{v} + \frac{1-u}{1-v} = \frac{v-u}{v(1-v)} < 0.$$

The first sign follows from $u(1-v) - v(1-u) = u - v > 0$; the second follows from $v - u < 0$ and $v(1-v) > 0$. By equation 71, $u = \mathbb{P}_{\mathbf{0}}(E_i) \geq 1 - \delta > \delta \geq \mathbb{P}_{e_i}(E_i) = v$. Decreasing u to $1 - \delta$ and increasing v to δ can only decrease the binary relative entropy. Boundary cases follow by limits. Consequently,

$$\mathrm{KL}(\mathsf{L}_{\mathbf{0}}|\mathsf{L}_{e_i}) \geq \mathrm{kl}(1-\delta, \delta).$$

Combining this with equation 72 gives

$$\mathbb{E}_{\mathbf{0}}[N_i] \geq \frac{h+\beta}{8\eta^2}\,\mathrm{kl}(1-\delta,\delta). \tag{73}$$

Step 4: summing the information requirements. The informative row pairs are disjoint across gadgets, so $\sum_{i=1}^{k} N_i \leq T$ on every transcript. All expectations in equation 73 are taken under the same reference model. Therefore

$$T \geq \sum_{i=1}^{k} \mathbb{E}_{\mathbf{0}}[N_i] \geq \frac{k(h+\beta)}{8\eta^2}\,\mathrm{kl}(1-\delta,\delta).$$

Substituting $\eta^2 = 1024\varepsilon^2 h^4$ proves equation 65. Since $S \geq 4$, we have $k \geq 1$ and $S \leq 3k+3 \leq 6k$, so $k \geq S/6$. Also,

$$\mathrm{kl}(1-\delta,\delta) = (1-2\delta)\log\frac{1-\delta}{\delta} > 0$$

for $\delta \in (0, 1/2)$. For each fixed such δ, using $\beta = 2A\rho$ now gives equation 66. □

Rate comparison. The factor S in equation 66 comes from summing the query requirements of disjoint gadgets under one reference model. The dependence on $A\rho$ comes from the shared statewise ambiguity budget. For fixed δ and $\rho \geq h$, the inequality $|\mathcal{S}||\mathcal{A}|(h+\rho) \leq 2|\mathcal{S}||\mathcal{A}|\rho$ shows that the adaptive lower bound also implies

$$T = \Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3} + \frac{|\mathcal{S}||\mathcal{A}|\rho}{\varepsilon^2(1-\gamma)^4}\right). \tag{74}$$

For general $\rho < h$, the established adaptive bound is equation 66; an additional factor A in its ordinary-RL term would require a further lower-bound argument. The balanced experiment imposes a different allocation rule: its total sample count is the common per-pair count multiplied by $|\mathcal{S}||\mathcal{A}|$. That multiplication does not apply to adaptive queries.

F PROOF OF LEMMAS

Proof of Lemma 1. First, we introduce a useful result from previous works.

Theorem 6 (Shapiro (2017))**.** *Let P be a probability measure on $\mathcal{S}$ and $\rho > 0$. Then, it holds that*

$$\begin{aligned}&\inf\left\{\mathbb{E}_P[V(s')] : \mathbb{E}_{P_0}\left[\frac{1}{2}\left(\frac{P(s')}{P_0(s')}-1\right)^2\right] \le \rho\right\}\\ &= \sup_{\lambda\ge 0,\eta\in\mathbb{R}}\left\{-\mathbb{E}_{P_0}\left[\frac{\lambda}{2}\left(\frac{\eta - V(s')}{\lambda}+1\right)^2 - \frac{\lambda}{2}\right] - \lambda\rho + \eta\right\}.\end{aligned}$$

Then, according to Theorem 6, we can get the following result,

$$\begin{aligned}&-\mathbb{E}_{P_0}\left[\frac{\lambda}{2}\left(\frac{\eta - V(s')}{\lambda}+1\right)_+^2 - \frac{\lambda}{2}\right] - \lambda\rho + \eta\\ &= -\frac{1}{2\lambda}\mathbb{E}_{P_0}\left[(\eta - V(s') + \lambda)_+^2\right] - \lambda(\rho - \frac{1}{2}) + \eta\\ &\overset{(a)}{=} -\frac{1}{2\lambda}\mathbb{E}_{P_0}\left[(\tilde{\eta} - V(s'))_+^2\right] - \lambda(\rho + \frac{1}{2}) + \tilde{\eta},\end{aligned}$$

where (a) set $\tilde{\eta} = \eta + \lambda$.

Next, since $\lambda \ge 0$, $\rho + \frac{1}{2} > 0$ and $\mathbb{E}_{P_0}\left[(\tilde{\eta} - V(s'))_+^2\right] > 0$, we can get

$$\begin{aligned}&\sup_{\lambda\ge 0,\tilde{\eta}\in\mathbb{R}}\left\{-\frac{1}{2\lambda}\mathbb{E}_{P_0}\left[(\tilde{\eta} - V(s'))_+^2\right] - \lambda(\rho + \frac{1}{2}) + \tilde{\eta}\right\}\\ &\overset{(a)}{=} \sup_{\tilde{\eta}\in\mathbb{R}}\left\{\sqrt{2\rho+1}\sqrt{\mathbb{E}_{P_0}\left[(\tilde{\eta} - V(s'))_+^2\right]} + \tilde{\eta}\right\},\end{aligned}$$

where (a) follows from that $x + y \ge 2\sqrt{xy}$ for $x, y \ge 0$.

□

Proof of Lemma 5. We first write

$$X - m\mathbb{E}[X] = X - \mathbb{E}[X] + (1-m)\mathbb{E}[X].$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[(X - m\mathbb{E}[X])^2\right] &= \mathbb{E}\left[(X - \mathbb{E}[X] + (1-m)\mathbb{E}[X])^2\right]\\ &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + 2(1-m)\mathbb{E}[X]\,\mathbb{E}\left[X - \mathbb{E}[X]\right]\\ &\quad + (1-m)^2\left(\mathbb{E}[X]\right)^2.\end{aligned}$$

Since

$$\mathbb{E}\left[X - \mathbb{E}[X]\right] = 0$$

and

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathrm{Var}(X),$$

we obtain

$$\mathbb{E}\left[(X - m\mathbb{E}[X])^2\right] = \mathrm{Var}(X) + (m-1)^2\left(\mathbb{E}[X]\right)^2.$$

□

Proof of Lemma 2. According to the dual reformulation (Fenchel duality), we can get

$$\inf_{P_s \in \mathcal{P}_s} \sum_{a \in \mathcal{A}} \pi(a|s)\, \mathbb{E}_{P_0(s'|s,a)}[V(s')]$$
$$= \sup_{\lambda \geq 0,\, \eta \in \mathbb{R}^{|\mathcal{A}|}} \left[-\lambda \sum_{a \in \mathcal{A}} \mathbb{E}_{P_{0,s,a}} \left(f^* \left(\frac{\eta_a - \pi(a|s) V(s')}{\lambda} \right) \right) - \lambda |\mathcal{A}| \rho + \sum_{a \in \mathcal{A}} \eta_a \right],$$

where λ is a Lagrange multiplier (constraint budget ρ), and η_a are dual variables.

Finally, Optimizing over λ, we can conclude that

$$\inf_{P_s \in \mathcal{P}_s} \sum_{a \in \mathcal{A}} \pi(a|s)\, \mathbb{E}_{P_0(s'|s,a)}[V(s')]$$
$$= \sup_{\eta \in \mathbb{R}^{|\mathcal{A}|}} \left\{ -\sqrt{|\mathcal{A}|(2\rho+1)} \left(\sum_{a \in \mathcal{A}} \mathbb{E}_{P_{0,s,a}} \left[(\eta_a - \pi(a|s) V(s))_+^2 \right] \right)^{1/2} + \sum_{a \in \mathcal{A}} \eta_a \right\}$$

□

Proof of Lemma 6. Fix P, $\widehat{P}$, V, and $\rho \geq 0$, and recall that

$$g(t, \eta) = \eta - \sqrt{2\rho + 1} \left(\mathbb{E}_{tP + (1-t)\widehat{P}} \big[(\eta - V)_+^2 \big] \right)^{1/2}, \qquad \eta^*(t) \in \arg\max_{\eta \in \mathbb{R}} g(t, \eta).$$

For every fixed η, the map

$$t \longmapsto \mathbb{E}_{tP + (1-t)\widehat{P}} \big[(\eta - V)_+^2 \big] = t\, \mathbb{E}_P \big[(\eta - V)_+^2 \big] + (1-t)\, \mathbb{E}_{\widehat{P}} \big[(\eta - V)_+^2 \big]$$

is nonnegative and affine. Since $x \mapsto -\sqrt{x}$ is convex on $[0, \infty)$, $t \mapsto g(t, \eta)$ is convex. Thus, for any $t_0, t_1 \in [0, 1]$, $\lambda \in [0, 1]$, and $\eta \in \mathbb{R}$,

$$g\big((1-\lambda)t_0 + \lambda t_1, \eta\big) \leq (1-\lambda) g(t_0, \eta) + \lambda g(t_1, \eta).$$

Taking the supremum over η on both sides and using optimality of the Lagrange parameters gives

$$\begin{aligned} & g\Big((1-\lambda)t_0 + \lambda t_1, \eta^*\big((1-\lambda)t_0 + \lambda t_1\big)\Big) \\ &= \sup_{\eta \in \mathbb{R}} g\big((1-\lambda)t_0 + \lambda t_1, \eta\big) \\ &\leq \sup_{\eta \in \mathbb{R}} \big\{ (1-\lambda) g(t_0, \eta) + \lambda g(t_1, \eta) \big\} \\ &\leq (1-\lambda) \sup_{\eta \in \mathbb{R}} g(t_0, \eta) + \lambda \sup_{\eta \in \mathbb{R}} g(t_1, \eta) \\ &= (1-\lambda) g(t_0, \eta^*(t_0)) + \lambda g(t_1, \eta^*(t_1)). \end{aligned}$$

Therefore, $t \mapsto g(t, \eta^*(t))$ is convex on $[0, 1]$. The proof does not require uniqueness, continuity, or differentiability of $\eta^*(t)$. □

Proof of Lemma 7. If $\eta^* \geq \frac{2}{1-\gamma}$, $\mathbb{E}_{P_{0,s,a}(t)} \left[(\eta^*(t) - V(s))_+ \right] \geq \frac{1}{1-\gamma}$.

Then, the term $0 \leq \frac{(\eta^*(t) - \tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)} [(\eta^*(t) - V(s))_+]} \leq 2$. Therefore,

$$\rho = \mathbb{E} \left[\left(\frac{(\eta^*(t) - \tilde{v})_+}{\mathbb{E}_{P_{0,s,a}(t)} \left[(\eta^*(t) - V(s))_+ \right]} - 1 \right)^2 \right] \leq 1.$$

□

Proof of Lemma 25. First suppose that $X_1, \ldots, X_N$ are mutually independent. Then all NT variables $X_{n,t}$ are independent, and

$$T \left(\widehat{H}_T - \sum_{n=1}^{N} u_n \right) = \sum_{t=1}^{T} \sum_{n=1}^{N} (X_{n,t} - u_n).$$

Because $u_n \in [0, a]$, each centered summand satisfies

$$\mathbb{E}[X_{n,t} - u_n] = 0, \qquad |X_{n,t} - u_n| \le a, \qquad \mathbb{E}[(X_{n,t} - u_n)^2] = \sigma_n.$$

The sum of their variances is therefore $T \sum_{n=1}^{N} \sigma_n$. Applying the two-sided Bernstein inequality at threshold $T\varepsilon$ gives

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{H}_T - \sum_{n=1}^{N} u_n\right| \ge \varepsilon\right) \\
&\quad = \mathbb{P}\left(\left|\sum_{t=1}^{T}\sum_{n=1}^{N}(X_{n,t} - u_n)\right| \ge T\varepsilon\right) \\
&\quad \le 2\exp\left(-\frac{T^2\varepsilon^2}{2T\sum_{n=1}^{N}\sigma_n + \frac{2}{3}aT\varepsilon}\right) \\
&\quad = 2\exp\left(-\frac{T\varepsilon^2}{2\sum_{n=1}^{N}\sigma_n + \frac{2}{3}a\varepsilon}\right).
\end{aligned}$$

Taking the complement proves equation 22. The range parameter is a, rather than Na, because Bernstein's inequality is applied to the individual component observations.

To obtain the confidence bound, solving the quadratic inequality that makes the last exponential bound at most δ gives the threshold

$$\frac{a\log(2/\delta)}{3T} + \sqrt{\frac{2\log(2/\delta)}{T}\sum_{n=1}^{N}\sigma_n + \frac{a^2\log^2(2/\delta)}{9T^2}}.$$

Using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$ bounds this threshold by the right-hand side of equation 23. Rearranging

$$\frac{T\varepsilon^2}{2\sum_{n=1}^{N}\sigma_n + \frac{2}{3}a\varepsilon} \ge \log\frac{2}{\delta}$$

gives equation 24.

Now allow dependence among the components. The T sums $\sum_{n=1}^{N} X_{n,t}$ remain independent copies of H, and $0 \le H \le Na$. Their common variance satisfies

$$\begin{aligned}
\mathrm{Var}(H) &= \sum_{n=1}^{N}\sigma_n + 2\sum_{1\le n<m\le N}\mathrm{Cov}(X_n, X_m) \\
&\le \sum_{n=1}^{N}\sigma_n + 2\sum_{1\le n<m\le N}\sqrt{\sigma_n\sigma_m} \\
&= \left(\sum_{n=1}^{N}\sqrt{\sigma_n}\right)^2,
\end{aligned}$$

where the inequality follows from Cauchy–Schwarz. Bernstein's inequality applied to these T sums therefore gives

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{H}_T - \sum_{n=1}^{N} u_n\right| \ge \varepsilon\right) \\
&\quad \le 2\exp\left(-\frac{T\varepsilon^2}{2\,\mathrm{Var}(H) + \frac{2}{3}Na\varepsilon}\right) \\
&\quad \le 2\exp\left(-\frac{T\varepsilon^2}{2\left(\sum_{n=1}^{N}\sqrt{\sigma_n}\right)^2 + \frac{2}{3}Na\varepsilon}\right).
\end{aligned}$$

Taking the complement proves equation 25.

If $\sigma_n = 0$ for every n, then $X_n = u_n$ almost surely, so $\widehat{H}_T = \sum_{n=1}^{N} u_n$ almost surely. All the stated conclusions also hold in this case. □

Proof. If $V = 0$, every X_n is constant almost surely, so $\widehat{V}_T = 0$ almost surely and both claims follow. Assume henceforth that $V > 0$, and put $m = \lfloor T/2 \rfloor$.

Step 1: represent the estimator by disjoint pairs. For each n, expanding the squared differences gives

$$\begin{aligned}\sum_{1\le i<j\le T}(X_{n,i}-X_{n,j})^2 &= T\sum_{t=1}^{T}X_{n,t}^2-\left(\sum_{t=1}^{T}X_{n,t}\right)^2\\ &= T\sum_{t=1}^{T}(X_{n,t}-\overline{X}_{n,T})^2.\end{aligned}$$

Therefore,

$$\widehat{V}_T=\frac{1}{T(T-1)}\sum_{n=1}^{N}\sum_{1\le i<j\le T}(X_{n,i}-X_{n,j})^2.$$

Let $\mathfrak{S}_T$ denote the set of permutations of $\{1,\dots,T\}$. For each $\tau\in\mathfrak{S}_T$, define

$$Z_\tau \triangleq \frac{1}{2m}\sum_{n=1}^{N}\sum_{k=1}^{m}(X_{n,\tau(2k-1)}-X_{n,\tau(2k)})^2.$$

For each fixed k, averaging over τ makes $(\tau(2k-1),\tau(2k))$ uniform over the $T(T-1)$ ordered pairs of distinct indices. Hence

$$\begin{aligned}\frac{1}{T!}\sum_{\tau\in\mathfrak{S}_T}Z_\tau &= \frac{1}{2m}\sum_{n=1}^{N}\sum_{k=1}^{m}\frac{2}{T(T-1)}\sum_{1\le i<j\le T}(X_{n,i}-X_{n,j})^2\\ &= \widehat{V}_T.\end{aligned}$$

This identity holds for both even and odd T.

Step 2: bound the exponential moment using the range. Fix $\tau\in\mathfrak{S}_T$ and define

$$Y_{n,k}\triangleq\frac{1}{2}(X_{n,\tau(2k-1)}-X_{n,\tau(2k)})^2.$$

The variables $Y_{n,k}$ are mutually independent because their underlying observations are disjoint and all component samples are independent. Moreover,

$$0\le Y_{n,k}\le\frac{a^2}{2},\qquad \mathbb{E}[Y_{n,k}]=\frac{1}{2}\left(2\mathbb{E}[X_n^2]-2u_n^2\right)=\sigma_n.$$

For every $\lambda\in\mathbb{R}$ and $y\in[0,a^2/2]$, convexity of the exponential function gives

$$e^{\lambda y/m}\le 1+\frac{2y}{a^2}\left(e^{\lambda a^2/(2m)}-1\right).$$

Taking expectations and using $1+z\le e^z$ yields

$$\mathbb{E}[e^{\lambda Y_{n,k}/m}]\le\exp\left\{\frac{2\sigma_n}{a^2}\left(e^{\lambda a^2/(2m)}-1\right)\right\}.$$

Since $Z_\tau = m^{-1}\sum_{n,k}Y_{n,k}$, independence implies

$$\begin{aligned}\mathbb{E}[e^{\lambda Z_\tau}] &= \prod_{n=1}^{N}\prod_{k=1}^{m}\mathbb{E}[e^{\lambda Y_{n,k}/m}]\\ &\le\exp\left\{\frac{2mV}{a^2}\left(e^{\lambda a^2/(2m)}-1\right)\right\}.\end{aligned}$$

Applying Jensen's inequality to the permutation average from Step 1 gives, for every $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[e^{\lambda \widehat{V}_T}] &\le \frac{1}{T!} \sum_{\tau \in \mathfrak{S}_T} \mathbb{E}[e^{\lambda Z_\tau}] \\ &\le \exp\left\{ \frac{2mV}{a^2} \left(e^{\lambda a^2/(2m)} - 1 \right) \right\}.\end{aligned} \tag{75}$$

Step 3: apply the exponential Markov inequality. For $x > V$ and $\lambda > 0$, equation 75 implies

$$\mathbb{P}(\widehat{V}_T \ge x) \le \exp\left\{ -\lambda x + \frac{2mV}{a^2} \left(e^{\lambda a^2/(2m)} - 1 \right) \right\}.$$

Choosing $\lambda = (2m/a^2)\log(x/V) > 0$ gives

$$\mathbb{P}(\widehat{V}_T \ge x) \le \exp\left\{ -\frac{2m}{a^2} \left(x \log \frac{x}{V} - x + V \right) \right\}, \qquad x > V. \tag{76}$$

For $0 < x < V$, choose $\lambda = (2m/a^2)\log(x/V) < 0$. On the event $\{\widehat{V}_T \le x\}$, this sign gives $e^{\lambda \widehat{V}_T} \ge e^{\lambda x}$. The same Markov argument therefore yields

$$\mathbb{P}(\widehat{V}_T \le x) \le \exp\left\{ -\frac{2m}{a^2} \left(x \log \frac{x}{V} - x + V \right) \right\}, \qquad 0 < x < V. \tag{77}$$

At $x = 0$, letting $\lambda \to -\infty$ in equation 75 gives

$$\mathbb{P}(\widehat{V}_T = 0) \le e^{-2mV/a^2}.$$

Step 4: convert to square-root deviations. For every $x \ge 0$,

$$x \log \frac{x}{V} - x + V \ge (\sqrt{x} - \sqrt{V})^2, \tag{78}$$

where $0 \log 0 = 0$. To verify this for $x > 0$, put $y = \sqrt{x/V}$. The difference between the two sides equals

$$2Vy\,(y \log y - y + 1).$$

The derivative of $y \log y - y + 1$ is $\log y$, which is negative on $(0, 1)$ and positive on $(1, \infty)$. Its minimum is therefore its value 0 at $y = 1$, proving nonnegativity. For $x = 0$, both sides of equation 78 equal V.

Taking $x = (\sqrt{V} + \varepsilon)^2$ in equation 76 now gives

$$\mathbb{P}\left(\sqrt{\widehat{V}_T} - \sqrt{V} \ge \varepsilon \right) \le e^{-2m\varepsilon^2/a^2}.$$

For $0 < \varepsilon < \sqrt{V}$, taking $x = (\sqrt{V} - \varepsilon)^2$ in equation 77 gives

$$\mathbb{P}\left(\sqrt{V} - \sqrt{\widehat{V}_T} \ge \varepsilon \right) \le e^{-2m\varepsilon^2/a^2}.$$

For $\varepsilon = \sqrt{V}$, the same inequality follows from the bound at $x = 0$. For $\varepsilon > \sqrt{V}$, the lower-tail event is empty.

A union bound on the two tails yields

$$\mathbb{P}\left(\left| \sqrt{\widehat{V}_T} - \sqrt{V} \right| \ge \varepsilon \right) \le 2e^{-2m\varepsilon^2/a^2}.$$

Taking the complement proves equation 26. Finally, choose $\varepsilon = a\sqrt{\log(2/\delta)/(2m)}$ and use $2m \ge T - 1$ to obtain equation 27. □

Empirical variances normalized by $1/T$. Define

$$\widehat{V}_T^{\mathrm{emp}} \triangleq \sum_{n=1}^{N} \frac{1}{T} \sum_{t=1}^{T} (X_{n,t} - \overline{X}_{n,T})^2.$$

Then $\widehat{V}_T^{\mathrm{emp}} = (T-1)\widehat{V}_T/T$. Consequently,

$$\left| \sqrt{\widehat{V}_T^{\mathrm{emp}}} - \sqrt{V} \right| \le \sqrt{\frac{T-1}{T}} \left| \sqrt{\widehat{V}_T} - \sqrt{V} \right| + \left(1 - \sqrt{\frac{T-1}{T}} \right) \sqrt{V}.$$

Thus, for $\varepsilon > (1 - \sqrt{(T-1)/T})\sqrt{V}$,

$$\begin{aligned} &\mathbb{P}\left(\left| \sqrt{\widehat{V}_T^{\mathrm{emp}}} - \sqrt{V} \right| < \varepsilon \right) \\ &\quad \ge 1 - 2\exp\left\{ -\frac{2\lfloor T/2 \rfloor T}{a^2(T-1)} \left[\varepsilon - \left(1 - \sqrt{\frac{T-1}{T}} \right) \sqrt{V} \right]^2 \right\}. \end{aligned} \tag{79}$$

In particular, with probability at least $1 - \delta$,

$$\begin{aligned} \left| \sqrt{\widehat{V}_T^{\mathrm{emp}}} - \sqrt{V} \right| &\le a \sqrt{\frac{(T-1)\log(2/\delta)}{2T\lfloor T/2 \rfloor}} + \left(1 - \sqrt{\frac{T-1}{T}} \right) \sqrt{V} \\ &\le a \sqrt{\frac{\log(2/\delta)}{T}} + \frac{a\sqrt{N}}{2T}. \end{aligned} \tag{80}$$

The last inequality follows from $2\lfloor T/2 \rfloor \ge T - 1$, $1 - \sqrt{1 - 1/T} \le 1/T$, and

$$V \le \frac{Na^2}{4}, \qquad \mathrm{Var}(X_n) \le \mathbb{E}[(X_n - a/2)^2] \le \frac{a^2}{4}.$$