ROBUST-POWER-HOOT: MONTE CARLO TREE SEARCH FOR CONTINUOUS STOCHASTIC MDPS UNDER MODEL AMBIGUITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study Monte Carlo Tree Search (MCTS) for continuous-action stochastic MDPs under model ambiguity, where the model used for planning may differ from the execution environment. Existing methods typically address at most two of three challenges: continuous-action exploration, stochastic value aggregation, and distributional robustness. We propose ROBUST-POWER-HOOT, which combines HOO-style continuous-action search, polynomial exploration bonuses, power-mean backups, and conservative distributionally robust Bellman updates. Exact robust optimization at every backup is costly in large search trees; instead, we use closed-form upper bounds on robust penalties to compute conservative lower estimates of robust values. We prove that the root estimate converges at rate $O(n^{-\zeta})$, where $\zeta \in (0, 1/2)$, to the projected finite-horizon value induced by these lower-bound updates. We decompose the gap to the exact robust objective into finite-horizon truncation, state abstraction, lower-bound approximation, and finite-sample errors, and identify regimes in which the penalty bounds are tight. Across seven continuous-control benchmarks with controlled planning–execution mismatch, the method often degrades less as execution noise increases.

1 INTRODUCTION

Monte Carlo Tree Search (MCTS) is a standard tool for online planning in complex sequential decision problems, from game playing (Silver et al., 2016; Schrittwieser et al., 2020) to robotics and control (Kartal et al., 2019; Dam et al., 2022). It selects actions under a limited simulation budget using a simulator or learned model, which may differ from the environment in which the chosen action is executed.

This paper studies H-step MCTS planning in continuous-action stochastic MDPs under model ambiguity. The planner has access to a nominal one-step model, but the execution environment may differ from this nominal model in its reward or transition laws. Three practical challenges arise simultaneously. First, **continuous actions** are unavoidable in many control problems, such as torque control, thrust allocation, and steering; naive discretization is often brittle and scales poorly with dimension. Second, **rewards and transitions are stochastic**, so rollout values are noisy and hard max backups can produce unstable estimates. Third, **model misspecification** can dominate the planning error: an action that looks best in the nominal simulator may perform poorly when the dynamics or rewards shift at deployment.

Existing MCTS methods address parts of this problem, but not the full combination. HOO/HOOT and POLY-HOOT adaptively search continuous actions (Bubeck et al., 2011; Mansley et al., 2011; Mao et al., 2020). Stochastic-Power-HOOT adds power-mean backups for stochastic continuous-action search (Dam, 2025), but does not model planning–execution mismatch. Robust-Power-UCT uses distributionally robust Bellman updates under model ambiguity (Dam et al., 2025), but is restricted to discrete actions. A gap therefore remains: continuous-action MCTS that is statistically stable under stochastic rollouts and explicitly plans for conservative worst-case performance.

Our approach. ROBUST-POWER-HOOT combines HOO-style action search, polynomial exploration bonuses, power-mean backups, and conservative robust Bellman updates. Exact robust backups

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.

require

$$\inf_{P \in \mathcal{U}_D(Q,\rho)} \mathbb{E}_P[Z] = \mathbb{E}_Q[Z] - \Delta_{D,\rho}(Q; Z),$$

where $\Delta_{D,\rho}$ is the robust penalty. Instead of computing this penalty at every edge update, we use closed-form upper bounds on $\Delta_{D,\rho}$ and form conservative lower estimates of robust expectations. HOO treats these lower-bound robust values as payoffs, while exploration remains governed by the standard polynomial statistical bonus and geometric optimism.

Scope of the guarantee. The finite-sample theorem targets the quantity estimated by the algorithm: the projected finite-horizon value induced by the lower-bound Bellman operator. It is not a direct convergence theorem to the exact infinite-horizon robust objective. We use the decomposition

$$V^{\star}_{\mathrm{rob}}(s_0) \to V^{\mathrm{rob}}_0(s_0) \to V^{\mathrm{rob},\psi}_0(s_0) \to V^{\mathrm{ub},\psi}_0(s_0) \to \mathbb{E}[\widehat{V}_n(v_0)],$$

corresponding to truncation, projection, lower-bound approximation, and statistical estimation. The main theorem controls the last step; the other gaps are bounded separately.

Contributions.

- **Algorithm.** We introduce ROBUST-POWER-HOOT, combining HOO-style continuous-action search, polynomial exploration, power-mean backups, and conservative robust Bellman updates.
- **Efficient robust backups.** For TV, χ^2, and W_1 ambiguity sets, we derive closed-form penalty upper bounds that yield conservative lower values, quantify their approximation error, and identify tightness regimes.
- **Finite-sample theory.** We prove root convergence to the projected finite-horizon lower-bound target $V^{\mathrm{ub},\psi}_0$ at rate $O(n^{-\zeta})$, with $\zeta = \alpha_0/\beta_0 \in (0, 1/2)$. Empirical plug-in penalties are handled by separate concentration results and enter through the effective exponents.
- **Empirical evaluation and exact-backup comparison.** On seven controlled-mismatch benchmarks, the robust variants often degrade less as execution noise increases, with task- and radius-dependent gains. We also compare the closed-form robust updates with exact inner optimization, both at the backup level and, for W_1, in separate planner runs (tables H.7 and H.8).

2 RELATED WORK

Robust and distributionally robust MDPs. Robust MDPs model uncertainty through ambiguity sets over rewards or transitions and admit dynamic programming under rectangularity (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). Distributionally robust MDPs define these sets around nominal models using divergences or transport distances (Xu & Mannor, 2012; Rahimian & Mehrotra, 2022; Mohajerin Esfahani & Kuhn, 2018; Ho et al., 2022). Related RL work studies robust learning with generative models, distributionally robust Q-learning, and model mismatch (Panaganti & Kalathil, 2022; Liu et al., 2022; Roy et al., 2017; Pinto et al., 2017). Our focus is online lookahead planning, where robust updates must be computed many times inside an MCTS tree.

Robust and risk-aware tree search. Bayes-adaptive MCTS plans in a belief-augmented state space (Guez et al., 2013), while robust and risk-aware tree-search methods use adversarial, tail-risk, or epistemic-uncertainty objectives (Sharma et al., 2019; Hayes et al., 2021; Oren et al., 2025; Zhang et al., 2025). Most closely related is Robust-Power-UCT, which uses exact distributionally robust Bellman backups for reward and transition ambiguity (Dam et al., 2025). It is designed for discrete actions. ROBUST-POWER-HOOT instead handles continuous actions and replaces exact per-backup robust optimization with closed-form upper bounds on robust penalties.

Continuous-action MCTS and optimistic partitioning. Continuous-action MCTS is commonly handled by progressive widening, continuous UCT, function approximation, or optimistic partitioning (Couëtoux et al., 2011; Auger et al., 2013; Yee et al., 2016; Lee et al., 2020; Kim et al., 2020). The optimistic-partitioning family includes HOO, HOOT, and POLY-HOOT (Bubeck et al., 2011; Mansley et al., 2011; Mao et al., 2020). ROBUST-POWER-HOOT builds on this framework, but passes conservative lower estimates of robust edge values to HOO rather than nominal return estimates.

Backups under stochasticity. In stochastic MDPs, hard max backups can amplify rollout noise. Generalized-mean and power-mean backups reduce this instability and have finite-sample analyses in stochastic MCTS (Dam et al., 2020; Dam, 2025; Dam et al., 2024). We combine power-mean aggregation with conservative lower bounds on robust expectations. With fixed penalty bounds, the robustness step changes the payoff estimates and constants, while the HOO/power-mean exponent recursion stays the same. Empirical plug-in bounds require separate concentration control; if their tails are slower, they enter the final rate through the effective exponents.

3 Problem and Robust Planning Target

3.1 Robust continuous stochastic MDPs

We consider a robust stochastic MDP with continuous actions in the infinite-horizon discounted setting, where model ambiguity is specified locally at each state-action pair. The state space $(S, \mathcal{B}(S))$ is Borel with metric d and finite diameter, the action space $A \subseteq [0,1]^m$ is nonempty and compact, rewards lie in $[0, R_{\max}]$, and $\gamma \in (0,1)$.

Definition 1 (Robust continuous stochastic MDP)**.** *A robust continuous stochastic MDP is a tuple*

$$\mathcal{M} = (S, A, \{\mathcal{P}_{s,a}\}, \{\mathcal{R}_{s,a}\}, \gamma),$$

where, for each $(s,a) \in S \times A, \mathcal{P}_{s,a} \subseteq \mathcal{M}_1(S), \mathcal{R}_{s,a} \subseteq \mathcal{M}_1([0, R_{\max}])$ are nonempty transition and reward ambiguity sets, respectively.

Here $\mathcal{M}_1(K)$ denotes Borel probability measures on K. The execution laws satisfy $P(\cdot \mid s,a) \in \mathcal{P}_{s,a}$ and $\nu(\cdot \mid s,a) \in \mathcal{R}_{s,a}$. The planner samples a nominal one-step model $(P^o_{s,a}, \nu^o_{s,a})$: a query at (s,a) returns $(R, S') \sim (\nu^o_{s,a}, P^o_{s,a})$. Following the distributionally robust optimization framework (Rahimian & Mehrotra, 2022), we define local ambiguity sets by divergence constraints around the nominal model. Given a divergence D and radii $\rho_P, \rho_R > 0$, we set

$$\mathcal{P}_{s,a} = \{P \in \mathcal{M}_1(S) : D(P \| P^o_{s,a}) \le \rho_P\}, \tag{3.1}$$

$$\mathcal{R}_{s,a} = \{\nu \in \mathcal{M}_1([0, R_{\max}]) : D(\nu \| \nu^o_{s,a}) \le \rho_R\}. \tag{3.2}$$

We focus on three standard choices: total variation (TV), chi-square (χ^2), and Wasserstein-1 (W_1).

3.2 Robust Bellman operators and exact robust value

For a stationary policy π, define

$$V^\pi_{\mathrm{rob}}(s) := \inf_{(\nu,P)\in\mathfrak{R}\times\mathfrak{P}} \mathbb{E}^\pi_{\nu,P}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, S_0 = s\right], \qquad V^\star_{\mathrm{rob}}(s) := \sup_\pi V^\pi_{\mathrm{rob}}(s),$$

where $\mathfrak{P}$ and $\mathfrak{R}$ are rectangular families of transition and reward laws (Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). Rectangularity allows the adversary to choose $P(\cdot \mid s,a) \in \mathcal{P}_{s,a}$ and $\nu(\cdot \mid s,a) \in \mathcal{R}_{s,a}$ locally at each state-action pair.

For any bounded measurable $V : S \to \mathbb{R}$, the robust Bellman backup is

$$r^{\mathrm{rob}}(s,a) := \inf_{\nu\in\mathcal{R}_{s,a}} \mathbb{E}_\nu[R],$$

$$(\mathcal{T}^{\mathrm{rob}}_Q V)(s,a) := r^{\mathrm{rob}}(s,a) + \gamma \inf_{P\in\mathcal{P}_{s,a}} \mathbb{E}_P[V(S')], \tag{3.3}$$

$$(\mathcal{T}^{\mathrm{rob}}_V V)(s) := \sup_{a\in A} (\mathcal{T}^{\mathrm{rob}}_Q V)(s,a). \tag{3.4}$$

Under standard measurability and compactness assumptions, $\mathcal{T}^{\mathrm{rob}}_V$ is a γ-contraction with fixed point $V^\star_{\mathrm{rob}}$; see Appendix A.1.

For planning depth H and rollout baseline V_{π_0}, define

$$V^{\mathrm{rob}}_H := V_{\pi_0}, \qquad V^{\mathrm{rob}}_h := \mathcal{T}^{\mathrm{rob}}_V V^{\mathrm{rob}}_{h+1}, \quad 0 \le h < H. \tag{3.5}$$

Then $\left\|V^\star_{\mathrm{rob}} - V^{\mathrm{rob}}_0\right\|_\infty \le \gamma^H \left\|V^\star_{\mathrm{rob}} - V_{\pi_0}\right\|_\infty$.

3.3 PROJECTED BIN-LEVEL PLANNING TARGET

In stochastic continuous-state MDPs, exact successor states may rarely recur, causing an exact-state tree to fragment visits across many children and undermining the repeated sampling needed for recursive value estimation. POLY-HOOT avoids this issue through deterministic transitions, while Stochastic-Power-HOOT assumes finitely many possible successors (Mao et al., 2020; Dam, 2025). We therefore aggregate successors into depth-dependent finite state cells, following prior MCTS state-abstraction approaches (Hostetler et al., 2014; Sokota et al., 2021). This abstraction affects only states: each abstract node still runs HOO over the original continuous action space A, while the cell diameters ε_h quantify the resulting projection error.

Definition 2 (Depth-dependent abstraction)**.** *For each depth $h \in \{0, \dots, H\}$, let $\psi_h : S \to \mathcal{C}_h$ be measurable with finite codebook $\mathcal{C}_h$. Define*

$$\varepsilon_h := \sup\{d(s, s') : \psi_h(s) = \psi_h(s')\}, \qquad \varepsilon := \max_{0 \le h \le H} \varepsilon_h.$$

A tree node is $v = (h, c)$, $c \in \mathcal{C}_h$.

For each cell choose a fixed representative $s_{h,c} \in \psi_h^{-1}(c)$ and set $(\Pi_h V)(s) := V(s_{h,\psi_h(s)})$. The finite-sample theorem is stated for this fixed-representative process, whose nominal successor-bin law is

$$P^{o,\psi}_{h,c,a}(c') := P^{o}_{s_{h,c},a}(\psi_{h+1}(S') = c'), \qquad c' \in \mathcal{C}_{h+1}.$$

Thus concentration is over finite successor-bin alphabets and the target is non-random. In theorem mode, Algorithm 1 uses $s^{\mathrm{rep}}(v) = s_{h,c}$. The practical first-visit variant freezes the first state reaching a new node, so there is no continuing within-node drift and each node–action law is stationary conditional on the creation history. The representatives and target are nevertheless data-dependent; the theorem therefore applies directly only to fixed representatives. Appendix A.3 gives corresponding bounds for covering variants with controlled cell or merge radii; these radii were not recorded in the present logs. Define

$$V_H^{\mathrm{rob},\psi} := \Pi_H V_{\pi_0}, \qquad V_h^{\mathrm{rob},\psi} := \Pi_h \mathcal{T}_V^{\mathrm{rob}} V_{h+1}^{\mathrm{rob},\psi}, \quad 0 \le h \le H-1. \tag{3.6}$$

For $v = (h, c)$, write $\widetilde{V}^{\mathrm{rob},\psi}(v) := V_h^{\mathrm{rob},\psi}(s_{h,c})$.

Proposition 1 (Projection bias)**.** *If V_h^{rob} is L_h-Lipschitz for each h, then*

$$\left\| V_0^{\mathrm{rob},\psi} - V_0^{\mathrm{rob}} \right\|_\infty \le \sum_{h=0}^{H} \gamma^h L_h \varepsilon_h \le \frac{\max_h L_h}{1-\gamma} \varepsilon.$$

4 ROBUST-POWER-HOOT

This section describes ROBUST-POWER-HOOT. The key idea is to replace the inner robust optimization in exact robust Bellman backups by conservative lower bounds, and then use these lower bounds as payoffs inside a continuous-action HOO search tree.

4.1 CONSERVATIVE ROBUST BACKUPS

The exact robust Bellman backup requires the inner worst-case expectation $\inf_{P \in \mathcal{U}_D(Q,\rho)} \mathbb{E}_P[Z]$, which exact robust MCTS methods (Dam et al., 2025) compute by solving a robust optimization problem during backup. This is costly in continuous-action trees, where many edges are updated repeatedly. We instead replace this inner problem by a conservative lower bound.

Let $Z \ge 0$ be bounded and let Q be the nominal law. Define the exact robust expectation and the corresponding robust penalty by

$$\sigma_{D,\rho}(Q; Z) := \inf_{P \in \mathcal{U}_D(Q,\rho)} \mathbb{E}_P[Z], \tag{4.1}$$

$$\Delta_{D,\rho}(Q; Z) := \mathbb{E}_Q[Z] - \sigma_{D,\rho}(Q; Z). \tag{4.2}$$

Here $\Delta_{D,\rho}(Q;Z)$ is the decrease from the nominal expectation to the worst-case expectation. Suppose that we have a computable upper bound $\overline{\Delta}_{D,\rho}(Q;Z) \geq \Delta_{D,\rho}(Q;Z)$. We then use the conservative lower bound

$$\underline{\sigma}_{D,\rho}(Q;Z) := \left[\mathbb{E}_Q[Z] - \overline{\Delta}_{D,\rho}(Q;Z)\right]_+. \tag{4.3}$$

Equivalently, this subtracts the truncated penalty upper bound $\Delta^{\mathrm{tr}}_{D,\rho}(Q;Z) := \mathbb{E}_Q[Z] - \underline{\sigma}_{D,\rho}(Q;Z) = \overline{\Delta}_{D,\rho}(Q;Z) \wedge \mathbb{E}_Q[Z]$. The positive part is needed because Z is nonnegative: if the penalty upper bound exceeds the nominal mean, the strongest nonnegative lower bound implied by this information is zero.

Proposition 2 (Conservative lower bound)**.** *If $Z \geq 0$ and $\overline{\Delta}_{D,\rho}(Q;Z) \geq \Delta_{D,\rho}(Q;Z)$, then*

$$0 \leq \underline{\sigma}_{D,\rho}(Q;Z) \leq \sigma_{D,\rho}(Q;Z).$$

Thus $\underline{\sigma}_{D,\rho}$ is a conservative lower bound on the exact robust expectation.

4.2 Closed-form upper bounds on robust penalties

We distinguish the population penalty bound that defines the planning target from the quantity computed by the implementation. The lower-bound Bellman operator uses a population upper bound $\overline{\Delta}_{D,\rho}(Q;Z)$ on the robust penalty. In deterministic mode, this bound is a fixed conservative envelope and introduces no sampling error. In plug-in mode, the implementation uses an empirical estimate $\widehat{\overline{\Delta}}_{D,\rho,t}$ from edge statistics; the finite-sample theorem applies only when this estimate concentrates around the corresponding population bound (finite-support for abstract transitions). The plug-in conditions, including reward-side conditions, are stated in Appendix B, and the basic inequalities are proved in Appendix A.2.

For a bounded random variable $0 \leq Z \leq M_Z$ under a nominal law Q, the deterministic bounds are

$$\overline{\Delta}^{\mathrm{det}}_{\mathrm{TV},\rho}(Q;Z) := \rho\,\Omega_Z, \qquad \overline{\Delta}^{\mathrm{det}}_{\chi^2,\rho}(Q;Z) := \sqrt{\rho}\,S_Z, \qquad \overline{\Delta}^{\mathrm{det}}_{W_1,\rho}(Q;Z) := \rho\,L_Z, \tag{4.4}$$

where $\mathrm{osc}(Z) := \sup Z - \inf Z$ on the relevant support, $\sigma_Q(Z)$ is the standard deviation under Q, and $\mathrm{Lip}(Z)$ is the Lipschitz seminorm; $\Omega_Z \geq \mathrm{osc}(Z)$, $S_Z \geq \sigma_Q(Z)$, and $L_Z \geq \mathrm{Lip}(Z)$ are valid envelopes. If the population quantities are available, one may take $\Omega_Z = \mathrm{osc}(Z)$ and $S_Z = \sigma_Q(Z)$; otherwise, the boundedness envelopes $\Omega_Z = M_Z$ and $S_Z = M_Z/2$ are valid.

For an empirical finite-support law $\widehat{Q}_t = t^{-1}\sum_{i=1}^t \delta_{x_i}$, the implementation uses

$$\widehat{\overline{\Delta}}_{\mathrm{TV},\rho,t}(Z) := \rho\,\mathrm{osc}_t(Z), \qquad \widehat{\overline{\Delta}}_{\chi^2,\rho,t}(Z) := \sqrt{\rho}\,\widehat{\sigma}_t(Z), \qquad \widehat{\overline{\Delta}}_{W_1,\rho,t}(Z) := \rho\,\widehat{L}_t(Z). \tag{4.5}$$

These are valid upper bounds for the corresponding empirical finite-support ambiguity sets. When used to estimate population bounds, they require the support-recovery or boundary-mass concentration conditions in Appendix B. For population W_1 ambiguity on the original metric space, we instead use a deterministic global Lipschitz envelope, not an empirical support seminorm.

4.3 Lower-bound robust Bellman operator

For each state-action pair, let $\overline{\Delta}_R(s,a)$ and $\overline{\Delta}_P(s,a;V)$ be upper bounds on the reward and transition robust penalties. Define

$$\ell^{\mathrm{ub}}_R(s,a) := \left[\mathbb{E}_{\nu^o_{s,a}}[R] - \overline{\Delta}_R(s,a)\right]_+, \tag{4.6}$$

$$\ell^{\mathrm{ub}}_P(s,a;V) := \left[\mathbb{E}_{P^o_{s,a}}[V(S')] - \overline{\Delta}_P(s,a;V)\right]_+, \qquad 0 \leq V \leq V_{\max}. \tag{4.7}$$

By Proposition 2, these are conservative lower bounds on the exact robust reward and transition expectations. They induce the lower-bound Bellman operators

$$(\mathcal{T}^{\mathrm{ub}}_Q V)(s,a) := \ell^{\mathrm{ub}}_R(s,a) + \gamma \ell^{\mathrm{ub}}_P(s,a;V), \tag{4.8}$$

$$(\mathcal{T}^{\mathrm{ub}}_V V)(s) := \sup_{a\in A}(\mathcal{T}^{\mathrm{ub}}_Q V)(s,a). \tag{4.9}$$

Thus $\mathcal{T}^{\mathrm{ub}}_Q$ and $\mathcal{T}^{\mathrm{ub}}_V$ are conservative surrogates of the exact robust Bellman operators.

The induced finite-horizon recursion is

$$V_H^{\mathrm{ub}} := V_{\pi_0}, \qquad V_h^{\mathrm{ub}} := \mathcal{T}_V^{\mathrm{ub}} V_{h+1}^{\mathrm{ub}}, \quad 0 \le h \le H-1, \tag{4.10}$$

and its projected counterpart is

$$V_H^{\mathrm{ub},\psi} := \Pi_H V_{\pi_0}, \qquad V_h^{\mathrm{ub},\psi} := \Pi_h \mathcal{T}_V^{\mathrm{ub}} V_{h+1}^{\mathrm{ub},\psi}, \quad 0 \le h \le H-1. \tag{4.11}$$

For $v = (h, c)$, set $\widetilde{V}^{\mathrm{ub},\psi}(v) := V_h^{\mathrm{ub},\psi}(s_{h,c}), \widetilde{Q}^{\mathrm{ub},\psi}(v, a) := (\mathcal{T}_Q^{\mathrm{ub}} V_{h+1}^{\mathrm{ub},\psi})(s_{h,c}, a)$.

Proposition 3 (One-sided conservatism)**.** *For every* $0 \le V \le V_{\max}$*,*

$$(\mathcal{T}_Q^{\mathrm{ub}} V)(s,a) \le (\mathcal{T}_Q^{\mathrm{rob}} V)(s,a), \qquad (\mathcal{T}_V^{\mathrm{ub}} V)(s) \le (\mathcal{T}_V^{\mathrm{rob}} V)(s).$$

Proposition 4 (Lower-bound approximation error)**.** *Let*

$$\eta_h^\psi := \sup_{c \in \mathcal{C}_h,\, a \in A} \left[(\mathcal{T}_Q^{\mathrm{rob}} - \mathcal{T}_Q^{\mathrm{ub}}) V_{h+1}^{\mathrm{ub},\psi} \right](s_{h,c}, a).$$

Then $\left\| V_0^{\mathrm{rob},\psi} - V_0^{\mathrm{ub},\psi} \right\|_\infty \le \sum_{h=0}^{H-1} \gamma^h \eta_h^\psi$*. Moreover,*

$$0 \le \eta_h^\psi \le \bar{\eta}_h^\psi := \sup_{c \in \mathcal{C}_h,\, a \in A} \left\{ \Delta_R^{\mathrm{tr}}(s_{h,c}, a) + \gamma \Delta_P^{\mathrm{tr}}(s_{h,c}, a; V_{h+1}^{\mathrm{ub},\psi}) \right\},$$

where Δ_R^{tr} *and* Δ_P^{tr} *are the truncated reward and transition penalty upper bounds. For the penalty envelopes used here,*

$$\bar{\eta}_h^\psi \le \begin{cases} \rho_R R_{\max} + \gamma \rho_P V_{\max}, & \textit{TV}, \\ \frac{1}{2}\left(\sqrt{\rho_R} R_{\max} + \gamma \sqrt{\rho_P} V_{\max} \right), & \chi^2, \\ \rho_R + \gamma \rho_P L_{h+1}^\psi, & W_1, \end{cases}$$

with $L_{h+1}^\psi \ge \mathrm{Lip}(V_{h+1}^{\mathrm{ub},\psi})$ *for* W_1.

Thus the lower-bound gap is controlled by the ambiguity radii and the looseness of the penalty envelopes. At depth h, the gap vanishes when the truncated reward and transition penalty bounds coincide with the exact robust penalties for all representatives and actions. Appendix A.4 identifies such equality regimes for finite-support TV, finite-support χ^2, and W_1 ambiguity.

4.4 Search procedure

The search tree is defined over abstract nodes $v = (h, c)$. Each node maintains a local HOO tree over the continuous action space. A HOO leaf selects an action; the simulator then returns a reward and next state, which is mapped to its depth-$(h+1)$ bin. Each edge stores empirical reward and successor-bin statistics; its value is made conservative by subtracting robust-penalty upper bounds.

Sample-based robust edge estimates. At a bin node $v = (h, c)$ with representative $\bar{s}(v)$ and action a, samples are

$$(r_j, s_j') \sim (\nu^o(\cdot \mid \bar{s}(v), a), P^o(\cdot \mid \bar{s}(v), a)), \qquad v_j' = (h+1, \psi_{h+1}(s_j')).$$

Define $\bar{r}_t = \frac{1}{t}\sum_{j=1}^t r_j, \overline{V}_t = \frac{1}{t}\sum_{j=1}^t \widehat{V}(v_j')$. Let $\widehat{\overline{\Delta}}_{R,t}$ and $\widehat{\overline{\Delta}}_{P,t}$ denote the reward and transition penalty upper bounds used by the implementation, either fixed deterministic envelopes or plug-in estimates satisfying Appendix B. The robust edge estimate is

$$\widehat{Q}_t(v,a) = \left[\bar{r}_t - \widehat{\overline{\Delta}}_{R,t} \right]_+ + \gamma \left[\overline{V}_t - \widehat{\overline{\Delta}}_{P,t} \right]_+ . \tag{4.12}$$

This empirical counterpart of (4.8) subtracts the truncated penalties $\widehat{\overline{\Delta}}_{R,t} \wedge \bar{r}_t$ and $\widehat{\overline{\Delta}}_{P,t} \wedge \overline{V}_t$. Since rewards and child values are nonnegative, $0 \le \widehat{Q}_t(v,a) \le R_{\max} + \gamma V_{\max} \le V_{\max}$ for $V_{\max} = R_{\max}/(1-\gamma)$.

HOO selection and power-mean value backup. The HOO layer uses $\widehat{Q}$ as the payoff estimate and applies the standard statistical and geometric optimism:

$$U_{k,i}(t) = \hat{f}_{k,i}(t) + C t^{b/\beta} T_{k,i}(t)^{-\alpha/\beta} + \nu_1 \rho^k. \tag{4.13}$$

Algorithm 1 ROBUST-POWER-HOOT

Main loop

Require: Root state s_0; budget n; horizon H; baseline V_{π_0}; abstractions $\{\psi_h\}_{h=0}^{H}$; ambiguity, HOO, and power-mean parameters.

Initialize an empty abstract search tree.
for $m = 1, \dots, n$ **do**
  SIMULATE$(s_0, 0, m)$.
end for
$v_0 \leftarrow (0, \psi_0(s_0))$.
$\ell^\star \leftarrow \arg\max_{\ell:\, T_{v_0,\ell}>0} \widehat{Q}(v_0, \ell)$.
return $\widehat{a}_n \leftarrow a(v_0, \ell^\star)$.

Recursive simulation

function SIMULATE(s, h, m)
$v \leftarrow (h, \psi_h(s))$.
if v is new **then**
  Initialize local HOO tree/statistics at v.
  Set $s^{\mathrm{rep}}(v) \leftarrow s_{h,\psi_h(s)}$ (theorem) or $s^{\mathrm{rep}}(v) \leftarrow s$ (first visit).
end if
if $h = H$ or s is terminal **then**
  return LEAFVALUE(v, s).
end if
$\ell \leftarrow$ SELECTHOO(v, m); $a \leftarrow a(v, \ell)$.
Draw (r, s') from the nominal model at $(s^{\mathrm{rep}}(v), a)$.
SIMULATE$(s', h+1, m)$.
$v' \leftarrow (h+1, \psi_{h+1}(s'))$.
Update edge statistics and $\widehat{Q}(v, \ell)$ by (4.12).
Update $\widehat{V}(v)$ by (4.14) and the HOO indices by (4.13).
return $\widehat{V}(v)$.
end function

The index uses the conservative edge estimate in (4.12). The bin value is a power mean over active HOO leaves:

$$\widehat{V}(v) = \left(\sum_{\ell \in \mathcal{L}(v)} \frac{T_{v,\ell}}{T_v} [\widehat{Q}(v, \ell)]^p \right)^{1/p}, \qquad T_v = \sum_{\ell \in \mathcal{L}(v)} T_{v,\ell}. \tag{4.14}$$

Because (4.12) keeps $\widehat{Q}(v, \ell) \in [0, V_{\max}]$, this backup is well-defined for every finite $p \geq 1$. The count and normalization conventions are specified in Appendix C.

5 THEORETICAL ANALYSIS

Proof roadmap. At each depth, we first propagate child-value concentration through the robust Q backup. We then combine the standard HOO concentration bound with the edge-estimation errors in the power-mean backup. Backward induction and integration of the root tail give the expected-value rate. Appendices G and F give the details.

Penalty-bound convention. We use deterministic penalty envelopes. Appendix B treats plug-in estimates through their effective concentration exponents.

Definition 3 (Polynomial concentration)**.** *For $\alpha, \beta > 0$, write $\widehat{u}_n \xrightarrow[\alpha,\beta]{n\to\infty} u$ if there exists $c > 0$ such that, for all $n \geq 1$ and all $\varepsilon > n^{-\alpha/\beta}$,*

$$\mathbb{P}(|\widehat{u}_n - u| > \varepsilon) \leq c n^{-\alpha} \varepsilon^{-\beta}.$$

Assumption 1 (HOO geometry)**.** *At each abstract node v, the action space admits a binary HOO partition $\{\mathcal{P}_{k,i}\}$ under dissimilarity ℓ, with diameter at most $\nu_1 \rho^k$, an inner ball of radius $\nu_2 \rho^k$, and disjoint inner balls at each depth.*

Assumption 2 (Lower-bound action-value smoothness)**.** *For each abstract node v, let $f_v(a) := \widetilde{Q}^{\mathrm{ub},\psi}(v, a)$ and $f_v^\star := \sup_a f_v(a)$. Then, for all $a, a' \in A$,*

$$f_v^\star - f_v(a') \leq f_v^\star - f_v(a) + \max\{f_v^\star - f_v(a), \ell(a, a')\}.$$

Assumption 3 (Abstraction and penalty-bound regimes)**.** *The robust and lower-bound lookahead values in the approximation chain are L_h-Lipschitz at each depth. Each visited edge uses either deterministic penalty upper bounds or plug-in estimates satisfying the concentration conditions in Appendix B.*

Proposition 5 (Leaf-baseline concentration). *Fix a depth-H bin node $v_H = (H, c)$. Under the leaf routine, for every $\beta_H > 0$ and every $0 < \alpha_H \le \beta_H/2$,*

$$\widehat{V}_n(v_H) \xrightarrow[\alpha_H,\beta_H]{n\to\infty} \widetilde{V}^{\mathrm{ub},\psi}(v_H).$$

Lemma 1 (Lower-bound Q-backup concentration). *Fix (v, a) at depth h. Suppose the successor-bin estimators satisfy $\widehat{V}_n(v') \xrightarrow[\alpha,\beta]{n\to\infty} \widetilde{V}^{\mathrm{ub},\psi}(v')$ uniformly over successor bins, with $\beta > 0$ and $2\alpha \le \beta$, and assume the successor-count condition in Assumption 5. Suppose also that the reward and transition penalty-bound estimates are deterministic, or satisfy $\widehat{\overline{\Delta}}_{R,t} \xrightarrow[\alpha,\beta]{n\to\infty} \overline{\Delta}_R, \widehat{\overline{\Delta}}_{P,t} \xrightarrow[\alpha,\beta]{n\to\infty} \overline{\Delta}_P$, where (α, β) is the effective pair after incorporating any plug-in bottleneck from Appendix B. Then the positive-part backup* (4.12) *satisfies*

$$\widehat{Q}_t(v, a) \xrightarrow[\alpha,\beta]{n\to\infty} \widetilde{Q}^{\mathrm{ub},\psi}(v, a).$$

Theorem 1 (Node-level power-mean HOO concentration). *Under Assumptions 1 to 3 and the leaf routine of Proposition 5, choose Enhanced-HOO parameters with*

$$\alpha(1 - \alpha/\beta) \le b < \alpha, \qquad b > 3, \qquad \rho^{\bar{H}} \le n^{-\alpha/\beta},$$

and a finite $p \ge 1$ with $\alpha \le \beta/\max\{2, p\}$. Assume also the standard HOO conditions in Assumption 4. Then with

$$\beta_v = \tfrac{b-3}{2}, \qquad \alpha_v = \frac{1 - b/\alpha}{1 + d' + \beta/\alpha} \cdot \tfrac{b-3}{2},$$

where d' is a packing exponent above the near-optimality dimension of (f_v, ℓ),

$$\widehat{V}_n(v) \xrightarrow[\alpha_v,\beta_v]{n\to\infty} \widetilde{V}^{\mathrm{ub},\psi}(v), \qquad \widetilde{V}^{\mathrm{ub},\psi}(v) = \sup_{a\in A} \widetilde{Q}^{\mathrm{ub},\psi}(v, a).$$

Use the effective incoming pair for plug-in penalties. The target is the fixed-representative projected value.

Theorem 2 (Root convergence to the projected lower-bound target). *Assume the depthwise hypotheses of Theorem 5. Let $v_0 = (0, \psi_0(s_0))$ and (α_0, β_0) be the resulting root exponents. If $\beta_0 > 1$, then*

$$\left|\mathbb{E}[\widehat{V}_n(v_0)] - \widetilde{V}^{\mathrm{ub},\psi}(v_0)\right| = O(n^{-\zeta}), \qquad \zeta := \frac{\alpha_0}{\beta_0} \in (0, 1/2). \tag{5.1}$$

Corollary 1 (Approximation chain to the exact robust value). *Combining the finite-horizon truncation bound, Proposition 1, Proposition 4, and Theorem 2, the following holds with $\zeta = \alpha_0/\beta_0 \in (0, 1/2)$:*

$$\left|\mathbb{E}[\widehat{V}_n(v_0)] - V^\star_{\mathrm{rob}}(s_0)\right| \le \gamma^H \left\|V^\star_{\mathrm{rob}} - V_{\pi_0}\right\|_\infty + \sum_{h=0}^{H} \gamma^h L_h \varepsilon_h + \sum_{h=0}^{H-1} \gamma^h \eta^\psi_h + C_\zeta n^{-\zeta}. \tag{5.2}$$

6 Experiments

We evaluate Robust-Power-HOOT on seven continuous-control benchmarks. Planning uses a low-noise simulator, whereas execution uses larger perturbations, isolating a controlled stochastic mismatch rather than arbitrary structural errors such as contact-model changes, delays, or parameter drift.

Setup. The tasks are Hopper ($m = 3$), Walker2d ($m = 6$), Ant ($m = 8$), Humanoid ($m = 17$), CartPole, Pendulum, and Acrobot. Gaussian noise is injected into actions, dynamics, and observations: $\sigma_{\mathrm{plan}} = 0.02$–$0.05$ during planning and $\sigma_{\mathrm{exec}} \in \{0.1, 0.2, 0.3, 0.4\}$ during execution. Rewards are shifted to be nonnegative for power-mean backups. Baselines include HOOT, POLY-HOOT, Progressive Widening, Stochastic-Power-HOOT, and discretized UCT (Mansley et al., 2011; Mao et al., 2020; Couëtoux et al., 2011; Dam, 2025; Kocsis & Szepesvári, 2006). All methods use $\gamma = 0.99$, evaluation horizon $T = 150$, and 500 simulations per planning step. Curves show mean $\pm$ standard error over five seeds. Robust variants use one fixed radius per divergence and environment family; Appendix H gives the configurations and reward preprocessing.

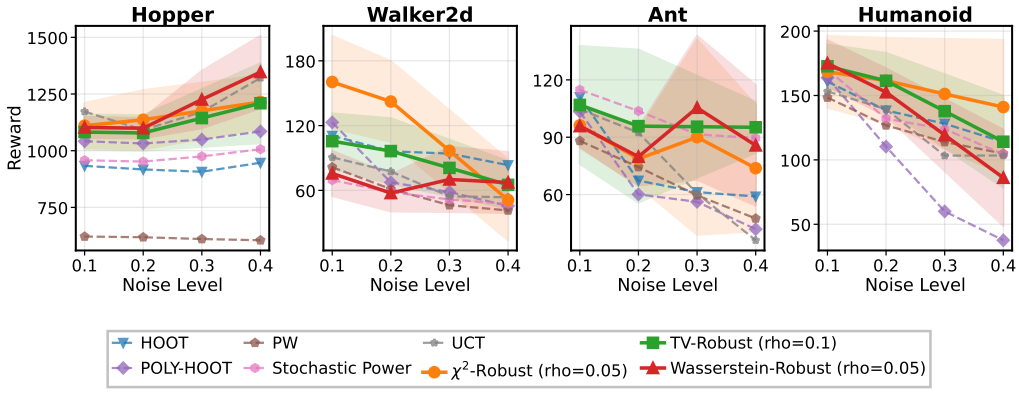

Figure 1: MuJoCo controlled-mismatch results (five-seed mean ± standard error).

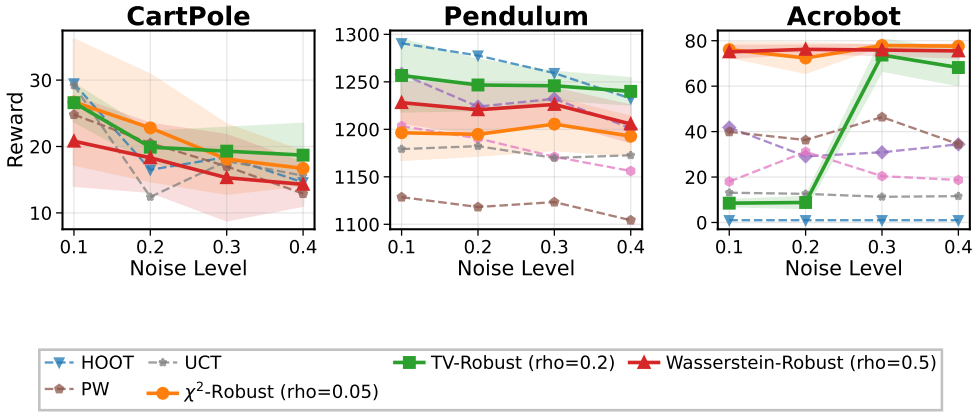

Figure 2: Classic-control controlled-mismatch results (five-seed mean ± standard error).

Results. Figures 1 and 2 show that the gains vary by task. The robust variants are strongest on Hopper, Humanoid, and Acrobot and remain comparatively stable on Ant; Walker2d, CartPole, and Pendulum are mixed. When the closed-form and exact penalties are evaluated on the same logged backups, the reported relative gap is 0–1% for χ^2; TV reaches 29% on CartPole and 11% on Hopper (Table H.7). In separate W_1 planner runs, replacing the closed-form backup by the exact backup changes returns in both directions on MuJoCo; the planner with exact backups has higher sample means on CartPole, Pendulum, and Acrobot (Table H.8). Over the radius sweep, χ^2 has the smallest worst-case return swing on Pendulum and MuJoCo (Table H.10). A single validation radius is shared within each environment family; these radii are not calibrated physical uncertainty radii. Appendix H.1 gives the remaining diagnostics.

7 Discussion and Limitations

Robust-Power-HOOT combines continuous-action optimistic search, power-mean backups, and conservative robust Bellman updates. It estimates a projected finite-horizon target rather than the exact infinite-horizon robust value; Corollary 1 separates truncation, abstraction, approximation, and finite-sample errors. The comparison at individual backups and the separate planner runs do not establish equivalence to exact robust planning. The theory uses fixed representatives; the first-visit implementation instead has data-dependent representatives whose gap is unquantified without cell or merge radii. The statistical result otherwise uses standard HOO assumptions and the successor-count conditions in the appendix. Finally, five seeds and controlled stochastic mismatch do not establish universal dominance or robustness to arbitrary structural misspecification.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval, manuscript organization and editing, proof checking, experimental design, and construction and checking of numerical diagnostics. The authors checked cited references against their sources and manually reviewed every AI-assisted claim, derivation, proof, and numerical conclusion against the stated assumptions and definitions. Numerical diagnostics supplemented rather than replaced the mathematical arguments. The authors take full responsibility for all text, claims, proofs, experiments, and reported results.

REPRODUCIBILITY STATEMENT

The experimental protocol is described in Section 6 and Appendix H, including the environments, planning and execution noise, $\gamma = 0.99$, evaluation horizon $T = 150$, and 500 simulations per planning step for all methods. Per-method planning depths and HOO parameters, the fixed ambiguity radii, and the baseline configurations are given in Tables H.3, H.4, and H.5. Results are averaged over five seeds and reported with standard error. Code and configurations will be released upon acceptance.

REFERENCES

David Auger, Adrien Couëtoux, and Olivier Teytaud. Continuous upper confidence trees with polynomial exploration – consistency. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (eds.), *Machine Learning and Knowledge Discovery in Databases*, volume 8188 of *Lecture Notes in Computer Science*, pp. 194–209, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-40988-2_13. URL https://doi.org/10.1007/978-3-642-40988-2_13.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(46):1655–1695, 2011. URL https://www.jmlr.org/papers/v12/bubeck11a.html.

Adrien Couëtoux, Jean-Baptiste Hoock, Nataliya Sokolovska, Olivier Teytaud, and Nicolas Bonnard. Continuous upper confidence trees. In Carlos A. Coello Coello (ed.), *Learning and Intelligent Optimization*, volume 6683 of *Lecture Notes in Computer Science*, pp. 433–445, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-25566-3_32. URL https://doi.org/10.1007/978-3-642-25566-3_32.

Tuan Dam, Pascal Klink, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Generalized mean estimation in Monte-Carlo tree search. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2397–2404, 2020. doi: 10.24963/ijcai.2020/332. URL https://www.ijcai.org/proceedings/2020/332.

Tuan Dam, Georgia Chalvatzaki, Jan Peters, and Joni Pajarinen. Monte-carlo robot path planning. *IEEE Robotics and Automation Letters*, 7(4):11213–11220, 2022. doi: 10.1109/LRA.2022.3199674.

Tuan Dam, Odalric-Ambrym Maillard, and Emilie Kaufmann. Power mean estimation in stochastic Monte-Carlo tree search. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 894–918. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/dam24a.html.

Tuan Quang Dam. Power mean estimation in stochastic continuous Monte-Carlo tree search. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12344–12376. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/dam25b.html.

Tuan Quang Dam, Kishan Panaganti, Brahim Driss, and Adam Wierman. Online robust reinforcement learning through Monte-Carlo planning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12314–12343. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/dam25a.html.

Arthur Guez, David Silver, and Peter Dayan. Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, 48(1): 841–883, October 2013. ISSN 1076-9757. doi: 10.1613/jair.4117. URL https://www.jair.org/index.php/jair/article/view/10853.

Conor F. Hayes, Mathieu Reymond, Diederik M. Roijers, Enda Howley, and Patrick Mannion. Distributional monte carlo tree search for risk-aware and multi-objective reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pp. 1530–1532, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Robust ϕ-divergence MDPs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32680–32693, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d2fe3a5711a6d488da9e9a78b84ee24c-Abstract-Conference.html.

Jesse Hostetler, Alan Fern, and Tom Dietterich. State aggregation in monte carlo tree search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1):2446–2452, 2014. doi: 10.1609/aaai.v28i1.9066. URL https://ojs.aaai.org/index.php/AAAI/article/view/9066.

Garud Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129.

Bilal Kartal, Pablo Hernandez-Leal, and Matthew E. Taylor. Action guidance with MCTS for deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 15(1):153–159, 2019. doi: 10.1609/aiide.v15i1.5238. URL https://ojs.aaai.org/index.php/AIIDE/article/view/5238.

Beomjoon Kim, Kyungjae Lee, Sungbin Lim, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Monte carlo tree search in continuous spaces using voronoi optimistic optimization with regret bounds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):9916–9924, 2020. doi: 10.1609/aaai.v34i06.6546. URL https://ojs.aaai.org/index.php/AAAI/article/view/6546.

Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou (eds.), *Machine Learning: ECML 2006*, volume 4212 of *Lecture Notes in Computer Science*, pp. 282–293. Springer, 2006. doi: 10.1007/11871842_29. URL https://doi.org/10.1007/11871842_29.

Jongmin Lee, Wonseok Jeon, Geon-Hyeong Kim, and Kee-Eung Kim. Monte-carlo tree search in continuous action spaces with value gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4561–4568, 2020. doi: 10.1609/aaai.v34i04.5885. URL https://ojs.aaai.org/index.php/AAAI/article/view/5885.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13623–13643. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22a.html.

Chris Mansley, Ari Weinstein, and Michael L. Littman. Sample-based planning for continuous action markov decision processes. *Proceedings of the International Conference on Automated Planning and Scheduling*, 21(1):335–338, 2011. doi: 10.1609/icaps.v21i1.13484. URL https://ojs.aaai.org/index.php/ICAPS/article/view/13484.

Weichao Mao, Kaiqing Zhang, Qiaomin Xie, and Tamer Başar. POLY-HOOT: Monte-Carlo planning in continuous space MDPs with non-asymptotic analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4549–4559, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/30de24287a6d8f07b37c716ad51623a7-Abstract.html.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018. doi: 10.1007/s10107-017-1172-1. URL https://doi.org/10.1007/s10107-017-1172-1.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Yaniv Oren, Viliam Vadocz, Matthijs T. J. Spaan, and Wendelin Böhmer. Epistemic monte carlo tree search. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bcbd670951a6dede2123961de19e5ea3-Abstract-Conference.html.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9582–9602. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2817–2826. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pinto17a.html.

Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:4, 2022. doi: 10.5802/ojmo.15. URL https://ojmo.centre-mersenne.org/articles/10.5802/ojmo.15/.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3043–3052, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/84c6494d30851c63a55cdb8cb047fadd-Abstract.html.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, December 2020. doi: 10.1038/s41586-020-03051-4. URL https://doi.org/10.1038/s41586-020-03051-4.

Apoorva Sharma, James Harrison, Matthew Tsao, and Marco Pavone. Robust and adaptive planning under model uncertainty. *Proceedings of the International Conference on Automated Planning and Scheduling*, 29(1):410–418, 2019. doi: 10.1609/icaps.v29i1.3505. URL https://ojs.aaai.org/index.php/ICAPS/article/view/3505.

David Silver, Aja Huang, Chris J. Maddison, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016. doi: 10.1038/nature16961. URL https://doi.org/10.1038/nature16961.

Samuel Sokota, Caleb Y. Ho, Zaheen Ahmad, and J. Zico Kolter. Monte carlo tree search with iteratively refining state abstractions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18698–18709, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9b0ead00a217ea2c12e06a72eec4923f-Abstract.html.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, February 2013. ISSN 0364-765X. doi: 10.1287/moor.1120.0566. URL https://doi.org/10.1287/moor.1120.0566.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, May 2012. ISSN 0364-765X. doi: 10.1287/moor.1120.0540. URL https://doi.org/10.1287/moor.1120.0540.

Timothy Yee, Viliam Lisý, and Michael Bowling. Monte carlo tree search in continuous action spaces with execution uncertainty. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 690–696. AAAI Press, 2016. URL https://www.ijcai.org/Proceedings/16/Papers/104.pdf.

Zuyuan Zhang, Arnob Ghosh, and Tian Lan. Tail-risk-safe monte carlo tree search under PAC-level guarantees, 2025. URL https://arxiv.org/abs/2508.05441.

Supplementary Material

Monte Carlo Tree Search for Continuous Stochastic MDPs Under Model Ambiguity

A Operator properties and approximation error . . . 15
A.1 Robust Bellman contraction . . . 15
A.2 Upper bounds on robust penalties . . . 16
A.3 Projection and abstraction errors . . . 17
A.4 Conservatism and lower-bound approximation error . . . 18
A.5 Finite-horizon lower-bound operator . . . 20
B Plug-in penalty estimates. . . . 21
B.1 Definitions and boundary assumptions . . . 21
B.2 Basic concentration and stability bounds . . . 22
B.3 Concentration rates . . . 23
C Local HOO search . . . 24
C.1 Local edge estimates . . . 24
D HOO geometry . . . 25
E Standard Enhanced-HOO concentration. . . . 25
F Power-mean concentration. . . . 26
G MCTS convergence . . . 28
G.1 Parameter schedule . . . 28
G.2 Sampling and one-step backups . . . 29
G.3 Leaf and edge concentration . . . 30
G.4 Backward induction and root convergence . . . 31
H Detailed experimental setup . . . 33
H.1 Additional diagnostic analyses . . . 35

NOTATION

Symbol	Meaning
S, A, γ	State space, compact continuous action space, and discount factor.
$P^o_{s,a}, \nu^o_{s,a}$	Nominal transition/reward laws queried by the planner.
$\mathcal{P}_{s,a}, \mathcal{R}_{s,a}$	Divergence ambiguity sets around nominal laws.
$\mathcal{T}^{\mathrm{rob}}_Q, \mathcal{T}^{\mathrm{rob}}_V$	Exact robust Bellman operators.
$\mathcal{T}^{\mathrm{ub}}_Q, \mathcal{T}^{\mathrm{ub}}_V$	Bellman operators formed from conservative lower bounds on robust expectations.
$\Delta_{D,\rho}$	Exact robust penalty, i.e., the decrease from the nominal expectation to the worst-case expectation.
$\overline{\Delta}_{D,\rho}$	Upper bound on the robust penalty.
$\Delta^{\mathrm{tr}}_{D,\rho}$	Truncated penalty upper bound $\overline{\Delta}_{D,\rho} \wedge \mathbb{E}_Q Z$.
$\underline{\sigma}_{D,\rho}$	Conservative lower bound $[\mathbb{E}_Q Z - \overline{\Delta}_{D,\rho}(Q; Z)]_+$.
H, V_{π_0}	Planning depth and terminal rollout/baseline value.
$\psi_h, \mathcal{C}_h, \varepsilon_h$	Depth-h abstraction, finite codebook, and cell resolution.
$v = (h, c), s_{h,c}$	Abstract node and fixed cell representative used in the theorem.
$\bar{s}(v)$	First-visit representative used by the practical implementation variant; the theorem uses fixed $s_{h,c}$.
$\widetilde{V}^{\mathrm{ub},\psi}, \widetilde{Q}^{\mathrm{ub},\psi}$	Projected finite-horizon lower-bound targets estimated by the algorithm.
$\widehat{V}, \widehat{Q}$	Empirical value and edge estimates used in MCTS.

Table .1: Core notation.

The appendix follows the proof in the order used by the main theorem. We first compare the exact robust operator with its conservative approximation, then analyze empirical penalty estimates. Sections C–F develop the local HOO and power-mean analysis, and Section G combines these results by backward induction. Section H gives the full experimental setup and diagnostics.

A OPERATOR PROPERTIES AND APPROXIMATION ERROR

A.1 ROBUST BELLMAN CONTRACTION

Proof of contraction of $\mathcal{T}^{\mathrm{rob}}_V$. Rectangularity means that the adversary chooses reward and transition laws locally at each state-action pair (Nilim & El Ghaoui, 2005; Wiesemann et al., 2013). The one-step robust update therefore separates as

$$\inf_{\nu \in \mathcal{R}_{s,a},\, P \in \mathcal{P}_{s,a}} \{\mathbb{E}_\nu[R] + \gamma \mathbb{E}_P[V(S')]\} = \inf_{\nu \in \mathcal{R}_{s,a}} \mathbb{E}_\nu[R] + \gamma \inf_{P \in \mathcal{P}_{s,a}} \mathbb{E}_P[V(S')],$$

which is the robust Q-operator in (3.3). The value operator then takes the best action,

$$(\mathcal{T}^{\mathrm{rob}}_V V)(s) = \sup_{a \in A} (\mathcal{T}^{\mathrm{rob}}_Q V)(s, a).$$

Let V, W be bounded measurable functions. The robust reward term does not depend on the continuation value, so for every (s, a),

$$\begin{aligned} \left|(\mathcal{T}^{\mathrm{rob}}_Q V)(s,a) - (\mathcal{T}^{\mathrm{rob}}_Q W)(s,a)\right| &= \gamma \left| \inf_{P \in \mathcal{P}_{s,a}} \mathbb{E}_P[V] - \inf_{P \in \mathcal{P}_{s,a}} \mathbb{E}_P[W] \right| \\ &\le \gamma \sup_{P \in \mathcal{P}_{s,a}} \mathbb{E}_P |V - W| \\ &\le \gamma \left\| V - W \right\|_\infty . \end{aligned}$$

Hence

$$\left\| \mathcal{T}^{\mathrm{rob}}_Q V - \mathcal{T}^{\mathrm{rob}}_Q W \right\|_\infty \le \gamma \left\| V - W \right\|_\infty .$$

Since $q \mapsto \sup_a q(a)$ is nonexpansive in sup norm,

$$\left\| \mathcal{T}^{\mathrm{rob}}_V V - \mathcal{T}^{\mathrm{rob}}_V W \right\|_\infty \le \gamma \left\| V - W \right\|_\infty .$$

This proves that $\mathcal{T}^{\mathrm{rob}}_V$ is a γ-contraction. Banach's fixed-point theorem gives a unique fixed point, and the standard robust dynamic-programming assumptions identify that fixed point with $V^\star_{\mathrm{rob}}$. □

Lemma 2 (Monotonicity)**.** *If $V \le W$ pointwise, then $\mathcal{T}_V^{\mathrm{rob}} V \le \mathcal{T}_V^{\mathrm{rob}} W$.*

Proof. For every feasible transition law $P \in \mathcal{P}_{s,a}$, $V \le W$ implies $\mathbb{E}_P[V(S')] \le \mathbb{E}_P[W(S')]$. Taking the infimum over P, adding the same reward term, and taking the supremum over actions preserves the order. □

A.2 Upper bounds on robust penalties

Proof of Proposition 2. Since $Z \ge 0$,

$$\sigma_{D,\rho}(Q; Z) = \inf_{P \in \mathcal{U}_D(Q,\rho)} \mathbb{E}_P[Z] \ge 0.$$

Therefore

$$\Delta_{D,\rho}(Q; Z) = \mathbb{E}_Q[Z] - \sigma_{D,\rho}(Q; Z) \le \mathbb{E}_Q[Z].$$

Since $\overline{\Delta}_{D,\rho}(Q; Z) \ge \Delta_{D,\rho}(Q; Z)$, we have

$$\overline{\Delta}_{D,\rho}(Q; Z) \wedge \mathbb{E}_Q[Z] \ge \Delta_{D,\rho}(Q; Z).$$

Hence

$$\underline{\sigma}_{D,\rho}(Q; Z) = \mathbb{E}_Q[Z] - \big(\overline{\Delta}_{D,\rho}(Q; Z) \wedge \mathbb{E}_Q[Z]\big) \le \mathbb{E}_Q[Z] - \Delta_{D,\rho}(Q; Z) = \sigma_{D,\rho}(Q; Z).$$

The lower bound is immediate from the positive part. □

Lemma 3 (Standard upper bounds on robust penalties)**.** *Let $0 \le Z \le M_Z$.*

1. *If $D_{\mathrm{TV}}(P, Q) \le \rho$, then*
$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le \rho \operatorname{osc}(Z).$$
2. *If $D_{\chi^2}(P\|Q) \le \rho$, then*
$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le \sqrt{\rho}\, \sigma_Q(Z).$$
3. *If $W_1(P, Q) \le \rho$ and $L_Z \ge \mathrm{Lip}(Z)$, then*
$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le \rho L_Z.$$

The same bounds hold for empirical finite-support ambiguity sets by taking $Q = \widehat{Q}_t$ and computing the span, standard deviation, or Lipschitz seminorm on the observed support. For a population W_1 ball on the original metric space, the empirical support seminorm is not enough; the analysis uses a deterministic global Lipschitz envelope.

Proof. For total variation, let $\mu = Q - P$. Since μ has total mass zero,

$$\int Z\, d(Q - P) = \int (Z - c)\, d\mu \le \|Z - c\|_\infty \, \|\mu\|_{\mathrm{TV}}$$

for any constant c. With the convention $D_{\mathrm{TV}}(P, Q) = \frac{1}{2}\|\mu\|_{\mathrm{TV}}$, choose $c = (\sup Z + \inf Z)/2$. Then $\|Z - c\|_\infty = \operatorname{osc}(Z)/2$, and hence

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le D_{\mathrm{TV}}(P, Q) \operatorname{osc}(Z) \le \rho \operatorname{osc}(Z).$$

On a finite support S, the same argument uses $\operatorname{osc}_S(Z)$. With $Q = \widehat{Q}_t$ on the observed support, this gives the plug-in bound $\rho \operatorname{osc}_t(Z)$.

For χ^2, let $w = dP/dQ$. Pearson χ^2 feasibility gives

$$\mathbb{E}_Q[(w - 1)^2] \le \rho.$$

Since $\mathbb{E}_Q[1 - w] = 0$,

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] = \mathbb{E}_Q[(1 - w)(Z - \mathbb{E}_Q Z)].$$

By Cauchy–Schwarz,

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le \big(\mathbb{E}_Q[(1 - w)^2]\big)^{1/2} \big(\mathbb{E}_Q[(Z - \mathbb{E}_Q Z)^2]\big)^{1/2} \le \sqrt{\rho}\, \sigma_Q(Z).$$

Taking $Q = \widehat{Q}_t$ gives the empirical finite-support bound $\sqrt{\rho}\,\widehat{\sigma}_t(Z)$. Since $0 \le Z \le M_Z$, Popoviciu's inequality gives $\sigma_Q(Z) \le M_Z/2$.

For W_1, Kantorovich–Rubinstein duality gives if Z is L_Z-Lipschitz, then

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \le L_Z W_1(P, Q) \le \rho L_Z.$$

For a finite-support Wasserstein ambiguity set, L_Z can be the finite-support Lipschitz seminorm. For a population Wasserstein ball on the original metric space, the empirical support seminorm is not enough because the adversary may move mass outside the observations; one needs a deterministic global envelope $L_Z \ge \mathrm{Lip}(Z)$. □

A.3 Projection and abstraction errors

Projection bias.

Proof of Proposition 1. Let $\delta_h = \left\| V_h^{\mathrm{rob},\psi} - V_h^{\mathrm{rob}} \right\|_\infty$. At H, $\delta_H \le L_H \varepsilon_H$. For $h < H$,

$$\begin{aligned}
\delta_h &= \left\| \Pi_h \mathcal{T}_V^{\mathrm{rob}} V_{h+1}^{\mathrm{rob},\psi} - \mathcal{T}_V^{\mathrm{rob}} V_{h+1}^{\mathrm{rob}} \right\|_\infty \\
&\le \left\| \Pi_h \mathcal{T}_V^{\mathrm{rob}} V_{h+1}^{\mathrm{rob},\psi} - \Pi_h \mathcal{T}_V^{\mathrm{rob}} V_{h+1}^{\mathrm{rob}} \right\|_\infty + \left\| \Pi_h V_h^{\mathrm{rob}} - V_h^{\mathrm{rob}} \right\|_\infty \\
&\le \gamma \delta_{h+1} + L_h \varepsilon_h,
\end{aligned}$$

using nonexpansiveness of Π_h and the robust Bellman contraction. Unrolling gives the result. □

Refinement radius and abstraction error. The main theorem uses a fixed abstraction. The next bound records what changes if an implementation refines the abstraction by merging a new successor only with a nearby representative.

Lemma 4 (Refinement radius controls abstraction error)**.** *Let (S, d) be a metric space and let $\phi : S \to S$ be an abstraction map satisfying*

$$d(s, \phi(s)) \le r \qquad \textit{for all states in the support under consideration.}$$

If $V : S \to [0, V_{\max}]$ is L-Lipschitz, then

$$|\mathbb{E}_{S' \sim P}[V(S')] - \mathbb{E}_{S' \sim P}[V(\phi(S'))]| \le Lr.$$

Moreover, for $z(s) = V(s)$ and $\tilde{z}(s) = V(\phi(s))$,

$$|\mathrm{osc}(z) - \mathrm{osc}(\tilde{z})| \le 2Lr, \qquad |\sigma_P(z) - \sigma_P(\tilde{z})| \le 2Lr.$$

For the span and standard-deviation penalty bounds, respectively,

$$\left| \underline{\sigma}_{\mathrm{TV},\rho}(P; V) - \underline{\sigma}_{\mathrm{TV},\rho}(P; V \circ \phi) \right| \le (1 + 2\rho) Lr,$$

and

$$\left| \underline{\sigma}_{\chi^2,\rho}(P; V) - \underline{\sigma}_{\chi^2,\rho}(P; V \circ \phi) \right| \le (1 + 2\sqrt{\rho}) Lr.$$

Proof. Lipschitzness gives $|V(s) - V(\phi(s))| \le Lr$, and the mean bound follows by taking expectations. For the oscillation,

$$|\sup z - \sup \tilde{z}| \le \|z - \tilde{z}\|_\infty, \qquad |\inf z - \inf \tilde{z}| \le \|z - \tilde{z}\|_\infty,$$

so $|\mathrm{osc}(z) - \mathrm{osc}(\tilde{z})| \le 2Lr$. For the standard deviation, write $\sigma_P(z) = \|z - \mathbb{E}_P z\|_{L^2(P)}$. Then

$$|\sigma_P(z) - \sigma_P(\tilde{z})| \le \|(z - \mathbb{E}_P z) - (\tilde{z} - \mathbb{E}_P \tilde{z})\|_{L^2(P)} \le \|z - \tilde{z}\|_\infty + |\mathbb{E}_P(z - \tilde{z})| \le 2Lr.$$

Finally, $x \mapsto [x]_+$ is 1-Lipschitz. Combining the mean perturbation with the TV penalty-bound perturbation $\rho|\mathrm{osc}(z) - \mathrm{osc}(\tilde{z})|$ gives $(1+2\rho)Lr$; combining it with the χ^2 penalty-bound perturbation $\sqrt{\rho}|\sigma_P(z) - \sigma_P(\tilde{z})|$ gives $(1 + 2\sqrt{\rho})Lr$. □

Corollary 2 (Error under a refining abstraction)**.** *Suppose a refinement-based implementation has radius $r_{h,n}$ at depth h, meaning every merged successor is within distance $r_{h,n}$ of its representative. Under the Lipschitz assumptions used in Proposition 1, the abstraction component in the total error bound can be replaced by a term of the form*

$$\sum_{h=0}^{H} \gamma^h \bar{L}_h r_{h,n},$$

where $\bar{L}_h$ includes the stability constants from Lemma 4. If $r_{h,n} \to 0$ at every depth, this component vanishes. The statement controls approximation bias; a finite-sample analysis of a random adaptive abstraction requires additional arguments.

A.4 Conservatism and lower-bound approximation error

One-sided conservatism.

Proof of Proposition 3. Fix a bounded nonnegative continuation value V. For the reward term, define the exact reward robust penalty

$$\Delta_R(s,a) = \mathbb{E}_{\nu^o_{s,a}}[R] - \inf_{\nu \in \mathcal{R}_{s,a}} \mathbb{E}_\nu[R].$$

By assumption, $\overline{\Delta}_R(s,a) \geq \Delta_R(s,a)$. By Proposition 2,

$$\ell^{\mathrm{ub}}_R(s,a) = \left[\mathbb{E}_{\nu^o_{s,a}}[R] - \overline{\Delta}_R(s,a)\right]_+ \leq \inf_{\nu \in \mathcal{R}_{s,a}} \mathbb{E}_\nu[R] = r^{\mathrm{rob}}(s,a).$$

The same argument applied to $Z = V(S')$ under the nominal transition law gives

$$\ell^{\mathrm{ub}}_P(s,a;V) \leq \inf_{P \in \mathcal{P}_{s,a}} \mathbb{E}_P[V(S')].$$

Adding the two inequalities with discount factor γ yields

$$(\mathcal{T}^{\mathrm{ub}}_Q V)(s,a) = \ell^{\mathrm{ub}}_R(s,a) + \gamma \ell^{\mathrm{ub}}_P(s,a;V) \leq (\mathcal{T}^{\mathrm{rob}}_Q V)(s,a).$$

Taking the supremum over $a \in A$ preserves the inequality, so $(\mathcal{T}^{\mathrm{ub}}_V V)(s) \leq (\mathcal{T}^{\mathrm{rob}}_V V)(s)$ for every state s. □

Approximation error of the lower-bound operator. For a fixed continuation value and ambiguity set, the backup gap is the difference between the exact robust penalty and its truncated upper bound.

For a bounded nonnegative random variable Z under nominal law Q, recall

$$\Delta_{D,\rho}(Q;Z) = \mathbb{E}_Q Z - \inf_{P \in \mathcal{U}_D(Q,\rho)} \mathbb{E}_P Z, \qquad \Delta^{\mathrm{tr}}_{D,\rho}(Q;Z) = \overline{\Delta}_{D,\rho}(Q;Z) \wedge \mathbb{E}_Q Z.$$

The exact robust expectation is $\mathbb{E}_Q Z - \Delta_{D,\rho}(Q;Z)$, while the conservative lower bound is $\mathbb{E}_Q Z - \Delta^{\mathrm{tr}}_{D,\rho}(Q;Z)$. Hence their difference is

$$\Delta^{\mathrm{tr}}_{D,\rho}(Q;Z) - \Delta_{D,\rho}(Q;Z) \geq 0.$$

It vanishes when the truncated upper bound equals the exact robust penalty.

For a state-action pair (s,a), let

$$\mu_R(s,a) = \mathbb{E}_{\nu^o_{s,a}}[R], \qquad \mu_P(s,a;V) = \mathbb{E}_{P^o_{s,a}}[V(S')].$$

Define the exact reward and transition robust penalties

$$\Delta_R(s,a) = \mu_R(s,a) - \inf_{\nu \in \mathcal{R}_{s,a}} \mathbb{E}_\nu[R], \qquad \Delta_P(s,a;V) = \mu_P(s,a;V) - \inf_{P \in \mathcal{P}_{s,a}} \mathbb{E}_P[V(S')],$$

and the truncated penalty upper bounds

$$\Delta^{\mathrm{tr}}_R(s,a) = \overline{\Delta}_R(s,a) \wedge \mu_R(s,a), \qquad \Delta^{\mathrm{tr}}_P(s,a;V) = \overline{\Delta}_P(s,a;V) \wedge \mu_P(s,a;V).$$

Then

$$\begin{aligned}(\mathcal{T}^{\mathrm{rob}}_Q V)(s,a) - (\mathcal{T}^{\mathrm{ub}}_Q V)(s,a) = \Delta^{\mathrm{tr}}_R(s,a) - \Delta_R(s,a) \\ + \gamma\{\Delta^{\mathrm{tr}}_P(s,a;V) - \Delta_P(s,a;V)\}. \end{aligned} \tag{A.1}$$

So the one-step error is just the reward penalty overestimate plus the discounted transition penalty overestimate.

Proof of Proposition 4. For the projected recursion, define

$$\eta_h^\psi := \sup_{c\in\mathcal{C}_h, a\in A}\Big[\Delta_R^{\mathrm{tr}}(s_{h,c},a) - \Delta_R(s_{h,c},a) + \gamma\{\Delta_P^{\mathrm{tr}}(s_{h,c},a;V_{h+1}^{\mathrm{ub},\psi}) - \Delta_P(s_{h,c},a;V_{h+1}^{\mathrm{ub},\psi})\}\Big].$$

By (A.1), this is equivalently

$$\eta_h^\psi = \sup_{c\in\mathcal{C}_h, a\in A}\Big[(\mathcal{T}_Q^{\mathrm{rob}}V_{h+1}^{\mathrm{ub},\psi})(s_{h,c},a) - (\mathcal{T}_Q^{\mathrm{ub}}V_{h+1}^{\mathrm{ub},\psi})(s_{h,c},a)\Big],$$

which is the definition used in the main text.

Fix a representative $s_{h,c}$. The contraction of $\mathcal{T}_V^{\mathrm{rob}}$ gives

$$(\mathcal{T}_V^{\mathrm{rob}}V_{h+1}^{\mathrm{rob},\psi})(s_{h,c}) \le (\mathcal{T}_V^{\mathrm{rob}}V_{h+1}^{\mathrm{ub},\psi})(s_{h,c}) + \gamma\left\|V_{h+1}^{\mathrm{rob},\psi} - V_{h+1}^{\mathrm{ub},\psi}\right\|_\infty.$$

The definition of η_h^ψ gives

$$(\mathcal{T}_V^{\mathrm{rob}}V_{h+1}^{\mathrm{ub},\psi})(s_{h,c}) \le (\mathcal{T}_V^{\mathrm{ub}}V_{h+1}^{\mathrm{ub},\psi})(s_{h,c}) + \eta_h^\psi.$$

The projection Π_h only evaluates these inequalities at cell representatives, so

$$V_h^{\mathrm{rob},\psi} \le V_h^{\mathrm{ub},\psi} + \gamma\left\|V_{h+1}^{\mathrm{rob},\psi} - V_{h+1}^{\mathrm{ub},\psi}\right\|_\infty + \eta_h^\psi.$$

The terminal values agree, $V_H^{\mathrm{rob},\psi} = V_H^{\mathrm{ub},\psi} = \Pi_H V_{\pi_0}$. If $V_{h+1}^{\mathrm{ub},\psi} \le V_{h+1}^{\mathrm{rob},\psi}$, then

$$\Pi_h\mathcal{T}_V^{\mathrm{ub}}V_{h+1}^{\mathrm{ub},\psi} \le \Pi_h\mathcal{T}_V^{\mathrm{rob}}V_{h+1}^{\mathrm{ub},\psi} \le \Pi_h\mathcal{T}_V^{\mathrm{rob}}V_{h+1}^{\mathrm{rob},\psi}.$$

Thus $V_h^{\mathrm{ub},\psi} \le V_h^{\mathrm{rob},\psi}$ by backward induction. This uses monotonicity of $\mathcal{T}_V^{\mathrm{rob}}$, not of the surrogate $\mathcal{T}_V^{\mathrm{ub}}$. Therefore

$$\left\|V_h^{\mathrm{rob},\psi} - V_h^{\mathrm{ub},\psi}\right\|_\infty \le \gamma\left\|V_{h+1}^{\mathrm{rob},\psi} - V_{h+1}^{\mathrm{ub},\psi}\right\|_\infty + \eta_h^\psi.$$

Unrolling from $h = 0$ gives

$$\left\|V_0^{\mathrm{rob},\psi} - V_0^{\mathrm{ub},\psi}\right\|_\infty \le \sum_{h=0}^{H-1}\gamma^h\eta_h^\psi.$$

The unprojected version is the same proof with arbitrary states $s \in S$ in place of representatives $s_{h,c}$. □

Bounds for the three ambiguity sets. Because the exact robust penalties are nonnegative,

$$0 \le \eta_h^\psi \le \bar{\eta}_h^\psi, \qquad \bar{\eta}_h^\psi := \sup_{c,a}\left[\Delta_R^{\mathrm{tr}}(s_{h,c},a) + \gamma\Delta_P^{\mathrm{tr}}(s_{h,c},a;V_{h+1}^{\mathrm{ub},\psi})\right].$$

TV. For TV, the bound is

$$\Delta_{\mathrm{TV},\rho}(Q;Z) \le \rho\,\mathrm{osc}(Z).$$

Since $R \in [0, R_{\max}]$ and $0 \le V_{h+1}^{\mathrm{ub},\psi} \le V_{\max}$,

$$\bar{\eta}_h^\psi \le \rho_R R_{\max} + \gamma\rho_P V_{\max}.$$

On a finite support with masses p_i and values z_i, let $A_{\max} = \arg\max_i z_i$ and $A_{\min} = \arg\min_i z_i$. If

$$\rho \le \min\{p(A_{\max}), 1 - p(A_{\min})\},$$

the TV adversary can move exactly ρ mass from maximizers to minimizers without saturating the simplex constraints. In that case

$$\Delta_{\mathrm{TV},\rho}(p;z) = \rho\,\mathrm{osc}(z),$$

so the TV upper bound is tight.

χ^2. For χ^2, the bound is

$$\Delta_{\chi^2,\rho}(Q;Z) \le \sqrt{\rho}\,\sigma_Q(Z).$$

Since any $Z \in [0, M]$ satisfies $\sigma(Z) \le M/2$,

$$\bar{\eta}_h^{\psi} \le \sqrt{\rho_R}\frac{R_{\max}}{2} + \gamma\sqrt{\rho_P}\frac{V_{\max}}{2}.$$

On a finite support, equality in Cauchy–Schwarz is attained by the tilted weights

$$w_i = 1 - \frac{\sqrt{\rho}}{\sigma_p(z)}(z_i - \bar{z}), \qquad \bar{z} = \sum_i p_i z_i,$$

with zero penalty when $\sigma_p(z) = 0$. If these weights are nonnegative, the unconstrained optimizer is feasible and

$$\Delta_{\chi^2,\rho}(p;z) = \sqrt{\rho}\,\sigma_p(z).$$

A sufficient condition is

$$\rho \le \frac{\sigma_p(z)^2}{\|z - \bar{z}\|_\infty^2},$$

with the convention that the condition is vacuous when $\sigma_p(z) = 0$.

W_1. For Wasserstein ambiguity, Kantorovich–Rubinstein duality gives

$$\Delta_{W_1,\rho}(Q;Z) \le \rho\mathrm{Lip}(Z) \le \rho L_Z.$$

For rewards under the absolute-value metric, $r \mapsto r$ is 1-Lipschitz, so the reward bound is at most ρ_R. If $L_{h+1}^{\psi} \ge \mathrm{Lip}(V_{h+1}^{\mathrm{ub},\psi})$ is a valid transition-side envelope, then

$$\bar{\eta}_h^{\psi} \le \rho_R + \gamma\rho_P L_{h+1}^{\psi}.$$

The bound is tight when the envelope is tight, $L_Z = \mathrm{Lip}(Z)$, and the Wasserstein adversary can spend its transport budget along pairs attaining the Lipschitz slope of Z. More generally,

$$\Delta_{W_1,\rho}^{\mathrm{tr}}(Q;Z) - \Delta_{W_1,\rho}(Q;Z) \le \rho(L_Z - \mathrm{Lip}(Z)) + \{\rho\mathrm{Lip}(Z) - \Delta_{W_1,\rho}(Q;Z)\}.$$

The first term is the looseness of the envelope. The second is the looseness of the Lipschitz bound for this support geometry. Combining the envelopes gives

$$\left\|V_0^{\mathrm{rob},\psi} - V_0^{\mathrm{ub},\psi}\right\|_\infty \le \sum_{h=0}^{H-1} \gamma^h \bar{\eta}_h^{\psi}.$$

Thus, with fixed horizon and bounded envelopes, the lower-bound approximation error goes to zero as $\rho_R, \rho_P \to 0$.

A.5 Finite-horizon role of $\mathcal{T}_V^{\mathrm{ub}}$

The finite-horizon recursion does not require a fixed point of $\mathcal{T}_V^{\mathrm{ub}}$. If the transition penalty bound satisfies

$$|\overline{\Delta}_P(s,a;V) - \overline{\Delta}_P(s,a;W)| \le L_B \|V - W\|_\infty,$$

then the 1-Lipschitz property of $x \mapsto [x]_+$ gives

$$|\ell_P^{\mathrm{ub}}(s,a;V) - \ell_P^{\mathrm{ub}}(s,a;W)| \le (1 + L_B)\|V - W\|_\infty.$$

Thus

$$\left\|\mathcal{T}_V^{\mathrm{ub}}V - \mathcal{T}_V^{\mathrm{ub}}W\right\|_\infty \le \gamma(1 + L_B)\|V - W\|_\infty.$$

This bound guarantees a contraction when $\gamma(1 + L_B) < 1$. The finite-horizon theorem does not require this sufficient condition.

B PLUG-IN PENALTY ESTIMATES

The projected Bellman target uses population upper bounds $\overline{\Delta}_R$ and $\overline{\Delta}_P$. These bounds may be fixed in advance or estimated from samples. This section gives conditions under which the empirical estimates retain polynomial concentration around their population counterparts.

All statements concern one visited edge and the fixed representative used in the theorem. Transition penalties are defined on the finite successor alphabet induced by the state abstraction; reward penalties concern the one-dimensional reward law at the same edge. If a plug-in estimate has a slower rate, the corresponding exponents are used in the HOO recursion.

B.1 DEFINITIONS AND BOUNDARY ASSUMPTIONS

Transition penalties. At depth h, the successor bin lies in the finite codebook $\mathcal{C}_{h+1}$. For a fixed edge (v, a), write

$$p_u := \mathbb{P}^o(\psi_{h+1}(S') = u \mid s^{\mathrm{rep}}(v), a), \qquad S_P := \{u : p_u > 0\}, \qquad p_{\min} := \min_{u \in S_P} p_u.$$

The transition plug-in results assume $p_{\min} > 0$, which gives exponential support recovery on the reachable finite successor support. Let

$$V_u := \widetilde{V}^{\mathrm{ub},\psi}(u), \qquad \widehat{V}_t(u)$$

denote the population child value and the child estimate read after t samples of the parent edge. Assumption 5 below formalizes the relation between this parent-edge clock and the child's local sample count. The empirical transition penalties are

$$\widehat{\overline{\Delta}}^{\mathrm{TV}}_{P,t} = \rho_P \operatorname{osc}_{\widehat{S}_t}(\widehat{V}_t), \qquad \widehat{\overline{\Delta}}^{\chi^2}_{P,t} = \sqrt{\rho_P}\, \sigma_{\widehat{p}_t}(\widehat{V}_t),$$

where $\widehat{p}_t$ is the empirical successor-bin histogram and $\widehat{S}_t = \operatorname{supp}(\widehat{p}_t)$. For finite-support Wasserstein ambiguity, successor bins use the representative-state metric

$$d_\psi(u, u') := d(s_{h+1,u}, s_{h+1,u'}),$$

and

$$\widehat{\overline{\Delta}}^{W_1}_{P,t} = \rho_P \max_{\substack{u,u' \in \widehat{S}_t \\ d_\psi(u,u') > 0}} \frac{|\widehat{V}_t(u) - \widehat{V}_t(u')|}{d_\psi(u, u')},$$

with value zero if there is no positive-distance pair.

Reward penalties. The deterministic TV and W_1 reward envelopes

$$\overline{\Delta}^{\mathrm{TV}}_R = \rho_R R_{\max}, \qquad \overline{\Delta}^{W_1}_R = \rho_R$$

are always admissible. If the implementation uses an empirical TV reward span, then the reward law must have enough mass near both support endpoints. Let

$$r_- := \inf \operatorname{supp}(\nu^o_{s^{\mathrm{rep}}(v),a}), \qquad r_+ := \sup \operatorname{supp}(\nu^o_{s^{\mathrm{rep}}(v),a}).$$

We assume a polynomial boundary-mass condition: there exist $c_-, c_+ > 0$, $\kappa_-, \kappa_+ > 0$, and $\varepsilon_0 > 0$ such that, for all $0 < \varepsilon \le \varepsilon_0$,

$$\mathbb{P}(R \le r_- + \varepsilon) \ge c_- \varepsilon^{\kappa_-}, \qquad \mathbb{P}(R \ge r_+ - \varepsilon) \ge c_+ \varepsilon^{\kappa_+}.$$

Let $\kappa_R := \max\{\kappa_-, \kappa_+\}$. Endpoint atoms are a stronger special case and give exponential endpoint recovery. Without this boundary-mass condition, we use the deterministic envelope $\rho_R R_{\max}$ instead of the empirical TV span.

B.2 Basic concentration and stability bounds

Lemma 5 (Exponential tails imply polynomial tails). *Let X_t be a nonnegative random variable. Suppose that, for constants $A, a, q > 0$,*

$$\mathbb{P}(X_t > \varepsilon) \le A \exp\{-at\varepsilon^q\} \qquad \text{for all } t \ge 1,\ \varepsilon > 0.$$

Then, for every $\beta > 0$ and every $0 < \alpha \le \beta/q$, there is a constant C such that

$$\mathbb{P}(X_t > \varepsilon) \le Ct^{-\alpha}\varepsilon^{-\beta}, \qquad \varepsilon > t^{-\alpha/\beta}.$$

Proof. For any $y > 0$, $e^{-ay^q}y^\beta$ is bounded by a finite constant. With $y = t^{1/q}\varepsilon$,

$$\exp\{-at\varepsilon^q\} \le Ct^{-\beta/q}\varepsilon^{-\beta} \le Ct^{-\alpha}\varepsilon^{-\beta},$$

because $\alpha \le \beta/q$. □

Lemma 6 (Finite-alphabet histogram concentration). *Let $S_1, \dots, S_t$ be i.i.d. on a finite set $[M]$ with law p, and let $\widehat{p}_t$ be the empirical histogram. Then*

$$\mathbb{P}(\|\widehat{p}_t - p\|_1 > u) \le 2M \exp\{-2tu^2/M^2\}.$$

If $p_{\min} := \min_{i \in \mathrm{supp}(p)} p_i > 0$, then

$$\mathbb{P}(\mathrm{supp}(\widehat{p}_t) \ne \mathrm{supp}(p)) \le M \exp\{-tp_{\min}\}.$$

Proof. The first bound follows from coordinate-wise Hoeffding and a union bound. The second follows because an atom $i \in \mathrm{supp}(p)$ is missing with probability $(1 - p_i)^t \le e^{-tp_{\min}}$, followed by a union bound over the support. □

For $x \in \mathbb{R}^M$ and $q \in \Delta_M$, write

$$\mathrm{osc}_S(x) := \max_{i \in S} x_i - \min_{i \in S} x_i, \qquad \sigma_q(x) := \left(\sum_i q_i (x_i - q^\top x)^2 \right)^{1/2}.$$

For a finite metric support (S, d_ψ), define

$$\mathrm{Lip}_S(x) := \max_{\substack{i,j \in S \\ d_\psi(i,j) > 0}} \frac{|x_i - x_j|}{d_\psi(i,j)},$$

with value zero if there is no positive-distance pair.

Lemma 7 (Stability of plug-in functionals). *For $u, v \in [0, V_{\max}]^M$ and $q \in \Delta_M$,*

$$|\mathrm{osc}_S(u) - \mathrm{osc}_S(v)| \le 2\|u - v\|_\infty, \qquad |\sigma_q(u) - \sigma_q(v)| \le 2\|u - v\|_\infty.$$

If

$$d_{\min}(S) := \min\{d_\psi(i,j) : i, j \in S,\ i \ne j,\ d_\psi(i,j) > 0\} > 0,$$

then

$$|\mathrm{Lip}_S(u) - \mathrm{Lip}_S(v)| \le \frac{2}{d_{\min}(S)} \|u - v\|_\infty.$$

Moreover, for $p, q \in \Delta_M$ and $v \in [0, V_{\max}]^M$,

$$|\sigma_q(v) - \sigma_p(v)| \le \sqrt{3}\, V_{\max} \|q - p\|_1^{1/2}.$$

Proof. The span and Lipschitz bounds follow directly from the sup-norm perturbation of the value vector. For the standard deviation in the value argument, use $\sigma_q(u) = \|u - q^\top u\|_{L^2(q)}$ and the triangle inequality. For the distribution argument,

$$|q^\top (v \odot v) - p^\top (v \odot v)| \le V_{\max}^2 \|q - p\|_1,$$
$$|(q^\top v)^2 - (p^\top v)^2| \le 2V_{\max}^2 \|q - p\|_1.$$

The variance difference is therefore at most $3V_{\max}^2 \|q - p\|_1$. Taking square roots and using $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x - y|}$ gives the stated bound. □

B.3 CONCENTRATION RATES

Proposition 6 (Plug-in penalty concentration)**.** *Assume the successor codebook is finite,* $p_{\min} > 0$*, and the child-value estimators satisfy*

$$\widehat{V}_t(u) \xrightarrow[\alpha,\beta]{n\to\infty} V_u$$

uniformly over the reachable successor support. Then the empirical penalties below concentrate around their population targets at the rates shown in the table.

Estimate	Condition	Resulting exponents
Transition TV span	finite successor support and $p_{\min} > 0$	preserves (α, β)
Transition finite-support W_1 seminorm	finite abstract support with positive minimum pair distance, or zero-seminorm convention	preserves (α, β)
Transition χ^2 standard deviation	finite successor support	preserves (α, β) if $4\alpha \le \beta$; otherwise use $(\min\{\alpha, \beta/4\}, \beta)$
Reward mean	bounded rewards $R \in [0, R_{\max}]$	any pair with $2\alpha \le \beta$
Reward χ^2 standard deviation	bounded rewards $R \in [0, R_{\max}]$	preserves (α, β) if $4\alpha \le \beta$; otherwise use $(\min\{\alpha, \beta/4\}, \beta)$
Reward TV span	endpoint atoms, or polynomial boundary mass with exponent κ_R	exponential with endpoint atoms; otherwise preserves only if $\kappa_R\alpha \le \beta$, else use $(\min\{\alpha, \beta/\kappa_R\}, \beta)$
Reward W_1	absolute-value reward metric	deterministic envelope ρ_R

For a population W_1 ball on the original state space, the analysis uses a deterministic global envelope $L_Z \ge \mathrm{Lip}(Z)$; an empirical finite-support seminorm is not sufficient.

Proof. For transition TV, support recovery gives $\widehat{S}_t = S_P$ with exponentially high probability. On that event, the empirical and population spans are computed over the same support, so

$$\left|\rho_P\mathrm{osc}_{\widehat{S}_t}(\widehat{V}_t) - \rho_P\mathrm{osc}_{S_P}(V)\right| \le 2\rho_P\|\widehat{V}_t - V\|_{\infty,S_P}.$$

A union bound over the finite support gives the same polynomial order as the child values.

The finite-support W_1 case is the same argument with the finite-support Lipschitz seminorm and the minimum positive distance in Lemma 7. If the abstract support has no positive-distance pair, both seminorms are zero by convention.

For transition χ^2, add and subtract $\sigma_{\widehat{p}_t}(V)$:

$$|\sigma_{\widehat{p}_t}(\widehat{V}_t) - \sigma_p(V)| \le |\sigma_{\widehat{p}_t}(\widehat{V}_t) - \sigma_{\widehat{p}_t}(V)| + |\sigma_{\widehat{p}_t}(V) - \sigma_p(V)|.$$

The first term inherits child-value concentration. The second term is bounded by $\sqrt{3}V_{\max}\|\widehat{p}_t - p\|_1^{1/2}$, whose tail is $\exp(-ct\varepsilon^4)$ by Lemma 6. Lemma 5 gives the $1/4$ exponent-ratio bottleneck.

For rewards, the sample mean has Hoeffding tails. The empirical χ^2 standard deviation is controlled by applying Hoeffding to the first and second moments and using $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$, which again gives an $\exp(-ct\varepsilon^4)$ tail. For the empirical TV reward span, define $\widehat{\Omega}_t := \max_{i\le t} R_i - \min_{i\le t} R_i$ and $\Omega := r_+ - r_-$. If $\Omega - \widehat{\Omega}_t > \delta$, then either no sample falls in $[r_-, r_- + \delta/2]$, or no sample falls in $[r_+ - \delta/2, r_+]$. Therefore, for $\rho_R > 0$,

$$\mathbb{P}\Big(\Big|\rho_R\widehat{\Omega}_t - \rho_R\Omega\Big| > \varepsilon\Big) \le \exp\left\{-tc_-\left(\frac{\varepsilon}{2\rho_R}\right)^{\kappa_-}\right\} + \exp\left\{-tc_+\left(\frac{\varepsilon}{2\rho_R}\right)^{\kappa_+}\right\}.$$

This inequality uses the boundary-mass assumption for $\varepsilon/(2\rho_R) \le \varepsilon_0$. For larger ε, monotonicity of the tail and boundedness of the span extend the bound with adjusted constants. If $\rho_R = 0$, the penalty is identically zero. Lemma 5 then gives polynomial concentration with exponent ratio at most $1/\kappa_R$. Endpoint atoms give exponential recovery with no ε-power bottleneck. Finally, $r \mapsto r$ is 1-Lipschitz under the absolute-value reward metric, so the W_1 reward penalty is the deterministic envelope ρ_R. □

Remark 1 (Rate summary). Transition TV and finite-support W_1 preserve the child-value exponents after support recovery. Empirical χ^2 penalties preserve them when $4\alpha \le \beta$; empirical TV reward spans preserve them when $\kappa_R\alpha \le \beta$. Otherwise the slower pair in Proposition 6 is used. Deterministic envelopes add no sampling error.

C LOCAL HOO SEARCH

At each abstract MCTS node, the planner runs a local HOO search over the continuous action space. This section fixes the local notation and gives the update used by the algorithm.

C.1 LOCAL EDGE ESTIMATES

Fix an abstract node v and an instantiated HOO arm (h, i) with representative action $A_{h,i}$. After k direct pulls of this action, the planner computes

$$\widehat{f}_{h,i,k} := \widehat{Q}_k(v, A_{h,i}) = \left[\bar{r}_k - \widehat{\overline{\Delta}}_{R,k}\right]_+ + \gamma \left[\frac{1}{k}\sum_{j=1}^{k} \widehat{V}(v'_j) - \widehat{\overline{\Delta}}_{P,k}\right]_+ .$$

The estimate is recomputed from the current edge statistics and child values. It is not the average $k^{-1}\sum_{j=1}^{k} \widehat{Q}_j(v, A_{h,i})$ of earlier versions of the backup.

Let $N_{h,i}(t)$ denote the number of direct pulls of $A_{h,i}$ during the first t visits to the local MCTS node. In Algorithms C.1–C.3, $T_{h,i} = N_{h,i}$, so the confidence denominator is the sample size used by $\widehat{Q}_{T_{h,i}}$. A path-aggregated HOO implementation may instead use descendant counts together with the corresponding cell statistic; in either case, the count and statistic must refer to the same observations. For the active records $\mathcal{L}_t$, write

$$T(t) = \sum_{(h,i)\in\mathcal{L}_t} T_{h,i}(t), \qquad w_{h,i}(t) = \frac{T_{h,i}(t)}{T(t)}.$$

The local round t is the number of visits to node v, not the global simulation index in Algorithm 1.

Algorithm C.1 Robust Power-Mean HOO

Require: $\alpha, \beta, b, C, p, \nu_1, \rho$; max depth $\bar{H}$; rounds n
1: Initialize HOO tree $\mathcal{T} \leftarrow \{(0,1)\}$; set child B-values to $+\infty$.
2: Initialize edge statistics for every instantiated HOO arm.
3: **for** $t = 1$ to n **do**
4: $(h, i) \leftarrow$ QUERY(t).
5: Pull action $A_{h,i}$ and observe one simulator outcome (r, v').
6: UPDATEEDGE(h, i, r, v') and recompute $\widehat{f}_{h,i,T_{h,i}} = \widehat{Q}_{T_{h,i}}(v, A_{h,i})$.
7: UPDATEINDICES(t) using (C.1).
8: **end for**
9: **Return** $\widehat{f}_n(p) = \left(\sum_{(h,i)\in\mathcal{L}_n} w_{h,i}(n)\widehat{f}^{\,p}_{h,i,T_{h,i}(n)}\right)^{1/p}$.

The confidence index is

$$U_{h,i}(t) = \widehat{f}_{h,i,T_{h,i}(t)} + C\, t^{b/\beta} T_{h,i}(t)^{-\alpha/\beta} + \nu_1\rho^h. \tag{C.1}$$

Robustness changes the payoff estimate $\widehat{f}_{h,i,T} = \widehat{Q}_T$ but introduces no additional HOO bonus.

Algorithm C.2 Query(t)

$(h, i) \leftarrow (0, 1)$; $P_t \leftarrow \{(h, i)\}$.
while $(h, i) \in \mathcal{T}$ **do**
  if $B_{h+1,2i-1} > B_{h+1,2i}$ **then**
    $(h, i) \leftarrow (h + 1, 2i - 1)$.
  else
    $(h, i) \leftarrow (h + 1, 2i)$.
  end if
  $P_t \leftarrow P_t \cup \{(h, i)\}$.
end while
if $h \leq \bar{H}$ **then**
  Pick any arm $A_{h,i} \in \mathcal{P}_{h,i}$; add (h, i) to $\mathcal{T}$.
  Initialize its edge statistics and set its children's B-values to $+\infty$.
  return (h, i).
else
  $(h, i) \leftarrow (h - 1, \lceil i/2 \rceil)$; **return** (h, i).
end if

Algorithm C.3 UpdateIndices(t)

1: For every instantiated arm (u, j) with $T_{u,j} > 0$, set

$$U_{u,j} \leftarrow \widehat{f}_{u,j,T_{u,j}} + C\, t^{b/\beta} T_{u,j}^{-\alpha/\beta} + \nu_1 \rho^u.$$

2: Process the tree from the leaves upward and set

$$B_{u,j} \leftarrow \min\big\{U_{u,j},\ \max\{B_{u+1,2j-1}, B_{u+1,2j}\}\big\}.$$

Remark 2 (Connection to the MCTS proof). The MCTS induction below establishes concentration of $\widehat{Q}_k(v, a)$ at a fixed action. The local HOO layer is then handled under the standard concentration conditions stated in Assumption 4; the power-mean step is proved separately in Theorem 1.

D HOO GEOMETRY

Let $f^\star = \sup_{a \in A} f(a)$ and define the gap of a HOO cell by

$$\Delta_{h,i} := f^\star - \sup_{a \in \mathcal{P}_{h,i}} f(a), \qquad I_h := \{(h, i) : \Delta_{h,i} \leq 2\nu_1 \rho^h\}.$$

The next two statements are standard consequences of the HOO geometry and smoothness assumptions.

Lemma 1 (Cell smoothness; Lemma 3 in Bubeck et al. (2011))**.** *Under Assumptions 1 and 2, if $\Delta_{h,i} \leq c\nu_1\rho^h$ for $c \geq 0$, then every action in $\mathcal{P}_{h,i}$ is $\max\{2c, c + 1\}\nu_1\rho^h$-optimal.*

Lemma 2 (Near-optimal cells; Theorem 6 in Bubeck et al. (2011))**.** *For every packing exponent d' strictly above the near-optimality dimension, there is a constant C such that*

$$|I_h| \leq C(\nu_2 \rho^h)^{-d'} \qquad \textit{for all } h \geq 0.$$

E STANDARD ENHANCED-HOO CONCENTRATION

The local search analysis uses the following standard concentration result.

Theorem 1 (Enhanced-HOO concentration; Mao et al. (2020))**.** *Suppose the bounded payoff process and the HOO update satisfy the hypotheses of Theorem 2 in Mao et al. (2020). In the notation used here, let the fixed-action tail exponents be (α, β) and let b be the polynomial exploration parameter. If*

$$\alpha(1 - \alpha/\beta) \leq b < \alpha, \qquad b > 3, \qquad \rho^{\bar{H}} < n^{-\alpha/\beta},$$

then the arithmetic HOO estimate has polynomial concentration around $f^\star$ with

$$\beta' = \frac{b-3}{2}, \qquad \alpha' = \frac{1 - b/\alpha}{1 + d' + \beta/\alpha}\,\beta',$$

where d' is any packing exponent above the near-optimality dimension.

When the cutoff $\bar{H}$ depends on the planning budget, we require the concentration constant in Theorem 1 to remain uniform over the chosen cutoffs.

Assumption 4 (Standard HOO conditions)**.** *At every MCTS node, the local HOO process satisfies the geometry, smoothness, and concentration hypotheses of Theorem 1. Enumerate action records in their order of creation, and let $\mathcal{L}_n$ be the active set after n local visits. For record j, let $N_j(n)$ be its sample count and let*

$$T(n) = \sum_{j \in \mathcal{L}_n} N_j(n), \qquad w_j = \frac{N_j(n)}{T(n)}.$$

We assume the following bounds hold uniformly over nodes and cutoffs:

1. *The edge errors $e_{j,k} = \widehat{f}_{j,k} - f(A_j)$ have (α, β) tails, conditional on the information available when A_j is created.*
2. *The arithmetic estimate $A_n = \sum_{j \in \mathcal{L}_n} w_j \widehat{f}_{j,N_j(n)}$ satisfies $A_n \xrightarrow[\alpha',\beta']{n\to\infty} f^\star$, with (α', β') from Theorem 1.*
3. *For constants $K_0 > 0$, $0 < c_0 \le 1$, and $0 \le \kappa < 1$, $|\mathcal{L}_n| \le K_0 n^\kappa$ and $T(n) \ge c_0 n$. If*

$$a_E := \alpha(1-\kappa) - 1, \qquad b_E := \beta + p(\alpha - 1), \qquad \zeta := \alpha'/\beta',$$

then $a_E > 0$ and $\zeta b_E \le a_E$.

The first two items are the standard HOO concentration conditions. The third controls the additional power-mean error as the local tree grows.

F Power-mean concentration

We use the power-mean decomposition of Dam (2025) and make the dependence on a growing active set explicit. The goal is to compare the power mean returned by Algorithm C.1 with the arithmetic HOO estimate in Assumption 4.

Lemma 1 (Error at an adaptive sample count)**.** *Suppose $\alpha > 1$ and*

$$\mathbb{P}(|e_k| > u) \le C k^{-\alpha} u^{-\beta} \qquad \text{for } u > k^{-\alpha/\beta}.$$

For any integer-valued random variable $N \ge 1$, any deterministic $m \ge 1$, and any $u > 0$,

$$\mathbb{P}(|e_N| > u,\ N \ge m) \le C_\alpha u^{-\beta} m^{1-\alpha}.$$

No independence between N and the error process is required.

Proof. After increasing C if necessary, the same inequality holds below the stated threshold because its right-hand side is then at least one. A union bound over $k \ge m$ gives

$$\mathbb{P}(|e_N| > u,\ N \ge m) \le C u^{-\beta} \sum_{k=m}^{\infty} k^{-\alpha} \le C_\alpha u^{-\beta} m^{1-\alpha}.$$

□

Lemma 2 (Power-mean comparison)**.** *Let $w_j \ge 0$ with $\sum_j w_j = 1$, let $x_j \ge 0$, and let $0 \le q_j \le f^\star$. For finite $p \ge 1$, define*

$$M_p = \Big(\sum_j w_j x_j^p\Big)^{1/p}, \qquad A = \sum_j w_j x_j, \qquad E_p = \Big(\sum_j w_j |x_j - q_j|^p\Big)^{1/p}.$$

Then

$$|M_p - f^\star| \le \max\{|A - f^\star|, E_p\}.$$

Proof. If $M_p \le f^\star$, the power-mean inequality gives $A \le M_p$, and hence $f^\star - M_p \le f^\star - A$. If $M_p > f^\star$, Minkowski's inequality gives

$$M_p \le \Big(\sum_j w_j q_j^p\Big)^{1/p} + E_p \le f^\star + E_p.$$

The argument is pointwise and therefore also applies when the weights are random. □

Lemma 3 (Weighted estimation error)**.** *Under item 1 of Assumption 4, suppose $\alpha > 1$ and define*

$$E_{p,n} = \left(\sum_{j \in \mathcal{L}_n} \frac{N_j(n)}{T(n)} |e_{j,N_j(n)}|^p\right)^{1/p}.$$

For every finite $p \ge 1$ and $u > 0$,

$$\mathbb{P}(E_{p,n} > u) \le C_E n^{-a_E} u^{-b_E}, \qquad a_E = \alpha(1-\kappa) - 1, \quad b_E = \beta + p(\alpha - 1).$$

The active set and the sample counts may be adaptive.

Proof. Let M bound the edge errors and let $K = \min\{n, \lceil K_0 n^\kappa \rceil\}$. The claim is trivial for $u > M$, so assume $0 < u \le M$ and set

$$m = \left\lfloor \frac{c_0 n u^p}{2KM^p} \right\rfloor.$$

Records with $N_j(n) \le m$ contribute at most $u^p/2$ to $E_{p,n}^p$. Thus $E_{p,n} > u$ implies that some record has both $N_j(n) > m$ and $|e_{j,N_j(n)}| > u/2^{1/p}$. Condition on the information available when that record is created, apply Lemma 1, and then average over the conditioning. A union bound over at most K records gives, for $m \ge 1$,

$$\mathbb{P}(E_{p,n} > u) \le CKu^{-\beta} m^{1-\alpha} \le C' K^\alpha n^{1-\alpha} u^{-\{\beta + p(\alpha-1)\}}.$$

If $m = 0$, then $u^p < 2KM^p/(c_0 n)$. Hence

$$K^\alpha n^{1-\alpha} u^{-b_E} \ge c\, n \left(\frac{K}{n}\right)^{1-\beta/p} \ge c,$$

where the last inequality follows from $1 \le K \le n$ and $\beta/p > 0$; increasing the constant therefore covers this case as well. Finally, $K^\alpha n^{1-\alpha} \le C'' n^{-a_E}$. □

Theorem 1 (Power-mean concentration)**.** *Assume the HOO geometry and smoothness conditions, Assumption 4, and*

$$\alpha(1 - \alpha/\beta) \le b < \alpha, \qquad b > 3, \qquad \rho^{\bar{H}} < n^{-\alpha/\beta}.$$

For a finite $p \ge 1$, suppose $\alpha \le \beta / \max\{2, p\}$. Then

$$\widehat{f}_n(p) \xrightarrow[\alpha', \beta']{n \to \infty} f^\star, \qquad \alpha' = \frac{1 - b/\alpha}{1 + d' + \beta/\alpha} \frac{b-3}{2}, \qquad \beta' = \frac{b-3}{2}.$$

Proof. Apply Lemma 2 with $q_j = f(A_j)$ and the active-set weights w_j. The arithmetic term has the standard HOO tail by item 2 of Assumption 4. Lemma 3 bounds the weighted estimation error with exponents (a_E, b_E). Since $b_E \ge \beta'$ and $a_E \ge (\alpha'/\beta') b_E$, for $u > n^{-\alpha'/\beta'}$,

$$n^{-a_E} u^{-b_E} \le n^{-\alpha'} u^{-\beta'}.$$

A union bound completes the proof. □

As in Dam et al. (2025), the robust Bellman estimate is recomputed from the current child values. Once the edge estimate has the required fixed-action tail, the HOO and power-mean arguments above apply without an additional robustness bonus.

G MCTS CONVERGENCE

We now combine the preceding results by backward induction on the planning depth. The leaf estimate supplies the base case. At an internal node, concentration of the child values yields concentration of each robust Q-estimate; the local HOO result then propagates this bound through action selection and the power-mean backup.

G.1 PARAMETER SCHEDULE

We first state the schedule for deterministic penalty envelopes. With plug-in penalties, the exponents from Proposition 6 are substituted before the schedule is applied. Here ρ denotes the geometric HOO parameter, not an ambiguity radius.

At depth $h < H$, choose the HOO cutoff so that

$$\rho^{\bar{H}_h(n)} < n^{-\alpha_{h+1}/\beta_{h+1}}. \tag{G.1}$$

The polynomial bonus satisfies

$$3 < b_{h+1} < \alpha_{h+1}, \qquad \alpha_{h+1}\left(1 - \frac{\alpha_{h+1}}{\beta_{h+1}}\right) \le b_{h+1}. \tag{G.2}$$

This is Theorem 2 of Mao et al. (2020) after the change of notation in Theorem 1. For a finite power $p \ge 1$, we use the sufficient condition

$$\alpha_{h+1} \le \frac{\beta_{h+1}}{\max\{2, p\}}. \tag{G.3}$$

This condition also implies $2\alpha_{h+1} \le \beta_{h+1}$, as required by the one-step concentration argument.

Let d'_h be any packing exponent above the near-optimality dimension at depth h. The HOO recursion is

$$\beta_h = \frac{b_{h+1} - 3}{2}, \qquad \alpha_h = \frac{1 - b_{h+1}/\alpha_{h+1}}{1 + d'_h + \beta_{h+1}/\alpha_{h+1}} \frac{b_{h+1} - 3}{2}. \tag{G.4}$$

The leaf exponents satisfy

$$2\alpha_H \le \beta_H, \qquad \beta_H > 1, \tag{G.5}$$

and root convergence in expectation additionally requires

$$\beta_0 > 1. \tag{G.6}$$

The latter is equivalent to $b_1 > 5$. The root rate is $\zeta = \alpha_0/\beta_0 \in (0, 1/2)$.

Component	Condition / consequence
HOO cutoff	$\rho^{\bar{H}_h(n)} < n^{-\alpha_{h+1}/\beta_{h+1}}$, with constants uniform over the chosen cutoffs.
Polynomial bonus	$3 < b_{h+1} < \alpha_{h+1}$ and $\alpha_{h+1}(1 - \alpha_{h+1}/\beta_{h+1}) \le b_{h+1}$.
Power mean	Finite $p \ge 1$ and $\alpha_{h+1} \le \beta_{h+1}/\max\{2, p\}$.
Depth recursion	$\beta_h = (b_{h+1} - 3)/2$ and $\alpha_h/\beta_h = (1 - b_{h+1}/\alpha_{h+1})/(1 + d'_h + \beta_{h+1}/\alpha_{h+1})$.
Local tree growth	Under Assumption 4, $\lvert\mathcal{L}_n\rvert \le K_0 n^\kappa$, $T(n) \ge c_0 n$, and $\zeta[\beta + p(\alpha - 1)] \le \alpha(1 - \kappa) - 1$.
Leaf / root	$2\alpha_H \le \beta_H$, $\beta_H > 1$; $\beta_0 > 1$ gives mean error $O(n^{-\alpha_0/\beta_0})$ at the root.

Table G.2: Conditions used in the depthwise induction.

The conditions above are jointly feasible at any finite planning depth. The following construction gives one explicit choice.

Proposition 7 (A feasible schedule for finite p)**.** *Fix a finite depth $H \geq 1$, a finite $p \geq 1$, and packing exponents $d'_h \geq 0$. Choose $q_H \geq \max\{2, p\}$ and define backwards*

$$q_h = q_{h+1}^2 + (1 + d'_h)q_{h+1}.$$

Starting from $\beta_0 = 2$ and $\alpha_0 = 2/q_0$, define forwards

$$b_{h+1} = 2\beta_h + 3, \qquad \alpha_{h+1} = \frac{b_{h+1}}{1 - 1/q_{h+1}}, \qquad \beta_{h+1} = q_{h+1}\alpha_{h+1}.$$

These parameters satisfy (G.2)–(G.6)*. In a binary HOO tree, let $D_\rho = \log 2/\log(1/\rho)$. If $q_H \geq \max\{8, p, 2D_\rho\}$ and the minimal cutoff below is used, the local tree-growth condition in Assumption 4 also holds.*

Proof. The construction gives $\beta_h/\alpha_h = q_h$ and $q_h \geq q_{h+1} \geq \max\{2, p\}$. Moreover, $b_{h+1} = \alpha_{h+1}(1 - 1/q_{h+1})$ and $b_{h+1} \geq 7$. Finally,

$$\frac{1 - b_{h+1}/\alpha_{h+1}}{1 + d'_h + \beta_{h+1}/\alpha_{h+1}} = \frac{1/q_{h+1}}{1 + d'_h + q_{h+1}} = \frac{1}{q_h},$$

which verifies the recursion and the bonus and power conditions.

For the growth condition, choose

$$\bar{H}_h(n) = 1 + \left\lceil \frac{\log n}{q_{h+1}\log(1/\rho)} \right\rceil.$$

The binary tree then contains at most $8n^{D_\rho/q_{h+1}}$ records. Write $q = q_{h+1}$ and $\kappa = D_\rho/q \leq 1/2$. Since the incoming exponents satisfy $\beta = q\alpha$ and $\alpha \geq 7$,

$$\frac{\alpha(1-\kappa) - 1}{\beta + p(\alpha - 1)} \geq \frac{5}{28q} \geq \frac{1}{q^2} \geq \frac{1}{q^2 + (1 + d'_h)q} = \frac{\alpha_h}{\beta_h}.$$

The middle inequality uses $q \geq 8$. In the direct-edge expansion above, an internal record contributes only its initial pull outside the active-leaf sum. Since the total number of records is at most $8n^\kappa$, we have $T(n) \geq n - 8n^\kappa \geq n/2$ for all sufficiently large n; the finitely many smaller budgets are absorbed into the constants. □

G.2 Sampling and one-step backups

Successor counts.

Assumption 5 (Sampling at successor counts)**.** *Fix a parent edge after n local samples. Let $\mathcal{C}_n = \sigma(N_1^n, \ldots, N_M^n)$ be its successor histogram, and let $K_m(n)$ be the sample count of the child estimate read by that edge. If a bounded child process has uniform (α, β) tails at every fixed local count, then on $N_m^n > 0$,*

$$\mathbb{P}\left(|\widehat{V}_{m,K_m(n)} - V_m| > u \mid \mathcal{C}_n\right) \leq C(N_m^n)^{-\alpha}u^{-\beta}, \qquad u > 0.$$

The constant is uniform over the edges and child processes used in the induction. A child with $N_m^n = 0$ has zero weight in the successor average.

The conditioning is only on the successor histogram, not on a routing history that may already reveal child returns.

Lemma 13 (Independent local sampling)**.** *Assumption 5 holds when each parent edge has separate child copies with fixed representatives, independent simulator streams across subtrees, and child policies driven only by their own local visits and data. In this construction, $K_m(n) = N_m^n$.*

Proof. Preassign an infinite stream of one-step simulator outcomes to each parent edge and, independently, all random variables used in each child subtree. Pausing a child between visits does not change its first k local updates. Hence the process $(\widehat{V}_{m,k})_{k\geq 1}$ is independent of the parent successor histogram at fixed n. Conditional on $N_m^n = k > 0$, the readout has the same law as $\widehat{V}_{m,k}$, and the fixed-count tail bound applies. □

If an implementation shares a child statistic across several parent edges, Assumption 5 must be verified for that shared process.

G.3 LEAF AND EDGE CONCENTRATION

Proof of Proposition 5. Let $\mu_H = \widetilde{V}^{\mathrm{ub},\psi}(v_H)$. If LEAFVALUE returns the baseline exactly, then $\widehat{V}_n(v_H) = \mu_H$. Otherwise,

$$\widehat{V}_n(v_H) = \frac{1}{n}\sum_{j=1}^{n} Y_j, \qquad 0 \le Y_j \le V_{\max}, \qquad \mathbb{E}[Y_j] = \mu_H,$$

where the Y_j are i.i.d. rollout returns conditional on the fixed leaf representative. Hoeffding's inequality gives

$$\mathbb{P}\left(|\widehat{V}_n(v_H) - \mu_H| > \varepsilon\right) \le 2\exp\{-2n\varepsilon^2/V_{\max}^2\}.$$

Lemma 5 with $q = 2$ converts this bound to

$$\mathbb{P}\left(|\widehat{V}_n(v_H) - \mu_H| > \varepsilon\right) \le C_H n^{-\alpha_H}\varepsilon^{-\beta_H}, \qquad \varepsilon > n^{-\alpha_H/\beta_H},$$

for every $\beta_H > 0$ and $0 < \alpha_H \le \beta_H/2$. □

Lemma 14 (Concentration under random successor counts)**.** *Let $S_1, \dots, S_n$ be i.i.d. successor-bin indices on $[M]$ with law p, and let $N_m^n = \#\{i \le n : S_i = m\}$. Suppose bounded child processes satisfy (α, β) concentration at each fixed local count, with $2\alpha \le \beta$, and satisfy Assumption 5. Define*

$$\widehat{T}_n = \sum_{m=1}^{M} \frac{N_m^n}{n}\widehat{V}_{m,K_m(n)}, \qquad T = \sum_{m=1}^{M} p_m V_m.$$

Then $\widehat{T}_n \xrightarrow[\alpha,\beta]{n\to\infty} T$. A bounded i.i.d. reward mean satisfies the same bound.

Proof. Decompose

$$|\widehat{T}_n - T| \le \underbrace{\left|\sum_m (N_m^n/n - p_m)V_m\right|}_{A_n} + \underbrace{\sum_m \frac{N_m^n}{n}|\widehat{V}_{m,K_m(n)} - V_m|}_{B_n}.$$

The first term is the deviation of the mean of the bounded i.i.d. variables V_{S_i}, so Hoeffding's inequality and Lemma 5 give the required tail.

For B_n, condition on $\mathcal{C}_n$. If $B_n > u$, then for some m with $N_m^n > 0$,

$$|\widehat{V}_{m,K_m(n)} - V_m| > \frac{nu}{MN_m^n}.$$

Assumption 5 therefore yields

$$\begin{aligned}\mathbb{P}(B_n > u \mid \mathcal{C}_n) &\le CM^\beta u^{-\beta} n^{-\beta} \sum_{m:N_m^n>0} (N_m^n)^{\beta-\alpha} \\ &\le CM^{\beta+1} n^{-\alpha} u^{-\beta},\end{aligned}$$

since $\beta \ge \alpha$ and $N_m^n \le n$. Taking expectations and combining the two terms proves the claim. The argument does not require a uniform lower bound on the positive successor probabilities. □

Lemma 15 (Positive part)**.** *Let $\widehat{\mu}_n \xrightarrow[\alpha,\beta]{n\to\infty} \mu$ and $\widehat{\overline{\Delta}}_n \xrightarrow[\alpha,\beta]{n\to\infty} \overline{\Delta}$. Then*

$$\widehat{L}_n = [\widehat{\mu}_n - \widehat{\overline{\Delta}}_n]_+ \quad \xrightarrow[\alpha,\beta]{n\to\infty} \quad L = [\mu - \overline{\Delta}]_+.$$

If the penalty bound is deterministic, only concentration of $\widehat{\mu}_n$ is needed.

Proof. Because $x \mapsto [x]_+$ is 1-Lipschitz,

$$|\widehat{L}_n - L| \le |\widehat{\mu}_n - \mu| + |\widehat{\overline{\Delta}}_n - \overline{\Delta}|.$$

For $\varepsilon > 2n^{-\alpha/\beta}$, use a union bound at threshold $\varepsilon/2$. For $n^{-\alpha/\beta} < \varepsilon \le 2n^{-\alpha/\beta}$, the desired inequality follows from $1 \le 2^\beta n^{-\alpha}\varepsilon^{-\beta}$. □

Lemma 16 (Concentration of the robust Q-estimate)**.** *Fix an edge (v, a) at depth h. Let $\mu_R = \mathbb{E}[X_1]$ and $\mu_P = \sum_m p_m V_m$. Let $\overline{\Delta}_R, \overline{\Delta}_P$ be the population penalty bounds in the projected target, and let $\widehat{\overline{\Delta}}_{R,n}, \widehat{\overline{\Delta}}_{P,n}$ be the estimates used by the algorithm. Assume:*

1. *the child values satisfy fixed-count concentration and Assumption 5;*
2. *the two penalty estimates are deterministic or satisfy $\widehat{\overline{\Delta}}_{R,n} \xrightarrow[\alpha,\beta]{n\to\infty} \overline{\Delta}_R$ and $\widehat{\overline{\Delta}}_{P,n} \xrightarrow[\alpha,\beta]{n\to\infty} \overline{\Delta}_P$;*
3. *$2\alpha \le \beta$.*

Here (α, β) denotes the exponents used after accounting for any plug-in penalty estimate. Define

$$\widehat{Q}_n^{\mathrm{ub}}(v, a) = [\bar{X}_n - \widehat{\overline{\Delta}}_{R,n}]_+ + \gamma[\widehat{T}_n - \widehat{\overline{\Delta}}_{P,n}]_+,$$

and

$$\widetilde{Q}^{\mathrm{ub},\psi}(v, a) = [\mu_R - \overline{\Delta}_R]_+ + \gamma[\mu_P - \overline{\Delta}_P]_+.$$

Then

$$\widehat{Q}_n^{\mathrm{ub}}(v, a) \xrightarrow[\alpha,\beta]{n\to\infty} \widetilde{Q}^{\mathrm{ub},\psi}(v, a).$$

Proof. Lemma 14 gives concentration of $\bar{X}_n$ around μ_R and of $\widehat{T}_n$ around μ_P. By assumption, the penalty estimates satisfy the same tail bound. Apply Lemma 15 to the reward and transition terms, then combine them with a union bound. □

Corollary 3 (Penalty bounds covered by Lemma 16)**.** *The lemma applies in each of the following cases.*

1. ***Deterministic envelopes.*** *Use fixed bounds such as*

$$\overline{\Delta}_R^{\mathrm{TV}} = \rho_R R_{\max}, \quad \overline{\Delta}_R^{\chi^2} = \sqrt{\rho_R} R_{\max}/2, \quad \overline{\Delta}_R^{W_1} = \rho_R,$$

   and

$$\overline{\Delta}_P^{\mathrm{TV}} = \rho_P V_{\max}, \quad \overline{\Delta}_P^{\chi^2} = \sqrt{\rho_P} V_{\max}/2, \quad \overline{\Delta}_P^{W_1} = \rho_P L_V.$$

   These bounds introduce no sampling error.
2. ***Plug-in estimates.*** *Use empirical TV, χ^2, or finite-support W_1 penalties under Proposition 6, using the exponents stated there.*
3. ***Original-space W_1.*** *For a population Wasserstein ball on the continuous state space, use a deterministic global Lipschitz envelope rather than an empirical support seminorm.*

G.4 Backward induction and root convergence

Theorem 5 (Depthwise MCTS concentration)**.** *Assume the HOO geometry and lower-bound action-value smoothness conditions in the main paper, together with Assumptions 4 and 5 at each internal node. Assume all constants are uniform over representatives and budget-dependent cutoffs. At every depth, let the parameters satisfy the schedule above and let the penalty bounds be covered by Corollary 3.*

Let (α_h, β_h) be defined by

$$\beta_h = \frac{b_{h+1} - 3}{2}, \qquad \alpha_h = \frac{1 - b_{h+1}/\alpha_{h+1}}{1 + d'_h + \beta_{h+1}/\alpha_{h+1}} \frac{b_{h+1} - 3}{2},$$

with leaf exponents supplied by Proposition 5. Then every visited node $v = (h, c)$ satisfies

$$\widehat{V}_n(v) \xrightarrow[\alpha_h,\beta_h]{n\to\infty} \widetilde{V}^{\mathrm{ub},\psi}(v).$$

For $h < H$, every instantiated action also satisfies

$$\widehat{Q}_n(v, a) \xrightarrow[\alpha_{h+1},\beta_{h+1}]{n\to\infty} \widetilde{Q}^{\mathrm{ub},\psi}(v, a),$$

where a slower plug-in rate replaces $(\alpha_{h+1}, \beta_{h+1})$ before the recursion is applied.

Proof. We argue by backward induction. At depth H, Proposition 5 gives the required bound for every fixed-representative leaf.

Now fix $h < H$ and suppose all child estimators have $(\alpha_{h+1}, \beta_{h+1})$ concentration at their local sample counts. Assumption 5 transfers these bounds to the counts read by a parent edge; Lemma 13 verifies this condition for independent local child copies. Lemma 16 then gives, for every instantiated edge,

$$\widehat{Q}_n(v,a) \xrightarrow[\alpha_{h+1},\beta_{h+1}]{n\to\infty} \widetilde{Q}^{\mathrm{ub},\psi}(v,a).$$

When a plug-in estimate yields slower exponents, they replace $(\alpha_{h+1}, \beta_{h+1})$ in this application of Lemma 16.

Define the local payoff function

$$f_v(a) = \widetilde{Q}^{\mathrm{ub},\psi}(v,a), \qquad f_v^\star = \sup_{a\in A} f_v(a) = \widetilde{V}^{\mathrm{ub},\psi}(v).$$

Theorem 1, under the standard HOO conditions in Assumption 4, now yields

$$\widehat{V}_n(v) \xrightarrow[\alpha_h,\beta_h]{n\to\infty} f_v^\star.$$

This is the induction claim at depth h. Iterating to the root proves the theorem. □

Root convergence in expectation.

Lemma 17 (Polynomial tails imply mean convergence)**.** *Let $0 \le X_n \le M$ and suppose that, for some $\alpha > 0$, $\beta > 1$, and $c > 0$,*

$$\mathbb{P}(X_n > \varepsilon) \le cn^{-\alpha}\varepsilon^{-\beta}, \qquad \varepsilon > n^{-\alpha/\beta}.$$

Then $\mathbb{E}[X_n] \le Cn^{-\alpha/\beta}$ for a constant C depending only on c, β, M.

Proof. Let $a_n = n^{-\alpha/\beta}$. If $a_n \ge M$, then $\mathbb{E}[X_n] \le a_n$. Otherwise,

$$\begin{aligned}\mathbb{E}[X_n] &= \int_0^M \mathbb{P}(X_n > \varepsilon)\, d\varepsilon \\ &\le a_n + cn^{-\alpha}\int_{a_n}^M \varepsilon^{-\beta}\, d\varepsilon \\ &\le \left(1 + \frac{c}{\beta - 1}\right) n^{-\alpha/\beta}.\end{aligned}$$

□

Theorem 6 (Root convergence in expectation to the projected lower-bound target)**.** *Let $v_0 = (0, \psi_0(s_0))$ and let (α_0, β_0) be the root exponents from Theorem 5. If $\beta_0 > 1$, then*

$$\left|\mathbb{E}[\widehat{V}_n(v_0)] - \widetilde{V}^{\mathrm{ub},\psi}(v_0)\right| = O(n^{-\zeta}), \qquad \zeta = \alpha_0/\beta_0.$$

Proof. Set $\mu_0 = \widetilde{V}^{\mathrm{ub},\psi}(v_0)$ and $E_n = |\widehat{V}_n(v_0) - \mu_0|$. The root case of Theorem 5 gives

$$\mathbb{P}(E_n > \varepsilon) \le c_0 n^{-\alpha_0}\varepsilon^{-\beta_0}, \qquad \varepsilon > n^{-\alpha_0/\beta_0}.$$

Since $0 \le E_n \le V_{\max}$, Lemma 17 gives $\mathbb{E}[E_n] = O(n^{-\alpha_0/\beta_0})$. Finally,

$$\left|\mathbb{E}[\widehat{V}_n(v_0)] - \mu_0\right| \le \mathbb{E}[E_n].$$

□

Remark 3 (Scope of the root result). The theorem controls the root value of the projected lower-bound problem. A guarantee for the returned action would additionally require uniform control over the instantiated root actions and an action-space approximation argument.

H DETAILED EXPERIMENTAL SETUP

This appendix provides implementation details for the controlled-mismatch experiments, including the planning depth, HOO configuration, baseline settings, ambiguity-radius protocol, and reward preprocessing. The theoretical abstraction in Section 3 is fixed-representative; the experimental planner uses the practical representative convention described below.

Evaluation statistics. Each reported curve is computed over five independent random seeds. For each method, environment, and execution-noise level, we report the mean episodic return across seeds. Shaded regions in Figures 1, 2, and the ablation figures denote $\pm$ one standard error of the mean, computed as the empirical standard deviation across seeds divided by $\sqrt{5}$.

Continuous-action classic-control variants. CartPole and Acrobot use one-dimensional continuous-action wrappers: CartPole applies a force in $[-10, 10]$ N and Acrobot applies a torque in $[-1, 1]$. Pendulum uses its native continuous torque action.

MCTS depth and HOO partition depth. The MCTS planning horizon H and the local HOO partition depth $\bar{H}$ are distinct. For ROBUST-POWER-HOOT, the environment-specific horizons are Hopper ($H = 50$), Walker2d ($H = 50$), Ant ($H = 40$), Humanoid ($H = 30$), and $H = 50$ on CartPole, Pendulum, and Acrobot. The broad comparison in the main curves used $H = 100$ for the closest nominal baseline, Stochastic-Power-HOOT; the matched-depth analysis in Appendix H.1 uses the robust planner's environment-specific H. All methods use 500 simulations per planning step.

Method/configuration	MCTS horizon H	HOO depth $\bar{H}$
Robust-Power-HOOT	environment-specific	10
Stochastic-Power-HOOT, broad comparison	100	10
Stochastic-Power-HOOT, matched rerun	environment-specific	10
HOOT	environment-specific	10
POLY-HOOT	environment-specific	8
Progressive Widening	environment-specific	–
UCT-Discretize	environment-specific	–

Table H.3: MCTS-horizon and HOO partition-depth configurations. The environment-specific MuJoCo horizons are $50/50/40/30$ for Hopper/Walker2d/Ant/Humanoid and 50 on classic control.

State abstraction and representatives. The finite-sample theorem assumes a fixed finite abstraction ψ_h and fixed representatives $s_{h,c}$. The experimental implementation uses the common first-visit representative variant: when a new abstract node is created, the first state that reaches the node is stored as $\bar{s}(v)$ and all later simulator queries at that node use the same frozen representative. Consequently, conditional on the node-creation history, each existing node–action edge has a stationary one-step sampling law; however, the representatives and the induced planning target are random. The theorem therefore applies directly only to the fixed-representative idealization. If an online covering or quantization map is used, its merge radius $r_{h,n}$ can be inserted into Corollary 2; smaller radii reduce representative bias but increase the number of nodes. The logged runs do not contain cell or merge radii, so the numerical distance between the first-visit and fixed-representative targets remains unmeasured.

Dimension-adaptive HOO parameters. HOO-based methods use parameters scaled with the action dimension m:

m	α_{HOO}	ξ	η	ρ_{HOO}	ν_1
1	5	20	0.5	2^{-2}	4
2–3	3	15	0.4	$2^{-1.5/m}$	$3m$
4–8	2	10	0.3	$2^{-1/m}$	$2m$
> 8	1.5	8	0.25	$2^{-0.8/m}$	$1.8m$

Table H.4: Dimension-adaptive HOO parameters. α_{HOO} controls exploration decay, ξ scales the polynomial bonus, η adjusts the bonus exponent, ρ_{HOO} determines the refinement rate, and ν_1 sets the covering constant.

The notation α_{HOO} in Table H.4 is an implementation hyperparameter and is distinct from the polynomial-tail exponent α used in the convergence analysis. The HOO tree depth is fixed at $\bar{H} = 10$ for HOOT, Stochastic-Power-HOOT, and the robust variants, and at $\bar{H} = 8$ for POLY-HOOT. These finite-budget settings are distinct from the growing cutoff used in the asymptotic analysis.

Baseline method configurations. All baseline methods use the same evaluation protocol (low-noise planning, high-noise execution) with method-specific implementation details:

- *HOOT* (Mansley et al., 2011): Standard hierarchical optimistic optimization with the dimension-adaptive parameters from Table H.4. Uses optimistic bonuses without power-mean backups.
- *POLY-HOOT* (Mao et al., 2020): Polynomial exploration bonuses with the environment-specific MCTS depth and HOO depth $\bar{H} = 8$. Implements the enhanced bonus structure that provides non-asymptotic convergence guarantees.
- *Stochastic-Power-HOOT* (Dam, 2025): Combines power-mean backups with polynomial bonuses. The broad comparison uses $H = 100$ and $\bar{H} = 10$; the matched-depth analysis uses the environment-specific H with the same $\bar{H}$. This is the closest non-robust baseline to our method.
- *UCT-Discretize*: Standard UCT applied to discretized action spaces. Grid size is 10 actions for $m = 1$, and approximately $10^{1/m}$ actions per dimension for $m > 1$ to keep the branching factor tractable. The UCB exploration constant is $c = 10$.
- *Progressive Widening (PW)* (Couëtoux et al., 2011): Action children are expanded as a function of the parent visit count N. We use $K = \max(5, \lfloor 0.5\sqrt{N} \rfloor)$ for $m \leq 3$ and $K = \max(3, \lfloor 0.3\sqrt{N} \rfloor)$ for $m > 3$.

Noise configuration and evaluation protocol. Stochasticity is injected into actions, dynamics, and observations:

$$
\begin{aligned}
a_{\text{actual}} &= a_{\text{planned}} + \mathcal{N}(0, \sigma^2_{\text{action}} I), \\
s_{t+1} &= f(s_t, a_{\text{actual}}) + \mathcal{N}(0, \sigma^2_{\text{dynamics}} I), \\
o_t &= s_t + \mathcal{N}(0, \sigma^2_{\text{obs}} I),
\end{aligned}
$$

with $\sigma_{\text{action}} = \sigma_{\text{dynamics}} = \sigma_{\text{obs}} = \sigma$. Planning uses a low-noise simulator with $\sigma_{\text{plan}} \in \{0.02, 0.03, 0.04, 0.05\}$, while execution uses $\sigma_{\text{exec}} \in \{0.1, 0.2, 0.3, 0.4\}$. This creates a controlled stochastic mismatch. It does not certify that the execution law lies inside the TV, χ^2, or Wasserstein ambiguity ball used by the robust backup.

Ambiguity radius selection. In the main figures, each divergence type uses a single radius ρ across all environments within the same family. This avoids per-environment tuning in the main comparison:

Environment Family	χ^2	**TV**	**Wasserstein**
MuJoCo (all four environments)	0.05	0.1	0.05
Classic control (all three environments)	0.05	0.2	0.5

Table H.5: Fixed ambiguity radii used in the main results. The radii are validation hyperparameters, not calibrated physical uncertainty radii.

For an ablation, we sweep $\rho \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ for each divergence-environment pair. Figure H.1 shows the MuJoCo sweep, and Table H.6 reports the best-performing radii under this additional tuning.

Environment	χ^2	TV	Wasserstein
Hopper	0.05	0.1	0.05
Walker2d	0.2	0.2	0.1
Ant	0.05	0.1	0.05
Humanoid	0.05	0.5	0.5
CartPole	0.05	0.2	0.5
Pendulum	0.05	0.05	0.05
Acrobot	0.05	0.2	0.1

Table H.6: Per-environment tuned ambiguity radii ρ that maximize worst-case performance within the tested sweep. These values are reported as an ablation, not used as the main comparison protocol.

Key observations from the radius sweep. The benefit of radius tuning varies by task. Table H.10 measures return sensitivity, whereas Table H.7 measures penalty tightness. On Pendulum and MuJoCo, χ^2 has the smallest worst-case return swing across the tested radii, although Walker2d remains sensitive. TV varies substantially on Hopper, and Acrobot's low returns amplify the relative swings. Per-environment tuning also requires additional validation.

Clipping and penalty diagnostics. For every robust run, the implementation records the fraction of reward and transition backups for which the positive part clips to zero,

$$\mathbf{1}\{\bar{r}_t \le \widehat{\overline{\Delta}}_{R,t}\}, \qquad \mathbf{1}\{\overline{V}_t \le \widehat{\overline{\Delta}}_{P,t}\},$$

as well as the ratios $\widehat{\overline{\Delta}}_{R,t}/(\bar{r}_t + 10^{-8})$ and $\widehat{\overline{\Delta}}_{P,t}/(\overline{V}_t + 10^{-8})$. These diagnostics are not used by the planner, but they identify overly pessimistic radii or loose deterministic envelopes.

H.1 Additional diagnostic analyses

The following diagnostics combine offline analyses of logged backups with separate planner comparisons. They quantify penalty tightness, clipping, radius sensitivity, depth fairness, and the value of non-myopic search under the evaluation protocol described below.

Exact robust penalty comparison. At each logged backup, we compare the exact robust penalty on the empirical support with the closed-form penalty capped by the largest decrease available on that support. This support cap is used only for the finite-support diagnostic; the planner itself uses the positive-part backup in Eq. (4.12). The exact inner problems are solved using the finite-support TV solution, a χ^2 quadratic program, and the Wasserstein Kantorovich dual. For each positive capped penalty, we report

$$g = \frac{\Delta^{\mathrm{cap}}_{\mathrm{closed}} - \Delta_{\mathrm{exact}}}{\Delta^{\mathrm{cap}}_{\mathrm{closed}}}.$$

This comparison is made at individual logged backups under the approximate planner's visitation distribution. It does not estimate the full gap to the exact infinite-horizon robust value or compare two independently constructed search trees.

Environment	TV	χ^2	Wasserstein
Acrobot	$\approx 0\%$	$\approx 0\%$	saturated
CartPole	29%	1%	saturated
Pendulum	3%	0%	saturated
Hopper	11%	0%	23%
Walker2d	4%	0%	8%
Ant	3%	0%	5%
Humanoid	9%	0%	2%

Table H.7: Relative gap between the support-capped closed-form penalty and the exact finite-support penalty. "Saturated" denotes a Wasserstein correction pinned to the empirical mean-minus-minimum ceiling; numeric gaps are not reported for these entries.

The reported χ^2 relative penalty gaps are at most 1% across the evaluated environments. TV and Wasserstein are also within 9% on Walker2d, Ant, and Humanoid, whereas Hopper is less tight for TV and Wasserstein.

Environment	Closed-form	Exact backup
Hopper	1189.6 ± 36.1	1152.7 ± 35.6
Walker2d	70.5 ± 10.1	82.1 ± 11.9
Ant	78.4 ± 9.9	82.5 ± 12.1
Humanoid	137.4 ± 11.7	134.5 ± 12.8
CartPole	7.4 ± 0.3	19.1 ± 1.9
Pendulum	1112.1 ± 5.1	1194.2 ± 11.6
Acrobot	29.5 ± 6.0	39.3 ± 5.8
MountainCar	14.8 ± 6.9	6.2 ± 4.8

Table H.8: Return of planners using exact and closed-form robust backups under W_1 ambiguity.

Table H.8 compares separate W_1 planner runs using exact and closed-form backups. The return differences have mixed signs on MuJoCo. The planner with exact backups has higher sample means on CartPole, Pendulum, and Acrobot, but not on MountainCar. Thus the effect of the approximation varies by task; the table does not establish equivalence between the two planners. MountainCar is included only as an additional diagnostic task and uses the continuous-action variant with acceleration in $[-1, 1]$.

Penalty inactivity, cap binding, and zero clipping. A penalty is inactive when its sample spread statistic is zero. For the finite-support diagnostic, the support cap binds when the raw closed-form bound exceeds the largest decrease available in the observed samples. This event is distinct from clipping the planner's positive-part backup to zero, so the rates below should not be read as zero-clipping rates. Table H.9 pools reward and transition backups.

	Penalty inactive			Support cap binds		
Environment	TV	χ^2	W_1	TV	χ^2	W_1
Hopper	0.4%	0.7%	0.0%	0.1%	0.1%	0.0%
Walker2d	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Ant	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Humanoid	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CartPole	0.2%	1.6%	0.1%	0.3%	0.0%	99.8%
Pendulum	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Acrobot	3.4%	22.5%	22.1%	0.0%	0.4%	78.1%

Table H.9: Penalty inactivity and support-cap binding rates at the fixed radii used in the main comparison.

The support cap binds in at most 0.1% of MuJoCo backups, compared with 78–100% for classic-control Wasserstein. On Acrobot, χ^2 and Wasserstein penalties are also inactive in roughly 22% of backups, consistent with repeated values in the sparse-reward samples.

Radius sensitivity. Let $\bar{J}(\rho)$ denote return averaged over the four execution-noise levels. Table H.10 reports $(\max_\rho \bar{J} - \min_\rho \bar{J})/|\operatorname{mean}_\rho \bar{J}|$ over $\rho \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$.

Environment	χ^2	TV	Wasserstein
Pendulum	4%	7%	4%
Hopper	7%	77%	11%
CartPole	18%	25%	15%
Humanoid	20%	23%	18%
Ant	28%	22%	22%
Walker2d	32%	40%	50%
Acrobot	158%	154%	121%

Table H.10: Relative performance swing over the ambiguity-radius sweep. Near-floor Acrobot returns inflate all relative ratios.

Across Pendulum and the four MuJoCo tasks, χ^2 varies by 4–32% and has the smallest worst-case swing of the three sets. This is a comparative stability statement, not radius insensitivity.

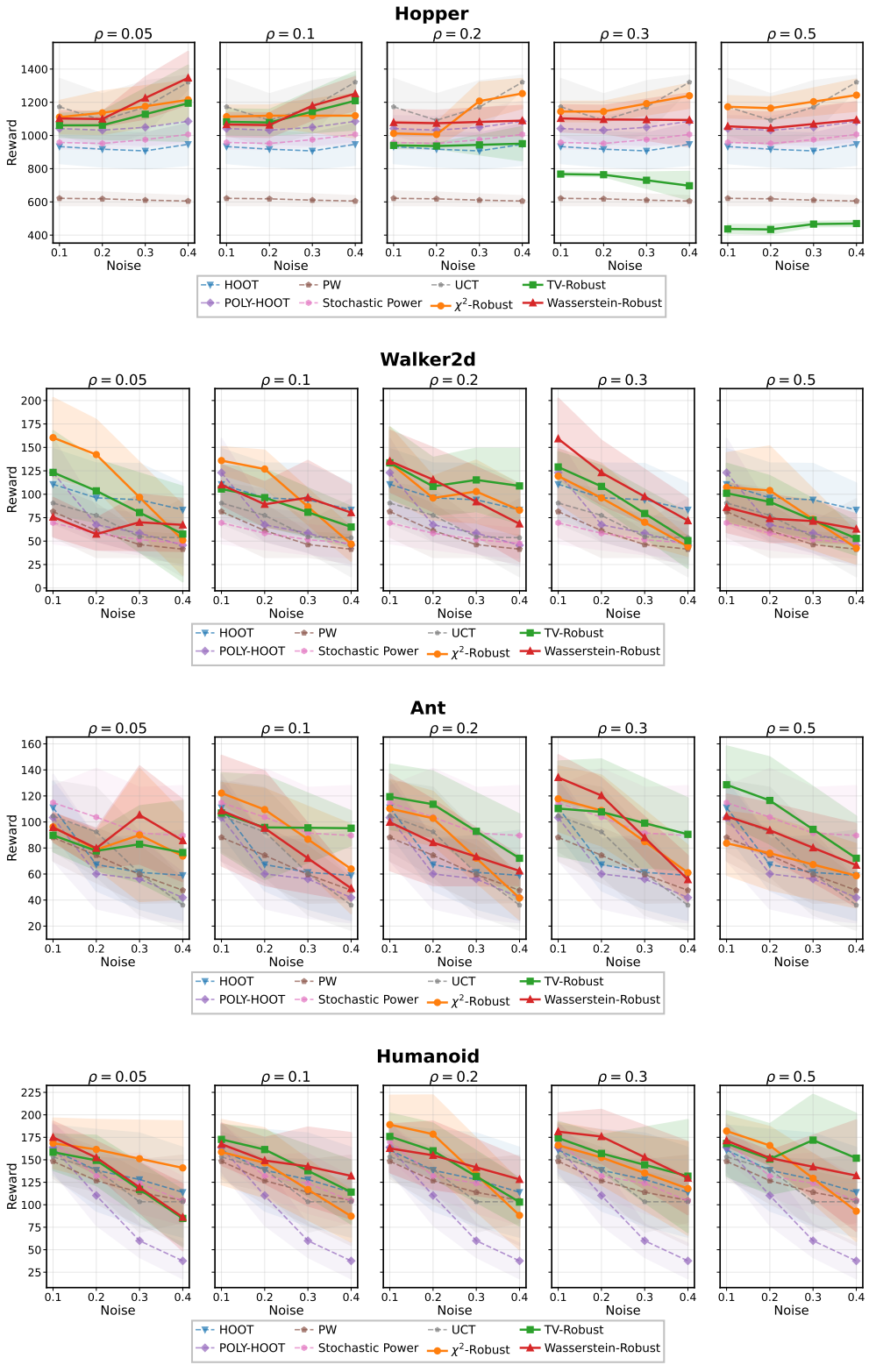

Figure H.1: **Sensitivity to ambiguity radius ρ on MuJoCo tasks.** From top to bottom: Hopper, Walker2d, Ant, and Humanoid. Each row shows performance across five values of ρ. The fixed radii in Table H.5 provide a conservative main protocol, while per-environment tuning can improve some tasks at the cost of additional validation.

Matched-depth comparison. We rerun Stochastic-Power-HOOT with the robust planner's environment-specific depth, keeping the planning model, execution-noise protocol, evaluation horizon, five seeds, and the 500-simulation budget fixed. Each seed is first averaged over execution noise $0.10/0.20/0.30/0.40$, and Table H.11 reports the mean over five seeds.

Environment	H	Nominal, matched	Robust, matched	Nominal, $H = 100$	Gain
Hopper	50	1102	1145	886	3.9%
Walker2d	50	65	86	92	32.3%
Ant	40	70	87	50	24.3%
Humanoid	30	137	153	175	11.7%

Table H.11: Closest nominal baseline under matched MCTS depth and simulation budget. The $H = 100$ column shows that nominal performance is not monotone in planning horizon.

Robust mean returns are higher at matched depth and simulation budget. However, the nominal $H = 100$ means exceed the matched-depth robust means on Walker2d and Humanoid; the comparison does not establish an advantage over every depth choice.

Nominal and robust myopic controls. Table H.12 compares a nominal short-horizon rollout-greedy planner, the same planner with a TV robust penalty, and the full method under the same mismatch protocol. The full method uses χ^2 on Hopper, Ant, CartPole, and Pendulum and Wasserstein on Acrobot; hence this is not a matched-divergence component ablation.

Environment	Planner	0.10	0.20	0.30	0.40	Avg.
Hopper	Myopic nominal	790	758	702	601	713
	Myopic robust (TV)	800	757	705	600	716
	Robust-Power-HOOT	1113	1119	1119	1119	1118
Ant	Myopic nominal	97	75	46	78	74
	Myopic robust (TV)	105	75	46	76	76
	Robust-Power-HOOT	122	110	87	64	96
Acrobot	Myopic nominal	63	66	26	12	42
	Myopic robust (TV)	56	74	52	16	50
	Robust-Power-HOOT	75	76	76	76	76
CartPole	Myopic nominal	71	65	63	42	60
	Myopic robust (TV)	73	60	46	62	60
	Robust-Power-HOOT	24	22	19	16	20
Pendulum	Myopic nominal	1420	1405	1252	1264	1335
	Myopic robust (TV)	1380	1364	1404	1293	1360
	Robust-Power-HOOT	1174	1166	1217	1211	1192

Table H.12: Myopic controls. The full planner has higher average returns on Hopper, Acrobot, and Ant, but lower average returns on CartPole and Pendulum.

Computational requirements. All experiments were run on a CPU cluster with Intel Xeon E5-2680 v4 CPUs (2.4 GHz). Per-seed runtimes varied across environments and uncertainty types, ranging from roughly 10 minutes for low-dimensional classic-control tasks to roughly 3 hours for high-dimensional MuJoCo tasks. Exact inner optimization is used for offline diagnostics and the comparison with exact backups; our planner uses the closed-form updates.

Reward preprocessing. Power-mean backups are well-defined for nonnegative edge values; we shift and scale rewards for numerical stability:

$$r'(s,a) = \frac{r(s,a) - r_{\min}}{r_{\max} - r_{\min}} R_{\text{scale}} + R_{\text{shift}}. \tag{H.1}$$

Here $r_{\min} < r_{\max}$ are environment-specific bounds, $R_{\text{scale}} > 0$ sets the reward scale, and $R_{\text{shift}} \geq 0$ ensures nonnegativity. All methods use the same transformation, and reported returns are on this transformed reward scale.