EFFECTIVE-RESISTANCE EXPLORATION IN MONTE-CARLO GRAPH SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Monte-Carlo graph search shares samples across different paths that reach the same state. This sharing also changes what remains worth exploring: an action may have been selected only a few times even though its successor states have already been extensively explored through other paths. We study how exploration should respond to this shared evidence in stochastic planning. Motivated by effective resistance, we introduce an exploration rule that combines an action's own visit count with the evidence pooled at its successor states. Visits through one path can therefore reduce exploration pressure along other paths leading to the same state, while each action retains a separate incentive for local exploration. Building on graph-based Power-UCT, we establish conditions under which this rule preserves the $O(n^{-1/2})$ convergence rate of the root value estimate for a fixed planning horizon, where n is the number of simulations. We also extend the approach to entropy-regularized graph search and quantify how sample sharing reduces an upper bound on cumulative exploration bonuses. The resulting framework makes exploration explicitly responsive to evidence collected across the search graph.

1 INTRODUCTION

Monte-Carlo Tree Search (MCTS) allocates simulations by balancing estimated action values against exploration bonuses (Coulom, 2007; Kocsis & Szepesvári, 2006; Browne et al., 2012). In stochastic planning, different trajectories often reach the same state. Graph-Based Stochastic Power-UCT (GS-Power-UCT) (Tran et al., 2026) shares their statistics by merging equal state–depth pairs. It combines this representation with power-mean backups and polynomial exploration, retaining the fixed-horizon $O(n^{-1/2})$ root convergence rate of Stochastic-Power-UCT (Dam et al., 2024b).

Sharing estimates raises a further question: *how should exploration respond to evidence collected through other paths?* An action selected only a few times may lead to successors already explored through other parents. Its selection count does not distinguish these successors from rarely visited ones. Shared successor visits should reduce the incentive to explore them again, while each action still needs samples to estimate its own reward and transition distribution.

We address this question with an exploration bonus motivated by effective resistance (ER). In an electrical network, parallel routes reduce resistance (Doyle & Snell, 1984; Klein & Randić, 1993); in a search graph, incoming paths pool evidence at their common successor. Our bonus combines an inverse local action count with an empirical average of inverse pooled successor counts. Visits through one parent can therefore reduce another parent's child bonus without changing that parent's local count. A recursive resistance provides the structural interpretation; the implemented rule requires only the counts already stored by graph search.

We develop this rule in two settings. The first adds the ER exploration bonus to GS-Power-UCT. The second brings convex-regularized MCTS (Dam et al., 2021; Xiao et al., 2019) to the same graph representation, covering maximum entropy, Tsallis entropy, and relative entropy with a fixed reference policy. Both retain *historical backups*: a parent averages the child value estimates used on its past visits, even as other parents continue to update those children. This dependence is the main obstacle in the analysis.

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.

1. **Exploration from shared successor evidence.** We introduce an ER bonus with local-action and shared-successor terms, give its inverse counts exact one-layer effective-resistance interpretations, characterize how sample sharing reduces the bonus, and express its cumulative sum through weighted averages of the sharing factors. For the same trajectories, merging nodes also reduces the resulting upper bound.
2. **Convergence with history-dependent exploration.** We prove a bandit result that allows the additional bonus to depend on all past observations, provided it is bounded by a constant times the inverse action count. Together with concentration bounds for historical graph backups, this preserves $O(n^{-1/2})$ expected root-value error for GS-Power-UCT-ER at a fixed horizon.
3. **Convex-regularized graph search.** We establish concentration bounds that hold for all later visits when each action has an explicit lower bound on its sampling probability. These yield root action-recommendation guarantees and bounds on regularization bias. For maximum-entropy and fixed-reference relative-entropy regularization, positive action probabilities give exponential concentration without logarithmic factors.

The graph representation and power-mean baseline are inherited from Tran et al. (2026). Our additions concern exploration, concentration analysis, and the regularized graph extension. The experiments in section 6 will assess whether ER improves finite-budget planning.

2 Shared graph statistics and historical backups

Consider a discounted Markov decision process (MDP) with discount $\gamma \in (0, 1]$, rewards in $[0, R_{\max}]$, between one and K actions at each internal node, and at most M possible successors for each state–action pair. Fix a planning horizon H and terminal function V_0. The finite-horizon optimal targets satisfy $V^*(s, H) = V_0(s)$ and

$$Q^*(s, h, a) = \mathbb{E}[R + \gamma V^*(S', h + 1) \mid s, a], \qquad V^*(s, h) = \max_a Q^*(s, h, a). \tag{1}$$

Our main guarantees concern these finite-horizon values, or their regularized counterparts, on a finite reachable state–depth graph.

Following GS-Power-UCT (Tran et al., 2026), each node is a state–depth pair $v = (s, h)$. We abbreviate $Q^*(v, a) = Q^*(s, h, a)$ and $V^*(v) = V^*(s, h)$. The root notation s_0 also denotes $(s_0, 0)$. Every edge advances the depth by one, so the search graph is a layered directed acyclic graph (DAG), even when the underlying MDP has cycles. Equal states at different depths remain distinct because their remaining-horizon values can differ. Write T_v for the number of visits to node v across all its parents, $T_{v,a}$ for its action-selection count, and $N_{v,a,w}$ for the number of observed transitions from (v, a) to child w. For every non-root node,

$$T_w = \sum_{(v,a):v\to w} N_{v,a,w} \geq N_{v,a,w}, \qquad \widehat{p}(w \mid v, a) = N_{v,a,w}/T_{v,a}. \tag{2}$$

The root count equals the number of completed simulations.

We use a full-depth formulation: every simulation reaches depth H, including when a new node is encountered. This makes explicit the terminal-evaluation model used in the proofs; the expanding-boundary implementation in Tran et al. (2026) instead evaluates newly expanded boundary nodes. After a simulation, only its realized path is updated, in reverse order. At the i-th selection of (v, a), let (R_i, S_i) be the fresh reward–successor pair, with S_i denoting the successor node including its depth, and $\widetilde{T}_i$ the child's pooled count *after* its update on that simulation. The stored action estimate is

$$\widehat{Q}_{v,a,m} = \frac{1}{m}\sum_{i=1}^{m}\big(R_i + \gamma \widehat{V}_{S_i,\widetilde{T}_i}\big). \tag{3}$$

Old targets are retained when another parent later updates the child. The algorithms differ in their action selection and node-value backup.

For representation comparisons, give the tree and graph the same simulated trajectories. If φ maps tree nodes to their shared state–depth node, then

$$T^{\mathcal{G}}_v = \sum_{u\in\varphi^{-1}(v)} T^{\mathcal{T}}_u, \tag{4}$$

with the same identity for action and transition counts. This comparison fixes the trajectories; independently run tree and graph planners can choose different trajectories.

3 EFFECTIVE-RESISTANCE EXPLORATION

Assign each observed transition conductance $N_{v,a,w}$, or resistance $1/N_{v,a,w}$. A recursive series–parallel construction combines each parent's resistance with its connecting edge and then combines incoming branches in parallel. It recovers path resistance on trees and decreases under additional counts or incoming routes. Its definition and structural results are in section B. The resulting $\widetilde{\mathcal{R}}$ is the resistance of a path-unrolled network; it is distinct from ordinary electrical resistance on the original graph. Neither convergence proof requires a comparison between the two.

The implemented bonus. We use a local count-based rule motivated by this resistance construction. For an initialized action, define

$$B^{\mathrm{ER}}_{v,a} = \underbrace{\frac{C_2}{T_{v,a}}}_{\text{local action evidence}} + \underbrace{C_3 \sum_{w:N_{v,a,w}>0} \frac{N_{v,a,w}}{T_{v,a}} \frac{1}{T_w}}_{\text{pooled successor evidence}}, \qquad C_2, C_3 \geq 0. \tag{5}$$

The local-action term encourages sampling of the reward and transition distribution. The successor term averages inverse pooled child counts under the empirical transition distribution. Only stored counts are needed, with no graph Laplacian solve or recursive resistance calculation.

The inverse counts in both channels admit exact one-layer effective-resistance interpretations (Proposition 3). Shorting all realized successors of (v, a) leaves parallel outgoing conductances summing to $T_{v,a}$, and shorting all parents of w leaves incoming conductances summing to T_w by eq. (2). Thus $B^{\mathrm{ER}}_{v,a} = C_2 R^{\mathrm{out}}_1(v, a) + C_3 \sum_{w:N_{v,a,w}>0} \widehat{p}(w \mid v, a)\, R^{\mathrm{in}}_1(w)$, where $R^{\mathrm{out}}_1(v, a) = 1/T_{v,a}$ and $R^{\mathrm{in}}_1(w) = 1/T_w \leq \widetilde{\mathcal{R}}(w)$. The bonus combines an outgoing resistance with an empirical average of incoming resistances. The incoming resistance is obtained from eq. (24) by setting all parent resistances to zero. Under historical backups, visits to w through other parents do not change $\widehat{Q}_{v,a}$ immediately; the successor term reflects the pooled sampling effort available to the child estimates used in future targets (see also Remark 2).

Proposition 1 (Bonus bound and effect of sample sharing)**.** *For $T_{v,a} \geq 1$, the bonus satisfies*

$$0 \leq B^{\mathrm{ER}}_{v,a} \leq \frac{C_+}{T_{v,a}}, \qquad C_+ = C_2 + C_3 M. \tag{6}$$

On a tree, its successor term is $C_3 M^{\mathrm{real}}_{v,a}/T_{v,a}$, where $M^{\mathrm{real}}_{v,a}$ counts observed successors. On a graph, each child contribution is multiplied by the sharing factor $N_{v,a,w}/T_w \in (0, 1]$.

The count identity eq. (2) gives this upper bound directly; the full proof is in section B.4. At fixed local counts, additional visits to w from other parents reduce its contribution. For instance, with one observed successor, $T_{v,a} = 10$, and $C_2 = C_3 = 1$, increasing T_w from 10 to 100 reduces the bonus from 0.2 to 0.11 while retaining the local-action term. The convergence theorem in section 5.1 allows this dependence on visits from other parents.

Cumulative exploration bonus. Let $\mathrm{Bud}(n)$ denote the sum of ER bonuses at selected actions over n simulations. Summing eq. (6) gives a logarithmic upper bound for each visited state–action pair. If $\mathcal{E}^{\mathrm{act}}_n$ is the set of these pairs, then

$$\mathrm{Bud}(n) \leq \mathrm{Led}(n) := C_+ \sum_{(v,a):T_{v,a}\geq 1} (1 + \log T_{v,a}) \leq C_+ |\mathcal{E}^{\mathrm{act}}_n| (1 + \log n). \tag{7}$$

Initialization selections pay zero in this accounting. For a pair with $T_{v,a} \geq 2$, its total paid bonus equals $\mathcal{H}_{T_{v,a}-1}\big(C_2 + C_3 \sum_w \bar{\omega}_{v,a,w}\big)$, where $\mathcal{H}_m$ is the m-th harmonic number and $\bar{\omega}_{v,a,w} \in [0, 1]$ is a weighted average of the sharing factors over the pair's selections, with weights proportional to the inverse pre-selection action count. A pair selected only once pays zero. Holding the pair's own selections and observed successors fixed, additional visits to w through other parents can only

decrease this average. Under eq. (4), $\mathrm{Led}^{\mathcal{G}} \leq \mathrm{Led}^{\mathcal{T}}$; the reduction for each set of merged nodes is derived in section B.5. This orders the upper bounds Led; it does not establish the same ordering for actual cumulative bonuses or planning errors.

4 ALGORITHMS

Both algorithms use the graph and historical backup in section 2, initialize every action once, and compute bonuses from the history available before selection.

4.1 GS-POWER-UCT-ER

We extend the state–depth version of GS-Power-UCT (Tran et al., 2026). Each node $v = (s, h)$ pools visits from all paths reaching state s at depth h, so its parents share the same counts and value estimates. All visits to v have the same remaining horizon $H - h$ and finite-horizon value target. Fix $1 \leq p < \infty$, $C_1 > 0$, and $C_2, C_3 \geq 0$. The graph initially contains only the root, with all statistics initialized to zero; its statistics persist between simulations, and each simulation finishes before the next begins.

Each simulation continues to depth H, including through newly created nodes, and then updates its sampled path in reverse order.

Algorithmic parameters. As in GS-Power-UCT, the exploration rule is parameterized by sequences $\{b_h\}_{h=1}^{H}$, $\{\alpha_h\}_{h=0}^{H}$, and $\{\beta_h\}_{h=0}^{H}$ satisfying Table 1. The pair (α_h, β_h) describes the concentration exponents at depth h. Action estimates at depth h inherit their exponents from the children at depth $h + 1$, so selection uses $(b_{h+1}, \alpha_{h+1}, \beta_{h+1})$. The table records the conditions for the concentration argument; theorem 1 gives an explicit choice that yields the $O(n^{-1/2})$ root error rate.

Table 1: Sufficient conditions for the GS-Power-UCT-ER parameters; all must hold, with (2)–(4) for every $0 \leq h < H$. The power exponent is fixed with $1 \leq p < \infty$.

(1) $0 < 2\alpha_H \leq \beta_H$; $\quad \beta_0 > 1$	(2) $2 < b_{h+1} < \alpha_{h+1} \leq \beta_{h+1}/2$
(3) $1 - b_{h+1}/\alpha_{h+1} \leq \alpha_{h+1}/\beta_{h+1}$	(4) $\alpha_h = (b_{h+1} - 1)(1 - b_{h+1}/\alpha_{h+1})$; $\quad \beta_h = b_{h+1} - 1$

Selection and simulation. We retain the state–depth lookup and shared statistics of GS-Power-UCT (Tran et al., 2026). At $v = (s, h)$ with $h < H$ and $n = T_v$ completed visits, select untried actions first in a fixed order; otherwise select

$$a_{n+1} \in \operatorname*{argmax}_a \left\{ \widehat{Q}_{v,a,T_{v,a}(n)} + C_1 \frac{n^{b_{h+1}/\beta_{h+1}}}{T_{v,a}(n)^{\alpha_{h+1}/\beta_{h+1}}} + B^{\mathrm{ER}}_{v,a} \right\}, \tag{8}$$

with fixed tie-breaking. The parameters satisfy Table 1, and the bonus in eq. (5) uses only completed search history. Initialization selections pay no bonus. Sample a fresh reward–successor pair and reuse or create the successor node, initializing new statistics to zero. Following section 2, continue through newly created nodes to depth H, where exact values or fresh bounded nonnegative samples of mean V_0 are averaged at the shared terminal node. Thus evaluation occurs at depth H, rather than at first expansion.

Backup and recommendation. Update the sampled path in reverse order, incrementing its node, action, and transition counts. Use the historical action backup in eq. (3), with each child's value read after its update on this simulation, and recompute

$$\widehat{V}_{v,n} = \left(\sum_{a:T_{v,a}(n)>0} \frac{T_{v,a}(n)}{n} \widehat{Q}^{p}_{v,a,T_{v,a}(n)} \right)^{1/p}, \tag{9}$$

where n is the updated node count. Old targets remain in the action averages; other parents are not updated. After n simulations, return $\widehat{V}_{s_0,n}$ and, for the guarantee in Corollary 8, a most-visited root

action $a_n^{\mathrm{rec}} \in \operatorname{argmax}_a T_{s_0,a}(n)$ with fixed tie-breaking. Setting $C_2 = C_3 = 0$ recovers the corresponding full-depth GS-Power-UCT rules. The parameter choice in theorem 1 gives exploration exponents $1/4$ and $1/2$ at every depth.

Full-state variants. The same bonus can be used when nodes are keyed only by physical state, as in GS-Power-UCT-F and GS-Power-UCT-F+. Such merging can mix targets with different remaining horizons. Section G states the additional conditions for these variants. It gives direct $O(n^{-1/2})$ guarantees for finite acyclic episodic models with exact zero terminal values, under the stated assumptions and a horizon covering every episode. For general cyclic merging, these conditions are substantive: Proposition 8 gives an example with a slower rate.

4.2 GS-E3W-ER: CONVEX-REGULARIZED GRAPH SEARCH

Extended Empirical Exponential Weight (E3W) combines a regularized policy with uniform exploration (Dam et al., 2021). GS-E3W-ER retains this structure and adds two features: *shared statistics* at equal state–depth nodes and *ER bonuses* in the action-selection scores. It uses the full-depth simulations and historical path updates of section 2. Equal states at different depths remain separate, preserving a common remaining horizon at each shared node.

Regularized policy and value. Let $\mathcal{A}_v$ be the actions at v, $K_v = |\mathcal{A}_v|$, and Δ_{K_v} their probability simplex. Fix $\tau > 0$ and a convex regularizer Ω_v throughout the search. As in E3W, define

$$F_v(\mathbf{Q}) = \max_{\pi \in \Delta_{K_v}} \{\langle \pi, \mathbf{Q}\rangle - \tau\Omega_v(\pi)\}, P_v(\mathbf{Q}) \in \operatorname*{argmax}_{\pi \in \Delta_{K_v}} \{\langle \pi, \mathbf{Q}\rangle - \tau\Omega_v(\pi)\}. \tag{10}$$

Here P_v selects a regularized policy and F_v gives its regularized value. Assume maximizers exist and use a fixed measurable choice for ties. The value map is 1-Lipschitz in the sup norm.

Shared estimates and historical backups. All parents of $v = (s, h)$ use the same counts and action estimates. At the i-th selection of (v, a), let S_i be the observed child and $\widetilde{T}_i$ its pooled count after its update on that simulation. The stored estimates are

$$\widehat{Q}_{v,a,m} = \frac{1}{m}\sum_{i=1}^{m}\big(R_i + \gamma \widehat{V}_{S_i,\widetilde{T}_i}\big), \widehat{\mathbf{Q}}_{v,n} = \big(\widehat{Q}_{v,a,T_{v,a}(n)}\big)_{a\in\mathcal{A}_v}, \widehat{V}_{v,n} = F_v(\widehat{\mathbf{Q}}_{v,n}). \tag{11}$$

Counts and estimates are updated only along the sampled path, in reverse order; old targets remain in the action averages when other parents update a shared child. When an internal node $v = (s, h)$ is created, set $d = H - h$ and initialize its counts to zero and each action estimate in $J_d = [\gamma l_{d-1}, R_{\max} + \gamma u_{d-1}]$. The endpoints l_d, u_d and action intervals are defined in eqs. (78) and (79). Initialize the node value by applying F_v to this vector. Untried actions retain their initial entries in $\widehat{\mathbf{Q}}_{v,n}$.

ER-modified action selection. Try each action once, in a fixed order over the first K_v pooled visits. Before local visit $k > K_v$, compute the bonus in eq. (5), with (C_2, C_3) replaced by fixed coefficients (c_2, c_3), $c_2, c_3 \geq 0$, from completed search history, and sample an action from

$$\mathbf{Q}^{\mathrm{ER}}_{v,k} = \widehat{\mathbf{Q}}_{v,k-1} + B^{\mathrm{ER}}_{v,\cdot}, \tag{12}$$

$$\pi^{\mathrm{ER}}_{v,k}(a) = (1 - \lambda_{v,k})[P_v(\mathbf{Q}^{\mathrm{ER}}_{v,k})]_a + \lambda_{v,k}/K_v, \tag{13}$$

where $\lambda_{v,k} = \min\{1, \epsilon_\lambda K_v/\log(2 + k)\}$ and $\epsilon_\lambda > 0$. The uniform component ensures $\pi^{\mathrm{ER}}_{v,k}(a) \geq \eta/\log(2 + k)$, with $\eta = \min\{\epsilon_\lambda, 1/K\}$. Visits through another parent can reduce the successor bonus even when this action's own count is unchanged. The policy uses $P_v(\widehat{\mathbf{Q}} + B^{\mathrm{ER}})$, whereas the value backup uses $F_v(\widehat{\mathbf{Q}})$: ER affects selection only. Setting $c_2 = c_3 = 0$ gives GS-E3W without ER under the same graph and simulation conventions.

Targets and regularizer choices. The fixed regularized targets satisfy

$$Q^*_\Omega(v, a) = \mathbb{E}[R + \gamma V^*_\Omega(S') \mid v, a], \qquad V^*_\Omega(v) = F_v(Q^*_\Omega(v, \cdot)),$$

with $V^*_\Omega(s,H) = V_0(s)$. The choices $\Omega(\pi) = \sum_a \pi_a \log \pi_a$ and $\Omega(\pi) = (\|\pi\|_2^2 - 1)/2$ give MENTS (softmax) and TENTS (sparsemax), respectively. Fixed-reference KL uses $\Omega_v(\pi) = \mathrm{KL}(\pi\|\rho_v)$ with $\rho_v(a) > 0$, giving weights proportional to $\rho_v(a)e^{Q_a/\tau}$. Each reference is fixed and shared across all visits to its node; moving-reference RENTS is outside the analysis.

After n simulations, recommend $a_n^{\mathrm{rec}} \in \mathrm{argmax}_a \widehat{\mathbf{Q}}_{s_0,n}(a)$ with fixed tie-breaking. Terminal estimates use the raw pooled sample mean projected onto $[0, V_{\max}]$, as specified in eq. (71); exact evaluations are also allowed. Full-state variants and their additional conditions are given in section G.4.

5 Convergence guarantees

We first establish root-value convergence for GS-Power-UCT-ER, then give value-estimation and action-recommendation guarantees for GS-E3W-ER. Throughout this section, the planning horizon and algorithmic parameters are fixed. The targets are the finite-horizon values defined in sections 2 and 4.2.

5.1 Convergence of GS-Power-UCT-ER

Assumptions and parameter choice. Assume a finite reachable state–depth graph, fresh simulator samples, and bounded nonnegative rewards and terminal evaluations. Every internal node with multiple actions has a unique optimal action and positive suboptimal gaps. The precise conditions are given in Assumptions 2 and 3. For any $\beta_0 > 1$, choose the parameters in Table 1 as

$$\begin{aligned} \beta_{h+1} &= 4(\beta_h + 1), \qquad b_{h+1} = \beta_{h+1}/4 \quad (0 \le h < H), \\ \alpha_h &= \beta_h/2 \qquad\qquad\qquad\qquad\quad (0 \le h \le H). \end{aligned} \tag{14}$$

This choice satisfies the table's conditions and gives $b_{h+1}/\beta_{h+1} = 1/4$ and $\alpha_{h+1}/\beta_{h+1} = 1/2$. Thus the polynomial exploration term is $C_1 n^{1/4}/\sqrt{T_{v,a}(n)}$ at every depth.

Theorem 1 (Convergence rate of GS-Power-UCT-ER)**.** *Under Assumptions 2 and 3, run GS-Power-UCT-ER with the parameter choice eq.* (14)*, fixed finite $p \ge 1$, $C_1 > 0$, and fixed finite $C_2, C_3 \ge 0$. There is a finite constant C_H, independent of n, such that*

$$\mathbb{E}\left[\left|\widehat{V}_{s_0,n} - V^*(s_0,0)\right|\right] \le \frac{C_H}{\sqrt{n}}, \qquad n \ge 1. \tag{15}$$

The same bound holds when the ER bonus is replaced by any bonus determined before selection and satisfying $0 \le \xi_{v,a} \le C_+/T_{v,a}$ for a fixed finite C_+.

The ER bonus therefore preserves the fixed-horizon $O(n^{-1/2})$ root-value convergence rate of GS-Power-UCT. The constant may depend on the horizon, action gaps, power exponent, and exploration coefficients; the result does not establish an ER speedup. Theorem 4 states the stronger maximal concentration result, and section D gives the proof. When $H \ge 1$, a most-visited root recommendation has error probability $O(n^{-\beta_0})$ for sufficiently large n (Corollary 8).

Corollary 1 (Conditional convergence under full-state merging)**.** *Suppose GS-Power-UCT-F-ER satisfies all premises of Proposition 7, including its fixed-target concentration and sequential-feedback conditions. Let $\bar{V}_H(s_0)$ be its effective root target and suppose $|\bar{V}_H(s_0) - V^*(s_0,0)| \le \delta_H^{\mathrm{merge}}$. Then, for a finite constant C_H^{F} independent of n,*

$$\mathbb{E}\left[\left|\widehat{V}^{\mathrm{F,ER}}_{n,H} - V^*(s_0,0)\right|\right] \le \underbrace{\frac{C_H^{\mathrm{F}}}{\sqrt{n}}}_{\textit{estimation error}} + \underbrace{\delta_H^{\mathrm{merge}}}_{\textit{merging bias}}. \tag{16}$$

The premises concern the actual historical action estimates; a bound on cross-depth discrepancy alone does not establish them. For F+, a growing horizon also requires control of the truncation error and horizon-dependent constants (Corollary 10). A direct $O(n^{-1/2})$ guarantee holds for finite acyclic episodic models with exact zero terminal values and a sufficient search horizon (Proposition 9).

5.2 VALUE ESTIMATION IN GS-E3W-ER

Assumption 1 (Fixed-regularizer model)**.** *Fix $H \geq 1$ and use the finite reachable state–depth graph, fresh bounded reward–successor samples, full-depth simulations, and reverse path updates specified in section 2. Terminal observations have conditional mean $V_0(w) \in [0, V_{\max}]$ and conditional sub-Gaussian variance proxy σ_0^2. The temperature and regularizers are fixed, the maximizers in eq.* (10) *exist, and $\omega_- \leq \Omega_v(\pi) \leq \omega_+$ for common finite bounds. Initialize action estimates and store terminal means as specified in section 4.2. No action-gap assumption is required.*

Write $\ell(n) = \log(2+n)$. The target V_Ω^* is the regularized finite-horizon value, which can differ from V^*.

Theorem 2 (Root-value convergence of GS-E3W-ER)**.** *Under Assumption 1, with fixed horizon $H \geq 1$, there is a finite constant $C_{\Omega,H}$, independent of n, such that*

$$\mathbb{E}\left[\left|\widehat{V}_{s_0,n} - V_\Omega^*(s_0)\right|\right] \leq C_{\Omega,H}\sqrt{\frac{\ell(n)^H}{n}}, \qquad n \geq 1. \tag{17}$$

Thus general fixed-regularizer search, including TENTS, has expected root error $O(\log(2+n)^{H/2}/\sqrt{n})$. After n visits to a node, the uniform-exploration component ensures at least order $n/\log(2+n)$ selections of each of its actions with high probability. Transferring action estimates to node values introduces one logarithmic factor per depth in the concentration bound. Theorem 5 gives the full node- and action-level statements; section E gives their proof.

Corollary 2 (Sharper rate for MENTS and fixed-reference KL)**.** *Under Assumption 1, suppose GS-E3W-ER uses MENTS or KL relative to fixed references with $\rho_v(a) \geq m > 0$, at fixed temperature $\tau > 0$ and fixed finite ER coefficients. There is a finite constant $\bar{C}_{\Omega,H}$, independent of n, such that*

$$\mathbb{E}\left[\left|\widehat{V}_{s_0,n} - V_\Omega^*(s_0)\right|\right] \leq \frac{\bar{C}_{\Omega,H}}{\sqrt{n}}, \qquad n \geq 1. \tag{18}$$

For these regularizers, bounded action scores and fixed temperature give every action a positive sampling-probability lower bound that does not decrease with n. This removes the logarithmic factors without changing the algorithm. The constants can be large for small temperatures or small reference probabilities. The corresponding concentration bounds are in Corollary 5 and section E.3.

Full-state variants. Proposition 10 and Corollary 11 give conditional bounds separating estimation, merging, and truncation errors. For a finite acyclic episodic model of maximum length D, with regularization stopped at exact-zero terminals and $H \geq D$, the expected-error rates are $O(\log(2+n)^{D/2}/\sqrt{n})$ in general and $O(n^{-1/2})$ for MENTS and fixed-reference KL (Proposition 11).

5.3 ROOT RECOMMENDATION AND REGULARIZATION BIAS

For GS-E3W-ER, recommend $a_n^{\mathrm{rec}} \in \operatorname{argmax}_a \widehat{\mathbf{Q}}_{s_0,n}(a)$ using the estimates without ER bonuses and fixed tie-breaking. Let A_H, S_H and $\bar{A}_H, \bar{S}_H$ denote the budget-independent concentration constants in theorem 5 and Corollary 5, respectively.

Corollary 3 (Root action recommendation)**.** *Under Assumption 1, suppose the root has at least two actions (with only one, the recommendation error is zero) and $Q_\Omega^*(s_0,\cdot)$ has a unique maximizing action a_Ω^*, with gap $\Delta_\Omega = \min_{a \neq a_\Omega^*}[Q_\Omega^*(s_0,a_\Omega^*) - Q_\Omega^*(s_0,a)] > 0$. Then, for every $n \geq 1$,*

$$\mathbb{P}(a_n^{\mathrm{rec}} \neq a_\Omega^*) \leq A_H \exp\!\big(-n\Delta_\Omega^2/(9S_H^2\ell(n)^H)\big), \tag{19}$$

$$\mathbb{P}(a_n^{\mathrm{rec}} \neq a_\Omega^*) \leq \bar{A}_H \exp\!\big(-n\Delta_\Omega^2/(9\bar{S}_H^2)\big) \quad \textit{under Corollary 2}. \tag{20}$$

These guarantees identify the best action for the regularized continuation values. To compare with the unregularized optimum, we separate estimation error from regularization bias. For $\delta \in (0,1)$, define the statistical error radii

$$\Psi_n(\delta) = S_H\sqrt{\frac{\ell(n)^H}{n}\log\frac{A_H}{\delta}}, \qquad \bar{\Psi}_n(\delta) = \bar{S}_H\sqrt{\frac{1}{n}\log\frac{\bar{A}_H}{\delta}}. \tag{21}$$

Set $G_d = \sum_{j=0}^{d-1} \gamma^j$ with $G_0 = 0$, $w_\Omega = \max\{|\omega_-|, |\omega_+|\}$, and $\Lambda_H = \gamma\tau(\omega_+ - \omega_-)G_{H-1}$, an upper bound on the width of the continuation regularization bias across root actions.

Corollary 4 (Error relative to the unregularized optimum)**.** *Under Assumption 1, for fixed $n \geq 1$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, both*

$$|\widehat{V}_{s_0,n} - V^*(s_0, 0)| \leq \underbrace{\Psi_n(\delta)}_{\text{estimation error}} + \underbrace{\tau w_\Omega G_H}_{\text{regularization bias}}, \tag{22}$$

$$V^*(s_0, 0) - Q^*(s_0, 0, a_n^{\text{rec}}) \leq 2\Psi_n(\delta) + \Lambda_H. \tag{23}$$

Under Corollary 2, both bounds hold with $\Psi_n(\delta)$ replaced by $\bar{\Psi}_n(\delta)$.

If the unregularized root has at least two actions and a unique best action a^*, with gap $\Delta^* > \Lambda_H$ to the second-best action, then a^* is also the unique maximizer of $Q^*_\Omega(s_0, \cdot)$, and the recommendation bounds hold with Δ_Ω replaced by $\Delta^* - \Lambda_H$. The regularization-bias terms need not vanish at fixed temperature. The ER bonus contributes no additional backup-bias term because it is used only for selection. A zero concentration scale above denotes an exact estimate, for which the corresponding error probability is zero. Sections C and E.4 give the stronger bounds valid for all later visits, the signed bias bounds, and the proofs of these root statements.

6 Numerical Experiments

We evaluate finite-budget planning on four stochastic domains using 16–2048 simulations per decision and 1,000 seeds per configuration and budget. Environment descriptions and configurations are provided in Appendix H.

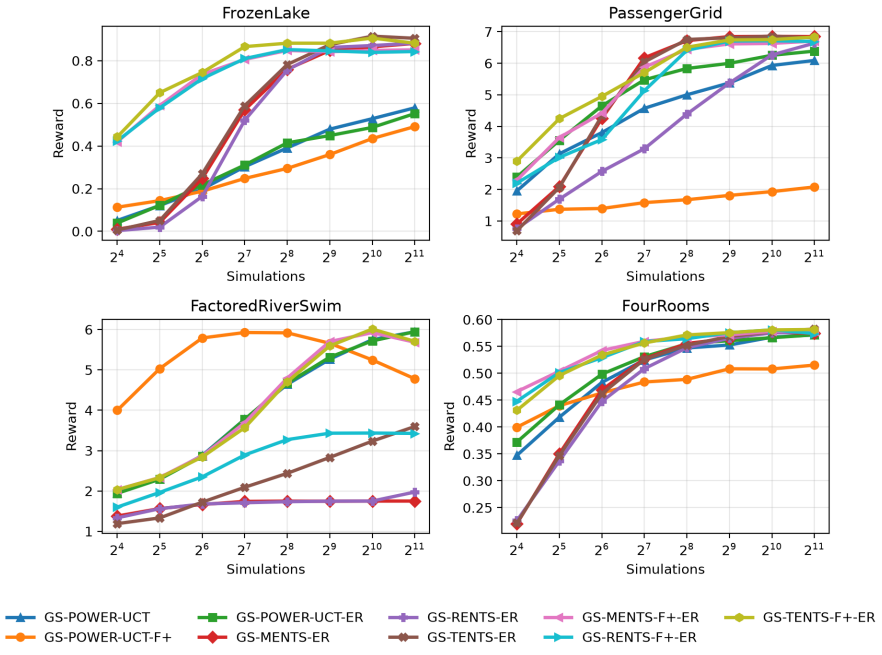

Figure 1: Mean episode return versus simulations per decision on four stochastic domains (1,000 seeds per configuration). Higher is better; curves show sample means without uncertainty bands.

Results. Figure 1 highlights the strong performance of the regularized F+-ER variants at small simulation budgets in FrozenLake, PassengerGrid, and FourRooms. These methods achieve high returns earlier than their depth-indexed counterparts and remain competitive as the budget increases. The matched comparison between GS-Power-UCT-ER and GS-Power-UCT also favors ER consistently in PassengerGrid. Performance remains domain dependent: GS-Power-UCT-F+ leads at small budgets in FactoredRiverSwim, while the full-state MENTS and TENTS extensions become competitive at larger budgets.

Discussion. The early gains support the practical value of combining shared graph statistics with exploration informed by pooled successor evidence. They are consistent with our motivation: evidence collected through one path can reduce exploration pressure along other paths to the same successor, while retaining local action exploration. The F+-ER comparisons evaluate the combined effects of representation, regularization, and exploration; their strong finite-budget performance complements the convergence analysis.

7 Related work and scope

Graph search and shared statistics. Transposition handling and DAG search (Childs et al., 2008; Saffidine et al., 2012; Leurent & Maillard, 2020; Czech et al., 2021) establish the motivation for sharing information across search paths. GS-Power-UCT (Tran et al., 2026) combines state–depth merging with stochastic power-mean estimation and is the direct starting point of this work. We retain its graph representation and historical path backups, and use pooled successor evidence to guide exploration. Our maximal analysis controls the historical estimates generated by these updates, including their dependence on adaptive pooled counts.

Power-mean backups. Power-UCT uses generalized means to interpolate between arithmetic averaging and maximization (Dam et al., 2020). Polynomial exploration provides a basis for nonasymptotic MCTS analysis (Shah et al., 2022), and Stochastic-Power-UCT extends power-mean estimation to stochastic transitions (Dam et al., 2024b). Stochastic-Power-HOOT further studies continuous stochastic MDPs using hierarchical exploration (Dam, 2025). Our finite-action result addresses shared graph statistics and permits history-dependent structural bonuses within an inverse-count envelope, while preserving the fixed-horizon $O(n^{-1/2})$ expected root-value error rate.

Regularized search. Maximum-entropy planning (Xiao et al., 2019) and convex-regularized MCTS (Dam et al., 2021) connect search policies with regularized MDP operators (Geist et al., 2019). The unified framework of Dam et al. (2024a) relates value backup and regularized exploration through α-divergence; sparsemax supplies the policy map for the quadratic Tsallis regularizer used by TENTS (Martins & Astudillo, 2016; Lee et al., 2018). Optimal-transport Bellman smoothing also supports second-order planning estimators and oracle-complexity guarantees (Dam, 2026). Our fixed-regularizer analysis gives concentration and greedy recommendation bounds for historical graph backups. For MENTS and fixed-reference KL, the stated bounded-value assumptions yield uniform full support and remove logarithmic factors from these bounds.

Resistance and graph structure. Electrical-network resistance (Doyle & Snell, 1984; Lyons & Peres, 2016) motivates our recursive construction over incoming paths from the root. For linear estimation from uncorrelated noisy relative measurements, the error covariance of a best linear unbiased node estimate relative to a fixed reference equals effective resistance when edge resistances encode measurement-noise covariances (Barooah & Hespanha, 2008). Remark 2 gives an idealized mean-squared-error interpretation of our two-channel bonus. Graph geometry also informs spectral prioritized sweeping, which schedules value updates under changing rewards and fixed transition dynamics (Pham & Dam, 2026). Here, the recursive construction motivates the local two-channel sampling bonus. Its guarantees use the inverse local-count envelope and, for regularized search, lower bounds on action-sampling probabilities.

8 Conclusion

Shared successor evidence can inform exploration as well as value estimation. The proposed ER bonus makes that dependence explicit while retaining local action exploration. Building on GS-Power-UCT, concentration bounds for historical backups and a bandit theorem allowing history-dependent bonuses preserve the fixed-horizon $O(n^{-1/2})$ expected root-value error rate. The same graph framework supports fixed convex regularization, with explicit concentration, recommendation, and bias bounds. MENTS and fixed-reference KL additionally admit exponential concentration without logarithmic factors under the stated bounded-value assumptions. The experimental protocol tests whether these mechanisms improve planning at finite simulation budgets.

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

Algorithms are specified in section 4; assumptions and full proofs are in sections D, E and E.3, with structural details in section B. Full-state extensions and their additional premises are in section G. The evaluation protocol is in section 6. Experimental results and the final implementation remain to be completed.

REFERENCES

Prabir Barooah and João P. Hespanha. Estimation from relative measurements: Electrical analogy and large graphs. *IEEE Transactions on Signal Processing*, 56(6):2181–2193, 2008.

Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012. doi: 10.1109/TCIAIG.2012.2186810. URL https://doi.org/10.1109/TCIAIG.2012.2186810.

Benjamin E. Childs, James H. Brodeur, and Levente Kocsis. Transpositions and move groups in Monte Carlo tree search. In *2008 IEEE Symposium on Computational Intelligence and Games*, pp. 389–395. IEEE, 2008. doi: 10.1109/CIG.2008.5035667. URL https://doi.org/10.1109/CIG.2008.5035667.

Rémi Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. In *Computers and Games (CG 2006)*, volume 4630 of *Lecture Notes in Computer Science*, pp. 72–83. Springer, 2007. doi: 10.1007/978-3-540-75538-8_7. URL https://doi.org/10.1007/978-3-540-75538-8_7.

Johannes Czech, Patrick Korus, and Kristian Kersting. Improving AlphaZero using Monte-Carlo graph search. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 31, pp. 103–111, 2021.

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. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/332. URL https://www.ijcai.org/proceedings/2020/332.

Tuan Dam, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. A unified perspective on value backup and exploration in Monte-Carlo tree search. *Journal of Artificial Intelligence Research*, 81: 511–577, 2024a. doi: 10.1613/jair.1.16019. URL https://www.jair.org/index.php/jair/article/view/16019.

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

Tuan Q. Dam, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Convex regularization in Monte-Carlo tree search. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2365–2375. PMLR, 2021. URL https://proceedings.mlr.press/v139/dam21a.html.

Tuan Quang Dam. Power mean estimation in stochastic continuous Monte-Carlo tree search. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12344–12376. PMLR, 2025. URL https://proceedings.mlr.press/v267/dam25b.html.

Tuan Quang Dam. Second-order smooth planning with optimal-transport Bellman smoothing. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2609.06484.

Peter G. Doyle and J. Laurie Snell. *Random Walks and Electric Networks*. Mathematical Association of America, 1984. URL https://math.dartmouth.edu/~doyle/docs/walks/walks.pdf.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

D. J. Klein and M. Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12:81–95, 1993. doi: 10.1007/BF01164627. URL https://doi.org/10.1007/BF01164627.

Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *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.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse Markov decision processes with causal sparse Tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018. doi: 10.1109/LRA.2018.2800085. URL https://doi.org/10.1109/LRA.2018.2800085.

Edouard Leurent and Odalric-Ambrym Maillard. Monte-Carlo graph search: The value of merging similar states. In *Proceedings of the 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pp. 577–592. PMLR, 2020. URL https://proceedings.mlr.press/v129/leurent20a.html.

Russell Lyons and Yuval Peres. *Probability on Trees and Networks*, volume 42 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2016. ISBN 978-1-107-16015-6. doi: 10.1017/9781316672815. URL https://rdlyons.pages.iu.edu/prbtree/.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623. PMLR, 2016. URL https://proceedings.mlr.press/v48/martins16.html.

Hung Pham and Tuan Dam. Spectral prioritized sweeping in nonstationary reinforcement learning. arXiv preprint arXiv:2609.06186, 2026. URL https://arxiv.org/abs/2609.06186.

Abdallah Saffidine, Tristan Cazenave, and Jean Méhat. UCD: Upper confidence bound for rooted directed acyclic graphs. *Knowledge-Based Systems*, 34:26–33, 2012. doi: 10.1016/j.knosys.2011.11.014. URL https://doi.org/10.1016/j.knosys.2011.11.014.

Devavrat Shah, Qiaomin Xie, and Zhi Xu. Nonasymptotic analysis of Monte Carlo tree search. *Operations Research*, 70(6):3234–3260, 2022. doi: 10.1287/opre.2021.2239. URL https://doi.org/10.1287/opre.2021.2239.

Tung Tran, Viet Bao Mai, Hoang Ta, and Tuan Dam. Graph-based stochastic Power-UCT: Monte-Carlo graph search with power mean estimation. arXiv preprint arXiv:2609.19956, 2026. URL https://arxiv.org/abs/2609.19956.

Chenjun Xiao, Ruitong Huang, Jincheng Mei, Dale Schuurmans, and Martin Müller. Maximum entropy Monte-Carlo planning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9520–9528. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/7ffb4e0ece07869880d51662a2234143-Abstract.html.

Appendix of "Effective-Resistance Exploration in Monte-Carlo Graph Search"

Table of Contents

A Notation **15**

B Structural resistance and exploration-bonus bounds **16**

B.1 Recursive resistance and its structural properties 16

B.2 Unrolling identity and comparison with electrical resistance 17

B.3 One-layer resistances and the implemented bonus 18

B.4 The local bonus and its sharing factor 19

B.5 Cumulative exploration bonuses 20

C Technical forms of the convergence guarantees **23**

C.1 Power-mean search with a bounded additional bonus 23

C.2 Historical regularized concentration 23

C.3 Maximal root recommendation and bias bounds 24

D Part I: power-mean graph-search proofs **25**

D.1 Setup and terminal concentration 25

D.2 Concentration of historical graph Q-values. 28

D.3 Power-mean concentration with ER exploration 29

D.4 Depth induction and root convergence 33

E Part II: fixed-regularizer concentration proofs **34**

E.1 Setup and terminal concentration 34

E.2 Concentration of historical regularized Q-values. 37

E.3 Action sampling and regularized value concentration 40

E.4 Root concentration and action recommendation 44

F A conditional support property for Tsallis regularization **47**

G Extensions to full-state merging **48**

G.1 Counts and the full-state bonus 48

G.2 Conditional power-mean guarantees 48

G.3 Power-mean search on acyclic episodic models 52

G.4 Regularized full-state variants. 52

H Experimental Details **57**

H.1 Evaluation Protocol 57

H.2 Hyperparameter Selection 57

H.3 Benchmark Environments 57

I Extensions, limitations, and open problems **59**

APPENDIX CONTENTS

Section A summarizes the notation. Sections B, D and E give the structural results and the complete power-mean and fixed-regularizer proofs. In each part, auxiliary definitions and lemmas precede the results that use them. Section E.4 proves the main-text root corollaries. The unified sampling argument in section E.3 covers both the general and uniformly positive action-probability bounds within Part II. Section F gives the conditional Tsallis support result. Section G gives conditional full-state extensions and acyclic episodic guarantees for both power-mean and regularized search. Finally, section I discusses limitations and open problems.

A NOTATION

Symbol	Meaning
$v = (s, h)$, s_0	state–depth node; root node $(s_0, 0)$
$\mathcal{A}_v$, Δ_{K_v}	actions at v; probability distributions over them
K, K_v, M	maximum actions per node; actions at v; maximum successors per action
H, γ, τ	planning horizon; discount factor; regularization temperature
V^*, Q^*	optimal state-value and action-value functions
V^*_Ω, Q^*_Ω	corresponding regularized value functions
$\widehat{V}_{v,n}$, $\widehat{Q}_{v,a,m}$	estimates after n node visits or m selections of (v, a)
$\widehat{\mathbf{Q}}_{v,n}$	vector of current action-value estimates at v
T_v, $T_{v,a}$	pooled node visit count; action selection count
$N_{v,a,w}$, $N(e)$	observed transition count; its edge notation
$T_{s,h}$, $T_{s,h,a}$, $T^{s'}_{s,h,a}$	alternative notation for node, action, and transition counts
$\widetilde{T}_i$	pooled successor visit count after its update on sample i
$\mathcal{N}_h$, $\mathrm{Par}(v)$	depth-h layer; parent node–action pairs of v
φ, $k(v)$	tree-to-graph merge map; number of tree nodes merged into v
$\widetilde{\mathcal{R}}(v)$, $\mathcal{R}_{\mathrm{eff}}$, $\mathcal{T}_v$	recursive resistance; electrical effective resistance; path-unrolled network
$B^{\mathrm{ER}}_{v,a}$	exploration bonus in eq. (5)
R^{out}_1, R^{in}_1	one-layer resistances (Proposition 3)
$\mathcal{H}_m$, $\bar{\omega}_{v,a,w}$	harmonic number; weighted average of sharing factors
C_1; C_2, C_3; c_2, c_3	polynomial exploration coefficient; ER coefficients for power-mean and regularized search
C_+, c_+	bonus-bound constants $C_2 + C_3 M$ and $c_2 + c_3 M$
$\mathsf{MC}(\alpha, \beta; c)$	maximal polynomial concentration bound (Definition 2)
(α_h, β_h, b_h)	depth-dependent concentration and exploration parameters
ρ, χ	Part I ratios b/α and $1 - b/\alpha$ at one node
$A_a(n)$, $\mathcal{A}_a(n)$	action-count thresholds without and with the additional bonus
Δ_a, $\Delta_{\min}$	suboptimal action gap; minimum positive gap
Ω_v, F_v, P_v	regularizer; regularized value and policy maps
$\lambda_{v,k}$, ϵ_λ	uniform-exploration probability; its coefficient
$d = H - h$, $\ell(n) = \log(2 + n)$	remaining depth; logarithmic factor
η, p_d	coefficient of the decreasing action-probability bound; constant lower bound at depth d
A_d, S_d, B_d, T_d	node and action concentration constants; bars denote the sharper bounds
ν^2_d, W_d	conditional variance proxy for one-step targets; deterministic error bound
ω_-, ω_+, G_d	regularizer range bounds; discount sum $\sum_{j=0}^{d-1} \gamma^j$
Ψ_n, $\bar{\Psi}_n$; Λ_H	root confidence radii; width of the continuation regularization bias
$\mathcal{K}(z)$, $\Gamma_k(z)$	actions with positive probability under sparsemax at scores z; sparsemax gap sums

B STRUCTURAL RESISTANCE AND EXPLORATION-BONUS BOUNDS

We first define the recursive resistance on a fixed realized layered graph and prove its structural properties. We then prove the local bonus bound and the cumulative-bonus comparison used in the main text. Realized edges have positive counts $N(e)$ and resistance $1/N(e)$; the source is $s_0 = (s_0, 0)$. Where stated explicitly, the electrical comparison permits arbitrary positive edge conductances.

B.1 RECURSIVE RESISTANCE AND ITS STRUCTURAL PROPERTIES

Write $\mathcal{G}_t$ for the realized layered search graph after t completed simulations and $\mathcal{E}_t$ for its set of observed action-labeled edges (u, a, w), namely those with $N(u, a, w) > 0$. The realized graph is finite: at most tH distinct observed edges can arise from t completed simulations. Every realized non-root node is reachable from s_0 and has at least one incoming observed edge. All observed counts are finite and positive. Consequently, the recursion below is well defined and gives $0 < \widetilde{\mathcal{R}}(v) < \infty$ for each realized non-root node. Each observed edge carries conductance $N(e)$ and resistance $1/N(e)$. The construction below motivates the local bonus in eq. (5); the algorithms do not need to maintain its deep recursion.

Definition 1 (Recursive resistance)**.** On the layered DAG $\mathcal{G}_t$, define $\widetilde{\mathcal{R}}(s_0) := 0$ and, in topological (depth) order, for every $v = (s', h+1)$ with parent edge set $\mathrm{Par}(v) = \{(u, a) : (u, a, v) \in \mathcal{E}_t\}$,

$$\widetilde{\mathcal{R}}(v) := \left(\sum_{(u,a)\in \mathrm{Par}(v)} \left(\widetilde{\mathcal{R}}(u) + \frac{1}{N(u, a, v)} \right)^{-1} \right)^{-1}. \tag{24}$$

Each parent branch is the parent's resistance in series with the connecting edge; branches combine in parallel. One topological pass costs $\mathcal{O}(\sum_v |\mathrm{Par}(v)|) = \mathcal{O}(|\mathcal{E}_t|)$ total and $\mathcal{O}(|\mathrm{Par}(v)|)$ to evaluate one node from already computed parent values. Updating an edge count may require recomputing descendants.

Proposition 2 (Structural properties of $\widetilde{\mathcal{R}}$)**.** *For every node v of the layered DAG $\mathcal{G}_t$ with positive edge counts:* (i) (tree consistency) *if every non-root node on every path to v has a single parent edge, then $\widetilde{\mathcal{R}}(v) = \sum_{e\in P(v)} 1/N(e)$, the path resistance used by tree-level ER bonuses;* (ii) (min-path domination) $\widetilde{\mathcal{R}}(v) \le \min_{P:s_0\to v} \sum_{e\in P} 1/N(e)$*;* (iii) (monotonicity) $\widetilde{\mathcal{R}}(v)$ *is nonincreasing in every edge count and strictly decreases when a new parent edge with finite resistance is added to any ancestor of v or to v itself (a Rayleigh-type law);* (iv) (parallel division) *if v has $k \ge 1$ parent edges whose branches have equal resistance x, then $\widetilde{\mathcal{R}}(v) = x/k$.*

Proof. We prove the four claims in order, each by induction along the topological (depth) order, which is well defined because every edge advances the depth coordinate by exactly one.

(i) Tree consistency. If every non-root node on every $s_0 \to v$ path has a single parent edge, then for each such node w with unique parent (u, a) the parallel sum in eq. (24) has one term, so $\widetilde{\mathcal{R}}(w) = \widetilde{\mathcal{R}}(u) + 1/N(u, a, w)$. Unwinding from v back to s_0, where $\widetilde{\mathcal{R}}(s_0) = 0$, gives $\widetilde{\mathcal{R}}(v) = \sum_{e\in P(v)} 1/N(e)$ for the unique path $P(v)$.

(ii) Min-path domination. For positive reals $x_1, \dots, x_k$ the parallel combination satisfies $(\sum_i x_i^{-1})^{-1} \le \min_i x_i$, since $\sum_i x_i^{-1} \ge \max_i x_i^{-1}$. Apply this to eq. (24): for any parent edge $(u, a) \in \mathrm{Par}(v)$,

$$\widetilde{\mathcal{R}}(v) \le \widetilde{\mathcal{R}}(u) + \frac{1}{N(u, a, v)}. \tag{25}$$

By induction along depth, $\widetilde{\mathcal{R}}(u) \le \min_{P:s_0\to u} \sum_{e\in P} 1/N(e)$, so the right side of eq. (25) is at most the resistance of the best path to u extended by the edge to v. Minimizing over parent edges covers every $s_0 \to v$ path, giving the claim. The base case $\widetilde{\mathcal{R}}(s_0) = 0$ is immediate.

(iii) Monotonicity. Write the recursion as $\widetilde{\mathcal{R}}(v) = \Phi\big((\widetilde{\mathcal{R}}(u) + r_{(u,a,v)})_{(u,a)\in\mathrm{Par}(v)}\big)$ with $\Phi(x_1, \dots, x_k) = (\sum_i x_i^{-1})^{-1}$. Since $\partial\Phi/\partial x_i = (x_i^{-2})\Phi^2 > 0$, the map Φ is strictly increasing in each argument, and it is strictly decreasing under appending a new argument: for $x_{k+1} < \infty$,

$\Phi(x_1, \ldots, x_{k+1}) < \Phi(x_1, \ldots, x_k)$ because the sum of conductances strictly increases. Increasing an edge count $N(e)$ decreases r_e and hence, by monotone composition along the topological order, decreases (weakly) $\widetilde{\mathcal{R}}(w)$ for every descendant w of the edge and leaves all other values unchanged. Adding a new parent edge at v or at any ancestor of v strictly decreases the corresponding $\widetilde{\mathcal{R}}$ value and, again by monotone composition, weakly decreases $\widetilde{\mathcal{R}}(v)$; the decrease is strict when the modified ancestor lies on some $s_0 \to v$ path, since then its value enters at least one branch of every recursion step down to v with positive derivative.

(iv) Parallel division. If all k branches equal x, then $\widetilde{\mathcal{R}}(v) = (k/x)^{-1} = x/k$. □

B.2 Unrolling identity and comparison with electrical resistance

Theorem 3 (Unrolling identity and weighted-network comparison)**.** *Let $\mathcal{T}_v$ denote the* path-unrolled network *of v, obtained by expanding reverse partial, action-labeled paths from v into a tree and then identifying all copies of s_0 into one terminal. Each copied edge retains the full resistance $1/N(e)$. Then*

$$\widetilde{\mathcal{R}}(v) = \mathcal{R}_{\text{eff}}\big(s_0, v; \mathcal{T}_v\big) \qquad \textit{and} \qquad \widetilde{\mathcal{R}}(v) \le \min_{P: s_0 \to v} \sum_{e \in P} \frac{1}{N(e)}, \tag{26}$$

When $\mathcal{G}_t$ is a tree, $\mathcal{T}_v$ is its unique s_0-to-v path, and equality holds throughout.

More generally, the identity holds when $N(e)$ is any positive edge conductance. In this class of positive weighted DAGs, $\widetilde{\mathcal{R}}(v)$ and $\mathcal{R}_{\text{eff}}(s_0, v; \mathcal{G}_t)$ are incomparable: either can exceed the other. Example 1 gives a four-layer weighted instance with $\widetilde{\mathcal{R}}(v)/\mathcal{R}_{\text{eff}}(s_0, v; \mathcal{G}_t) > 200$. This weighted-network comparison does not assert that every conductance assignment can arise from completed simulation trajectories.

Proof of the identity $\widetilde{\mathcal{R}}(v) = \mathcal{R}_{\text{eff}}(s_0, v; \mathcal{T}_v)$. Define $\mathcal{T}_v$ formally. Its nodes are the reverse partial paths, with action labels retained,

$$\sigma = \big(v = x_0 \xleftarrow{a_0} x_1 \xleftarrow{a_1} \cdots \xleftarrow{a_{j-1}} x_j\big), \qquad (x_{i+1}, a_i, x_i) \in \mathcal{E}_t \text{ for all } i,$$

with root $\sigma_\emptyset = (v)$; the children of σ ending at x_j are its one-step extensions by the parent edges of x_j, and the copy of the edge (x_{j+1}, a_j, x_j) joining σ to that child carries resistance $1/N(x_{j+1}, a_j, x_j)$, the full resistance of the original edge. Leaves are the reverse paths terminating at s_0; all leaves are identified (shorted) into a single terminal, also written s_0. The realized graph is finite and every reverse path has length at most H, so there are only finitely many reverse partial paths. Thus $\mathcal{T}_v$ is finite.

For a node σ of $\mathcal{T}_v$ ending at $x \in \mathcal{G}$, let $\rho(\sigma)$ be the effective resistance in the subtree rooted at σ between σ and the identified terminal. Before the terminal leaves are identified, the reverse-path expansion is a tree. After identification, distinct child branches share only σ and the shorted terminal; hence their currents combine in parallel, and each branch is a series composition of its first edge and the rest of the branch:

$$\rho(\sigma) = \left(\sum_{\text{children } \sigma' \text{ of } \sigma} \big(r_{\sigma\sigma'} + \rho(\sigma')\big)^{-1} \right)^{-1}, \qquad \rho(\text{leaf}) = 0. \tag{27}$$

(The series and parallel reduction rules are exact on trees with a single shorted terminal; see Lyons & Peres (2016, Ch. 2).) The children of σ ending at x are in bijection with $\mathrm{Par}(x)$, and the resistance of the edge to the child extending by (u, a) is $1/N(u, a, x)$. Comparing eq. (27) with eq. (24) and inducting on the (finite) depth of the subtree gives $\rho(\sigma) = \widetilde{\mathcal{R}}(x)$ for every σ ending at x—in particular the value depends on σ only through its endpoint, which is exactly the content of the recursion. Taking $\sigma = \sigma_\emptyset$ yields $\mathcal{R}_{\text{eff}}(s_0, v; \mathcal{T}_v) = \rho(\sigma_\emptyset) = \widetilde{\mathcal{R}}(v)$.

The min-path bound restates Proposition 2(ii). When $\mathcal{G}$ is a tree the unrolling is the identity map on the ancestor subgraph of v, so all three quantities coincide. □

Example 1 ($\widetilde{\mathcal{R}}$ and $\mathcal{R}_{\text{eff}}$ are incomparable)**.** *(a) $\widetilde{\mathcal{R}}$ can be far larger for arbitrary positive weights.* Consider the four-layer weighted DAG with $\mathcal{N}_0 = \{s_0\}, \mathcal{N}_1 = \{a_1, a_2\}, \mathcal{N}_2 = \{b_1, b_2\}, \mathcal{N}_3 = \{v\}$ and edge resistances

s_0a_1	s_0a_2	a_1b_1	a_1b_2	a_2b_1	a_2b_2	b_1v	b_2v
100	0.01	0.1	0.01	0.01	100	100	0.01

Then $\widetilde{\mathcal{R}}(v) \approx 33.34$ while $\mathcal{R}_{\text{eff}}(s_0, v) \approx 0.1396$, a ratio above 238. The reason is transparent: the electrically cheapest route $s_0 \to a_2 \to b_1 \to a_1 \to b_2 \to v$ has total resistance $0.01 + 0.01 + 0.1 + 0.01 + 0.01 = 0.14$, but it traverses the edge a_1b_1 *backwards*, from depth 2 to depth 1. No sequence of simulations realizes such a route, and the forward recursion correctly refuses to credit it: every forward path through this DAG passes through a resistance-100 edge. These weights demonstrate the general network comparison, not a realizable visit-count history: at a_1, for example, the incoming conductance is 0.01 whereas the outgoing conductances sum to 110, contrary to the count conservation of completed trajectories.

(b) $\widetilde{\mathcal{R}}$ can be smaller even for realized counts. Let $s_0 \to u$ have count 2, hence resistance $1/2$. At u, two distinct actions each lead to v once, giving two parallel action-labeled edges of count 1 and resistance 1. These counts arise from two completed trajectories, one using each action. Then

$$\mathcal{R}_{\text{eff}}(s_0, v) = \frac{1}{2} + \frac{1}{2} = 1, \qquad \widetilde{\mathcal{R}}(v) = \left((\tfrac{1}{2} + 1)^{-1} + (\tfrac{1}{2} + 1)^{-1}\right)^{-1} = \frac{3}{4}.$$

The unrolling duplicates the shared prefix. Repeated observations of the same action–successor pair would instead be aggregated into one counted edge.

Remark 1 (What survives, and what the algorithms use). Example 1 rules out a universal ordering on arbitrary positive weighted DAGs. Its first example does not settle the ordering when conductances must satisfy all visit-count constraints. The structural properties and unrolling identity in Proposition 2 and theorem 3 hold for the realized count network. The Part I guarantee theorem 1 uses the envelope eq. (6), proved from the count identity in section B.4. The Part II guarantees theorem 5 and Corollaries 6 and 7 use the conditional sampling floor; their constants do not involve the ER coefficients. The resistance calculus supplies the interpretation and design rationale for the bonus; Proposition 3 makes the connection exact at one layer.

B.3 One-layer resistances and the implemented bonus

The implemented bonus uses resistances over a single layer of the count network rather than the full recursion.

Proposition 3 (One-layer resistance form of the bonus)**.** *Fix a completed search history.* (i) *For an initialized query (v, a), consider the network formed by its realized edges (v, a, w), each with resistance $1/N_{v,a,w}$, with all realized successors identified into one terminal. Its effective resistance between v and that terminal is $R_1^{\text{out}}(v, a) = 1/T_{v,a}$.* (ii) *For a non-root node w, consider the network formed by its incoming realized edges (u, a', w), with all parent nodes identified into one terminal. Its effective resistance between that terminal and w is $R_1^{\text{in}}(w) = 1/T_w$. This value is the right side of eq.* (24) *with every parent resistance $\widetilde{\mathcal{R}}(u)$ replaced by zero. Consequently, $R_1^{\text{in}}(w) \le \widetilde{\mathcal{R}}(w)$, with equality if and only if w lies at depth one.* (iii) *The bonus in eq.* (5) *satisfies*

$$B_{v,a}^{\text{ER}} = C_2 R_1^{\text{out}}(v, a) + C_3 \sum_{w: N_{v,a,w} > 0} \widehat{p}(w \mid v, a)\, R_1^{\text{in}}(w).$$

Proof. (i) Every completed selection of (v, a) produces exactly one observed transition, so $\sum_w N_{v,a,w} = T_{v,a}$. After the realized successors are identified, the edges (v, a, w) all join v to the same terminal and therefore combine in parallel. Parallel conductances add, so the effective conductance is $T_{v,a}$ and the resistance is $1/T_{v,a}$.

(ii) After the parents are identified, the incoming edges of w join the same two terminals in parallel, with total conductance $\sum_{(u,a') \in \text{Par}(w)} N_{u,a',w} = T_w$ by eq. (2). Setting $\widetilde{\mathcal{R}}(u) = 0$ in eq. (24) gives the same expression. Each branch in eq. (24) satisfies $(\widetilde{\mathcal{R}}(u) + 1/N_{u,a',w})^{-1} \le N_{u,a',w}$ because $\widetilde{\mathcal{R}}(u) \ge 0$, so the parallel sum is at most T_w and $\widetilde{\mathcal{R}}(w) \ge 1/T_w$. If w lies at depth one, its only parent is the root, whose resistance is zero, and equality holds. Otherwise every parent u is a non-root node, and $\widetilde{\mathcal{R}}(u) > 0$ by induction on depth, since a parallel combination of positive branch resistances is positive. Every branch inequality is then strict, and so is the final one.

(iii) Substitute $\widehat{p}(w \mid v,a) = N_{v,a,w}/T_{v,a}$ and parts (i)–(ii) into eq. (5). □

Remark 2 (A variance reading of the two channels). The following heuristic is not used in any proof. Consider a current-value backup $\widehat{Q}^{\mathrm{cur}}_{v,a} = \widehat{r}_{v,a} + \gamma \sum_w \widehat{p}(w \mid v,a)\widehat{V}_w$, where $\widehat{r}_{v,a}$ is the empirical mean reward and $\widehat{V}_w$ are the current child estimates. Its error splits as

$$\widehat{Q}^{\mathrm{cur}}_{v,a} - Q^*(v,a) = \underbrace{\widehat{r}_{v,a} + \gamma \sum_w \widehat{p}(w \mid v,a)V^*(w) - Q^*(v,a)}_{\text{local sampling error}} + \gamma \sum_w \widehat{p}(w \mid v,a)\big(\widehat{V}_w - V^*(w)\big).$$

At a deterministic local sample size m, the local sampling error has second moment at most ν^2/m for a suitable variance bound ν^2. For intuition, consider an idealized model in which child estimates use fixed sample sizes $t_w \geq 1$, are independent of the empirical transition frequencies, and satisfy $\mathbb{E}[(\widehat{V}_w - V^*(w))^2] \leq c/t_w$. Jensen's inequality and $(x+y)^2 \leq 2x^2 + 2y^2$ then give

$$\mathbb{E}[(\widehat{Q}^{\mathrm{cur}}_{v,a} - Q^*(v,a))^2] \leq \frac{2\nu^2}{m} + 2\gamma^2 c \sum_w \frac{P(w \mid v,a)}{t_w}.$$

Replacing the unknown transition probabilities by empirical frequencies and the fixed sample sizes by pooled counts suggests the two count channels in eq. (5). This substitution is heuristic: the actual counts and child estimates are adaptive and dependent. The related electrical analogy for best linear unbiased estimation uses uncorrelated relative-measurement errors, an anchored reference, and edge resistances equal to measurement-noise covariances (Barooah & Hespanha, 2008). The algorithms analyzed here retain historical targets, so updating a shared child does not change its other parents' stored action estimates. The successor channel summarizes the pooled sampling effort available to future backup targets; it is not asserted to equal their estimation variance.

B.4 The local bonus and its sharing factor

Proof of Proposition 1. Fix an initialized query (v,a), with $T_{v,a} \geq 1$, and write $N_{v,a,w}$ for the number of its completed transitions to w. Let $\mathcal{W}_{v,a} = \{w : N_{v,a,w} > 0\}$ be its realized successor set. The count T_w pools completed visits to w over all incoming queries, and $|\mathcal{W}_{v,a}| \leq M$.

We verify the counting inequality, the bonus envelope, and predictability in turn.

Step 1: Compare local transition counts with pooled child counts. Every completed transition from query (v,a) to successor w contributes one visit to w. The pooled count T_w includes all these visits, together with any visits arriving through other queries. Thus, on every realized search history,

$$T_w \geq N_{v,a,w}.$$

For $w \in \mathcal{W}_{v,a}$, we additionally have $N_{v,a,w} \geq 1$, so

$$T_w \geq N_{v,a,w} \geq 1.$$

In particular, every denominator appearing in the bonus is positive, and

$$0 \leq \frac{N_{v,a,w}}{T_w} \leq 1. \tag{28}$$

Restricting the sum to realized successors avoids any undefined expression involving an unvisited child with $T_w = 0$.

Step 2: Bound the pooled-successor contribution. Factor out the local action count:

$$B^{\mathrm{ER}}_{v,a} = \frac{1}{T_{v,a}} \left[C_2 + C_3 \sum_{w \in \mathcal{W}_{v,a}} \frac{N_{v,a,w}}{T_w} \right].$$

By (28), each term in the sum is at most one. Hence

$$\sum_{w \in \mathcal{W}_{v,a}} \frac{N_{v,a,w}}{T_w} \leq |\mathcal{W}_{v,a}| \leq M.$$

Therefore,

$$B_{v,a}^{\mathrm{ER}} \le \frac{C_2 + C_3|\mathcal{W}_{v,a}|}{T_{v,a}} \le \frac{C_2 + C_3 M}{T_{v,a}}.$$

All coefficients and counts are nonnegative, so $B_{v,a}^{\mathrm{ER}} \ge 0$ as well. This proves (6).

Step 3: Verify predictability. Let $\mathcal{F}^-$ denote the information available in the complete search history immediately before the action is selected. The counts $T_{v,a}$, $N_{v,a,w}$, and T_w, as well as the realized successor set $\mathcal{W}_{v,a}$, are all $\mathcal{F}^-$-measurable. Since the bonus is a deterministic function of these quantities, it is also $\mathcal{F}^-$-measurable. Thus it is known before the next action and its simulator outcome are generated.

Setting the additional exploration term to $\xi_{v,a} = B_{v,a}^{\mathrm{ER}}$ consequently gives a predictable perturbation satisfying

$$0 \le \xi_{v,a} \le \frac{C_+}{T_{v,a}}, \qquad C_+ = C_2 + C_3 M.$$

These are the perturbation conditions used in the Part I analysis.

Step 4: Identify the sharing factor. For each realized successor, its contribution can be written as

$$\frac{N_{v,a,w}}{T_{v,a}} \frac{1}{T_w} = \frac{1}{T_{v,a}} \underbrace{\frac{N_{v,a,w}}{T_w}}_{\omega_w \in (0,1]}. \tag{29}$$

On a tree, the child has a unique incoming action edge, so $T_w = N_{v,a,w}$ and every sharing factor equals one. Summing gives the exact successor term $C_3 M_{v,a}^{\mathrm{real}}/T_{v,a}$. On a graph, ω_w is this query's share of the pooled child visits. It is less than one precisely when another incoming query has contributed a visit to w. This proves the remaining assertions of Proposition 1. □

Remark 3 (The bonus is not a function of the local count). Fix s, h, a and a value $m \ge 1$ of the local count, and consider two times $t < t'$ with $T_{s,h,a}(t) = T_{s,h,a}(t') = m$. Then $B_{s,h,a}^{\mathrm{ER}}(t') \le B_{s,h,a}^{\mathrm{ER}}(t)$ with strict inequality when $C_3 > 0$ and some realized successor $(s', h+1)$ received, during $(t, t']$, a transition from a query other than (s, h, a), since $T_{s',h+1}$ strictly increased while the numerators $T_{s,h,a}^{s'}$ and the denominator $T_{s,h,a}$ did not. Thus the bonus can change while this action selection count remains fixed, because simulations through other parents update the shared successor counts. The bound in theorem 7 allows this dependence on the full search history.

B.5 Cumulative exploration bonuses

The exploration bound also controls the total ER bonus along the sampled trajectories. In this calculation, initialization contributes zero; each later selection contributes the implemented bonus evaluated at the positive action count immediately before selection.

Proposition 4 (Cumulative ER bonuses under state sharing)**.** *For $n \ge 1$ completed root simulations, let $\mathrm{Bud}(n)$ be the sum of the paid structural bonuses over all initialized action selections on those trajectories, and let $\mathcal{E}_n^{\mathrm{act}} = \{(v,a) : T_{v,a}(n) \ge 1\}$. Define the cumulative-bonus upper bound*

$$\mathrm{Led}(n) := C_+ \sum_{(v,a)\in\mathcal{E}_n^{\mathrm{act}}} \big(1 + \log T_{v,a}(n)\big).$$

Then, pathwise,

$$\mathrm{Bud}(n) \le \mathrm{Led}(n) \le C_+ \,|\mathcal{E}_n^{\mathrm{act}}|\,(1 + \log n). \tag{30}$$

Moreover, for an active pair with $q = T_{v,a}(n) \ge 2$, let $\omega_{v,a,w}(j)$ be the sharing factor $N_{v,a,w}/T_w$ evaluated immediately before the j-th selection of (v,a), set to zero if $N_{v,a,w} = 0$ at that time, and define the weighted sharing factor

$$\bar{\omega}_{v,a,w} = \frac{1}{\mathcal{H}_{q-1}} \sum_{j=2}^{q} \frac{\omega_{v,a,w}(j)}{j-1} \in [0,1], \qquad \mathcal{H}_m = \sum_{s=1}^{m} \frac{1}{s}.$$

Here and in the sums below, w ranges over the final realized successor set $\mathcal{W}_{v,a}(n) = \{w : N_{v,a,w}(n) > 0\}$. The zero convention before a successor is first observed makes this one finite index set valid for every selection in the sum. The ER bonus paid on (v, a) is then exactly $\mathcal{H}_{q-1}\big(C_2 + C_3 \sum_w \bar{\omega}_{v,a,w}\big)$. For $q = 1$, the paid bonus is zero; set $\mathcal{H}_0 = 0$ and $\bar{\omega}_{v,a,w} = 0$ so that the same identity holds. On a tree, each $\omega_{v,a,w}(j)$ indicates whether w has been realized. Holding the pair's own selections and observed successors fixed, additional visits to w from other queries can only decrease the corresponding ratios and their weighted average. Under the same-trajectory coupling eq. (4), group active tree state–action pairs into their graph transposition classes. For a class with positive tree counts $t_1, \ldots, t_k$ and pooled graph count $T = \sum_{i=1}^{k} t_i$, the upper-bound contributions satisfy

$$C_+(1 + \log T) \le C_+ \left(k + \sum_{i=1}^{k} \log t_i \right).$$

Consequently,

$$\mathrm{Led}^{\mathcal{T}}(n) - \mathrm{Led}^{\mathcal{G}}(n) \ge C_+ \sum_{\text{active classes}} \big(k - 1 - \log k\big) \ge 0.$$

This compares upper bounds under a common trajectory history; it does not order the actual cumulative bonuses.

Proof. Fix a realization of the first n completed root simulations. All arguments below are deterministic on this realization. We use the same envelope constant $C_+ \ge 0$ for the tree and graph representations.

We will first bound the cumulative bonus of one action. Fix an active state–action pair (v, a) and write

$$q = T_{v,a}(n) \ge 1.$$

Index its selections chronologically by $j = 1, \ldots, q$. The first selection initializes the action and is not included in the paid-bonus sum. Immediately before its j-th selection, for $j \ge 2$, its local action count equals $j - 1$. Therefore, the bonus envelope gives

$$0 \le B^{\mathrm{ER}}_{v,a}(j) \le \frac{C_+}{j-1}, \qquad j = 2, \ldots, q,$$

where $B^{\mathrm{ER}}_{v,a}(j)$ denotes the bonus evaluated immediately before that selection.

Consequently, the total structural bonus paid on this pair is at most

$$\sum_{j=2}^{q} B^{\mathrm{ER}}_{v,a}(j) \le C_+ \sum_{s=1}^{q-1} \frac{1}{s} = C_+ \mathcal{H}_{q-1},$$

where

$$\mathcal{H}_m = \sum_{s=1}^{m} \frac{1}{s}, \qquad \mathcal{H}_0 = 0.$$

This reindexing is by actual selections of (v, a), rather than by the global simulation clock. In particular, changes in pooled successor counts between two selections may change the bonus, but do not themselves contribute a payment.

For $q = 1$, the sum is empty and $\mathcal{H}_{q-1} = 0 \le 1 + \log q$. For $q \ge 2$, the standard integral comparison yields

$$\mathcal{H}_{q-1} = 1 + \sum_{s=2}^{q-1} \frac{1}{s} \le 1 + \int_1^{q-1} \frac{dx}{x} = 1 + \log(q-1) \le 1 + \log q.$$

Thus, for every active pair,

$$\sum_{j=2}^{T_{v,a}(n)} B^{\mathrm{ER}}_{v,a}(j) \le C_+\big(1 + \log T_{v,a}(n)\big).$$

For $q \geq 2$, eq. (29) with local count $j-1$ writes the bonus at the j-th selection as $B^{\mathrm{ER}}_{v,a}(j) = (j-1)^{-1}\big[C_2 + C_3 \sum_w \omega_{v,a,w}(j)\big]$. Summing over $j = 2, \ldots, q$ gives $C_2\mathcal{H}_{q-1} + C_3 \sum_w \sum_{j=2}^{q} \omega_{v,a,w}(j)/(j-1) = \mathcal{H}_{q-1}\big(C_2 + C_3 \sum_w \bar{\omega}_{v,a,w}\big)$. Since $0 \leq \omega_{v,a,w}(j) \leq 1$ by eq. (28), each $\bar{\omega}_{v,a,w}$ is a weighted average of numbers in $[0,1]$. For $q = 1$, the exact identity follows from the zero convention above. Summing over all active state–action pairs proves

$$\mathrm{Bud}(n) \leq C_+ \sum_{(v,a)\in\mathcal{E}_n^{\mathrm{act}}} \big(1 + \log T_{v,a}(n)\big) = \mathrm{Led}(n).$$

We will now bound the ledger by the number of active pairs. Because the search graph is depth-augmented and each simulation follows a trajectory of strictly increasing depth, a fixed state–depth node is visited at most once in one simulation. Hence a fixed state–action pair is selected at most once per root simulation. Its count therefore satisfies

$$1 \leq T_{v,a}(n) \leq n, \qquad (v,a) \in \mathcal{E}_n^{\mathrm{act}}.$$

By monotonicity of the logarithm,

$$\mathrm{Led}(n) \leq C_+ \sum_{(v,a)\in\mathcal{E}_n^{\mathrm{act}}} (1 + \log n) = C_+ \, |\mathcal{E}_n^{\mathrm{act}}| \, (1 + \log n).$$

Together with the preceding bound, this establishes (30).

Now we will compare one transposition class. Under the same-trajectory coupling eq. (4), consider one active graph state–action pair. Let its active tree copies have counts

$$t_1, \ldots, t_k \in \{1, 2, \ldots\}.$$

Inactive copies have count zero and contribute to neither ledger, so they are omitted. The coupling gives the pooled graph count

$$T = \sum_{i=1}^{k} t_i.$$

The tree and graph ledger contributions of this class are, respectively,

$$L_{\mathcal{T}} = C_+ \sum_{i=1}^{k} (1 + \log t_i), \qquad L_{\mathcal{G}} = C_+(1 + \log T).$$

Their difference is exactly

$$L_{\mathcal{T}} - L_{\mathcal{G}} = C_+ \left(k - 1 + \sum_{i=1}^{k} \log t_i - \log T \right).$$

Since every $t_i \geq 1$, for each i we have

$$t_i \leq \prod_{j=1}^{k} t_j.$$

Summing these k inequalities gives

$$T = \sum_{i=1}^{k} t_i \leq k \prod_{i=1}^{k} t_i.$$

Taking logarithms therefore yields

$$\log T \leq \log k + \sum_{i=1}^{k} \log t_i.$$

Substituting into the expression for the ledger difference gives

$$L_{\mathcal{T}} - L_{\mathcal{G}} \geq C_+\big(k - 1 - \log k\big).$$

Finally, $\log k \le k-1$ for every integer $k \ge 1$, so this difference is nonnegative. In particular,

$$C_+(1+\log T) \le C_+\left(k + \sum_{i=1}^{k} \log t_i\right).$$

To sum over all active classes. The active tree state–action pairs partition into transposition classes, each corresponding to one active graph state–action pair. Summing the preceding bound over these classes gives

$$\mathrm{Led}^{\mathcal{T}}(n) - \mathrm{Led}^{\mathcal{G}}(n) \ge C_+ \sum_{\text{active classes}} \big(k - 1 - \log k\big) \ge 0.$$

This comparison concerns the two ledger upper bounds evaluated on a common trajectory history. It does not imply an ordering of the actual paid bonuses $\mathrm{Bud}^{\mathcal{T}}(n)$ and $\mathrm{Bud}^{\mathcal{G}}(n)$, nor does it require independently run tree and graph algorithms to generate the same trajectories. □

C TECHNICAL FORMS OF THE CONVERGENCE GUARANTEES

This section states the maximal concentration guarantees underlying section 5, together with their root-error consequences. The complete concentration proofs are in sections D, E, E.3 and E.4.

C.1 POWER-MEAN SEARCH WITH A BOUNDED ADDITIONAL BONUS

We write $\mathsf{MC}(\alpha, \beta; c)$ for the bound

$$\mathbb{P}\left(\sup_{k \ge r} |Z_k - z| > \varepsilon\right) \le c r^{-\alpha} \varepsilon^{-\beta}, \qquad r \ge 1, \quad \varepsilon > 0,$$

where r is an integer and z is a fixed target. Set $L = \max\{1, V_{\max} + HR_{\max}\}$. For $0 < \gamma \le 1$, all finite-horizon power-mean targets and stored estimates lie in $[0, L]$, by the backward-induction argument in section D.1.

Theorem 4 (Maximal concentration with a bounded additional bonus)**.** *Under Assumptions 2 and 3, use the historical and power-mean backups of section 4.1, with fixed finite $p \ge 1$, $C_1 > 0$, and the parameter schedule eq.* (14)*. In eq.* (8)*, replace $B^{\mathrm{ER}}_{v,a}$ by any additional bonus $\xi_{v,a}$ determined by the history before selection and satisfying $0 \le \xi_{v,a} \le C_+/T_{v,a}$ for fixed finite C_+. The ER bonus qualifies with $C_+ = C_2 + C_3 M$. Every depth-h node satisfies $\mathsf{MC}(\alpha_h, \beta_h; c_v)$ and, for $h < H$, each action-estimate sequence satisfies $\mathsf{MC}(\alpha_{h+1}, \beta_{h+1}; c_{v,a})$ about its finite-horizon target. The constants are finite for the fixed instance and parameters and independent of the sample count and deviation threshold. In particular, for every integer $n \ge 1$ and $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{k \ge n} |\widehat{V}_{s_0,k} - V^*(s_0, 0)| > \varepsilon\right) \le c_0 n^{-\beta_0/2} \varepsilon^{-\beta_0}. \tag{31}$$

The backward induction in section D.4 proves this statement. Integrating the root tail gives eq. (15) with $C_H = 1 + c_0/(\beta_0 - 1)$, as detailed below.

C.2 HISTORICAL REGULARIZED CONCENTRATION

Let $d = H - h$ be remaining depth and $\ell(r) = \log(2 + r)$. Write Q^*_Ω, V^*_Ω for the fixed regularized targets. The next result separates the number of selections of one action from the node visits pooled across all parents.

Theorem 5 (Historical regularized concentration)**.** *Under Assumption 1, GS-E3W-ER satisfies, for every internal node of remaining depth $d \ge 1$, integer $r \ge 1$, and $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{n \ge r} \|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v, \cdot)\|_\infty > \epsilon\right) \le A_d \exp\left(-\frac{r\epsilon^2}{S_d^2 \ell(r)^d}\right), \tag{32}$$

$$\mathbb{P}\left(\sup_{m \ge r} |\widehat{Q}_{v,a,m} - Q^*_\Omega(v, a)| > \epsilon\right) \le B_d \exp\left(-\frac{r\epsilon^2}{T_d^2 \ell(r)^{d-1}}\right). \tag{33}$$

The first bound also holds for the node-value error. Terminals have logarithmic exponent zero, $A_0 = 2$, and $S_0^2 = 2\sigma_0^2$. The explicit recursive constants in section E.3 are independent of the simulation budget. A zero concentration scale means the corresponding estimate is exact.

Historical averaging preserves the child's logarithmic exponent because its pooled count dominates its transition count from any parent. The sampling-probability bound ensures that, after r node visits, every action has at least order $r/\ell(r)$ selections with high probability. Since counts only increase, the maximal action-value bounds control all later visits, adding one logarithmic factor per depth. The 1-Lipschitz property of F_v gives the value bound. Section E supplies the proof, bounded-value intervals, and variance bounds accounting for transition uncertainty. This argument uses only the lower bound on sampling probabilities, so its constants do not depend on the ER coefficients. The regularized policy itself gives the following sharper rate.

A sharper bound for MENTS and fixed-reference KL. The stored action vectors lie in the deterministic action intervals J_d of eq. (79), and the ER bonus is bounded by $c_+ = c_2 + c_3 M$ after initialization. Thus, at remaining depth d, the span of the perturbed scores $\widehat{\mathbf{Q}} + B^{\mathrm{ER}}$ is bounded by the budget-independent constant D_d in eq. (98). For fixed temperature, softmax assigns every action probability at least $e^{-D_d/\tau}/K$. Fixed-reference KL assigns probability at least $me^{-D_d/\tau}$ when $\rho_v(a) \geq m > 0$ uniformly. Mixing with the uniform distribution preserves the positive lower bound $p_d \leq 1/K$ of Proposition 5, which does not decrease with the visit count.

Corollary 5 (Sharper bounds for MENTS and fixed-reference KL)**.** *Under Assumption 1, at horizon $H \geq 1$, suppose GS-E3W-ER uses MENTS or KL relative to fixed references with $\rho_v(a) \geq m > 0$, at fixed temperature $\tau > 0$ and fixed finite coefficients $c_2, c_3 \geq 0$. There are finite constants $\bar{A}_H \geq 2$ and $\bar{S}_H \geq 0$, independent of the budget, such that, for every integer $n \geq 1$ and $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{t \geq n} \|\widehat{\mathbf{Q}}_{s_0,t} - Q^*_\Omega(s_0, \cdot)\|_\infty > \epsilon\right) \leq \bar{A}_H \exp\left(-\frac{n\epsilon^2}{\bar{S}_H^2}\right). \tag{34}$$

The same bound holds for the root-value error relative to $V^*_\Omega(s_0)$. For a fixed horizon and fixed parameters, the expected absolute error satisfies $\mathbb{E}|\widehat{V}_{s_0,n} - V^*_\Omega(s_0)| = O(n^{-1/2})$. Node- and action-level bounds, explicit constants, and the proof appear in section E.3. A zero concentration scale means an exact estimate.

No sampling-rule change is needed for this corollary. Its constants depend on the minimum action probability and may be large for small temperatures, small reference probabilities, or large ER coefficients. TENTS can assign zero probability in its sparse component, so it retains the general guarantee of theorem 5. A constant uniform mixture would also admit the sharper analysis, but is a different sampling rule.

C.3 Maximal root recommendation and bias bounds

Assume theorem 5 with $H \geq 1$. After t simulations, recommend $a_t^{\mathrm{rec}} \in \mathrm{argmax}_a \, \widehat{\mathbf{Q}}_{s_0,t}(a)$ from the unperturbed estimates, with fixed tie-breaking.

Corollary 6 (Root action recommendation)**.** *Suppose the root has at least two actions and $Q^*_\Omega(s_0, \cdot)$ has a unique maximizing action a^*_Ω, with gap*

$$\Delta_\Omega := \min_{a \neq a^*_\Omega} \left[Q^*_\Omega(s_0, a^*_\Omega) - Q^*_\Omega(s_0, a)\right] > 0.$$

For every integer $n \geq 1$,

$$\mathbb{P}(\exists t \geq n : a_t^{\mathrm{rec}} \neq a^*_\Omega) \leq A_H \exp\left(-\frac{n\Delta_\Omega^2}{9S_H^2 \ell(n)^H}\right). \tag{35}$$

With only one root action, the recommendation error is zero.

Corollary 7 (Root value error and regularization bias)**.** *Let V^*, Q^* be the unregularized finite-horizon targets with the same reward model and V_0. For fixed $n \geq 1$ and $\delta \in (0,1)$, put*

$$\Psi_n(\delta) := S_H \sqrt{\ell(n)^H \log(A_H/\delta)/n}, \quad G_d := \textstyle\sum_{j=0}^{d-1} \gamma^j \ (G_0 := 0), \quad \Lambda_H := \gamma\tau(\omega_+ - \omega_-) G_{H-1}. \tag{36}$$

The action-value confidence event $\mathcal{E}_{n,\delta} := \{\sup_{t\geq n} \|\widehat{\mathbf{Q}}_{s_0,t} - Q^*_\Omega(s_0,\cdot)\|_\infty \leq \Psi_n(\delta)\}$ *satisfies* $\mathbb{P}(\mathcal{E}_{n,\delta}) \geq 1-\delta$*, and on it, simultaneously for all* $t \geq n$*,*

$$|\widehat{V}_{s_0,t} - V^*_\Omega(s_0)| \leq \Psi_n(\delta), \tag{37}$$

$$-\Psi_n(\delta) - \tau\omega_+ G_H \leq \widehat{V}_{s_0,t} - V^*(s_0) \leq \Psi_n(\delta) - \tau\omega_- G_H, \tag{38}$$

$$V^*(s_0) - Q^*(s_0, a_t^{\mathrm{rec}}) \leq 2\Psi_n(\delta) + \Lambda_H. \tag{39}$$

If the unregularized root has at least two actions and a unique optimum a^* *with gap*

$$\Delta^* := \min_{a\neq a^*}[Q^*(s_0,a^*) - Q^*(s_0,a)] > \Lambda_H,$$

then $a^*_\Omega = a^*$ *and eq.* (35) *holds with gap* $\Delta^* - \Lambda_H$*.*

The radius is fixed at n for all later times. The bonus is absent from backups, so it contributes no extra bias term. Complete proofs are in section E.4.

Full-support specialization. Under Corollary 5, the same confidence and error guarantees hold on the corresponding action-value confidence event with $\Psi_n(\delta)$ replaced by

$$\bar{\Psi}_n(\delta) := \bar{S}_H \sqrt{\frac{1}{n}\log\frac{\bar{A}_H}{\delta}}. \tag{40}$$

Likewise, under the gap assumption of Corollary 6,

$$\mathbb{P}(\exists t \geq n : a_t^{\mathrm{rec}} \neq a^*_\Omega) \leq \bar{A}_H \exp\left(-\frac{n\Delta^2_\Omega}{9\bar{S}^2_H}\right). \tag{41}$$

These statements follow from eq. (34) by the same root arguments; see section E.4. The logarithmic factors disappear, but the different constants mean that the two bounds need not be ordered at a particular finite budget.

The root expected-error and recommendation statements follow from the proofs in sections D.4 and E.4.

D Part I: power-mean graph-search proofs

We follow the depth-induction structure of Stochastic-Power-UCT (Dam et al., 2024b) and GS-Power-UCT: terminal concentration is transferred to action values and then to power-mean node values. We make the historical graph backup explicit and isolate the effect of ER in the local action-selection argument.

D.1 Setup and terminal concentration

Assumption 2 (Finite-depth fresh simulator)**.** *Fix a horizon* $H \geq 0$*, discount* $\gamma \in (0,1]$*, and a finite reachable state–depth graph. Each internal node has between one and* $K < \infty$ *actions, and each action has at most* $M < \infty$ *successors. At a fixed query* (s,h,a)*, the simulator returns a fresh reward–successor pair drawn from the same conditional law, independently of the preceding search history. Reward and successor within a pair may be dependent. Rewards belong to* $[0, R_{\max}]$*. Terminal evaluations are exact or fresh independent observations in* $[0, V_{\max}]$ *with mean* $V_0(s)$*.*

For $0 \leq h \leq H$, define

$$L_h = V_{\max} + (H-h)R_{\max}, \qquad L = \max\{1, V_{\max} + HR_{\max}\}.$$

The fixed-horizon targets satisfy

$$V^*(s,H) = V_0(s), \qquad Q^*(v,a) = \mathbb{E}[R + \gamma V^*(S) \mid v,a], \qquad V^*(v) = \max_a Q^*(v,a).$$

Here S is the successor state–depth node.

Backward induction shows that targets and stored estimates at depth h lie in $[0, L_h]$. At depth H, terminal values and their sample averages belong to $[0, V_{\max}] = [0, L_H]$. If child values lie in $[0, L_{h+1}]$, then, for $0 < \gamma \leq 1$,

$$0 \leq R + \gamma \widehat{V} \leq R_{\max} + \gamma L_{h+1} \leq R_{\max} + L_{h+1} = L_h.$$

Historical arithmetic averages and power-mean backups preserve this interval. The same induction applies to the Bellman targets. Consequently, every target and stored estimate lies in the common interval $[0, L]$.

Assumption 3 (Positive action gaps)**.** *Every internal node with at least two actions has a unique optimal action, with $\Delta_{v,a} = V^*(v) - Q^*(v,a) > 0$ for each suboptimal action. Single-action nodes require no gap assumption.*

Counts and historical estimates. Simulations reach depth H and update only their realized path, in reverse depth order. Let $T_v, T_{v,a}, N_{v,a,w}$ count completed node visits, action selections, and transitions, respectively. At completed simulation boundaries, every non-root node satisfies

$$T_w = \sum_{(v,a):v \to w} N_{v,a,w}.$$

On the i-th selection of (v,a), let (R_i, S_i) be the fresh simulator response and $\widetilde{T}_i$ the successor's pooled count after its update on that simulation. The stored action estimate is

$$\widehat{Q}_{v,a,m} = \frac{1}{m} \sum_{i=1}^{m} \big(R_i + \gamma \widehat{V}_{S_i, \widetilde{T}_i}\big). \tag{42}$$

Old targets are retained when other parents subsequently update a shared child. After n visits to v, its value is

$$\widehat{V}_{v,n} = \left(\sum_{a: T_{v,a}(n) > 0} \frac{T_{v,a}(n)}{n} \widehat{Q}^p_{v,a,T_{v,a}(n)} \right)^{1/p}, \qquad 1 \leq p < \infty. \tag{43}$$

Untried actions are selected first, in a fixed order. Subsequently, at a node of depth h, write $(b, \alpha, \beta) = (b_{h+1}, \alpha_{h+1}, \beta_{h+1})$ and select

$$a_{n+1} \in \operatorname*{argmax}_a \left\{ \widehat{Q}_{v,a,T_{v,a}(n)} + C_1 n^{b/\beta} T_{v,a}(n)^{-\alpha/\beta} + \xi_{v,a}(n) \right\}. \tag{44}$$

Here $C_1 > 0$ and the additional bonus is determined by the complete history before selection, with

$$0 \leq \xi_{v,a}(n) \leq C_+ / T_{v,a}(n) \tag{45}$$

for a fixed finite C_+. By Proposition 1, ER satisfies this condition with $C_+ = C_2 + C_3 M$. The bonus may change between visits to v as other parents collect successor observations.

Global search histories. Index action selections, simulator responses, and terminal evaluations in chronological order by t, and write $(\mathcal{F}_t)$ for the complete search history after these operations. Selection and its simulator response are separate operations: the node at a selection and the query at a response are known before that operation's random outcome. Deterministic updates between operations are included in the ensuing history. The following elementary fact will establish the existence of local sample streams in both parts of the proof.

Lemma 1 (Repeated events with divergent conditional probabilities)**.** *Let $I_t \in \{0,1\}$ be $\mathcal{F}_t$-measurable and put $p_t = \mathbb{E}[I_t \mid \mathcal{F}_{t-1}]$. Then*

$$\mathbb{P}\left(\sum_{t \geq 1} p_t = \infty, \quad \sum_{t \geq 1} I_t < \infty \right) = 0.$$

Proof. Set $c_{\mathrm{B}} = 1 - e^{-1}$ and

$$\mathcal{M}_N = \exp\left(c_{\mathrm{B}} \sum_{t=1}^{N} p_t - \sum_{t=1}^{N} I_t \right).$$

Since $\mathbb{E}[e^{-I_t} \mid \mathcal{F}_{t-1}] = 1 - c_{\mathrm{B}} p_t \le e^{-c_{\mathrm{B}} p_t}$, this is a nonnegative supermartingale with $\mathcal{M}_0 = 1$. If, for some N, $\sum_{t \le N} p_t \ge b$ and $\sum_{t \le N} I_t \le j$, then $\mathcal{M}_N \ge e^{c_{\mathrm{B}} b - j}$. Ville's inequality gives

$$\mathbb{P}\left(\exists N : \sum_{t \le N} p_t \ge b, \ \sum_{t \le N} I_t \le j \right) \le e^{j - c_{\mathrm{B}} b}.$$

This maximal inequality follows by stopping the nonnegative supermartingale at its first crossing of the displayed level, capping the stopping time by a deterministic horizon, and then letting that horizon tend to infinity. On the event $\{\sum_t p_t = \infty, \sum_t I_t \le j\}$, such a crossing occurs for every $b > 0$. Let $b \to \infty$ and then take the union over integers $j \ge 0$ to prove the assertion. □

Well-defined visit sequences. Consider a perpetual continuation with the same fixed parameters. At an infinitely visited internal node, the first K_v local visits complete initialization. Suppose some action counts remain bounded. Each such integer-valued nondecreasing count is eventually constant, whereas every unbounded count tends to infinity. After dividing the selection index by $n^{b/\beta}$, the bounded estimates and additional bonuses vanish, because $b > 0$ and $0 \le \xi_a \le C_+$ after initialization. A bounded-count action retains the positive limit $C_1 T_a(\infty)^{-\alpha/\beta}$; an unbounded-count action has limit zero. The action set is finite, so eventually only actions from the bounded-count set could be selected. Their total count would then increase on every visit, contradicting boundedness. Thus every action at an infinitely visited node is selected infinitely often, pathwise.

For a fixed positive-probability edge (v, a, w), apply lemma 1 to the global response indicators of this edge. Their conditional probabilities are $\mathbf{1}\{\text{the current response query is } (v, a)\} P(w \mid v, a)$, since the query is chosen before its fresh response. Their sum diverges whenever (v, a) is selected infinitely often. Hence w is then visited infinitely often almost surely. The root is visited on every simulation. Induction over depth and a finite intersection over the reachable nodes, actions, and positive-probability edges establish infinite visitation of all reachable local streams. Here reachability uses positive-probability transitions. Every finite local sample time is therefore almost surely finite, without any concentration hypothesis.

Lemma 2 (Fresh observations at adaptive local times)**.** *Fix a query whose successive sample times are almost surely finite. Let $\mathcal{C}_i$ be the complete history immediately before its i-th response, and let $\mathcal{G}_i$ be the complete history immediately after that response, with $\mathcal{G}_0$ the initial history. Suppose the response's true-target observation X_i has conditional mean μ and conditional sub-Gaussian variance proxy ν^2 given the pre-response history. Then*

$$\mathbb{E}[X_i \mid \mathcal{G}_{i-1}] = \mu, \qquad \mathbb{E}[e^{\lambda(X_i - \mu)} \mid \mathcal{G}_{i-1}] \le e^{\lambda^2 \nu^2 / 2}, \qquad \lambda \in \mathbb{R}. \tag{46}$$

The same assertion applies to adaptively scheduled terminal observations under their conditional moment assumptions.

Proof. Let τ_i be the global response stage of the query's i-th sample. For each deterministic t, the event $\{\tau_i = t\}$ is known before the response at stage t: it specifies the selected query and the number of its earlier responses. Decomposing according to these events transfers the model's conditional mean and moment-generating-function bounds to $\mathcal{C}_i$. More explicitly, for any $A \in \mathcal{C}_i$, each $A \cap \{\tau_i = t\}$ belongs to the pre-response history at t. Applying the conditional bound there and summing over t gives

$$\mathbb{E}[\mathbf{1}_A e^{\lambda(X_i - \mu)}] \le e^{\lambda^2 \nu^2 / 2} \mathbb{P}(A).$$

Thus the conditional MGF bound holds given $\mathcal{C}_i$, and the conditional mean follows in the same way. The stopped histories satisfy $\mathcal{G}_{i-1} \subseteq \mathcal{C}_i \subseteq \mathcal{G}_i$. The tower property now yields

$$\mathbb{E}[e^{\lambda(X_i - \mu)} \mid \mathcal{G}_{i-1}] = \mathbb{E}\Big[\mathbb{E}[e^{\lambda(X_i - \mu)} \mid \mathcal{C}_i] \mid \mathcal{G}_{i-1}\Big] \le e^{\lambda^2 \nu^2 / 2},$$

and likewise for the conditional mean. Arbitrary observations and updates at other nodes between two samples are already included in these histories. No step conditions on a final realized sample count or on an event describing future visits. □

Definition 2 (Maximal polynomial concentration)**.** A sequence (Z_k) with deterministic target z satisfies $\mathsf{MC}(\alpha, \beta; c)$ if, for every integer $r \ge 1$ and $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{k \ge r} |Z_k - z| > \varepsilon \right) \le c r^{-\alpha} \varepsilon^{-\beta}. \tag{47}$$

Constants may be enlarged to at least one.

This form controls adaptive pooled counts by event inclusion: on $N \geq r$, $|Z_N - z| \leq \sup_{k \geq r} |Z_k - z|$. No independence between N and the estimates is needed.

Lemma 3 (Bounded fresh averages)**.** *Let $Y_i \in [0, L]$ be adapted to $(\mathcal{F}_i)$ and satisfy $\mathbb{E}[Y_i \mid \mathcal{F}_{i-1}] = \mu$ for a deterministic μ. Then $\overline{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$ satisfies*

$$\mathbb{P}\left(\sup_{n \geq r} |\overline{Y}_n - \mu| > \varepsilon\right) \leq 2 \exp(-2r\varepsilon^2/L^2). \tag{48}$$

Consequently, for $0 < 2\alpha \leq \beta$, it satisfies $\mathsf{MC}(\alpha, \beta; c_{\mathrm{iid}})$, where

$$c_{\mathrm{iid}} = \max\left\{1, 2\left(\frac{\beta L^2}{4e}\right)^{\beta/2}\right\}.$$

Proof. With $S_n = \sum_{i=1}^{n}(Y_i - \mu)$, conditional Hoeffding's lemma makes $\exp(\lambda S_n - n\lambda^2 L^2/8)$ a nonnegative supermartingale starting at one. Set $\lambda = 4\varepsilon/L^2$. If $S_n > n\varepsilon$ for some $n \geq r$, this supermartingale exceeds $\exp(2r\varepsilon^2/L^2)$. Ville's inequality gives upper-tail probability at most $\exp(-2r\varepsilon^2/L^2)$; applying the same argument to $-S_n$ proves eq. (48). Finally, the inequality $x^a e^{-x} \leq (a/e)^a$ for $x \geq 0$, $a > 0$, applied with $x = 2r\varepsilon^2/L^2$ and $a = \beta/2$, gives

$$2e^{-2r\varepsilon^2/L^2} \leq 2\left(\frac{\beta L^2}{4e}\right)^{\beta/2} r^{-\beta/2}\varepsilon^{-\beta} \leq c_{\mathrm{iid}} r^{-\alpha}\varepsilon^{-\beta}.$$

□

The lemma applies to pooled terminal evaluations. Exact terminal values have zero error. This provides the base case for the depth induction.

D.2 Concentration of historical graph Q-values

The graph transfer uses the same count comparison as GS-Power-UCT. We first record the averaging fact needed for historical child estimates, which need not be independent.

Lemma 4 (Averaging tail error envelopes)**.** *Suppose (E_k) satisfies $\mathbb{P}(\sup_{k \geq r} |E_k| > \varepsilon) \leq c r^{-\alpha}\varepsilon^{-\beta}$ with $0 < \alpha < \beta$. Set $Z_j = \sup_{k \geq j} |E_k|$ and $A_n = n^{-1} \sum_{j=1}^{n} Z_j$. Both sequences are nonincreasing almost surely. For every $\theta \in (\alpha/\beta, 1)$,*

$$\mathbb{P}\left(\sup_{n \geq r} A_n > \varepsilon\right) = \mathbb{P}(A_r > \varepsilon) \leq c D_{\alpha,\beta,\theta} r^{-\alpha}\varepsilon^{-\beta}, \tag{49}$$

where

$$D_{\alpha,\beta,\theta} = \frac{2^\alpha}{(1 - 2^{\theta-1})^\beta (1 - 2^{\alpha - \theta\beta})}.$$

Proof. The assumption implies $Z_1 < \infty$ almost surely. Since Z_j is nonincreasing, $Z_{n+1} \leq A_n$ and $A_{n+1} \leq A_n$. For $j = 0, \ldots, \lfloor \log_2 r \rfloor$, partition $\{1, \ldots, r\}$ into $I_j = \{k : r2^{-j-1} < k \leq r2^{-j}\}$ and put $r_j = \lfloor r2^{-j-1} \rfloor + 1$. Then $|I_j|/r \leq 2^{-j}$ and $A_r \leq \sum_j 2^{-j} Z_{r_j}$. If each $Z_{r_j} \leq \varepsilon(1 - 2^{\theta-1})2^{j\theta}$, this sum is at most ε, since $(1 - 2^{\theta-1}) \sum_{j \geq 0} 2^{j(\theta-1)} = 1$. Also $r_j > r2^{-j-1}$, so $r_j^{-\alpha} \leq 2^{\alpha(j+1)} r^{-\alpha}$. Applying the assumed tail bound at each r_j and taking a union bound therefore gives

$$\mathbb{P}(A_r > \varepsilon) \leq c 2^\alpha (1 - 2^{\theta-1})^{-\beta} r^{-\alpha}\varepsilon^{-\beta} \sum_{j \geq 0} 2^{j(\alpha - \theta\beta)}.$$

The geometric series converges because $\theta\beta > \alpha$, which proves the stated constant. □

Theorem 6 (Concentration of historical graph Q-values)**.** *Fix a query (v, a) with positive-probability successors $w_1, \dots, w_m$, where $1 \le m \le M$. Under Assumption 2 and the historical update eq.* (42)*, consider a perpetual continuation of the search and assume that (v, a) is selected infinitely often almost surely. Suppose each child's pooled value-estimate sequence satisfies $\mathsf{MC}(\alpha, \beta; c_j)$ about $V^*(w_j)$, where $0 < 2\alpha \le \beta$.*

Then the action-estimate sequence satisfies $\mathsf{MC}(\alpha, \beta; c_Q)$ about $Q^*(v, a)$. For any $\theta \in (\alpha/\beta, 1)$, one may take

$$c_Q = 2^\beta c_{\mathrm{iid}} + (2\gamma m)^\beta D_{\alpha,\beta,\theta} \sum_{j=1}^{m} c_j. \tag{50}$$

Proof. The infinite-visitation assumption ensures that every finite local sample time of (v, a) is almost surely finite, as required by lemma 2. Since each listed successor has positive transition probability, the fresh-simulator assumption and lemma 1 imply that each is visited infinitely often almost surely. Thus the action and pooled child sequences used below are well defined on the perpetual continuation. For GS-Power-UCT-ER, the required infinite visitation was already established by the well-defined visit-sequence argument, independently of any concentration conclusion.

Abbreviate $Q^* = Q^*(v, a)$ and $V_j^* = V^*(w_j)$. Adding and subtracting the true successor values yields

$$\widehat{Q}_{v,a,n} - Q^* = \underbrace{\frac{1}{n} \sum_{i=1}^{n} (R_i + \gamma V^*(S_i) - Q^*)}_{B_n} + \frac{\gamma}{n} \sum_{i=1}^{n} (\widehat{V}_{S_i, \widetilde{T}_i} - V^*(S_i)). \tag{51}$$

The true-target observations $R_i + \gamma V^*(S_i)$ are fresh, belong to $[0, L]$, and have conditional mean Q^* given the complete history immediately before the response. By lemma 2, this conditional-mean property remains valid in the local query filtration despite adaptive query times and updates through other parents. Thus lemma 3 controls $\sup_{n \ge r} |B_n|$ with constant c_{iid}.

For child w_j, define

$$Z_{j,\ell} = \sup_{k \ge \ell} |\widehat{V}_{w_j,k} - V_j^*|, \qquad A_{j,n} = \frac{1}{n} \sum_{\ell=1}^{n} Z_{j,\ell}.$$

On the ℓ-th transition from (v, a) to w_j, its updated pooled count $\widetilde{T}_{j,\ell}$ satisfies $\widetilde{T}_{j,\ell} \ge \ell$. Consequently, writing $N_j(n)$ for the number of these transitions,

$$\frac{1}{n} \sum_{i=1}^{n} |\widehat{V}_{S_i, \widetilde{T}_i} - V^*(S_i)| \le \frac{1}{n} \sum_{j=1}^{m} \sum_{\ell=1}^{N_j(n)} Z_{j,\ell} \le \sum_{j=1}^{m} A_{j,n} \le \sum_{j=1}^{m} A_{j,r}, \qquad n \ge r. \tag{52}$$

Here $N_j(n) \le n$, the errors are nonnegative, and each $A_{j,n}$ is nonincreasing. Hence

$$\sup_{n \ge r} |\widehat{Q}_{v,a,n} - Q^*| \le \sup_{n \ge r} |B_n| + \gamma \sum_{j=1}^{m} A_{j,r}.$$

Apply lemma 3 at threshold $\varepsilon/2$ and lemma 4 to each child at threshold $\varepsilon/(2\gamma m)$. A union bound gives eq. (50). □

All comparisons with pooled counts are pathwise. The transfer requires neither independence between child estimates and counts nor a lower bound on successor probabilities. It holds for any adapted selection rule satisfying the stated premises; the next subsection establishes the child concentration needed by induction.

D.3 Power-mean concentration with ER exploration

We retain the Power-UCT proof structure: control suboptimal action counts, then transfer action-estimation errors to the power-mean backup. The new selection argument allows an additional,

history-dependent inverse-count bonus. We give the maximal version needed by the historical graph-backup induction.

Consider $K_v \geq 2$ actions with arithmetic running means

$$\widehat{Q}_{a,s} = \frac{1}{s}\sum_{i=1}^{s} X_{a,i}, \qquad X_{a,i} \in [0, L], \qquad L > 0.$$

The streams are defined on a perpetual continuation of the search; they may be adaptive and dependent across actions. Each satisfies $\mathsf{MC}(\alpha, \beta; c)$ about a deterministic $Q_a \in [0, L]$, with a common finite $c > 0$. Let a^* be the unique maximizing action, $Q^* = Q_{a^*} > 0$, and $\Delta_a = Q^* - Q_a > 0$ for $a \neq a^*$. Initialize each action once. After t completed selections, choose an action maximizing

$$I_a(t) = \widehat{Q}_{a,T_a(t)} + \mathcal{B}(t, T_a(t)) + \xi_a(t), \qquad \mathcal{B}(t, s) = C_1 t^{b/\beta} s^{-\alpha/\beta},$$

where $C_1 > 0$ and the perturbation is determined before selection and satisfies $0 \leq \xi_a(t) \leq C_+/T_a(t)$ for a fixed $C_+ \geq 0$. Feedback is sequential: a selected action's update completes before the next selection at this node. The parameter conditions are

$$2 < b < \alpha \leq \beta/2, \qquad \chi := 1 - b/\alpha \leq \alpha/\beta. \tag{53}$$

Write $\rho = b/\alpha \in (0, 1)$, so that $\chi = 1 - \rho$, and $\beta' = b - 1 > 1$. The maximal concentration premise is supplied by the preceding historical graph-action result; no independence of the action estimates is assumed here.

Lemma 5 (Additive exploration threshold)**.** *For a suboptimal action, define*

$$A_a(n) = \left(\frac{2C_1}{\Delta_a}\right)^{\beta/\alpha} n^{b/\alpha}, \qquad \mathcal{A}_a(n) = \left\lceil \max\left\{2, A_a(n) + \frac{2\beta C_+}{\alpha \Delta_a}\right\}\right\rceil.$$

Every integer $s \geq \mathcal{A}_a(n)$ satisfies

$$\mathcal{B}(n, s) + \frac{C_+}{s} \leq \frac{\Delta_a}{2}. \tag{54}$$

In particular, $\alpha/\beta = 1/2$ gives the additive threshold shift $4C_+/\Delta_a$ before rounding.

Proof. Put $q = \alpha/\beta \in (0, 1]$, $\delta = \Delta_a/2$, and $A = A_a(n)$. Then

$$\mathcal{B}(n, s) = \delta(A/s)^q. \tag{55}$$

Concavity gives $1 - x^q \geq q(1 - x)$ for $x \in [0, 1]$. Thus, for $s \geq A$,

$$\delta - \mathcal{B}(n, s) \geq q\delta(s - A)/s. \tag{56}$$

The right-hand side is at least C_+/s whenever $s \geq A + C_+/(q\delta)$, which is the stated threshold. □

Lemma 6 (Suboptimal-action selection bound)**.** *Set $C_T = 2c/((b-1)C_1^{\beta})$. For every integer $n \geq K_v$ and integer $u \geq \mathcal{A}_a(n)$,*

$$\mathbb{P}(T_a(n) > u) \leq C_T(u - 1)^{-\beta'}. \tag{57}$$

Proof. The claim is immediate for $u \geq n$, so suppose $2 \leq u < n$. For deterministic $t, s \geq 1$, define

$$U^-_{x,t,s} = \widehat{Q}_{x,s} + \mathcal{B}(t, s), \qquad U^+_{x,t,s} = \widehat{Q}_{x,s} + \mathcal{B}(t, s) + C_+/s.$$

At the actual action count,

$$U^-_{x,t,T_x(t)} \leq I_x(t) \leq U^+_{x,t,T_x(t)}. \tag{58}$$

Both envelopes are nondecreasing in t at fixed s; the actual perturbation need not be. The maximal concentration premise implies the fixed-count bound, also with a non-strict deviation inequality by taking limits of thresholds. Consequently,

$$\mathbb{P}(U^-_{a^*,t,s} \leq Q^*) \leq cs^{-\alpha}\mathcal{B}(t, s)^{-\beta} = cC_1^{-\beta}t^{-b}. \tag{59}$$

If $s \geq \mathcal{A}_a(t)$, lemma 5 gives $\mathcal{B}(t,s) + C_+/s \leq \Delta_a/2$. Hence $U^+_{a,t,s} > Q^*$ implies $\widehat{Q}_{a,s} - Q_a > \Delta_a/2 \geq \mathcal{B}(t,s)$, and

$$\mathbb{P}(U^+_{a,t,s} > Q^*) \leq cC_1^{-\beta} t^{-b}. \tag{60}$$

If $T_a(n) > u$, consider the $(u+1)$st selection of action a. Because $u \geq 2$ and each action is selected only once during initialization, this selection occurs after initialization. In particular, the optimal action has already been selected. Just before this selection let the total count be t and the optimal-action count be j. Then $u \leq t \leq n-1$, $j \geq 1$, and $t \geq u + j$. If both $U^+_{a,t,u} \leq Q^*$ and $U^-_{a^*,u+j,j} > Q^*$ held, the index sandwich and monotonicity would give $I_a(t) \leq Q^* < I_{a^*}(t)$, contradicting selection of a. Therefore

$$\{T_a(n) > u\} \subseteq \bigcup_{t=u}^{n-1} \{U^+_{a,t,u} > Q^*\} \cup \bigcup_{j=1}^{n-u} \{U^-_{a^*,u+j,j} \leq Q^*\}. \tag{61}$$

Since $\mathcal{A}_a(t) \leq \mathcal{A}_a(n) \leq u$ for $t \leq n$, the two envelope bounds and a union bound yield

$$\mathbb{P}(T_a(n) > u) \leq cC_1^{-\beta} \left(\sum_{t=u}^{\infty} t^{-b} + \sum_{j=1}^{\infty} (u+j)^{-b} \right) \leq \frac{2c}{(b-1)C_1^{\beta}} (u-1)^{-(b-1)}.$$

The last inequality follows by integral comparison. This enumeration uses one optimal-action count at a time, with reference clock $u + j$; it introduces no additional factor of n. □

Lemma 7 (Maximal suboptimal fraction)**.** *Let $J = K_v - 1$ and $U_n = \sum_{a \neq a^*} T_a(n)$. Define*

$$A = 3 + \max_{a \neq a^*} \left\{ \left(\frac{2C_1}{\Delta_a} \right)^{\beta/\alpha} + \frac{2\beta C_+}{\alpha \Delta_a} \right\}, \qquad z_0 = \max\{2^{\rho+1} JA, 4J\},$$

$$D_T = \frac{JC_T(2J)^{\beta'}}{1 - 2^{-\beta'}}, \qquad \widetilde{D}_T = \max\{1, z_0^{\beta'}, D_T\}.$$

For every integer $r \geq K_v$ and $z > 0$,

$$\mathbb{P}\left(\sup_{n \geq r} \frac{U_n}{n} > z \right) \leq \widetilde{D}_T r^{-\chi\beta'} z^{-\beta'}. \tag{62}$$

If $z \geq z_0 r^{-\chi}$, the stronger bound

$$\mathbb{P}\left(\sup_{n \geq r} \frac{U_n}{n} > z \right) \leq D_T r^{-\beta'} z^{-\beta'} \tag{63}$$

holds.

Proof. Rounding and $n^\rho \geq 1$ give $\mathcal{A}_a(n) \leq An^\rho$. First suppose $z \geq z_0 r^{-\chi}$. For each $j \geq 0$, put $L_j = 2^j r$, $N_j = 2^{j+1} r$, $w_j = zL_j/J$, and $u_j = \lfloor w_j \rfloor$. Monotonicity of the action counts implies

$$\{\exists n \in [L_j, N_j) : U_n/n > z\} \subseteq \bigcup_{a \neq a^*} \{T_a(N_j) > u_j\}. \tag{64}$$

Indeed, such an n gives $U_{N_j} > zL_j$, so some suboptimal action has count greater than w_j. The choice of z_0 ensures

$$w_j \geq 4, \qquad \frac{w_j}{2AN_j^\rho} = \frac{zr^\chi}{2^{\rho+1}JA} 2^{j\chi} \geq 1.$$

Thus $u_j \geq w_j/2 \geq AN_j^\rho \geq \mathcal{A}_a(N_j)$ and $u_j - 1 \geq w_j/2$. Applying lemma 6 at the deterministic horizon N_j and summing over actions and dyadic blocks gives

$$\mathbb{P}\left(\sup_{n \geq r} U_n/n > z \right) \leq JC_T \sum_{j \geq 0} \left(\frac{z2^j r}{2J} \right)^{-\beta'} = D_T r^{-\beta'} z^{-\beta'}.$$

This proves eq. (63) and, since $r^{-\beta'} \leq r^{-\chi\beta'}$, proves eq. (62) in this range. For $z < z_0 r^{-\chi}$, its proposed right-hand side exceeds one because $\widetilde{D}_T \geq z_0^{\beta'}$. □

Lemma 8 (Pathwise power-mean reduction)**.** *For a fixed finite $p \geq 1$, let*

$$\widehat{V}_n = \left(\sum_{a:T_a(n)>0} \frac{T_a(n)}{n} \widehat{Q}^p_{a,T_a(n)} \right)^{1/p}, \qquad C_E = p(L/Q^*)^{p-1}, \qquad C_U = (p+1)L^p/(Q^*)^{p-1}.$$

For every $n \geq K_v$,

$$|\widehat{V}_n - Q^*| \leq C_E |\widehat{Q}_{a^*,n} - Q^*| + C_U U_n/n. \tag{65}$$

Proof. Put $s = T_{a^*}(n)$, so $n - s = U_n$. Since both estimates are prefix means of the same bounded stream, for $s < n$ we have

$$\widehat{Q}_{a^*,n} = \frac{s}{n}\widehat{Q}_{a^*,s} + \frac{n-s}{n}\overline{X}_{s+1:n}, \qquad \overline{X}_{s+1:n} = \frac{1}{n-s}\sum_{i=s+1}^{n} X_{a^*,i} \in [0, L].$$

Both averages lie in $[0, L]$, so $|\widehat{Q}_{a^*,s} - \widehat{Q}_{a^*,n}| \leq L(n-s)/n = LU_n/n$. The case $s = n$ is immediate. The identity holds simultaneously for every pair of indices and hence at the adaptive count s.

Subtracting $(Q^*)^p$ from the weighted power sum gives

$$\widehat{V}_n^p - (Q^*)^p = \frac{s}{n}\big(\widehat{Q}^p_{a^*,s} - (Q^*)^p\big) + \sum_{a \neq a^*} \frac{T_a(n)}{n}\big(\widehat{Q}^p_{a,T_a(n)} - (Q^*)^p\big).$$

The first difference is at most $pL^{p-1}|\widehat{Q}_{a^*,s} - Q^*|$ in absolute value, and each remaining difference is at most L^p. Therefore

$$|\widehat{V}_n^p - (Q^*)^p| \leq pL^{p-1}|\widehat{Q}_{a^*,s} - Q^*| + L^p U_n/n.$$

For $x \geq 0$, $y > 0$, and $p \geq 1$, $|x^p - y^p| \geq y^{p-1}|x - y|$. Indeed, scaling by y reduces this to $|t^p - 1| \geq |t - 1|$ for $t \geq 0$, which follows from $t^p \leq t$ on $[0, 1]$ and $t^p \geq t$ on $[1, \infty)$. This argument applies to every real finite $p \geq 1$. Apply the inequality with $y = Q^*$ and combine the preceding inequalities. The future prefix $\widehat{Q}_{a^*,n}$ is used only in the analysis; the algorithm does not access unobserved samples. □

D.3.1 Node concentration

Theorem 7 (Power-mean concentration with bounded additional exploration)**.** *Under the bounded-running-mean, fixed-target maximal concentration, initialization, selection, sequential-feedback, and parameter conditions above, $\widehat{V}_n$ satisfies $\mathsf{MC}(\alpha', \beta'; c')$ about Q^* for every fixed finite $p \geq 1$, where*

$$\alpha' = \chi\beta' = (b-1)(1 - b/\alpha), \qquad \beta' = b - 1. \tag{66}$$

With the constants of lemmas 7 and 8, one may take

$$c'_{\text{large}} = \max\{1, c(2C_E)^\beta + \widetilde{D}_T(2C_U)^{\beta'}\}, \qquad c' = \max\{c'_{\text{large}}, K_v^{\alpha'} \max\{1, L^{\beta'}\}\}.$$

Explicitly, for all integers $r \geq 1$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{n \geq r} |\widehat{V}_n - Q^*| > \epsilon\right) \leq c' r^{-\alpha'} \epsilon^{-\beta'}.$$

Proof. First let $r \geq K_v$. Taking suprema in eq. (65), using the optimal action's maximal concentration, and applying lemma 7 yields

$$\mathbb{P}\left(\sup_{n \geq r} |\widehat{V}_n - Q^*| > \epsilon\right) \leq c(2C_E)^\beta r^{-\alpha}\epsilon^{-\beta} + \widetilde{D}_T(2C_U)^{\beta'} r^{-\chi\beta'}\epsilon^{-\beta'}. \tag{67}$$

Here $\beta > \beta'$ and $\alpha \geq \chi\beta$. If $\epsilon \geq r^{-\chi}$, then

$$\frac{r^{-\alpha}\epsilon^{-\beta}}{r^{-\chi\beta'}\epsilon^{-\beta'}} = r^{-\alpha+\chi\beta'}\epsilon^{-(\beta-\beta')} \leq r^{\chi\beta-\alpha} \leq 1.$$

Thus eq. (67) has the required exponents. If $0 < \epsilon < r^{-\chi}$, the proposed bound exceeds one because $c' \geq 1$. This proves the result for $r \geq K_v$.

During initialization, omitted zero-count terms leave weights summing to one, so $\widehat{V}_n, Q^* \in [0, L]$. For $1 \leq r < K_v$, deviations with $\epsilon \geq L$ are impossible; for $\epsilon < L$, $c' \geq K_v^{\alpha'} L^{\beta'}$ makes the claimed bound at least one. □

Remark 4 (Scope, constants, and single-action nodes). The proof uses bounded arithmetic running means, not estimators recomputed between selections. Their prefix identity in lemma 8 permits every finite $p \geq 1$; no separate $p > 2$ condition is needed. The constants contain $(Q^*)^{1-p}$, which is finite at a multi-action node with nonnegative targets and positive gaps. Only the selection threshold has an additive C_+/Δ_a shift; the final concentration constant need not.

If $K_v = 1$, $\widehat{V}_n$ is the sole action's running mean, and its target may be zero. Under eq. (53), the same output exponents hold with $c' = \max\{1, c\}$: above $\epsilon = r^{-\chi}$ use $r^{-\alpha}\epsilon^{-\beta} \leq r^{-\chi\beta'}\epsilon^{-\beta'}$, and below it use the trivial probability bound. No power-mean reduction or gap assumption is needed in this case.

D.4 Depth induction and root convergence

Proof of theorems 1 and 4. Use the schedule in eq. (14). We prove two claims by backward induction. Each depth-h node satisfies $\mathsf{MC}(\alpha_h, \beta_h; c_v)$ about $V^*(v)$. For $h < H$, the historical action sequences satisfy $\mathsf{MC}(\alpha_{h+1}, \beta_{h+1}; c_{v,a})$ about $Q^*(v, a)$. All constants below are finite and independent of the sample count and deviation threshold.

Terminal depth. At $h = H$, the value is exact or a pooled average of fresh bounded evaluations. Since $2\alpha_H = \beta_H$, lemma 3 establishes the claimed node concentration.

From child values to action values. Suppose the node claim holds at depth $h+1$ and fix a depth-h node v. Every successor lies at depth $h+1$ and therefore has the same remaining horizon as its other visits. Since $2\alpha_{h+1} = \beta_{h+1}$, theorem 6 applies to every action, for example with $\theta = 3/4$. It gives the claimed action concentration, with a common finite constant $c_v^Q = \max_a c_{v,a}$.

From action values to the node value. The action estimates are arithmetic running means of bounded historical targets. At a multi-action node, Assumption 3 gives a unique optimum; nonnegative targets imply $V^*(v) > 0$. Moreover, the schedule gives

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

Thus theorem 7 applies and returns

$$\beta' = b_{h+1} - 1 = \beta_h, \qquad \alpha' = (b_{h+1} - 1)(1 - b_{h+1}/\alpha_{h+1}) = \beta_h/2 = \alpha_h.$$

For a single-action node, $\widehat{V}_{v,n} = \widehat{Q}_{v,a,n}$. The input bound with exponents $(\beta_{h+1}/2, \beta_{h+1})$ implies the desired bound with $(\beta_h/2, \beta_h)$: for $\varepsilon \geq r^{-1/2}$ compare $(\sqrt{r}\,\varepsilon)^{-\beta_{h+1}}$ and $(\sqrt{r}\,\varepsilon)^{-\beta_h}$; below this threshold use the trivial probability bound. No positivity condition is needed in the single-action case. This completes the induction.

Root convergence. After n simulations the root has exactly n visits. Its maximal bound is therefore

$$\mathbb{P}\left(\sup_{k \geq n} |\widehat{V}_{s_0,k} - V^*(s_0, 0)| > \varepsilon\right) \leq c_0 n^{-\beta_0/2} \varepsilon^{-\beta_0},$$

which proves eq. (31). Integrating the tail and using $\beta_0 > 1$ gives

$$\begin{aligned}\mathbb{E}|\widehat{V}_{s_0,n} - V^*(s_0, 0)| &\leq n^{-1/2} + c_0 n^{-\beta_0/2} \int_{n^{-1/2}}^{\infty} \varepsilon^{-\beta_0}\, d\varepsilon \\ &= \left(1 + \frac{c_0}{\beta_0 - 1}\right) n^{-1/2}.\end{aligned}$$

This is eq. (15) with $C_H = 1 + c_0/(\beta_0 - 1)$. The induction allows every predictable bonus satisfying eq. (45); the ER bonus is an instance by Proposition 1. Finally, $b_{h+1}/\beta_{h+1} = 1/4$ and $\alpha_{h+1}/\beta_{h+1} = 1/2$, so the implemented polynomial exploration powers are the same at every depth. □

Corollary 8 (Most-visited root recommendation)**.** *Under theorem 1, with $H \geq 1$, recommend a most-visited root action,*

$$a_n^{\mathrm{rec}} \in \operatorname*{argmax}_a T_{s_0,a}(n),$$

with fixed tie-breaking. For all sufficiently large n,

$$\mathbb{P}(a_n^{\mathrm{rec}} \neq a^*) = O(n^{-\beta_0}).$$

With one root action the error is zero.

Proof. If a suboptimal action is most visited, its count is at least n/K_{s_0}. For sufficiently large n, $u = \lfloor n/K_{s_0} \rfloor - 1$ exceeds every root threshold in lemma 5, since these thresholds are $O(\sqrt{n})$. Apply lemma 6 at this u and take a union bound over the suboptimal actions. Its exponent is $b_1 - 1 = \beta_0$. Increasing the starting budget if necessary ensures $u - 1 \geq n/(2K_{s_0})$. Consequently,

$$\mathbb{P}(a_n^{\mathrm{rec}} \neq a^*) \leq (K_{s_0} - 1) C_T (2K_{s_0})^{\beta_0} n^{-\beta_0},$$

where C_T uses the common root action-concentration constant. All factors in this bound are independent of n. □

The recommendation result concerns the most-visited rule. The ER analysis preserves the root convergence exponents; its constants may depend on the fixed horizon, gaps, power exponent, and exploration coefficients.

E Part II: fixed-regularizer concentration proofs

We follow the E3W proof structure: control action sampling, propagate value concentration from the leaves to the root, and apply a gap argument to the final recommendation. The additional graph argument controls the historical child estimates retained in each action mean. One proof covers both the default exploration lower bound and the constant lower bound available for MENTS and fixed-reference KL. The former gives exponential concentration with logarithmic corrections; the latter gives exponential concentration in the simulation budget.

E.1 Setup and terminal concentration

Assumption 4 (Fixed finite-horizon model)**.** *Fix $H \geq 1$, $\gamma \in (0, 1]$, and a finite reachable state–depth graph. Each internal node has $1 \leq K_v \leq K < \infty$ actions, and each action has at most $M < \infty$ successors. Simulations reach depth H and update only their realized path, in reverse order. Every simulator query (v, a) returns a fresh reward–successor pair from its fixed conditional law, independently of the preceding search history. Rewards satisfy $0 \leq R \leq R_{\max}$; reward and successor within a pair may be dependent.*

At a terminal node w, $V_0(w) \in [0, V_{\max}]$. The successive observations $Y_{w,i}$ have conditional mean $V_0(w)$ and satisfy

$$\mathbb{E}\left[e^{\lambda(Y_{w,i} - V_0(w))} \mid \mathcal{F}_{w,i-}\right] \leq e^{\lambda^2 \sigma_0^2/2}, \qquad \lambda \in \mathbb{R},$$

where $\mathcal{F}_{w,i-}$ contains the complete search history before the observation. Exact terminal evaluations are allowed.

Assumption 5 (Fixed regularizers)**.** *Fix $\tau > 0$ and a convex regularizer Ω_v at each internal node. These are held fixed throughout the search and obey common finite bounds $\omega_- \leq \Omega_v(\pi) \leq \omega_+$ on Δ_{K_v}. Define*

$$F_v(\mathbf{Q}) = \max_{\pi \in \Delta_{K_v}} \{\langle \pi, \mathbf{Q} \rangle - \tau \Omega_v(\pi)\}, \qquad P_v(\mathbf{Q}) \in \operatorname*{argmax}_{\pi \in \Delta_{K_v}} \{\langle \pi, \mathbf{Q} \rangle - \tau \Omega_v(\pi)\}.$$

Assume maximizers exist and use a fixed measurable choice for ties. Strong convexity ensures uniqueness but is unnecessary for the concentration argument. KL reference policies, when used, are fixed at their state–depth nodes and positive on every action.

The value map is monotone and satisfies

$$|F_v(\mathbf{Q}) - F_v(\mathbf{Q}')| \le \|\mathbf{Q} - \mathbf{Q}'\|_\infty, \tag{68}$$

$$\max_a Q_a - \tau\omega_+ \le F_v(\mathbf{Q}) \le \max_a Q_a - \tau\omega_-. \tag{69}$$

For eq. (68), the objectives at any fixed policy differ by at most $\|\mathbf{Q} - \mathbf{Q}'\|_\infty$; maximizing preserves this bound in both directions. The upper range bound follows from $\langle \pi, \mathbf{Q}\rangle \le \max_a Q_a$; the lower bound follows by evaluating the objective at a point mass on a maximizing action. These properties require no differentiability of P_v.

Write $d = H - h$ for remaining depth. The fixed regularized targets are

$$\begin{aligned} V^*_\Omega(w) &= V_0(w), && d = 0, \\ Q^*_\Omega(v,a) &= \mathbb{E}[R + \gamma V^*_\Omega(S) \mid v, a], && d \ge 1, \\ V^*_\Omega(v) &= F_v(Q^*_\Omega(v,\cdot)), && d \ge 1. \end{aligned} \tag{70}$$

The same physical state at different depths is represented by different nodes. Thus every visit to a shared node has the same target. No action-gap assumption is imposed until the recommendation result.

Terminal estimates and historical backups. At a terminal node, retain the raw sum of observations and store

$$\overline{Y}_{w,k} = \frac{1}{k}\sum_{i=1}^{k} Y_{w,i}, \qquad \widehat{V}_{w,k} = \Pi_{[0,V_{\max}]}(\overline{Y}_{w,k}). \tag{71}$$

Since $V_0(w)$ lies in the projection interval,

$$|\widehat{V}_{w,k} - V_0(w)| \le |\overline{Y}_{w,k} - V_0(w)|. \tag{72}$$

Projection is inactive for observations already in the interval. It is applied to the pooled raw mean, not separately to observations or recursively to previously projected means. Exact evaluations use $\widehat{V}_{w,k} = V_0(w)$.

On the i-th selection of (v,a), let (R_i, S_i) be the simulator response and $\widetilde{T}_i$ the child's pooled count after its update on that simulation. The historical action estimate is

$$\widehat{Q}_{v,a,m} = \frac{1}{m}\sum_{i=1}^{m}\big(R_i + \gamma \widehat{V}_{S_i,\widetilde{T}_i}\big). \tag{73}$$

After n completed visits to v, store

$$\widehat{\mathbf{Q}}_{v,n} = \big(\widehat{Q}_{v,a,T_{v,a}(n)}\big)_{a\in\mathcal{A}_v}, \qquad \widehat{V}_{v,n} = F_v(\widehat{\mathbf{Q}}_{v,n}). \tag{74}$$

An unvisited action uses its prescribed initial estimate. Old targets in eq. (73) are retained when a shared child changes. The ER bonus enters selection only.

Sampling rule. Initialize every action once in a fixed order. Before local visit $k > K_v$, sample from

$$\begin{aligned} \pi_{v,k} &= (1-\lambda_{v,k})P_v(\widehat{\mathbf{Q}}_{v,k-1} + B^{\mathrm{ER}}_{v,\cdot}) + \lambda_{v,k}\,\mathrm{Unif}(\mathcal{A}_v), \\ \lambda_{v,k} &= \min\{1, \epsilon_\lambda K_v/\log(2+k)\}, \qquad \epsilon_\lambda > 0. \end{aligned} \tag{75}$$

All scores and counts are determined by the complete history before selection. Put

$$\ell(t) = \log(2+t), \qquad \eta = \min\{\epsilon_\lambda, 1/K\} > 0. \tag{76}$$

Then

$$\pi_{v,k}(a) \ge \min\{1/K_v, \epsilon_\lambda/\ell(k)\} \ge \eta/\ell(k). \tag{77}$$

This lower bound remains valid when observations collected through other parents change the ER bonus. The general concentration argument uses only this bound, so it also permits any other predictable probability distribution in place of P_v in the mixture.

Bounded estimates and fresh-target noise. Define the node intervals $I_d = [l_d, u_d]$ and widths $W_d = u_d - l_d$ by

$$\begin{aligned} l_0 &= 0, & u_0 &= V_{\max}, & W_0 &= V_{\max}, \\ l_d &= \gamma l_{d-1} - \tau\omega_+, & u_d &= R_{\max} + \gamma u_{d-1} - \tau\omega_-, & & \\ W_d &= R_{\max} + \gamma W_{d-1} + \tau(\omega_+ - \omega_-), & d &\geq 1. & & \end{aligned} \tag{78}$$

At remaining depth $d \geq 1$, define the action interval

$$J_d = [\gamma l_{d-1}, R_{\max} + \gamma u_{d-1}]. \tag{79}$$

Initialize every action estimate in J_d and initialize the node by applying F_v to that vector.

These intervals are preserved pathwise. Terminal estimates belong to I_0 by projection. If child values lie in I_{d-1}, each historical target lies in J_d, as does its arithmetic mean; applying eq. (69) puts the node value in I_d. The targets satisfy the same interval induction by eq. (70). Consequently, including initialization,

$$|\widehat{V}_{v,n} - V^*_\Omega(v)| \leq W_d, \qquad \|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v,\cdot)\|_\infty \leq W_d. \tag{80}$$

The action interval has width $R_{\max} + \gamma W_{d-1} \leq W_d$.

The fresh true-target observations $X_i = R_i + \gamma V^*_\Omega(S_i)$ have conditional mean $Q^*_\Omega(v,a)$ given the complete pre-response history. A uniform conditional sub-Gaussian variance proxy at remaining depth d is

$$\nu_d^2 = \begin{cases} \gamma^2 W_{d-1}^2/4, & R = r(v,a) \text{ almost surely for each fixed } (v,a), \\ (R_{\max} + \gamma W_{d-1})^2/4, & \text{bounded reward samples.} \end{cases} \tag{81}$$

These choices follow from the range form of Hoeffding's lemma; any smaller valid uniform proxy may instead be used. In particular, stochastic transitions can contribute noise even with exact leaves. A known reward of the form $r(v,a,S_i)$ generally belongs to the bounded-reward case: knowing the function does not remove randomness from its sampled successor argument. The first case requires the reward itself to be constant at the fixed query.

As established independently of concentration in Remark 5, every reachable local sample time is almost surely finite. The global-history convention and lemma 2 therefore apply here as well. They transfer these pre-response conditional bounds to the local query filtration, and the terminal moment assumption to the local terminal filtration. This justifies the next lemma's use at adaptively chosen sample times without conditioning on a future visit count.

Lemma 9 (Maximal concentration of fresh averages)**.** *Let X_i be adapted to $(\mathcal{F}_i)$, with deterministic conditional mean μ and*

$$\mathbb{E}[e^{\lambda(X_i - \mu)} \mid \mathcal{F}_{i-1}] \leq e^{\lambda^2\nu^2/2}, \qquad \lambda \in \mathbb{R}. \tag{82}$$

For every integer $r \geq 1$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{m \geq r} \left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| > \epsilon\right) \leq 2\exp\left(-\frac{r\epsilon^2}{2\nu^2}\right). \tag{83}$$

When $\nu = 0$, every $X_i = \mu$ almost surely and the probability is zero.

Proof. For $\nu > 0$, set $S_m = \sum_{i=1}^m (X_i - \mu)$. For fixed $\lambda > 0$, $M_m = \exp(\lambda S_m - m\lambda^2\nu^2/2)$ is a nonnegative supermartingale starting at one by eq. (82). Choose $\lambda = \epsilon/\nu^2$. If $S_m > m\epsilon$ at any $m \geq r$, then $M_m > \exp(r\epsilon^2/(2\nu^2))$. Ville's inequality therefore gives upper-tail probability at most $\exp(-r\epsilon^2/(2\nu^2))$. To justify that inequality, stop the nonnegative supermartingale at its first crossing of $c > 0$: for every finite N, optional stopping at the bounded stopping time gives $c\mathbb{P}(\tau_c \leq N) \leq \mathbb{E}M_{\tau_c \wedge N} \leq 1$; then let $N \to \infty$. Applying the same argument to $-S_m$ and taking a union bound proves eq. (83). For $\nu = 0$, conditional Jensen's inequality and eq. (82) force equality in the strictly convex exponential inequality, so $X_i = \mu$ almost surely. □

Applying the lemma to the raw terminal means and using projection gives, for every terminal node w,

$$\mathbb{P}\left(\sup_{k \geq r} |\widehat{V}_{w,k} - V_0(w)| > \epsilon\right) \leq A_0 \exp(-r\epsilon^2/S_0^2), \qquad A_0 = 2, \quad S_0^2 = 2\sigma_0^2. \tag{84}$$

A zero scale denotes an exact error sequence. This terminal bound is the base case for both versions of the depth induction.

E.2 Concentration of historical regularized Q-values

The action estimate averages the child values used on its past selections, even when other parents subsequently update those children. We first show that averaging decreasing error envelopes preserves the logarithmic exponent in an exponential concentration bound. We then apply this fact to the historical backup. Throughout this subsection, $\ell(t) = \log(2+t)$, and $\kappa \geq 0$ denotes a logarithmic exponent.

Lemma 10 (Averaging exponential tail envelopes)**.** *Let $(E_k)_{k\geq 1}$ be real-valued random errors. Suppose that, for some $A \geq 1$, $S \geq 0$, and $\kappa \geq 0$, every integer $r \geq 1$ and every $\epsilon > 0$ satisfy*

$$\mathbb{P}\left(\sup_{k\geq r} |E_k| > \epsilon\right) \leq A \exp\left(-\frac{r\epsilon^2}{S^2 \ell(r)^\kappa}\right). \tag{85}$$

When $S = 0$, this hypothesis means that the error sequence is exact: $E_k = 0$ for all k almost surely. Define

$$Z_j = \sup_{k\geq j} |E_k|, \qquad \mathcal{A}_n = \frac{1}{n}\sum_{j=1}^{n} Z_j,$$

and the numerical constants

$$\begin{aligned} a_0 &= 1 - 2^{-1/4}, & \kappa_C &= \frac{2}{a_0^2}, \\ D_C &= \frac{1}{1 - \exp(-(\sqrt{2}-1))}, & \mathcal{C}(A) &= \max\{e, D_C A\}. \end{aligned} \tag{86}$$

Then Z_j and $\mathcal{A}_n$ are finite and nonincreasing almost surely, and

$$\mathbb{P}\left(\sup_{n\geq r} \mathcal{A}_n > \epsilon\right) = \mathbb{P}(\mathcal{A}_r > \epsilon) \leq \mathcal{C}(A) \exp\left(-\frac{r\epsilon^2}{\kappa_C S^2 \ell(r)^\kappa}\right). \tag{87}$$

For $S = 0$, both deviation probabilities vanish. No independence between the errors is required.

Proof. The case $S = 0$ follows immediately from exactness, so assume $S > 0$. By eq. (85) with $r = 1$,

$$\mathbb{P}(Z_1 > x) \leq A \exp\left(-\frac{x^2}{S^2 \ell(1)^\kappa}\right) \longrightarrow 0 \qquad (x \to \infty).$$

Thus $Z_1 < \infty$ almost surely. The defining suprema give $0 \leq Z_{j+1} \leq Z_j \leq Z_1$, so all finite averages exist. Moreover, $Z_{n+1} \leq Z_j$ for $j \leq n$ and hence

$$\mathcal{A}_{n+1} = \frac{n\mathcal{A}_n + Z_{n+1}}{n+1} \leq \mathcal{A}_n.$$

It remains to bound $\mathbb{P}(\mathcal{A}_r > \epsilon)$.

Fix $r \geq 1$, put $J = \lfloor \log_2 r \rfloor$, and partition $\{1, \ldots, r\}$ into the dyadic blocks

$$I_j = \{i : r2^{-j-1} < i \leq r2^{-j}\}, \qquad r_j = \lfloor r2^{-j-1} \rfloor + 1, \qquad 0 \leq j \leq J.$$

The indices i in this definition are integers. These blocks are nonempty and disjoint, and their union is $\{1, \ldots, r\}$. Their smallest elements satisfy $r_j \geq r2^{-j-1}$, while $|I_j|/r \leq 2^{-j}$. Since the envelopes decrease,

$$\mathcal{A}_r = \frac{1}{r}\sum_{j=0}^{J}\sum_{i\in I_j} Z_i \leq \sum_{j=0}^{J} \frac{|I_j|}{r} Z_{r_j} \leq \sum_{j=0}^{J} 2^{-j} Z_{r_j}. \tag{88}$$

Assign threshold $t_j = \epsilon a_0 2^{3j/4}$ to block j. The choice of a_0 ensures

$$\sum_{j=0}^{\infty} 2^{-j} t_j = \epsilon a_0 \sum_{j=0}^{\infty} 2^{-j/4} = \epsilon.$$

Consequently, if every $Z_{r_j} \le t_j$, then eq. (88) gives $\mathcal{A}_r \le \epsilon$. Taking the complementary event and applying a union bound gives

$$\mathbb{P}(\mathcal{A}_r > \epsilon) \le \sum_{j=0}^{J} \mathbb{P}(Z_{r_j} > t_j) \le A \sum_{j=0}^{J} \exp\left(-\frac{r_j t_j^2}{S^2 \ell(r_j)^\kappa}\right). \tag{89}$$

Because $r_j \le r$, ℓ is increasing, and $\kappa \ge 0$,

$$\frac{r_j t_j^2}{S^2 \ell(r_j)^\kappa} \ge \frac{a_0^2 r \epsilon^2}{2 S^2 \ell(r)^\kappa} 2^{j/2} = y 2^{j/2}, \qquad y := \frac{r\epsilon^2}{\kappa_C S^2 \ell(r)^\kappa}.$$

It follows that

$$\mathbb{P}(\mathcal{A}_r > \epsilon) \le A \sum_{j=0}^{\infty} e^{-y 2^{j/2}}.$$

For $y \ge 1$, Bernoulli's inequality yields $2^{j/2} = (1+\sqrt{2}-1)^j \ge 1 + j(\sqrt{2}-1)$, and therefore

$$\begin{aligned} \sum_{j=0}^{\infty} e^{-y 2^{j/2}} &\le e^{-y} \sum_{j=0}^{\infty} e^{-yj(\sqrt{2}-1)} \\ &= \frac{e^{-y}}{1 - e^{-y(\sqrt{2}-1)}} \le D_C e^{-y}. \end{aligned}$$

Thus the probability is at most $A D_C e^{-y}$ in this case. For $0 < y < 1$, the trivial probability bound gives $\mathbb{P}(\mathcal{A}_r > \epsilon) \le 1 \le e\, e^{-y}$. Combining both ranges proves eq. (87) with $\mathcal{C}(A) = \max\{e, D_C A\}$. Every step uses a pathwise comparison or a union bound, so dependence among the errors does not affect the conclusion. □

Lemma 11 (Historical graph-action transfer)**.** *Fix an internal query (v, a) at remaining depth $d \ge 1$, with possible successor nodes $w_1, \dots, w_J$, where $1 \le J \le M$. Suppose that their pooled value-estimate sequences satisfy, for common constants $A \ge 1$, $S \ge 0$, and $\kappa \ge 0$,*

$$\mathbb{P}\left(\sup_{k \ge r} |\widehat{V}_{w_j,k} - V_\Omega^*(w_j)| > \epsilon\right) \le A \exp\left(-\frac{r\epsilon^2}{S^2 \ell(r)^\kappa}\right) \tag{90}$$

for all j, integers $r \ge 1$, and $\epsilon > 0$. Here k counts visits from all parents, and $S = 0$ means that each child estimate is exact after every positive pooled count.

At selection i of (v, a), let (R_i, S_i) be the fresh reward–successor pair and $\widetilde{T}_i$ the successor's pooled count after its update on that simulation. Suppose that the true-target observations

$$X_i = R_i + \gamma V_\Omega^*(S_i)$$

have conditional mean $Q_\Omega^*(v, a)$ and satisfy the conditional sub-Gaussian assumption of lemma 9, with variance proxy $\nu_d^2 \ge 0$, relative to the complete history before the query. For the historical action estimate

$$\widehat{Q}_{v,a,m} = \frac{1}{m} \sum_{i=1}^{m} \big(R_i + \gamma \widehat{V}_{S_i, \widetilde{T}_i}\big),$$

the following bound holds for every integer $r \ge 1$ and $\epsilon > 0$:

$$\mathbb{P}\left(\sup_{m \ge r} |\widehat{Q}_{v,a,m} - Q_\Omega^*(v, a)| > \epsilon\right) \le B \exp\left(-\frac{r\epsilon^2}{T^2 \ell(r)^\kappa}\right), \tag{91}$$

where

$$B = 2 + M\mathcal{C}(A), \qquad T^2 = \max\{8\nu_d^2,\, 4\gamma^2 M^2 \kappa_C S^2\}. \tag{92}$$

A zero output scale means that the action stream is exact. The transfer preserves the logarithmic exponent κ and requires neither independence of child estimates and counts nor a positive lower bound on successor probabilities.

Proof. Write $Q^* = Q^*_\Omega(v,a)$ and $V^*_j = V^*_\Omega(w_j)$. Adding and subtracting the true successor value in every stored target gives the exact decomposition

$$\widehat{Q}_{v,a,m} - Q^* = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\big(R_i + \gamma V^*_\Omega(S_i) - Q^*\big)}_{\Phi_m} + \gamma \underbrace{\frac{1}{m}\sum_{i=1}^{m}\big(\widehat{V}_{S_i,\widetilde{T}_i} - V^*_\Omega(S_i)\big)}_{\Xi_m}. \tag{93}$$

In particular, the second term retains the actual historical counts $\widetilde{T}_i$; it does not replace past child values by current ones. Define

$$D_m = \frac{1}{m}\sum_{i=1}^{m} |\widehat{V}_{S_i,\widetilde{T}_i} - V^*_\Omega(S_i)|.$$

Then $|\Xi_m| \le D_m$ and $|\widehat{Q}_{v,a,m} - Q^*| \le |\Phi_m| + \gamma D_m$. By the conditional mean and sub-Gaussian assumptions, lemma 9 gives, when $\nu_d > 0$,

$$\mathbb{P}\left(\sup_{m\ge r}|\Phi_m| > \epsilon/2\right) \le 2\exp\left(-\frac{r\epsilon^2}{8\nu_d^2}\right). \tag{94}$$

For $\nu_d = 0$, the fresh-average error is identically zero.

For each successor, introduce its tail envelopes and their averages,

$$Z_{j,l} = \sup_{k\ge l}|\widehat{V}_{w_j,k} - V^*_j|, \qquad \mathcal{A}_{j,n} = \frac{1}{n}\sum_{l=1}^{n} Z_{j,l}.$$

Let $N_j(m) = \sum_{i=1}^{m}\mathbf{1}\{S_i = w_j\}$, and let $i_{j,l}$ index the l-th occurrence of w_j among the query's samples. On that occurrence, the child has already received at least l pooled updates. Consequently,

$$\widetilde{T}_{i_{j,l}} \ge l, \qquad |\widehat{V}_{w_j,\widetilde{T}_{i_{j,l}}} - V^*_j| \le Z_{j,l}.$$

Additional visits from other parents can only increase this count. Grouping errors by successor, using $N_j(m) \le m$ and nonnegativity, and then applying the monotonicity in lemma 10, we obtain, simultaneously for all $m \ge r$,

$$D_m \le \frac{1}{m}\sum_{j=1}^{J}\sum_{l=1}^{N_j(m)} Z_{j,l} \le \sum_{j=1}^{J}\mathcal{A}_{j,m} \le \sum_{j=1}^{J}\mathcal{A}_{j,r}. \tag{95}$$

Empty occurrence sums are zero. This inequality holds pathwise, even when occurrence counts and child estimates are dependent.

If the fresh-average supremum is at most $\epsilon/2$ and every $\mathcal{A}_{j,r} \le \epsilon/(2\gamma M)$, then eqs. (93) and (95) and $J \le M$ give $|\widehat{Q}_{v,a,m} - Q^*| \le \epsilon$ for every $m \ge r$. Thus

$$\left\{\sup_{m\ge r}|\widehat{Q}_{v,a,m} - Q^*| > \epsilon\right\} \subseteq \left\{\sup_{m\ge r}|\Phi_m| > \epsilon/2\right\} \cup \bigcup_{j=1}^{J}\{\mathcal{A}_{j,r} > \epsilon/(2\gamma M)\}.$$

For $S > 0$, apply lemma 10 to each child and combine with eq. (94). A union bound yields

$$\mathbb{P}\left(\sup_{m\ge r}|\widehat{Q}_{v,a,m} - Q^*| > \epsilon\right) \le 2\exp\left(-\frac{r\epsilon^2}{8\nu_d^2}\right) + M\mathcal{C}(A)\exp\left(-\frac{r\epsilon^2}{4\gamma^2M^2\kappa_C S^2\ell(r)^\kappa}\right). \tag{96}$$

If $\nu_d = 0$, omit the first term. If $S = 0$, all child envelopes vanish and the same argument applies with the second term omitted.

Finally, $\ell(r)^\kappa \ge 1$ for $r \ge 1$ and $\kappa \ge 0$. When $T > 0$, its definition implies

$$T^2\ell(r)^\kappa \ge 8\nu_d^2, \qquad T^2\ell(r)^\kappa \ge 4\gamma^2M^2\kappa_C S^2\ell(r)^\kappa.$$

Every nonzero contribution in eq. (96) is therefore bounded by its prefactor times $\exp(-r\epsilon^2/(T^2\ell(r)^\kappa))$. Adding the prefactors proves eq. (91). If $T = 0$, both $\nu_d = 0$ and $S = 0$; the decomposition then shows that every action estimate equals its target almost surely. The proof uses only $\widetilde{T}_{i_{j,l}} \ge l$ and $N_j(m) \le m$; no inverse successor-probability factor or conditioning on an adaptive count is introduced. □

Historical averaging consequently passes the child's logarithmic exponent unchanged to the action stream. The remaining step is to convert this action-count bound into a bound indexed by pooled node visits, using the sampling probabilities of GS-E3W-ER.

E.3 Action sampling and regularized value concentration

We now convert concentration in an action's own selection count into concentration in the node's pooled visit count. The same argument covers the default exploration guarantee and the stronger guarantee available for MENTS and fixed-reference KL. Throughout this subsection, $\ell(k) = \log(2+k)$ and $q \in \{0, 1\}$ is fixed. At a node with remaining depth d, the required sampling condition is

$$\pi_{v,k}(a) \geq \frac{\eta_d}{\ell(k)^q}, \qquad 0 < \eta_d \leq 1/K, \qquad k > K_v. \tag{97}$$

All probabilities are conditional on the complete search history immediately before selection. The first K_v visits select each action once. Here q describes the sampling floor, not a power-mean parameter or an action value.

The two sampling regimes. For every fixed regularizer considered here, the uniform component in eq. (75) gives

$$\pi_{v,k}(a) \geq \min\{1/K_v, \epsilon_\lambda/\ell(k)\} \geq \frac{\eta}{\ell(k)}, \qquad \eta = \min\{\epsilon_\lambda, 1/K\}.$$

Thus eq. (97) holds with $q = 1$ and $\eta_d = \eta$. This conclusion does not require the regularized policy component to have full support, and does not depend on the ER coefficients. For MENTS and fixed-reference KL, bounded scores give the stronger choice $q = 0$, as follows.

Proposition 5 (Constant action-probability floors)**.** *Let $c_+ = c_2 + c_3 M$ for the implemented ER bonus, with fixed finite $c_2, c_3 \geq 0$, and define*

$$D_d = R_{\max} + \gamma W_{d-1} + c_+. \tag{98}$$

For fixed-reference KL, put $\rho_{\min,d} = \min_{v:\, H-h(v)=d} \min_{a \in \mathcal{A}_v} \rho_v(a) > 0$. At any internal node of remaining depth d, after initialization, the actual sampling policy has every coordinate at least

$$p_d = \begin{cases} K^{-1} \exp(-D_d/\tau), & \textit{MENTS}, \\ \min\{K^{-1}, \rho_{\min,d} \exp(-D_d/\tau)\}, & \textit{fixed-reference KL}. \end{cases} \tag{99}$$

In particular, $p_d \in (0, 1/K]$ is independent of the simulation budget.

Proof. By eq. (79), the stored action estimates lie in $J_d = [\gamma l_{d-1}, R_{\max} + \gamma u_{d-1}]$, of width $R_{\max} + \gamma W_{d-1}$. After initialization, eq. (6) gives

$$0 \leq B^{\mathrm{ER}}_{v,a} \leq c_+/T_{v,a} \leq c_+.$$

Therefore the perturbed scores $x_a = \widehat{\mathbf{Q}}_{v,k-1}(a) + B^{\mathrm{ER}}_{v,a}$ satisfy $\max_a x_a - \min_a x_a \leq D_d$ before every selection. For MENTS,

$$[P_v(x)]_a = \frac{e^{x_a/\tau}}{\sum_b e^{x_b/\tau}} \geq \frac{e^{-D_d/\tau}}{K_v} \geq \frac{e^{-D_d/\tau}}{K}.$$

For fixed-reference KL, since $\sum_b \rho_v(b) = 1$,

$$[P_v(x)]_a = \frac{\rho_v(a) e^{x_a/\tau}}{\sum_b \rho_v(b) e^{x_b/\tau}} \geq \rho_v(a) e^{-D_d/\tau} \geq \rho_{\min,d} e^{-D_d/\tau}.$$

The uniform distribution assigns each action probability $1/K_v \geq 1/K$. Consequently, mixing it with either regularized policy preserves the corresponding lower bound in eq. (99), for every possible mixture weight in $[0, 1]$. The reference minimum is positive because the reachable graph is finite and each reference is fixed with full support. All these inequalities hold conditional on the pre-selection history. □

Existence of the local streams.

Remark 5 (Perpetual continuation). Extend the search indefinitely with the same horizon and parameters. Fix a node–action pair (v, a) and, at each global operation t, let I_t indicate a selection of a at v. The operation's node and local visit index are known before the action is drawn. After initialization, its conditional probability is therefore

$$\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] = \mathbf{1}\{\text{the current operation selects at } v\}\pi_{v,k_t}(a),$$

where k_t is the local visit index when that indicator is one. On the event that v is visited infinitely often, the sum of these conditional probabilities diverges: by eq. (97), it dominates, after finitely many terms, $\sum_k \eta_d/\ell(k)^q = \infty$, for both $q = 0$ and $q = 1$. Lemma 1 implies that a is then selected infinitely often, almost surely. This is an implication between events in the global filtration, not conditioning the samples on an event describing their future visitation.

For a positive-probability successor w, apply the same lemma to the global indicators of responses (v, a, w). Their conditional probabilities are the predictable query indicator times $P(w \mid v, a) > 0$, so infinitely many selections of (v, a) imply infinitely many visits to w, almost surely. The root is visited in every simulation. Induction over depth and a finite intersection over reachable nodes, actions, and positive-probability edges establish infinite visitation of all reachable local streams. Every finite local sample time is thus almost surely finite, and all future-count suprema are well defined. This argument uses only the sampling floor and the fresh-simulator model; it does not use a concentration conclusion.

Lemma 12 (Action counts in the local visit sequence)**.** *At a node v with $K_v \le K$ actions, suppose initialization and eq.* (97) *hold. Let $T_{v,a}(r)$ count selections of action a among the first r visits, including initialization. For every deterministic integer $r \ge 2K$,*

$$\mathbb{P}\left(\min_{a\in\mathcal{A}_v} T_{v,a}(r) < \frac{\eta_d r}{4\ell(r)^q}\right) \le K \exp\left(-\frac{\eta_d r}{16\ell(r)^q}\right). \tag{100}$$

For $q = 1$ this is the logarithmic-floor count bound; for $q = 0$ it is the constant-floor count bound.

Proof. Fix an action a. For $k > K_v$, write $I_{a,k} = \mathbf{1}\{a_k = a\}$ and $Z_{a,r} = \sum_{k=K_v+1}^{r} I_{a,k}$, so $T_{v,a}(r) = 1 + Z_{a,r}$. For $k \le r$, monotonicity of ℓ gives the deterministic floor

$$\pi_{v,k}(a) \ge u_r := \frac{\eta_d}{\ell(r)^q}.$$

Let $\mathcal{H}_{v,k-}$ contain the entire history before selection. Conditionally on this history, $I_{a,k}$ is Bernoulli, whence

$$\begin{aligned}\mathbb{E}[e^{-I_{a,k}} \mid \mathcal{H}_{v,k-}] &= 1 - (1 - e^{-1})\pi_{v,k}(a) \\ &\le \exp(-(1 - e^{-1})u_r).\end{aligned}$$

Every earlier indicator is measurable in $\mathcal{H}_{v,k-}$. Iterating conditional expectations therefore yields

$$\mathbb{E}[e^{-Z_{a,r}}] \le e^{-(1-e^{-1})\mu_r}, \qquad \mu_r = (r - K_v)u_r \ge \frac{\eta_d r}{2\ell(r)^q}.$$

By Markov's inequality and $1/2 - e^{-1} \ge 1/8$,

$$\mathbb{P}(Z_{a,r} < \mu_r/2) \le e^{\mu_r/2}\mathbb{E}[e^{-Z_{a,r}}] \le e^{-\mu_r/8}.$$

Since $\eta_d r/(4\ell(r)^q) \le \mu_r/2$ and $T_{v,a}(r) \ge Z_{a,r}$, the event that this action's count is below the claimed threshold is contained in $\{Z_{a,r} < \mu_r/2\}$. A union bound over at most K actions proves eq. (100). The argument allows arbitrary information from other graph nodes between visits to v. Its time r is deterministic in the local visit sequence, not a random count conditioned on a global budget. □

Lemma 13 (Maximal transfer from action counts to node visits)**.** *Use the initialization and sampling conditions of lemma 12. Suppose that for each action, every integer $s \ge 1$, and every $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{m\ge s} |\widehat{Q}_{v,a,m} - Q^*_\Omega(v, a)| > \epsilon\right) \le B \exp\left(-\frac{s\epsilon^2}{T^2\ell(s)^\kappa}\right), \qquad B \ge 2, \quad T \ge 0, \quad \kappa \ge 0. \tag{101}$$

Assume also a deterministic error cap $W \geq 0$ satisfying

$$\|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v,\cdot)\|_\infty \leq W \quad \textit{almost surely for every } n \geq 1,$$

including initialized coordinates of actions not yet selected. Then, for every integer $r \geq 1$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{n\geq r} \|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v,\cdot)\|_\infty > \epsilon\right) \leq A' \exp\left(-\frac{r\epsilon^2}{S'^2 \ell(r)^{\kappa+q}}\right), \tag{102}$$

where

$$A' = K(1+B), \qquad S'^2 = \eta_d^{-1} \max\{16W^2, 4T^2\}. \tag{103}$$

The same bound holds for the node value $\widehat{V}_{v,n} = F_v(\widehat{\mathbf{Q}}_{v,n})$ relative to $V^*_\Omega(v) = F_v(Q^*_\Omega(v,\cdot))$. A zero scale is interpreted as an exact corresponding error stream.

Proof. If $W = 0$, the vector error vanishes identically, as does the value error by eq. (68). Suppose henceforth that $W > 0$, so $S' > 0$.

First take $r \geq 2K$ and define

$$m_r = \left\lceil \frac{\eta_d r}{4\ell(r)^q} \right\rceil, \qquad \mathcal{V}_r = \left\{ T_{v,a}(r) \geq \frac{\eta_d r}{4\ell(r)^q} \text{ for every } a \right\}.$$

Because $\eta_d \leq 1/K \leq 1$ and $\ell(r) > 1$, we have $1 \leq m_r \leq r$. On $\mathcal{V}_r$, integer-valued counts and their monotonicity give $T_{v,a}(n) \geq m_r$ for every action and every $n \geq r$. Thus the vector-deviation event is contained in

$$\mathcal{V}_r^c \cup \bigcup_{a\in\mathcal{A}_v} \left\{ \sup_{m\geq m_r} |\widehat{Q}_{v,a,m} - Q^*_\Omega(v,a)| > \epsilon \right\}.$$

This event inclusion uses no independence of estimates and counts, and no conditioning of an estimate on its realized count. The count lemma, the action-stream premise, and a union bound give

$$\mathbb{P}\left(\sup_{n\geq r} \|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v,\cdot)\|_\infty > \epsilon\right) \leq Ke^{-\eta_d r/(16\ell(r)^q)} + KBe^{-m_r\epsilon^2/(T^2\ell(m_r)^\kappa)}.$$

Since $m_r \leq r$, $\kappa \geq 0$, and $m_r \geq \eta_d r/(4\ell(r)^q)$,

$$\frac{m_r}{\ell(m_r)^\kappa} \geq \frac{\eta_d r}{4\ell(r)^{\kappa+q}}.$$

Consequently,

$$\mathbb{P}\left(\sup_{n\geq r} \|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v,\cdot)\|_\infty > \epsilon\right) \leq Ke^{-\eta_d r/(16\ell(r)^q)} + KBe^{-\eta_d r\epsilon^2/(4T^2\ell(r)^{\kappa+q})}. \tag{104}$$

If $T = 0$, the action streams are exact after their first samples, and their failure contribution in these displays is zero.

For $0 < \epsilon \leq W$, eq. (103) implies

$$\frac{r\epsilon^2}{S'^2\ell(r)^{\kappa+q}} \leq \frac{\eta_d r}{16\ell(r)^q} \quad \text{and} \quad S'^2 \geq 4T^2/\eta_d.$$

The first inequality uses $\epsilon^2 \leq W^2$ and $\ell(r)^\kappa \geq 1$. Both exponentials in eq. (104) are therefore at most the exponential in eq. (102); their prefactors sum to A'. For $\epsilon > W$ the deviation event is empty.

It remains to cover $1 \leq r < 2K$. For $0 < \epsilon \leq W$,

$$\frac{r\epsilon^2}{S'^2\ell(r)^{\kappa+q}} \leq \frac{r\eta_d}{16} < \frac{1}{8}.$$

Since $A' = K(1+B) \geq 3$, the proposed bound exceeds $3e^{-1/8} > 1$ and is valid throughout initialization. Finally, the 1-Lipschitz property of F_v gives pathwise

$$|\widehat{V}_{v,n} - V^*_\Omega(v)| \leq \|\widehat{\mathbf{Q}}_{v,n} - Q^*_\Omega(v,\cdot)\|_\infty.$$

Taking suprema proves the value claim with the same constants. □

The deterministic error cap is used to absorb the probability of too few action selections into a tail bound quadratic in ϵ. This is why the terminal convention retains raw observations but projects their pooled mean before using it in backups. Zero observation noise alone does not eliminate the initialization error of an untried action; the $16W^2$ term accounts for that error.

A unified induction over the remaining depth. Recall W_d and ν_d^2 from eqs. (80) and (81), and $\mathcal{C}(\cdot)$ and κ_C from eq. (86). The constants in the following theorem use the chosen regime q and its depth-dependent floors η_d.

Theorem 8 (Historical regularized concentration under a sampling floor)**.** *Under Assumptions 4 and 5, use the initialized historical estimators and terminal convention of the preceding subsections. Suppose eq.* (97) *holds at every remaining depth $d \geq 1$ for one fixed $q \in \{0, 1\}$. Define the finite constants recursively by*

$$\begin{aligned} A_0^{(q)} &= 2, & (S_0^{(q)})^2 &= 2\sigma_0^2, \\ B_d^{(q)} &= 2 + M\mathcal{C}(A_{d-1}^{(q)}), & (T_d^{(q)})^2 &= \max\{8\nu_d^2, 4\gamma^2 M^2 \kappa_C (S_{d-1}^{(q)})^2\}, \\ A_d^{(q)} &= K(1 + B_d^{(q)}), & (S_d^{(q)})^2 &= \eta_d^{-1} \max\{16 W_d^2, 4(T_d^{(q)})^2\}, \quad 1 \leq d \leq H. \end{aligned} \tag{105}$$

For every node v of remaining depth d, integer $r \geq 1$, and $\epsilon > 0$, the value estimate satisfies

$$\mathbb{P}\left(\sup_{n \geq r} |\widehat{V}_{v,n} - V_\Omega^*(v)| > \epsilon\right) \leq A_d^{(q)} \exp\left(-\frac{r\epsilon^2}{(S_d^{(q)})^2 \ell(r)^{qd}}\right). \tag{106}$$

At every internal node, the current action vector satisfies

$$\mathbb{P}\left(\sup_{n \geq r} \|\widehat{\mathbf{Q}}_{v,n} - Q_\Omega^*(v, \cdot)\|_\infty > \epsilon\right) \leq A_d^{(q)} \exp\left(-\frac{r\epsilon^2}{(S_d^{(q)})^2 \ell(r)^{qd}}\right), \tag{107}$$

and each historical action stream satisfies

$$\mathbb{P}\left(\sup_{m \geq r} |\widehat{Q}_{v,a,m} - Q_\Omega^*(v, a)| > \epsilon\right) \leq B_d^{(q)} \exp\left(-\frac{r\epsilon^2}{(T_d^{(q)})^2 \ell(r)^{q(d-1)}}\right). \tag{108}$$

Here n is the pooled node count and m the action's own selection count. All constants are independent of the simulation budget, and a zero scale means the corresponding error stream is exact.

Proof. We prove the statements by induction on the remaining depth d. The deterministic interval bounds from the first subsection hold before invoking concentration: every historical target lies in J_d, averages and initialization remain in J_d, and applying F_v places node values in I_d. Thus eq. (80) provides the required error cap W_d at every internal node. Remark 5 establishes existence of all the local streams without assuming the induction conclusion.

First suppose the noise proxies are positive. At $d = 0$, the fresh average bound lemma 9 gives

$$\mathbb{P}\left(\sup_{n \geq r} |\overline{Y}_{v,n} - V_0(v)| > \epsilon\right) \leq 2 \exp\left(-\frac{r\epsilon^2}{2\sigma_0^2}\right).$$

Projection onto $[0, V_{\max}]$ cannot increase the error about $V_0(v)$, so the same bound holds for $\widehat{V}_{v,n}$. This is eq. (106) at $d = 0$; terminal nodes have no action-vector statement.

Assume the value bound holds at depth $d - 1$. Fix an action (v, a) at depth d and enumerate its finitely many successors. For each successor, the induction hypothesis has constants $A_{d-1}^{(q)}, S_{d-1}^{(q)}$ and logarithmic exponent $q(d-1)$. The fresh true-target samples

$$X_i = R_i + \gamma V_\Omega^*(S_i)$$

are conditionally centered about $Q_\Omega^*(v, a)$ after subtracting that mean, and have the valid sub-Gaussian proxy ν_d^2. At the j-th transition from (v, a) to any given successor, that successor's post-update pooled count is at least j. Thus every premise of the historical graph-action transfer lemma 11 holds. Applying it with logarithmic exponent $q(d-1)$ gives eq. (108), with exactly $B_d^{(q)}, T_d^{(q)}$ from eq. (105).

Apply lemma 13 to these action streams, with $\kappa = q(d-1)$, floor $\eta_d / \ell(k)^q$, and error cap W_d. Its output exponent is $\kappa + q = qd$, and its constants are $A_d^{(q)}, S_d^{(q)}$. This proves the current-vector bound eq. (107); the Lipschitz backup gives eq. (106). No independence among successor estimates

or between those estimates and their pooled counts is used. This completes the positive-proxy induction.

For possibly zero proxies, replace σ_0^2 and every ν_d^2 by their sums with $\zeta > 0$. These remain valid proxies, whereas the intervals, error caps, and action-probability floors are unchanged. The proved result applies with the perturbed recursion. All its scales are continuous in ζ, so taking $\zeta \downarrow 0$ establishes each displayed bound with the original scales. If a scale tends to zero, the probability of every positive deviation is zero. Taking a countable intersection over integer starting counts and positive rational tolerances gives exactness of that error stream almost surely. Finiteness of the recursion follows from finite depth, finite K, M, bounded intervals, and positive η_d. □

Recovering the two sets of constants. For the default mixture, take $q = 1$ and $\eta_d = \eta$ and write

$$(A_d, S_d, B_d, T_d) = (A_d^{(1)}, S_d^{(1)}, B_d^{(1)}, T_d^{(1)}).$$

This proves the general node and action concentration statements of theorem 5. For MENTS or fixed-reference KL, Proposition 5 permits $q = 0$ and $\eta_d = p_d$; write

$$(\bar{A}_d, \bar{S}_d, \bar{B}_d, \bar{T}_d) = (A_d^{(0)}, S_d^{(0)}, B_d^{(0)}, T_d^{(0)}). \tag{109}$$

Then every logarithmic power in the theorem is zero. At the root, $d = H$ and the local count equals the number of simulations, giving the concentration claims of Corollary 5. The expected-error and recommendation consequences are derived in the following subsection.

Corollary 9 (Explicit logarithmic powers)**.** *For any $\vartheta > 0$, node v of remaining depth d, and integer $r \geq 1$,*

$$\mathbb{P}\left(\sup_{n \geq r} |\widehat{V}_{v,n} - V_\Omega^*(v)| > \frac{\vartheta}{\ell(r)}\right) \leq A_d^{(q)} \exp\left(-\frac{\vartheta^2 r}{(S_d^{(q)})^2 \ell(r)^{qd+2}}\right). \tag{110}$$

Proof. Substitute the fixed tolerance $\epsilon = \vartheta/\ell(r)$ into eq. (106). The tolerance is held fixed at the starting count r throughout the supremum; this is not a claim with a moving threshold $\vartheta/\ell(n)$. □

The proof therefore uses one logarithmic power per depth only when the available sampling floor decays as $1/\ell(k)$. Under constant floors, historical averaging and node transfer introduce no logarithmic factor. The sharper constants depend on the ER bonus through D_d and p_d and may be large for small temperatures, small reference probabilities, or large ER coefficients. TENTS does not in general provide a constant floor: its sparse component may assign zero probability to some actions, and its default uniform-mixture weight decreases with the visit count. None of these concentration comparisons proves that adding ER improves finite-budget accuracy.

E.4 Root concentration and action recommendation

At the root node $(s_0, 0)$, abbreviated as s_0 below, the pooled local count equals the number of completed simulations. We apply the preceding theorem with remaining depth H. Keep $q = 1$ for the general sampling guarantee and $q = 0$ when constant action-probability floors are available. This gives both root results from the same argument.

Maximal root concentration and expected error. Define

$$Z_n = \sup_{t \geq n} \|\widehat{\mathbf{Q}}_{s_0,t} - Q_\Omega^*(s_0, \cdot)\|_\infty.$$

By theorem 8, for every $n \geq 1$ and $\epsilon > 0$,

$$\mathbb{P}(Z_n > \epsilon) \leq A_H^{(q)} \exp\left(-\frac{n\epsilon^2}{(S_H^{(q)})^2 \ell(n)^{qH}}\right). \tag{111}$$

The root-value error is bounded pathwise by the action-vector error:

$$\sup_{t \geq n} |\widehat{V}_{s_0,t} - V_\Omega^*(s_0)| \leq Z_n. \tag{112}$$

For $\delta \in (0,1)$, let

$$\Psi_n^{(q)}(\delta) = S_H^{(q)} \sqrt{\frac{\ell(n)^{qH}}{n} \log \frac{A_H^{(q)}}{\delta}}, \qquad \mathcal{E}_{n,\delta}^{(q)} = \{Z_n \le \Psi_n^{(q)}(\delta)\}. \tag{113}$$

Substitution in eq. (111) gives

$$\mathbb{P}(\mathcal{E}_{n,\delta}^{(q)}) \ge 1-\delta, \qquad |\widehat{V}_{s_0,t} - V_\Omega^*(s_0)| \le \Psi_n^{(q)}(\delta) \quad \text{for every } t \ge n \quad \text{on } \mathcal{E}_{n,\delta}^{(q)}. \tag{114}$$

The radius is fixed at the starting budget n for all later budgets t. For $q=1$ it is Ψ_n in eq. (36); for $q=0$ it is $\bar{\Psi}_n$ in eq. (40).

To obtain the expected-error bound, suppose first that $S_H^{(q)} > 0$ and put $a_n = (S_H^{(q)})^2 \ell(n)^{qH}/n$. Then

$$\mathbb{P}(Z_n^2 > u) \le \min\{1, A_H^{(q)} e^{-u/a_n}\}, \qquad u > 0.$$

Since $A_H^{(q)} \ge 1$, integration at $u_0 = a_n \log A_H^{(q)}$ yields

$$\begin{aligned}
\mathbb{E}Z_n^2 &= \int_0^\infty \mathbb{P}(Z_n^2 > u)\, du \\
&\le a_n \log A_H^{(q)} + A_H^{(q)} \int_{a_n \log A_H^{(q)}}^\infty e^{-u/a_n}\, du \\
&= a_n(1 + \log A_H^{(q)}).
\end{aligned}$$

By Cauchy–Schwarz and eq. (112),

$$\mathbb{E}\left[\sup_{t \ge n} |\widehat{V}_{s_0,t} - V_\Omega^*(s_0)|\right] \le S_H^{(q)} \sqrt{1 + \log A_H^{(q)}} \sqrt{\frac{\ell(n)^{qH}}{n}}. \tag{115}$$

For a zero scale, the errors vanish almost surely and all these statements hold directly, with radius zero.

In particular, theorem 2 follows by taking $q = 1$ and $C_{\Omega,H} = S_H\sqrt{1+\log A_H}$. For MENTS and fixed-reference KL, take $q = 0$ and $\bar{C}_{\Omega,H} = \bar{S}_H\sqrt{1+\log \bar{A}_H}$ to obtain Corollary 2 and the expected-error claim in Corollary 5. The exponential statements concern deviation probabilities at a fixed positive tolerance. Integrating them gives an expected-error rate of order $\ell(n)^{H/2}/\sqrt{n}$ in the general case and $1/\sqrt{n}$ under constant sampling floors.

Recommendation of the best regularized action. After t simulations, recommend an action using the stored estimates without the exploration bonus:

$$a_t^{\mathrm{rec}} \in \operatorname*{argmax}_a \widehat{\mathbf{Q}}_{s_0,t}(a), \tag{116}$$

with fixed tie-breaking. Assume the root has at least two actions and a unique maximizer a_Ω^* of $Q_\Omega^*(s_0,\cdot)$, with

$$\Delta_\Omega = \min_{a \ne a_\Omega^*} \left[Q_\Omega^*(s_0,a_\Omega^*) - Q_\Omega^*(s_0,a)\right] > 0.$$

If $Z_n \le \Delta_\Omega/3$, then for every $t \ge n$ and every $a \ne a_\Omega^*$,

$$\begin{aligned}
\widehat{\mathbf{Q}}_{s_0,t}(a_\Omega^*) - \widehat{\mathbf{Q}}_{s_0,t}(a) &\ge Q_\Omega^*(s_0,a_\Omega^*) - Q_\Omega^*(s_0,a) - 2Z_n \\
&\ge \Delta_\Omega/3 > 0.
\end{aligned}$$

Thus $a_t^{\mathrm{rec}} = a_\Omega^*$ for every $t \ge n$ on this event. Its complementary event and eq. (111) give

$$\mathbb{P}(\exists t \ge n : a_t^{\mathrm{rec}} \ne a_\Omega^*) \le A_H^{(q)} \exp\left(-\frac{n\Delta_\Omega^2}{9(S_H^{(q)})^2 \ell(n)^{qH}}\right). \tag{117}$$

For $q=1$, this proves Corollary 6 and eq. (35); for $q=0$, it proves eq. (41). An error at budget n is included in an error at some $t \ge n$, so Corollary 3 follows as well. If the root has only one action, its recommendation error is zero. If the concentration scale is zero, the stored action vector is exact and the error is again zero under the stated gap assumption.

The recommended action maximizes the regularized continuation action values. The sampling policy P_{s_0} may remain stochastic at fixed temperature; no convergence of the simulation policy to a point mass is needed for this recommendation result.

Comparison with the unregularized targets. Let V^*, Q^* denote the unregularized finite-horizon targets with the same rewards and terminal values. Write

$$G_d = \sum_{j=0}^{d-1} \gamma^j \quad (d \geq 1), \qquad G_0 = 0, \qquad \Lambda_H = \gamma\tau(\omega_+ - \omega_-)G_{H-1}. \tag{118}$$

Proposition 6 (Finite-horizon regularization bias)**.** *At every node v with remaining depth d,*

$$-\tau\omega_+ G_d \leq V^*_\Omega(v) - V^*(v) \leq -\tau\omega_- G_d. \tag{119}$$

For each action at an internal node,

$$-\gamma\tau\omega_+ G_{d-1} \leq Q^*_\Omega(v,a) - Q^*(v,a) \leq -\gamma\tau\omega_- G_{d-1}. \tag{120}$$

Proof. The value assertion holds at $d = 0$, where both targets equal V_0. Suppose it holds at remaining depth $d-1$. The reward terms cancel when comparing action targets, so

$$Q^*_\Omega(v,a) - Q^*(v,a) = \gamma\mathbb{E}[V^*_\Omega(S) - V^*(S) \mid v, a].$$

Applying the inductive bounds inside this expectation proves eq. (120). Since they are common to every action,

$$\begin{aligned}\max_a Q^*(v,a) - \gamma\tau\omega_+ G_{d-1} &\leq \max_a Q^*_\Omega(v,a) \\ &\leq \max_a Q^*(v,a) - \gamma\tau\omega_- G_{d-1}.\end{aligned}$$

Apply eq. (69) to the regularized action vector and use $V^*(v) = \max_a Q^*(v,a)$:

$$\begin{aligned}V^*_\Omega(v) &\geq V^*(v) - \gamma\tau\omega_+ G_{d-1} - \tau\omega_+ = V^*(v) - \tau\omega_+ G_d, \\ V^*_\Omega(v) &\leq V^*(v) - \gamma\tau\omega_- G_{d-1} - \tau\omega_- = V^*(v) - \tau\omega_- G_d.\end{aligned}$$

Here $G_d = 1 + \gamma G_{d-1}$. This closes the induction. □

For the regularizers in section 4.2, valid common bounds are

Regularizer	ω_-	ω_+
MENTS	$-\log K$	0
TENTS	$-(K-1)/(2K)$	0
Fixed-reference KL, $\rho_v(a) \geq m > 0$	0	$\log(1/m)$.

The MENTS bounds follow from the entropy range on a simplex. For TENTS, $1/K_v \leq \|\pi\|_2^2 \leq 1$ and $K_v \leq K$. For KL, nonnegativity gives the lower bound, while

$$\sum_a \pi_a \log \frac{\pi_a}{\rho_v(a)} \leq \sum_a \pi_a \log \frac{1}{\rho_v(a)} \leq \log(1/m)$$

gives the upper bound. The value-bias interval depends on the normalization of the regularizer as well as its range width.

Root value error and simple regret on one confidence event. On $\mathcal{E}^{(q)}_{n,\delta}$, write $\Psi = \Psi^{(q)}_n(\delta)$. For every $t \geq n$, adding eq. (119) to the estimation bound eq. (114) gives

$$-\Psi - \tau\omega_+ G_H \leq \widehat{V}_{s_0,t} - V^*(s_0) \leq \Psi - \tau\omega_- G_H. \tag{121}$$

In particular, with $w_\Omega = \max\{|\omega_-|, |\omega_+|\}$,

$$|\widehat{V}_{s_0,t} - V^*(s_0)| \leq \Psi + \tau w_\Omega G_H. \tag{122}$$

For the recommendation, define $\delta_a = Q^*_\Omega(s_0,a) - Q^*(s_0,a)$. By eq. (120), all δ_a lie in an interval of width Λ_H. Let a^* be any maximizer of the unregularized root action values. Since a^{rec}_t maximizes the stored vector, on the same confidence event,

$$Q^*_\Omega(s_0,a^*) - Q^*_\Omega(s_0,a^{\mathrm{rec}}_t) \leq \left[Q^*_\Omega(s_0,a^*) - \widehat{\mathbf{Q}}_{s_0,t}(a^*)\right] + \left[\widehat{\mathbf{Q}}_{s_0,t}(a^{\mathrm{rec}}_t) - Q^*_\Omega(s_0,a^{\mathrm{rec}}_t)\right] \leq 2\Psi.$$

It follows that, simultaneously for all $t \geq n$,

$$V^*(s_0) - Q^*(s_0, a_t^{\mathrm{rec}}) = Q^*_\Omega(s_0, a^*) - Q^*_\Omega(s_0, a_t^{\mathrm{rec}}) - \delta_{a^*} + \delta_{a_t^{\mathrm{rec}}} \leq 2\Psi + \Lambda_H. \tag{123}$$

Only continuation regularization enters Λ_H: the root penalty is applied after its action values have been formed. In particular, $\Lambda_H = 0$ when $H = 1$.

Finally, suppose the unregularized root has at least two actions and a unique maximizing action a^* with gap $\Delta^* > \Lambda_H$. For every $a \neq a^*$,

$$Q^*_\Omega(s_0, a^*) - Q^*_\Omega(s_0, a) = Q^*(s_0, a^*) - Q^*(s_0, a) + \delta_{a^*} - \delta_a \geq \Delta^* - \Lambda_H > 0.$$

Thus $a^*_\Omega = a^*$, and the regularized gap is at least $\Delta^* - \Lambda_H$. Substituting this lower bound in eq. (117) yields

$$\mathbb{P}(\exists t \geq n : a_t^{\mathrm{rec}} \neq a^*) \leq A_H^{(q)} \exp\left(-\frac{n(\Delta^* - \Lambda_H)^2}{9(S_H^{(q)})^2 \ell(n)^{qH}}\right). \tag{124}$$

Taking $q = 1$ proves all claims of Corollary 7; taking $q = 0$ gives their full-support specialization. Restricting the common confidence event to $t = n$ proves Corollary 4. At a fixed temperature, regularization bias need not vanish with the simulation budget. The ER bonus changes sampling but is absent from the historical targets and from F_v, so it contributes no additional target-bias term. All results here concern the fixed-horizon state–depth graph; they make no unconditional claim for cyclic full-state merging.

F A CONDITIONAL SUPPORT PROPERTY FOR TSALLIS REGULARIZATION

For Tsallis regularization, the regularized policy component is $\mathrm{sparsemax}(z)$ with scores $z = \widehat{\mathbf{Q}}/\tau$. The ER perturbation changes these scores by $\xi = B^{\mathrm{ER}}/\tau$. Sparsemax is invariant to a uniform score shift, so only differences between bonus coordinates can change its support. Write $z_{(1)} \geq \cdots \geq z_{(K)}$ for sorted scores and define $\Gamma_k(z) = \sum_{j=1}^{k}(z_{(j)} - z_{(k)})$. Let $\mathcal{K}(z)$ be the set of actions assigned positive probability by sparsemax. The following sufficient condition describes one support change.

Lemma 14 (Order-compatible anti-monotone perturbations enlarge the support)**.** *Let $z \in \mathbb{R}^K$ and $\xi \in \mathbb{R}^K_{\geq 0}$ satisfy: (a)* anti-monotonicity*, $z_i \geq z_j \Rightarrow \xi_i \leq \xi_j$; and (b)* order compatibility*, there is a permutation sorting both z and $z + \xi$ in nonincreasing order. Then $\Gamma_k(z + \xi) \leq \Gamma_k(z)$ for every k, hence the sparsemax supports satisfy $\mathcal{K}(z) \subseteq \mathcal{K}(z + \xi)$: every previously active action remains active. Condition (b) cannot be dropped: for $z = (10, 0)$ and the anti-monotone $\xi = (0, 20)$, the supports are $\{1\}$ and $\{2\}$, which are not nested.*

Proof of lemma 14. Choose a common sorting permutation as in (b), write $z_{(1)} \geq \cdots \geq z_{(K)}$, and recall that the sparsemax support is $\mathcal{K}(z) = \{(j) : 1 + jz_{(j)} > \sum_{i \leq j} z_{(i)}\}$, equivalently $\mathcal{K}(z) = \{(j) : \Gamma_j(z) < 1\}$ with the gap sums $\Gamma_j(z) := \sum_{i \leq j}(z_{(i)} - z_{(j)}) \geq 0$ (Martins & Astudillo, 2016; Lee et al., 2018). With this common permutation, $\Gamma_j(z + \xi)$ is computed with the same index sets and

$$\Gamma_j(z + \xi) = \sum_{i \leq j}\big(z_{(i)} + \xi_{(i)} - z_{(j)} - \xi_{(j)}\big) = \Gamma_j(z) + \sum_{i \leq j}\big(\xi_{(i)} - \xi_{(j)}\big).$$

By anti-monotonicity (a), $i \leq j$ implies $z_{(i)} \geq z_{(j)}$ hence $\xi_{(i)} \leq \xi_{(j)}$, so each summand is non-positive and $\Gamma_j(z + \xi) \leq \Gamma_j(z)$ for every j. Consequently $\Gamma_j(z) < 1$ implies $\Gamma_j(z + \xi) < 1$, i.e., $\mathcal{K}(z) \subseteq \mathcal{K}(z + \xi)$.

That (b) cannot be dropped is shown by $z = (10, 0)$ and $\xi = (0, 20)$, which is anti-monotone; here $\mathcal{K}(z) = \{1\}$ while $z + \xi = (10, 20)$ has $\mathcal{K}(z + \xi) = \{2\}$, and the two supports are not nested. Thus anti-monotonicity alone does not imply nested supports when rankings change. □

The count-based ER bonus is not guaranteed to be anti-monotone in the estimated action values or to preserve their ranking. Accordingly, lemma 14 is a conditional statement about the regularized policy component. After initialization, the full sampling policy in eq. (13) has full support because its uniform-mixture weight is positive. Support recovery and faster recovery under graph sharing require additional arguments beyond the concentration foundation in section 5.2.

G Extensions to full-state merging

Throughout this section, assume $0 < \gamma < 1$. The fixed-horizon state–depth results of sections D and E also allow $\gamma = 1$. The inverse-count ER envelope does not require a state–depth DAG. The missing issue for general full-state merging is concentration of the actual shared estimates. We separate the reusable statements from the additional premises needed for that issue. These premises cannot all be inferred from bounded rewards and a finite physical state space: Proposition 8 gives an exact cyclic example in which historical averaging has a slower rate even without sampling noise.

G.1 Counts and the full-state bonus

For full-state sharing, let $T_{s,a}$ count completed action backups, let $D_s = \sum_a T_{s,a}$ count completed decisions, and let T_s^{all} count all completed occurrences, including terminal evaluations. These counts need not agree. Internal power-mean backups use weights $T_{s,a}/D_s$. Terminal evaluations have separate statistics and do not overwrite an internal value estimate. The direct application of theorem 7 below requires the usual sequential feedback convention: at each state, an action backup completes before that state selects its next action. State revisits within one simulation require a separate delayed-feedback argument unless this convention is otherwise enforced. Proposition 7 explicitly includes this requirement.

Lemma 15 (Full-state ER envelope)**.** *Suppose $N_{s,a,w} \le T_w^{\mathrm{all}}$ on every completed history and there are at most M realized successors of (s,a). For $T_{s,a} \ge 1$, define*

$$B_{s,a}^{\mathrm{ER}} = \frac{C_2}{T_{s,a}} + C_3 \sum_{w: N_{s,a,w} > 0} \frac{N_{s,a,w}}{T_{s,a}} \frac{1}{T_w^{\mathrm{all}}}, \qquad C_2, C_3 \ge 0.$$

Then $0 \le B_{s,a}^{\mathrm{ER}} \le C_+/T_{s,a}$, where $C_+ = C_2 + C_3 M$. Computed before selection, this is an adapted perturbation of the type allowed by theorem 7.

Proof. For each realized successor, $N_{s,a,w}/T_w^{\mathrm{all}} \le 1$. Sum this inequality over at most M successors and divide by $T_{s,a}$. □

G.2 Conditional power-mean guarantees

Proposition 7 (Conditional fixed-horizon full-merge ER guarantee)**.** *Fix $H \ge 1$, and write $V_H(s,h)$ for the finite-horizon Bellman target with terminal function V_0 at depth H. Consider a full-state search using the preceding counts and terminal convention. Suppose its root decision process satisfies all the bounded-running-mean, initialization, selection, and feedback premises of theorem 7, with the ER bonus of lemma 15. In particular, its root action estimates obey*

$$\widehat{Q}_{a,m} = \frac{1}{m} \sum_{i=1}^{m} X_{a,i}, \qquad X_{a,i} \in [0, L].$$

Assume that there are deterministic effective action targets $\bar{Q}_H(s_0, a) \in [0, L]$ such that the actual ER-generated action streams satisfy $\mathsf{MC}(\alpha_+, \beta_+; c_H)$ about these targets, where

$$\beta_0 > 1, \qquad \beta_+ = 4(\beta_0 + 1), \qquad \alpha_+ = \beta_+/2, \qquad b = \beta_+/4.$$

For a root with multiple actions, assume the effective maximizer is unique with positive value and positive action gaps, as required by theorem 7. Put

$$\bar{V}_H(s_0) = \max_a \bar{Q}_H(s_0, a), \qquad |\bar{V}_H(s_0) - V_H(s_0, 0)| \le \delta_H^{\mathrm{merge}}$$

for a deterministic $\delta_H^{\mathrm{merge}} \geq 0$. Let $\widehat{V}_{n,H}^{\mathrm{F,ER}}$ be the root estimate after n completed simulations, and suppose each simulation includes a root decision. Then there is a finite c_H^{ER} such that

$$\mathbb{P}\left(\sup_{k\geq n} |\widehat{V}_{k,H}^{\mathrm{F,ER}} - V_H(s_0,0)| > \delta_H^{\mathrm{merge}} + \epsilon\right) \leq c_H^{\mathrm{ER}} n^{-\beta_0/2} \epsilon^{-\beta_0}. \tag{125}$$

Consequently,

$$\mathbb{E}|\widehat{V}_{n,H}^{\mathrm{F,ER}} - V_H(s_0,0)| \leq \delta_H^{\mathrm{merge}} + \frac{C_H^{\mathrm{F}}}{\sqrt{n}}, \qquad C_H^{\mathrm{F}} = 1 + \frac{c_H^{\mathrm{ER}}}{\beta_0 - 1}. \tag{126}$$

Proof. By lemma 15, the structural bonus satisfies the perturbation premise of theorem 7. That theorem gives maximal concentration for the root's decision-indexed power mean $\widehat{V}_t^{\mathrm{dec}}$ about $\bar{V}_H(s_0)$, with exponents

$$\beta' = b - 1 = \beta_0, \qquad \alpha' = (b-1)(1 - b/\alpha_+) = \beta_0/2.$$

For a single root action, $\widehat{V}_t^{\mathrm{dec}}$ is its running mean; its stronger input bound implies the displayed output exponents by splitting into $\sqrt{r}\epsilon \geq 1$ and $\sqrt{r}\epsilon < 1$ and enlarging the constant.

The number $D_{s_0}(k)$ of completed root decisions is at least k, even when additional root visits occur inside a simulation. Therefore, pathwise,

$$\sup_{k\geq n} |\widehat{V}_{k,H}^{\mathrm{F,ER}} - V_H(s_0,0)| \leq \delta_H^{\mathrm{merge}} + \sup_{t\geq n} |\widehat{V}_t^{\mathrm{dec}} - \bar{V}_H(s_0)|.$$

Apply the maximal decision-count bound to obtain eq. (125). No conditioning on the random count $D_{s_0}(k)$ is needed. Finally integrate the tail above $n^{-1/2}$:

$$\mathbb{E}|\widehat{V}_{n,H}^{\mathrm{F,ER}} - V_H(s_0,0)| \leq \delta_H^{\mathrm{merge}} + n^{-1/2} + c_H^{\mathrm{ER}} n^{-\beta_0/2} \int_{n^{-1/2}}^{\infty} \epsilon^{-\beta_0}\, d\epsilon,$$

which is eq. (126). □

Meaning of the premise. The fixed effective targets and their maximal concentration must be proved for the actual ER search. A bound with an additive error radius around a depth-specific target does not imply this premise. Neither do budget-dependent empirical Bellman fixed points provide the required fixed targets automatically. These are the effective-target premises of the local value step in Tran et al. (2026), strengthened to maximal concentration and imposed on the ER-generated streams. The proposition therefore gives a conditional extension, not an unconditional theorem for cyclic GS-Power-UCT-F-ER. Under its premises, the sufficient exploration threshold shift is $4C_+/(\bar{V}_H(s_0) - \bar{Q}_H(s_0,a))$ for each suboptimal effective action, by lemma 5.

Corollary 10 (Conditional budget-dependent-horizon guarantee)**.** *Suppose Proposition 7 applies to each fixed horizon used below. For budget n, start a fresh run with horizon H_n and keep that horizon fixed throughout its n simulations. Let V_∞^* be the optimal infinite-horizon discounted value and put $E_0 = \|V_0 - V_\infty^*\|_\infty < \infty$. Then*

$$\mathbb{E}|\widehat{V}_{n,H_n}^{\mathrm{F}+,\mathrm{ER}} - V_\infty^*(s_0)| \leq \frac{C_{H_n}^{\mathrm{F}}}{\sqrt{n}} + \delta_{H_n}^{\mathrm{merge}} + \gamma^{H_n} E_0. \tag{127}$$

For $n \geq 2$ and $H_n = \lceil \log n/(2\log(1/\gamma)) \rceil$, this becomes

$$\mathbb{E}|\widehat{V}_{n,H_n}^{\mathrm{F}+,\mathrm{ER}} - V_\infty^*(s_0)| \leq \frac{C_{H_n}^{\mathrm{F}} + E_0}{\sqrt{n}} + \delta_{H_n}^{\mathrm{merge}}.$$

For the displayed horizon, the bound establishes consistency if $C_{H_n}^{\mathrm{F}}/\sqrt{n} \to 0$ and $\delta_{H_n}^{\mathrm{merge}} \to 0$. An overall $O(n^{-1/2})$ bound follows if $C_{H_n}^{\mathrm{F}} = O(1)$ and $\delta_{H_n}^{\mathrm{merge}} = O(n^{-1/2})$; these are additional conditions.

Proof. The discounted Bellman operator is a γ-contraction, so $|V_{H_n}(s_0,0) - V_\infty^*(s_0)| \leq \gamma^{H_n} E_0$. Add this deterministic truncation bound to eq. (126). For the displayed horizon, $\gamma^{H_n} \leq n^{-1/2}$. □

Why the concentration premise is substantive. The next result analyzes the direct cyclic implementation that completes the forward trajectory before performing its historical backups in reverse order. It retains one internal estimate per physical state and separate terminal statistics. A cutoff occurrence uses the terminal evaluation; every other occurrence uses the internal value just computed for its successor on that trajectory. This implementation can have pending selections at the same state, and is not covered by the sequential-feedback premise of Proposition 7.

Proposition 8 (An exact cyclic example and its convergence rate)**.** *Consider one physical state, one action, deterministic reward* 1*, and a deterministic transition back to the same state. Fix* $0 < \gamma < 1$*, use exact cutoff value* $V_0 = 0$*, initialize the action estimate to zero, and use the preceding reverse historical backups. For horizon* $H \geq 2$*, let* $\widehat{Q}_m$ *be the shared action estimate after* m *completed action backups. The root estimate after* n *simulations is* $\widehat{V}_{n,H} = \widehat{Q}_{nH}$*. Define*

$$\lambda_H = 1 - \gamma + \frac{\gamma}{H}, \qquad \overline{Q}_H = \frac{1}{\lambda_H}, \qquad U = \frac{1}{1-\gamma}. \tag{128}$$

For each fixed $H \geq 2$*,*

$$0 < \overline{Q}_H - \widehat{V}_{n,H} = \Theta(n^{-\lambda_H}). \tag{129}$$

In particular, if $\lambda_H < 1/2$*, the historical action stream does not satisfy* $\mathsf{MC}(\beta/2, \beta; C)$ *about any deterministic target, for any* $\beta > 0$ *and finite* C*.*

There is also a finite constant C_γ*, independent of* H *and* n*, such that, for every* $H \geq 2$ *and* $n \geq 1$*,*

$$\frac{\gamma}{(1-\gamma)[H(1-\gamma)+\gamma]} \leq U - \widehat{V}_{n,H} \leq \frac{\gamma}{(1-\gamma)[H(1-\gamma)+\gamma]} + C_\gamma (n+1)^{-(1-\gamma)}. \tag{130}$$

Consequently, fresh runs with $H_n = \Theta(\log n)$ *satisfy* $U - \widehat{V}_{n,H_n} = \Theta(1/\log n)$*. All these conclusions hold for every finite power* $p \geq 1$ *and every ER coefficient choice, since there is only one action.*

Proof. Write $c = 1 - \gamma \in (0, 1)$. Every simulation has H internal action occurrences. Its first reverse backup uses the exact terminal value zero, so its target is 1. Each remaining backup uses the current internal estimate as its successor value, so its target is $1 + \gamma \widehat{Q}_{m-1}$. With a single action, the power mean equals that action estimate. Thus, exactly,

$$\widehat{Q}_m = \left(1 - \frac{a_m}{m}\right) \widehat{Q}_{m-1} + \frac{1}{m}, \qquad a_m = \begin{cases} 1, & m \equiv 1 \pmod{H}, \\ c, & \text{otherwise.} \end{cases} \tag{131}$$

The terminal evaluation is never substituted for the stored internal estimate; it is used only in the target at the cutoff occurrence.

For simulation n, set $t = H(n-1)$ and define

$$A_n = \left(1 - \frac{1}{t+1}\right) \prod_{j=2}^{H} \left(1 - \frac{c}{t+j}\right), \qquad w_{n,j} = \frac{1}{t+j} \prod_{k=j+1}^{H} \left(1 - \frac{c}{t+k}\right), \quad 1 \leq j \leq H.$$

An empty product equals one. These are respectively the coefficient of the previous simulation's estimate and the propagated coefficient of the j-th update's additive term. Put $b = 1 - c\overline{Q}_H = \gamma \overline{Q}_H / H > 0$. The identity $1 - \overline{Q}_H = -(H-1)b$ then gives, for $e_n = \overline{Q}_H - \widehat{Q}_{nH}$,

$$e_n = A_n e_{n-1} + D_n, \qquad D_n = b \sum_{j=2}^{H} (w_{n,1} - w_{n,j}). \tag{132}$$

Indeed, this follows by composing the H affine updates in eq. (131) around the constant $\overline{Q}_H$. The first update has drift $-(H-1)b$ and the other $H-1$ updates have drift b.

For $1 \leq j < H$,

$$\frac{w_{n,j+1}}{w_{n,j}} = \frac{t+j}{t+j+1-c} = \frac{t+j}{t+j+\gamma} < 1.$$

Thus $D_n > 0$. Since $A_1 = 0$, we have $e_1 = D_1 > 0$; and $A_n > 0$ for $n \geq 2$, so $e_n > 0$ for every n. For fixed H, expansion of the finite product gives

$$\log A_n = -\frac{\lambda_H}{n} + O(n^{-2}), \qquad D_n = O(n^{-2}).$$

For the second estimate, each $w_{n,j} = 1/(Hn) + O(n^{-2})$, so their leading terms cancel in every difference. Let $P_1 = 1$ and $P_n = \prod_{k=2}^{n} A_k$. Summing the logarithmic expansion and using convergence of $\sum_k k^{-2}$ gives $P_n = \Theta(n^{-\lambda_H})$. Iterating eq. (132) yields

$$e_n = P_n \left(e_1 + \sum_{k=2}^{n} \frac{D_k}{P_k} \right).$$

Because $0 < \lambda_H < 1$, the positive series is bounded by a constant times $\sum_{k \geq 2} k^{\lambda_H - 2} < \infty$, and its bracket is bounded below by $e_1 > 0$. This proves eq. (129) and convergence to $\overline{Q}_H$.

If $\lambda_H < 1/2$, take $r = nH$ and $\epsilon = e_n/2$ in a proposed $\mathsf{MC}(\beta/2, \beta; C)$ bound about $\overline{Q}_H$. The event has probability one already at the deterministic index r, whereas its proposed upper bound is

$$C(nH)^{-\beta/2}(e_n/2)^{-\beta} = O\Big(n^{-\beta(1/2-\lambda_H)}\Big) \longrightarrow 0.$$

No other fixed target can satisfy the bound, because the subsequence $\widehat{Q}_{nH}$ converges to $\overline{Q}_H$.

To control growing horizons, we need an estimate uniform in H. First, $A_n \leq \exp(-c/n)$ for $n \geq 2$, since

$$\frac{1}{t+1} + \sum_{j=2}^{H} \frac{c}{t+j} \geq \frac{1 + (H-1)c}{Hn} = \frac{\lambda_H}{n} \geq \frac{c}{n}.$$

Therefore

$$\prod_{j=k+1}^{n} A_j \leq \exp\left(-c \sum_{j=k+1}^{n} \frac{1}{j} \right) \leq \left(\frac{k+1}{n+1} \right)^{c}, \qquad 1 \leq k \leq n.$$

Second, the ratio formula for the weights and $1 - \prod_i (1 - x_i) \leq \sum_i x_i$ for $x_i \in [0, 1]$ give

$$0 < w_{n,1} - w_{n,j} \leq \frac{\gamma(j-1)}{(t+1)^2}, \qquad j \geq 2.$$

Since $\overline{Q}_H \leq U$, for $n \geq 2$ this implies

$$0 < D_n \leq \frac{\gamma^2 \overline{Q}_H (H-1)}{2(t+1)^2} \leq \frac{\gamma^2 U}{2H(n-1)^2}.$$

Also $e_1 \leq \overline{Q}_H \leq U$, because all estimates are nonnegative. Applying the product bound in the error recurrence, and using $H \geq 2$, yields

$$e_n \leq \frac{U}{(n+1)^c} \left(2^c + \frac{\gamma^2}{4} \sum_{k=2}^{\infty} \frac{(k+1)^c}{(k-1)^2} \right) = \frac{C_\gamma}{(n+1)^c}.$$

The series converges because $c < 1$. Finally,

$$U - \overline{Q}_H = \frac{\gamma}{(1-\gamma)[H(1-\gamma) + \gamma]}.$$

Adding e_n proves eq. (130). If $H_n = \Theta(\log n)$, this deterministic target discrepancy is $\Theta(1/\log n)$, whereas $n^{-(1-\gamma)} = o(1/\log n)$. This proves the last assertion. □

For example, $H = 3$ and $\gamma = 0.9$ give $\lambda_H = 0.4$, $\overline{Q}_H = 2.5$, and the finite-horizon target $V_H(s, 0) = 1 + 0.9 + 0.9^2 = 2.71$. The estimation error about the effective limit is $\Theta(n^{-0.4})$, and the merging bias relative to the finite-horizon target is 0.21. This choice of H and γ does not contradict Proposition 7 and Corollary 10: it violates their concentration premise and their sequential-feedback convention. It shows why those premises cannot simply be deleted for ordinary cyclic reverse backups. In particular, the logarithmic horizon controls Bellman truncation but does not by itself control the bias produced by pooling cutoff and internal backup targets.

G.3 POWER-MEAN SEARCH ON ACYCLIC EPISODIC MODELS

There is a direct extension when the reachable physical nonterminal graph is finite and acyclic. Assume the root is nonterminal and every trajectory reaches an absorbing terminal state in at most $D \geq 1$ transitions. At every terminal state z, take $V_0(z) = 0$ and use exact zero evaluation; rewards after absorption are zero. A terminal root has zero estimation error and needs no search.

Choose $H \geq D$. Simulations continue through newly created nonterminal nodes and stop at an absorbing terminal; in particular, expansion does not trigger an early evaluation of V_0. Equivalently, one may pad the trajectory with zero-valued terminal steps without additional internal action selections or backups. Use historical backups and the preceding decision/terminal counting convention. Assume the fresh-simulator, bounded-reward, initialization, and positive-gap premises of Part I, with targets given by the episodic optimal values. Fix $C_1 > 0$, $C_2, C_3 \geq 0$, and a finite power-mean exponent $p \geq 1$, independently of H. At each nonterminal state, select each action once in a fixed order. Thereafter select, with fixed tie-breaking,

$$a \in \operatorname*{argmax}_{a \in \mathcal{A}_s} \left\{ \widehat{Q}_{s,a,T_{s,a}} + C_1 \frac{D_s^{1/4}}{\sqrt{T_{s,a}}} + B_{s,a}^{\mathrm{ER}} \right\}, \tag{133}$$

using the completed counts and ER bonus from section G.1. Full-state sharing can then merge paths of different lengths without mixing value targets.

Proposition 9 (Full-state ER search on an acyclic episodic model)**.** *In this special case, the full-state root estimate satisfies* $\mathsf{MC}(\beta_0/2, \beta_0; c_D)$ *about* $V_\infty^*(s_0)$ *for every* $\beta_0 > 1$*, and its expected absolute error is* $O(n^{-1/2})$*. The constant may depend on the finite model and* D*, but can be chosen uniformly over* $H \geq D$*. Consequently, the fixed-budget tail and expected-error bounds apply to a fresh F+ run of* n *simulations whenever* $H_n \geq D$*. The maximal bound concerns later visits within a single fixed-horizon run; it is not asserted simultaneously over separately restarted runs.*

Proof. Let $d(s)$ be the maximum number of transitions from a nonterminal state s to a terminal, and set terminal ranks to zero. Every nonterminal transition reduces rank strictly. A state reached at trajectory depth h has $h + d(s) \leq D \leq H$, so no nonterminal continuation is cut off. Its value target is the same at every occurrence. There are no repeated nonterminal states within a trajectory, so the local action-feedback convention holds and the root count equals the simulation budget. The common bound $L = \max\{1, R_{\max}/(1-\gamma)\}$ is independent of H. Exact zero terminal values and nonnegative rewards put all targets and historical estimates in $[0, L]$: if child values are in this interval, then $R + \gamma\widehat{V}(w) \leq R_{\max} + \gamma L \leq L$, and both arithmetic and power means preserve the interval.

Use the Part I parameter schedule for maximum rank D, and assign state s the proof index $j(s) = D - d(s)$. Each child has proof index at least $j(s) + 1$. A bound with exponents $(\beta/2, \beta)$ implies one with $(\widetilde{\beta}/2, \widetilde{\beta})$ for $0 < \widetilde{\beta} \leq \beta$: when $\sqrt{r}\epsilon \geq 1$ compare the powers directly, and otherwise use probability at most one. Explicitly, writing $x = \sqrt{r}\epsilon$, if $x \geq 1$ then $cx^{-\beta} \leq \max\{1, c\}x^{-\widetilde{\beta}}$; if $x < 1$, then $1 \leq \max\{1, c\}x^{-\widetilde{\beta}}$. Thus child bounds can be weakened to the common next-index exponents when a transition skips ranks.

Exact zero terminal values provide the base case. The historical graph-action bound (theorem 6) and the enveloped bandit result (theorem 7) now apply successively in topological rank, just as in Part I. The exploration term in eq. (133) remains $C_1 D_s^{1/4}/\sqrt{T_{s,a}} + B_{s,a}^{\mathrm{ER}}$. This proves the root maximal bound. Tail integration gives the mean bound. If D is a strict upper bound on $d(s_0)$, first weaken the root's stronger exponents to $(\beta_0/2, \beta_0)$ by the same argument. Neither the induction nor its constants use H once $H \geq D$: all nonterminal continuations finish before the cutoff, the parameter schedule uses D, and the algorithmic coefficients are fixed. For F+, apply the fixed-budget consequence separately to each run with horizon $H_n \geq D$; no union over restarted runs is taken. □

G.4 REGULARIZED FULL-STATE VARIANTS

GS-E3W-F-ER shares one action-value vector and regularized value per physical state, across all arrival depths. Its horizon H is fixed within a run. GS-E3W-F+-ER uses a fresh run at each budget

n with horizon H_n, fixed throughout that run. Each physical state has one fixed regularizer Ω_s and, for KL regularization, one fixed full-support reference ρ_s. The temperature and these operators are the same at every arrival depth and across horizons. The maps F_s, P_s are defined by eq. (10) with this physical-state index.

G.4.1 Counts, terminal evaluations, and bounded initialization

Use the completed-backup counts in section G.1. In addition, let D_s^{sel} and $T_{s,a}^{\mathrm{sel}}$ count initiated decisions and action selections, including those whose backups are pending. Reserve each action on its first initiation, so the first K_s initiated decisions select distinct actions in a fixed order. Subsequently use eq. (13), with v replaced by s and $k = D_s^{\mathrm{sel}} + 1$. Thus every action has conditional probability at least $\eta/\ell(k)$, where $\ell(k) = \log(2+k)$ and $0 < \eta \le 1/K$. The estimates and ER bonus use completed backups. If $T_{s,a} = 0$, retain its initial estimate and set its bonus coordinate to zero until its first backup completes. If $T_{s,a} \ge 1$, use lemma 15 with coefficients (c_2, c_3); write $c_+ = c_2 + c_3 M$. This makes selection well defined even when a state is revisited before earlier backups finish.

Complete each simulation before starting the next. At the horizon cutoff, use the terminal evaluator V_0, with separate pooled terminal statistics and the projection in eq. (71) when needed. During reverse path updates, an action appends the target using the value just computed for its realized successor occurrence: the terminal evaluation at the cutoff, and the regularized internal backup elsewhere. Terminal evaluation never overwrites an internal value. Index each historical action stream by its completed-backup order, which can differ from selection order. Read recommendations and root estimates only after all backups of a simulation have finished.

There is a common bounded initialization that does not use arrival depth. Assume $|\Omega_s(\pi)| \le \Omega_{\max} < \infty$ on every action simplex, and choose $U > 0$ with

$$U \ge \max\left\{V_{\max}, \frac{R_{\max} + \tau\Omega_{\max}}{1-\gamma}\right\}, \qquad J = [-\gamma U, R_{\max} + \gamma U].$$

Initialize all action coordinates in J and the internal value by F_s. If child values lie in $[-U, U]$, every historical target lies in J and

$$F_s(J^{K_s}) \subseteq [-\gamma U - \tau\Omega_{\max}, R_{\max} + \gamma U + \tau\Omega_{\max}] \subseteq [-U, U].$$

Induction over completed updates, starting from bounded terminal estimates, proves these bounds even with cycles. Boundedness alone does not prove concentration about a fixed target.

G.4.2 Conditional fixed-horizon transfer

Proposition 10 (Conditional regularized full-state concentration). *Fix $H \ge 1$ and use the preceding conventions. Suppose the actual root action streams satisfy, for deterministic $\overline{\mathbf{Q}}_{\Omega,H} \in \mathbb{R}^{K_{s_0}}$, $\mathsf{B}_H \ge 2$, $\mathsf{T}_H > 0$, and $q_H \ge 0$,*

$$\mathbb{P}\left(\sup_{m \ge r} |\widehat{Q}_{s_0,a,m} - \overline{\mathbf{Q}}_{\Omega,H}(a)| > \epsilon\right) \le \mathsf{B}_H \exp\left(-\frac{r\epsilon^2}{\mathsf{T}_H^2 \ell(r)^{q_H}}\right) \tag{134}$$

for every action, integer $r \ge 1$, and $\epsilon > 0$. Let $\widehat{\mathbf{Q}}_{n,H}^{\mathrm{F}}$ be the root vector after n completed simulations and assume its error about $\overline{\mathbf{Q}}_{\Omega,H}$ is bounded by W_H almost surely, including during initialization. Put

$$\overline{V}_{\Omega,H} = F_{s_0}(\overline{\mathbf{Q}}_{\Omega,H}), \qquad \widehat{V}_{\Omega,n,H}^{\mathrm{F}} = F_{s_0}(\widehat{\mathbf{Q}}_{n,H}^{\mathrm{F}}),$$

and define

$$\mathsf{A}_H = K(1+\mathsf{B}_H), \qquad \mathsf{S}_H^2 = \eta^{-1}\max\{16\mathsf{W}_H^2, 4\mathsf{T}_H^2\}. \tag{135}$$

Suppose every simulation initiates a root decision. For every integer $n \ge 1$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{k \ge n} \|\widehat{\mathbf{Q}}_{k,H}^{\mathrm{F}} - \overline{\mathbf{Q}}_{\Omega,H}\|_\infty > \epsilon\right) \le \mathsf{A}_H \exp\left(-\frac{n\epsilon^2}{\mathsf{S}_H^2 \ell(n)^{q_H+1}}\right). \tag{136}$$

The same bound holds for the value error about $\overline{V}_{\Omega,H}$. Write $V_{\Omega,H}^*(s_0)$ for the finite-horizon regularized target with terminal function V_0. If $|\overline{V}_{\Omega,H} - V_{\Omega,H}^*(s_0)| \le \delta_{\Omega,H}^{\mathrm{merge}}$, then

$$\mathbb{P}\left(\sup_{k \ge n} |\widehat{V}_{\Omega,k,H}^{\mathrm{F}} - V_{\Omega,H}^*(s_0)| > \delta_{\Omega,H}^{\mathrm{merge}} + \epsilon\right) \le \mathsf{A}_H e^{-n\epsilon^2/(\mathsf{S}_H^2 \ell(n)^{q_H+1})}, \tag{137}$$

$$\mathbb{E}|\widehat{V}_{\Omega,n,H}^{\mathrm{F}} - V_{\Omega,H}^*(s_0)| \le \delta_{\Omega,H}^{\mathrm{merge}} + e_H(n), \tag{138}$$

where

$$e_H(n) = \mathsf{S}_H \sqrt{\frac{(1+\log \mathsf{A}_H)\ell(n)^{q_H+1}}{n}}. \tag{139}$$

No action-gap assumption is needed for these value bounds.

Proof. Let $I_{a,j}$ indicate action a at the j-th initiated root decision, and put $C_a(n) = \sum_{j=1}^{n} I_{a,j}$. The proof of lemma 12 uses only conditional selection probabilities and measurability of earlier indicators, so it applies to this initiated clock even if some values are awaiting backup. For $n \geq 2K$ it gives

$$\mathbb{P}\left(\min_a C_a(n) < \frac{\eta n}{4\ell(n)}\right) \leq K e^{-\eta n/(16\ell(n))}.$$

By the end of simulation n, at least n root decisions have been initiated and all their backups have completed. Therefore the completed root action counts at every later simulation boundary $k \geq n$ satisfy $T_{s_0,a}(k) \geq C_a(n)$, pathwise. Here $T_{s_0,a}(k)$ denotes the count at that boundary, rather than the count after k local decisions. Set $m_n = \lceil \eta n/[4\ell(n)] \rceil$. Except on the preceding failure event, every later root action estimate is read at a completed count at least m_n. Since $0 < \eta \leq 1/K \leq 1$ and $\ell(n) > 1$, we have $1 \leq m_n \leq n$. Hence

$$\frac{m_n}{\ell(m_n)^{q_H}} \geq \frac{\eta n}{4\ell(n)^{q_H+1}},$$

using $m_n \geq \eta n/[4\ell(n)]$ and $\ell(m_n)^{q_H} \leq \ell(n)^{q_H}$ separately. No monotonicity of $t/\ell(t)^{q_H}$ is required. An event inclusion and a union bound using eq. (134) give

$$\mathbb{P}\left(\sup_{k\geq n} \|\widehat{\mathbf{Q}}^{\mathrm{F}}_{k,H} - \overline{\mathbf{Q}}_{\Omega,H}\|_\infty > \epsilon\right) \leq K e^{-\eta n/(16\ell(n))} + K\mathsf{B}_H \exp\left(-\frac{\eta n \epsilon^2}{4\mathsf{T}_H^2 \ell(n)^{q_H+1}}\right).$$

For $0 < \epsilon \leq \mathsf{W}_H$, the constants in eq. (135) dominate both exponentials by the one in eq. (136); for larger ϵ the event is empty. If $1 \leq n < 2K$, then $n\epsilon^2/[\mathsf{S}_H^2 \ell(n)^{q_H+1}] \leq n\eta/16 < 1/8$, so the proposed right-hand side exceeds one. This proves the vector bound for all n.

The 1-Lipschitz property of F_{s_0} gives the value bound about $\overline{V}_{\Omega,H}$, and the triangle inequality adds the merging bias. For the mean bound, set $X_n = \sup_{k\geq n} |\widehat{V}^{\mathrm{F}}_{\Omega,k,H} - \overline{V}_{\Omega,H}|$ and $a_n = \mathsf{S}_H^2 \ell(n)^{q_H+1}/n$. Integrating $\mathbb{P}(X_n^2 > u) \leq \min\{1, \mathsf{A}_H e^{-u/a_n}\}$ at the split $u = a_n \log \mathsf{A}_H$ gives $\mathbb{E}X_n^2 \leq a_n(1+\log \mathsf{A}_H)$. Cauchy–Schwarz and the deterministic bias bound prove eq. (138). No estimate was conditioned on its adaptive count, and no ordering of selections and completions was assumed beyond completion of each simulation before the next begins. The action-stream concentration is an explicit premise in completed-backup order; this proof does not apply a fresh-sample martingale argument after reordering responses by reverse backups. □

What remains conditional. The premise eq. (134) concerns the actual ER-generated historical streams, in completion order, about fixed deterministic action targets. Neither bounded estimates nor an error tube around depth-specific targets establishes that premise. The cyclic historical-stream analysis is therefore still required for a general full-state convergence theorem. In contrast to the power-mean bandit step, the conditional transfer just proved requires no sequential feedback between repeated selections at the same state.

Greedy recommendation. If $\overline{\mathbf{Q}}_{\Omega,H}$ has a unique maximizer $\bar{a}_H$ with gap $\overline{\Delta}_H > 0$, recommend an action maximizing the unperturbed current vector. If all action errors are at most $\overline{\Delta}_H/3$, every estimated gap remains positive. Hence eq. (136) gives

$$\mathbb{P}(\exists k \geq n : a^{\mathrm{rec}}_{k,H} \neq \bar{a}_H) \leq \mathsf{A}_H \exp\left(-\frac{n\overline{\Delta}_H^2}{9\mathsf{S}_H^2 \ell(n)^{q_H+1}}\right). \tag{140}$$

The scalar value-bias bound does not identify $\bar{a}_H$ with the finite-horizon best action. A sufficient additional condition is $\|\overline{\mathbf{Q}}_{\Omega,H} - Q^*_{\Omega,H}(s_0,\cdot)\|_\infty \leq \zeta_H$ and a unique finite-horizon action gap $\Delta_{\Omega,H} > 2\zeta_H$. Each effective gap is then at least $\Delta_{\Omega,H} - 2\zeta_H$, which can replace $\overline{\Delta}_H$ in eq. (140).

G.4.3 GROWING HORIZONS

Corollary 11 (Conditional regularized F+ guarantee)**.** *Suppose Proposition 10 holds at each fixed horizon used below. Let the physical-state model be finite, with fixed bounded regularizers and a common temperature across runs. Denote its infinite-horizon regularized value by $V^*_{\Omega,\infty}$ and put $E_{\Omega,0} = \|V_0 - V^*_{\Omega,\infty}\|_\infty$. For a fresh run of n simulations with fixed horizon H_n,*

$$\mathbb{E}|\widehat{V}^{\mathrm{F}+}_{\Omega,n,H_n} - V^*_{\Omega,\infty}(s_0)| \le e_{H_n}(n) + \delta^{\mathrm{merge}}_{\Omega,H_n} + \gamma^{H_n} E_{\Omega,0}. \tag{141}$$

Thus the bound proves consistency if $H_n \to \infty$, $e_{H_n}(n) \to 0$, and $\delta^{\mathrm{merge}}_{\Omega,H_n} \to 0$. This bound yields an $O(n^{-1/2})$ rate when each term on the right has that order.

Proof. For bounded value functions V, V', the regularized Bellman operator $\mathcal{T}_\Omega$ satisfies $\|\mathcal{T}_\Omega V - \mathcal{T}_\Omega V'\|_\infty \le \gamma\|V - V'\|_\infty$, by the 1-Lipschitz property of F_s. The finite model and bounded regularizers ensure that this operator maps bounded functions to bounded functions. Its unique fixed point therefore obeys $\|\mathcal{T}^{H_n}_\Omega V_0 - V^*_{\Omega,\infty}\|_\infty \le \gamma^{H_n} E_{\Omega,0}$. Add this deterministic truncation bound to eq. (138). The same argument applies to a stopped episodic operator with a fixed zero terminal boundary. □

For $n \ge 2$, taking $H_n = \lceil \log n / [2\log(1/\gamma)] \rceil$ makes the truncation term at most $E_{\Omega,0}/\sqrt{n}$. It does not control the merging bias or the horizon dependence of $\mathsf{A}_H, \mathsf{S}_H, q_H$. In particular, a logarithmic exponent proportional to H_n can make $e_{H_n}(n)$ fail to vanish. The maximal statements above apply within one fixed-horizon run; they are not simultaneous statements over different fresh runs with changing horizons.

G.4.4 A PROVED ACYCLIC EPISODIC CASE

Proposition 11 (Regularized full-state search on an acyclic model)**.** *Assume the reachable physical nonterminal graph is finite and acyclic, and every trajectory from the nonterminal root reaches a terminal in at most $D \ge 1$ transitions. Use fresh bounded reward–successor samples, at most K actions and M successors, and full trajectories through newly created nonterminal nodes. At termination, use exact value zero and perform no further action selection or regularization. Let $H \ge D$, and fix the regularizers, references, temperature, ER coefficients, and $\epsilon_\lambda > 0$ independently of H. Use the common initialization interval J above. The target V^*_Ω is the optimal regularized episodic value with this zero terminal boundary.*

Put

$$W = 2U, \qquad \nu^2 = (R_{\max} + 2\gamma U)^2/4, \qquad \mathfrak{A}_0 = 2, \quad \mathfrak{S}_0^2 = 0.$$

For $1 \le d \le D$, define

$$\begin{aligned}
\mathfrak{B}_d &= 2 + M\mathcal{C}(\mathfrak{A}_{d-1}),\\
\mathfrak{T}_d^2 &= \max\{8\nu^2, 4\gamma^2 M^2 \kappa_C \mathfrak{S}_{d-1}^2\},\\
\mathfrak{A}_d &= \max\{\mathfrak{A}_{d-1}, K(1+\mathfrak{B}_d)\},\\
\mathfrak{S}_d^2 &= \max\{\mathfrak{S}_{d-1}^2, \eta^{-1}\max\{16W^2, 4\mathfrak{T}_d^2\}\}.
\end{aligned} \tag{142}$$

Here $\mathcal{C}$ and κ_C are the constants from eq. (86)*. For every integer $n \ge 1$ and $\epsilon > 0$, the root vector and value satisfy*

$$\mathbb{P}\left(\sup_{k \ge n} \|\widehat{\mathbf{Q}}^{\mathrm{F}}_{k,H} - Q^*_\Omega(s_0,\cdot)\|_\infty > \epsilon\right) \le \mathfrak{A}_D \exp\left(-\frac{n\epsilon^2}{\mathfrak{S}_D^2 \ell(n)^D}\right), \tag{143}$$

$$\mathbb{E}|\widehat{V}^{\mathrm{F}}_{\Omega,n,H} - V^*_\Omega(s_0)| \le \mathfrak{S}_D \sqrt{\frac{(1+\log\mathfrak{A}_D)\ell(n)^D}{n}}. \tag{144}$$

The maximal value tail has the same right-hand side as the vector tail. All these constants are independent of $H \ge D$.

For MENTS or fixed-reference KL, the logarithmic factors can be removed. Specifically, put $D_{\mathrm{sc}} = R_{\max} + 2\gamma U + c_+$ and

$$\underline{p} = \begin{cases} K^{-1}e^{-D_{\mathrm{sc}}/\tau}, & \textit{MENTS},\\ \min\{K^{-1}, \rho_{\min}e^{-D_{\mathrm{sc}}/\tau}\}, & \textit{fixed-reference KL}, \end{cases} \tag{145}$$

where $\rho_{\min} = \min_{s,a} \rho_s(a) > 0$ *over the finite nonterminal graph. Define barred constants by eq.* (142) *with* η *replaced by* $\underline{p}$*. Then the preceding bounds hold with barred constants and* $\ell(n)^D$ *replaced by* 1*, giving expected error* $O(n^{-1/2})$*. For a fresh F+ run of* n *simulations with* $H_n \geq D$*, the corresponding fixed-budget concentration and expected-error bounds also hold. The maximal bounds concern the continuation of each fixed-horizon run, not a supremum over separately restarted runs. No action gaps are required for value concentration.*

Proof. Let $d(s)$ be the maximum number of transitions from s to a terminal, with terminal rank zero. Every nonterminal transition reduces rank. If s is reached at trajectory depth h, then $h + d(s) \leq D \leq H$: otherwise a longest continuation from s would give a root trajectory longer than D. Thus no nonterminal continuation is cut off, and all visits to s share one regularized target. There are no repeated nonterminal states within a trajectory, so selection and completion order agree within each local stream. Every simulation visits the root once.

The preceding invariant-interval argument bounds all stored values. It bounds the true episodic targets as well, by backward induction from zero terminal values. Both the node error and action-vector error are at most $W = 2U$: the width of J is $R_{\max} + 2\gamma U \leq 2U$. The fresh true target $R + \gamma V^*_\Omega(S')$ lies in J, so the bounded-variable moment-generating-function bound gives the stated variance proxy ν^2. All reachable local streams exist under perpetual continuation: the mixture floor has divergent sum, so every action at an infinitely visited state is selected infinitely often, and fresh transitions visit each positive-probability successor infinitely often. Propagate this argument from the root through the finite acyclic graph, as in Remark 5.

Induct on rank. Terminal errors are zero. At a rank-d query, every child has rank $j \leq d - 1$. The constants in eq. (142) are nondecreasing by construction, and $\ell(r)^j \leq \ell(r)^{d-1}$. Consequently all child bounds fit the common parameters $(\mathfrak{A}_{d-1}, \mathfrak{S}_{d-1}, d - 1)$. The historical graph-action transfer lemma 11 gives action-stream parameters $(\mathfrak{B}_d, \mathfrak{T}_d, d - 1)$. The counting inequality used there still holds: the l-th occurrence of a child from a query sees its pooled post-update count at least l. At rank one, the child-error contribution vanishes and the fresh-average lemma directly gives the same stated action bound. Apply lemma 13 with error cap W to obtain vector and value concentration at rank d with logarithmic exponent d. This closes the induction. If the root rank is smaller than D, weaken its bound using the nondecreasing constants and $\ell(n) \geq 1$. Integrating the squared maximal value tail as in the proof of Proposition 10 gives eq. (144).

For the sharper result, after initialization the completed count of each action is at least one. Thus $0 \leq B^{\mathrm{ER}}_{s,a} \leq c_+$, and the score span is at most the width of J plus c_+, namely D_{sc}. The softmax and fixed-reference KL calculations in Proposition 5 give eq. (145). The uniform component has every coordinate at least $1/K$, so the same floor holds for their mixture. Repeat the rank induction with logarithmic exponent zero: historical graph transfer preserves zero, and lemma 13 transfers to node values without introducing a logarithm. This gives the barred-constant claims and their mean bound.

Neither induction uses H once $H \geq D$; the target, initialization, and algorithmic parameters are fixed, and all trajectories terminate before the cutoff. This proves uniformity in H. Applying the fixed-budget bound separately to each run with $H_n \geq D$ gives the stated F+ consequence, without a union over restarted runs. □

The action-vector bounds also give greedy recommendation guarantees by the gap argument in eq. (140), now about the true regularized episodic action values. The terminal convention is essential: zero rewards alone do not imply zero regularized continuation value. For example, at an artificial absorbing state with K_z actions, Shannon regularization gives $F_z(\mathbf{0}) = \tau \log K_z$. Stopping regularization at the terminal boundary avoids introducing such additional continuation rewards.

H Experimental Details

H.1 Evaluation Protocol

To evaluate the performance and scalability of each algorithm, we define a set of simulation budgets (number of iterations) tailored to the complexity of each environment. For every environment, each algorithm is evaluated across these simulation checkpoints to observe its performance profile.

Our evaluation follows a replanning paradigm. Specifically, at each time step of an episode, the agent executes a complete search process using its allotted simulation budget to determine the optimal action. This process is repeated for every step until a terminal state is reached. To ensure statistical significance and account for the stochastic nature of Monte Carlo-based methods, we conduct 1,000 independent runs for each combination of algorithm, environment, and simulation budget. The final performance is reported as the mean cumulative reward (or other environment-specific metric) over the corresponding evaluation runs.

H.2 Hyperparameter Selection

To ensure a fair comparison, we perform an extensive hyperparameter sweep for both the proposed method and the baselines using a grid search strategy. The search space for each algorithm is defined based on literature-standard ranges and preliminary runs.

The tuning process is conducted as follows:

- For each environment and algorithm, we evaluate candidate hyperparameter sets using 300 independent runs.
- These runs are performed across all predefined simulation budget checkpoints to ensure the robustness of the selected parameters.
- **Selection Criterion:** The optimal hyperparameter configuration is chosen based on the highest aggregate mean reward, calculated by averaging the performance across all simulation budgets within a specific environment.

By selecting parameters that perform well across different simulation budgets, we aim to identify configurations that are not only high-performing but also stable under varying computational constraints.

H.3 Benchmark Environments

Shared Evaluation Settings: The hyperparameters below correspond to the nine algorithms plotted in Figure 1. Each configuration is evaluated over $1{,}000$ runs at each simulation budget $n \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$ per decision, with discount factor $\gamma = 1$. Here, c is the exploration coefficient of the power-UCT variants, p controls their value backup, τ is the regularization temperature, and ϵ controls uniform exploration in the regularized policies. The ER coefficients c_2 and c_3 weight the local and successor bonus channels, respectively; both are zero for the non-ER methods. A dash denotes an inapplicable parameter.

H.3.1 FrozenLake

Environment Description: The FrozenLake environment consists of an 8×8 grid where the agent starts at coordinate $(0, 0)$ and aims to reach the goal state at $(7, 7)$. The state space is defined by the agent's (x, y) coordinates. The grid cells are categorized into Frozen (F), Holes (H), Start (S), and Goal (G). The action space consists of four discrete movements: *left*, *down*, *right*, and *up*. To introduce stochasticity (slippery mode), transitions are non-deterministic: the executed action results in the intended direction with a probability of $1/3$, or deviates to a perpendicular direction (left turn or right turn) with a probability of $1/3$ each. Boundary collisions result in the agent remaining in its current cell. The episode terminates if the agent reaches the goal, falls into a hole, or exhausts the maximum horizon maximum of $H = 200$. The reward function is strictly sparse, granting a reward of 1 only upon reaching the goal, and 0 for all other transitions.

Selected Hyperparameters:

Algorithm	c	p	τ	ϵ	c_2	c_3
GS-POWER-UCT	0.5	∞	–	–	0.0	0.0
GS-POWER-UCT-F+	1.25	∞	–	–	0.0	0.0
GS-POWER-UCT-ER	0.5	∞	–	–	1.0	1.0
GS-MENTS-ER	–	–	0.01	0.5	0.5	0.5
GS-RENTS-ER	–	–	0.01	0.5	0.1	0.1
GS-TENTS-ER	–	–	0.01	0.5	0.1	0.1
GS-MENTS-F+-ER	–	–	0.01	0.5	0.1	0.1
GS-RENTS-F+-ER	–	–	0.01	0.5	0.1	0.1
GS-TENTS-F+-ER	–	–	0.01	0.5	0.1	0.1

H.3.2 PASSENGER GRID

Environment Description: The Passenger Grid is formulated on a 7×6 layout. The state is represented as a tuple (x, y, M), where M is a bitmask tracking the pickup status of passengers. The agent starts at $(0, 0)$ and must deliver passengers to the drop-off goal at $(6, 0)$. Three passengers are initially located at $(1, 2)$, $(0, 5)$, and $(6, 4)$. The i-th bit of M flips to 1 when the agent occupies passenger i's initial cell. Actions include *left*, *down*, *right*, and *up*. The environment operates in a slippery mode where the effective action is sampled from $\{\text{turn-left}, \text{intended}, \text{turn-right}\}$ with probabilities $\{0.25, 0.5, 0.25\}$, respectively. Boundary collisions keep the agent stationary. Each transition increments the time step, and episodes terminate either upon reaching the drop-off cell or at horizon $H = 70$. The reward is delayed until the goal is reached and depends on the total number of collected passengers: 0, 1, 3, or 7 for collecting 0, 1, 2, or 3 passengers, respectively. All other transitions yield a reward of 0.

Selected Hyperparameters:

Algorithm	c	p	τ	ϵ	c_2	c_3
GS-POWER-UCT	0.5	∞	–	–	0.0	0.0
GS-POWER-UCT-F+	0.5	4.0	–	–	0.0	0.0
GS-POWER-UCT-ER	0.5	∞	–	–	2.0	2.0
GS-MENTS-ER	–	–	0.01	0.5	1.0	1.0
GS-RENTS-ER	–	–	0.01	0.5	0.1	0.1
GS-TENTS-ER	–	–	0.25	0.5	0.1	0.1
GS-MENTS-F+-ER	–	–	0.01	0.5	2.0	2.0
GS-RENTS-F+-ER	–	–	0.25	0.5	2.0	2.0
GS-TENTS-F+-ER	–	–	0.25	0.5	0.1	0.1

H.3.3 FACTORED RIVER SWIM

Environment Description: The Factored River Swim environment comprises 3 independent river-swim chains. The joint state captures the position of an agent on each chain, $P \in \{0, \ldots, 4\}^3$. All rivers are initialized at position 0, with position 4 as the goal. A joint action is represented as a 3-bit vector (8 possible actions), where the i-th bit indicates the behavior for river i: 0 for resting/swimming downstream (left) and 1 for swimming upstream (right). Action 0 deterministically moves the agent one step left (bounded at 0). Action 1 introduces stochastic transitions: at position 0, the agent stays with probability 0.4 and moves right with 0.6; at the goal (4), it falls back to 3 with probability 0.4 and stays with 0.6; at any intermediate state, it moves left (0.05), stays (0.60), or moves right (0.35). Episodes terminate at $H = 35$. Rewards are evaluated prior to position updates. For each river, selecting action 0 at position 0 yields 0.1, while selecting action 1 at position 4 yields 1.0. A synergistic bonus of 3.0 is granted if all rivers are simultaneously at position 4 and all action bits are set to 1. The total aggregated reward is normalized by a factor of 6.

Selected Hyperparameters:

Algorithm	c	p	τ	ϵ	c_2	c_3
GS-POWER-UCT	0.5	∞	–	–	0.0	0.0
GS-POWER-UCT-F+	1.5	∞	–	–	0.0	0.0
GS-POWER-UCT-ER	0.5	∞	–	–	0.1	0.1
GS-MENTS-ER	–	–	0.01	0.5	2.0	2.0
GS-RENTS-ER	–	–	0.25	0.5	2.0	2.0
GS-TENTS-ER	–	–	0.25	0.5	1.0	1.0
GS-MENTS-F+-ER	–	–	0.01	0.5	2.0	2.0
GS-RENTS-F+-ER	–	–	0.01	0.5	2.0	2.0
GS-TENTS-F+-ER	–	–	0.25	0.5	2.0	2.0

H.3.4 FOURROOMS

Environment Description: FourRooms is an 11×11 grid partitioned by a vertical wall at $x = 5$ and a horizontal wall at $y = 5$. Communication between rooms is facilitated by four door cells, one randomly placed on each of the four wall segments during reset. The agent's start and goal positions are uniformly sampled from non-wall and non-door cells. The agent executes four directional actions, subject to slippery dynamics identical to the Passenger Grid: effective actions are sampled from $\{\text{turn-left}, \text{intended}, \text{turn-right}\}$ with probabilities $\{0.25, 0.5, 0.25\}$. Collisions with walls or boundaries leave the agent in place. Episodes terminate upon reaching the goal or hitting the time limit $H = 50$. The reward is 0 at all steps except when the goal is reached, where a time-discounted reward is issued: $R = 1 - 0.9 \times (t/50)$, incentivizing faster navigation.

Selected Hyperparameters:

Algorithm	c	p	τ	ϵ	c_2	c_3
GS-POWER-UCT	0.5	∞	–	–	0.0	0.0
GS-POWER-UCT-F+	0.5	4.0	–	–	0.0	0.0
GS-POWER-UCT-ER	0.5	2.5	–	–	2.0	2.0
GS-MENTS-ER	–	–	0.01	0.5	0.1	0.1
GS-RENTS-ER	–	–	0.25	0.5	0.1	0.1
GS-TENTS-ER	–	–	0.01	0.5	0.1	0.1
GS-MENTS-F+-ER	–	–	0.25	0.5	0.5	0.5
GS-RENTS-F+-ER	–	–	0.25	0.5	0.1	0.1
GS-TENTS-F+-ER	–	–	0.25	0.5	0.5	0.5

I EXTENSIONS, LIMITATIONS, AND OPEN PROBLEMS

Interpretation of the guarantees. Part I preserves the fixed-horizon $n^{-1/2}$ expected-error rate, with problem- and horizon-dependent constants. Part II establishes concentration for historical backups under fixed regularization; section E.3 removes its logarithmic factors for MENTS and fixed-reference KL. Neither result proves that ER improves estimation accuracy over the corresponding method without ER. Likewise, the cumulative-bonus comparison orders upper bounds, not the actual bonuses or estimation errors. Proving a planning advantage from shared-successor exploration remains open. Other open questions include evolving reference policies, visitation frequencies, cumulative regret, and recovery of the Tsallis policy support. Section 6 tests performance at small simulation budgets.

Cycles and merging across depths. The recursive resistance $\widetilde{\mathcal{R}}$ is defined here by an acyclic recursion. Merging by state alone can create cycles, where this recursion is undefined, and can mix values for different remaining horizons. The inverse-count bonus remains well defined, with the envelope in lemma 15. The acyclic episodic cases in Propositions 9 and 11 avoid both problems. For general cyclic full-state search, the conditional results still require concentration for the actual historical streams and control of merging bias and horizon-dependent constants. The power-mean bandit argument also requires feedback before another selection at that state; repeated selections with pending backups need an additional argument. The regularized transfer in Proposition 10 permits pending backups through its separate initiation and completion counts, but retains the action-stream concentration premise. Other possible extensions include spectral resistance for cycles and a resistance that accounts for depth. The latter could retain counts by depth or increase $1/N(e)$ using the observed variation in depths contributing to a merged node. These extensions require separate analysis.

Choosing the coefficients. Part I allows any finite $C_2, C_3 \geq 0$; its sufficient exploration threshold increases by $4(C_2 + C_3 M)/\Delta_a$ at the chosen tuning. The general Part II constants do not depend on c_2, c_3, because the proof uses the probability lower bound from uniform exploration. The sharper constants in section E.3 can depend on these coefficients. Neither result identifies their best finite-budget values; the sensitivity sweep in section 6 studies that choice.

Logarithmic factors. At remaining depth d, the general Part II bounds contain $\log^{d-1}(2+n)$ when indexed by action selections and $\log^{d}(2+n)$ when indexed by node visits. A root error threshold $\vartheta/\log(2+n)$ gives $\log^{H+2}(2+n)$ in the denominator. These exponents are sufficient; their optimality is open. MENTS and fixed-reference KL avoid these factors through their constant lower bounds on action probabilities.

Beyond the local bonus. The bonus combines outgoing and incoming one-layer resistances (Proposition 3). The incoming resistance is obtained from the recursive construction by setting all parent resistances to zero. For an initialized query (v, a), a predictable recursive alternative can be formed from completed history as

$$D^{\mathcal{R}}_{v,a} = \sum_{w : N_{v,a,w} > 0} \widehat{p}(w \mid v, a)\big(\widetilde{\mathcal{R}}(w) - \widetilde{\mathcal{R}}(v)\big),$$

$$B^{\mathrm{clip}}_{v,a} = \max\left\{0, \min\left\{c_{\mathrm{R}} D^{\mathcal{R}}_{v,a}, \frac{C_+}{T_{v,a}}\right\}\right\}.$$

where $c_{\mathrm{R}} \geq 0$ is fixed. These quantities use only observed successors and counts available before selection; initialization selections contribute zero. Thus $B^{\mathrm{clip}}_{v,a}$ satisfies eq. (45), and theorems 1 and 7 apply. Whether clipping substantially affects planning, or an unclipped recursive bonus admits bounds whose constants scale as $p_{\min}^{-H}$, where $p_{\min}$ is the smallest positive transition probability, remains open.