NEARLY QUADRATIC QUERY COMPLEXITY FOR SMOOTH BELLMAN FIXED POINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Smooth Bellman equations arise in entropy-regularized planning and require estimating nonlinear continuation values over an infinite horizon. We give an explicit algorithm whose pointwise accuracy exponent approaches the mean-estimation limit. For bounded rewards, finitely many actions, and smooth backups with possibly signed gradients of ℓ_1-norm bounded by G, under $\beta G < 1$, RMS error ε costs at most $\varepsilon^{-2}\exp(O([\log(1+\log(1/\varepsilon))]^2+1))$ transition queries. The budget is deterministic, all structural parameters are fixed, and high-probability accuracy adds a factor $1+\log(1/\delta)$. The construction combines signed derivative corrections, a common action proposal, and uniform control of recursively coupled differences. A quantitative transformation $p \mapsto 1 + p/2$ controls the constants sufficiently to choose a finite, accuracy-dependent number of stages. For fixed entropy temperature, the resulting local regularized decision loss η costs η^{-1} queries up to a subpolynomial overhead and confidence factor. Proposal controls and a finite-state end-to-end experiment examine the coupling and allocation mechanisms. Exact quadratic value-estimation complexity for the full nonlinear class remains open in this analysis.

1 INTRODUCTION

Planning with a simulator requires estimating the consequences of actions before taking them. In a large or continuous state space, even evaluating one action at one state can require recursively estimating values at previously unseen successors. A single expectation has Monte Carlo cost $O(\varepsilon^{-2})$, but a nonlinear Bellman backup nests expectations inside a nonlinear function. It is therefore unclear whether the same accuracy exponent remains attainable when the recursion has infinite horizon. This question is particularly relevant to entropy-regularized reinforcement learning. Regularized Bellman operators provide a common mathematical language for soft value iteration and related policy updates (Geist et al., 2019); maximum-entropy objectives also underlie soft actor-critic (Haarnoja et al., 2018). At a fixed positive temperature, the finite-action entropy backup is smooth even when actions are nearly tied. We study the oracle cost of solving this smooth planning problem at a prescribed root. The setting assumes a generative model, rather than learning its dynamics from a single trajectory, and does not impose a neural function approximation architecture.

SmoothCruiser (Grill et al., 2019) uses a derivative correction to replace first-order proxy error by a quadratic Taylor remainder. Recursive multilevel fixed-point methods establish polynomial complexity under broader Lipschitz assumptions (Beck et al., 2023). We combine the smooth correction with a stable recursive coupling to obtain an explicit accuracy exponent approaching two, including all recursion and confidence costs in a deterministic budget. We construct an accuracy-dependent multilevel algorithm that estimates a prescribed root action value to absolute error at most ε, with probability at least $1-\delta$. Its deterministic worst-case transition-query budget $N(\varepsilon,\delta)$ satisfies

$$N(\varepsilon,\delta) \le \varepsilon^{-2}\exp\big(O\big([\log(1+\log(1/\varepsilon))]^2+1\big)\big)\,[1+\log(1/\delta)]. \tag{1}$$

All structural parameters, including the action count and smoothness bound, are fixed. Smoothness permits an accuracy exponent arbitrarily close to the mean-estimation limit despite the infinite-horizon recursion. The displayed overhead is subpolynomial in $1/\varepsilon$, but is not a fixed polylogarithmic factor. Exact $O(\varepsilon^{-2})$ complexity for the general nonlinear class remains open.

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.

Method	Setting / extra structure	Accuracy cost	Budget convention
SmoothCruiser	Smooth regularized planning	Fourth-power type	Deterministic queries; high probability
READ	Smooth nested expectations, fixed depth D	$O_D(\varepsilon^{-2})$	Expected cost; unbounded branching
Multilevel fixed points	Lipschitz contraction; infinite horizon, including hard max	$O(\varepsilon^{-p_B})$ with parameter-dependent $p_B > 2$	Deterministic primitive cost; RMS
This work	Smooth signed contraction; fixed structural parameters	Nearly quadratic; explicit bound (1)	Deterministic queries; RMS and high probability

Table 1: Comparison of stated guarantees, not a claim of incompatible method classes. READ's depth dependence is not absorbed when D grows with accuracy. Our oracle convention treats backup evaluation as free; Appendix A identifies a common subclass with the multilevel fixed-point framework.

Contributions. We establish three main results. First, we give an explicit algorithm for smooth Bellman fixed points on arbitrary measurable state spaces, with deterministic query complexity given by (1). At fixed structural parameters, its accuracy exponent approaches the quadratic mean-estimation limit, improving on SmoothCruiser's fourth-power guarantee. A lower bound of $\Omega(\varepsilon^{-2}\log(1/\delta))$ matches the leading accuracy exponent and confidence dependence, leaving a subpolynomial gap. Second, for entropy-regularized planning at fixed temperature, we obtain a prescribed-root decision with regularized loss at most η using η^{-1} queries up to subpolynomial and confidence factors. A two-action lower bound establishes the optimality of the leading accuracy exponent and confidence dependence. Third, we construct a transformation that improves any certified query-cost exponent $2 < p \leq 4$ to $1 + p/2$. A shared-proposal coupling accommodates signed derivatives, while uniform control of the constants permits an accuracy-dependent number of stages. Finite-state experiments examine the coupling mechanism and demonstrate an estimation-cost crossover between practical single-level and multilevel implementations.

2 Related work

Table 1 compares the relevant guarantees. We inherit the Taylor correction from SmoothCruiser (Grill et al., 2019), rather than claim the first recursive or high-probability planning estimator. Regularized MDP theory (Geist et al., 2019) supplies the entropy-planning interpretation; the connection to soft actor-critic (Haarnoja et al., 2018) concerns its objective, not a neural-training sample-complexity theorem. READ (Syed & Wang, 2023), Theorem 2.2 and Corollary 2.3, controls a fixed nesting depth. Bounded terminal quantities satisfy every finite moment requirement; in our setting, the substantive issue is the quantitative growth of variance and cost with depth, followed by a deterministic-budget conversion. Beck et al. (2023), Theorem 4.1 and Corollary 4.2, already treat infinite-horizon Lipschitz fixed points and hard-max Bellman equations with deterministic recursive costs. Smoothness supplies the additional ingredients used here: a quadratic Taylor residual and stable coupling of derivative corrections. Appendix A details the theorem and cost-convention comparison. Allocation follows standard multilevel Monte Carlo (Giles, 2015); a finite-tree initializer follows the sparse-sampling principle (Kearns et al., 1999).

3 Problem, ML interpretation, and main result

3.1 Oracle model and smoothness

Let $\mathcal{S}$ be a measurable state space and $\mathcal{A} = \{1, \ldots, K\}$, $K < \infty$. A query at any (s, a) returns (r, S') with $|r| \leq R$. Observations are fresh conditionally on their requested state-action pairs; r and

S' within an observation may be dependent. Known differentiable backups $F_s : \mathbb{R}^K \to \mathbb{R}$ satisfy

$$F_s(0) = 0, \qquad \|\nabla F_s(q)\|_1 \le G, \tag{2}$$

$$|F_s(q') - F_s(q) - \nabla F_s(q) \cdot (q' - q)| \le \tfrac{L}{2}\|q' - q\|_\infty^2. \tag{3}$$

Assume joint measurability of the functions, gradients, and oracle kernel. Fix $R, G, L > 0$ and $0 < \beta < 1$ with $\alpha := \beta G < 1$. All of (β, G, L, K, R) are independent of the target accuracy. The operator $(Tv)(s) = F_s((\mathbb{E}[r + \beta v(S') \mid s, a])_{a=1}^K)$ is an α-contraction on bounded measurable functions. Its unique bounded fixed point satisfies

$$\begin{aligned} V^\star(s) = F_s(Q_s^\star), \qquad Q^\star(s,a) = \mathbb{E}[r + \beta V^\star(S') \mid s, a], \\ |Q^\star| \le B_Q := R/(1-\alpha), \qquad |V^\star| \le B_V := GB_Q. \end{aligned} \tag{4}$$

We count individual transition queries; evaluating F_s, its gradient, and arithmetic are free in this oracle model. Every budget bounds all auxiliary calls on every realization. The target is pointwise root mean squared (RMS) error, uniformly over admissible models and prescribed roots. It is not a simultaneous confidence event over the entire state space. Note that clipping action-value estimates to $[-B_Q, B_Q]$ can only decrease their error.

3.2 Regularized planning and local decisions

Entropy-regularized planning provides a concrete instance of our framework in which accurate action-value estimates yield quantitative guarantees on decision quality at a prescribed state. The normalized entropy backup

$$F_\tau(q) = \tau \log\left(K^{-1}\sum_a e^{q_a/\tau}\right) = \max_{\pi \in \Delta_K}\{\pi \cdot q + \tau\mathcal{H}(\pi) - \tau \log K\}, \qquad \tau > 0, \tag{5}$$

has $G = 1$ and $L = 1/\tau$: its Hessian quadratic form is a variance divided by τ. Its maximizing policy is $\mathrm{softmax}(q/\tau)$. The normalization corresponds to a known constant reward shift. More generally, a known bounded $F_s(0) = b(s)$ can be removed by replacing V by $V - b$, F_s by $F_s - b(s)$, and r by $r + \beta b(S')$; the original Q is preserved. The root guarantee has a stronger consequence than a linear decision-loss bound. Put $q = Q_s^\star$, $\widehat{\pi} = \nabla F_\tau(\widehat{q})$, and $J_q(\pi) = \pi \cdot q + \tau\mathcal{H}(\pi) - \tau \log K$. The variational identity and Taylor's remainder give

$$\begin{aligned} F_\tau(q) - J_q(\widehat{\pi}) &= F_\tau(q) - F_\tau(\widehat{q}) - \nabla F_\tau(\widehat{q}) \cdot (q - \widehat{q}) \\ &= \tau\,\mathrm{KL}(\widehat{\pi}\|\pi_s^\star) \le \min\left\{2\varepsilon, \frac{\varepsilon^2}{2\tau}\right\} \quad \text{if } \|q - \widehat{q}\|_\infty \le \varepsilon. \end{aligned} \tag{6}$$

Write $\mathcal{L}(t) = 1 + \log_+(1/t)$; logarithms are natural unless indicated. Let $\eta > 0$ be a prescribed tolerance for the local regularized decision loss $F_\tau(Q_s^\star) - J_{Q_s^\star}(\widehat{\pi})$. By (6), achieving loss at most η requires only action-value accuracy $\varepsilon_\eta = \sqrt{2\tau\eta}$. Applying Theorem 1 below to all K actions with failure probability δ/K yields, for fixed K, τ and sufficiently small η,

$$N_{\mathrm{decision}}(\eta, \delta) \le \eta^{-1} \exp\big(O((\log \mathcal{L}(\eta))^2 + 1)\big)\,[1 + \log(1/\delta)]. \tag{7}$$

Appendix H gives the proof and explicit budget. Proposition 15 shows that the powers of η and $\log(1/\delta)$ are optimal already for two actions; only the subpolynomial accuracy overhead remains unmatched. The theorem also covers nonmonotone smooth backups, such as $F_s(q) = g\sin(q_1)$, which define contractive fixed-point problems but need not correspond to standard reward-maximizing planning objectives.

3.3 Main guarantee

Theorem 1. *Under* (2)–(3) *and* $\beta G < 1$*, there is an explicit algorithm estimating a prescribed* $Q^\star(s,a)$ *with RMS error at most* t*, for every* $0 < t \le 1$*, using at most*

$$N(t) \le t^{-2} \exp\big(O([\log \mathcal{L}(t)]^2 + 1)\big) \tag{8}$$

queries. For $\varepsilon \in (0, 1]$ *and* $\delta \in (0, 1)$*, an estimator satisfies*

$$\Pr(|\widehat{Q} - Q^\star| > \varepsilon) \le \delta, \qquad N(\varepsilon, \delta) \le \varepsilon^{-2} \exp\big(O([\log \mathcal{L}(\varepsilon)]^2 + 1)\big)\,[1 + \log(1/\delta)]. \tag{9}$$

The budgets are deterministic. Constants depend only on the fixed (β, G, L, K, R)*, uniformly over models and prescribed roots. The same orders hold for a prescribed* $V^\star(s)$*.*

Section 5 specifies the finite-stage choice and the quantitative bootstrap proposition. Appendix C.3 gives a closed numerical budget. The construction uses a single global cutoff, including at coarse accuracies; no stage-dependent "sufficiently small" qualification is hidden in the theorem.

4 A COMMON PROPOSAL FOR RECURSIVE DERIVATIVE CORRECTIONS

We now construct the recursive estimator and show how a shared action proposal couples adjacent accuracies, providing the variance control needed for the multilevel complexity improvement.

4.1 FROM AN RMS PROXY TO A LOW-BIAS SAMPLE

Suppose an available algorithm $\mathcal{A}_p$ has RMS error t and known deterministic cost bound $C_p t^{-p}\mathcal{L}(t)^b$, $2 < p \le 4$. Set $c = (1+\alpha)/(2\alpha) > 1$ and $d = (1-\alpha c)/L > 0$. At state s, construct an independent clipped anchor $\widetilde{q}$ and fine proxy $\widehat{q}_u$ with

$$\mathbb{E}\|\widetilde{q} - Q_s^\star\|_\infty^2 \le a_0^2, \qquad \mathbb{E}\|\widehat{q}_u - Q_s^\star\|_\infty^2 \le du, \qquad \pi_a = \frac{|\partial_a F_s(\widetilde{q})| + \lambda/K}{\|\nabla F_s(\widetilde{q})\|_1 + \lambda}. \tag{10}$$

The anchor always uses the fixed stage-zero algorithm, at tolerance $a_0/\sqrt{K}$ per coordinate. The fine proxy uses $\mathcal{A}_p$ at $\sqrt{du/K}$. Fixed positive a_0, λ, u_0 are chosen once in Appendix G, with $u_0 \le \min(1/2, K/d)$. For $A \sim \pi$, let $w_u(A) = \partial_A F_s(\widehat{q}_u)/\pi_A$. The recipe ensures

$$\mathbb{E}w_u(A)^2 \le M, \qquad k = (1+\xi)\beta^2 M, \qquad kc < 1 \tag{11}$$

for fixed $M > G^2$ and $\xi > 0$, uniformly over all stages and $u < u_0$. The positive proposal floor prevents division by a small random derivative; the coarse anchor keeps M close enough to G^2 for contraction. Let W be a bounded fixed-cost estimator with $|W| \le B_V$ and $\sup_s \mathbb{E}|W(s) - V^\star(s)| \le u_0$. A finite sparse-sampling tree supplies it. Define $Z_u = W$ for $u \ge u_0$; otherwise sample the proxies, $A \sim \pi$, and (r, S') at (s, A), and set

$$Z_u(s) = F_s(\widehat{q}_u) + w_u(A)[r + \beta Z_{cu}(S') - \widehat{q}_u(A)]. \tag{12}$$

The child uses fresh randomness. Its correction depth is bounded by $\lceil \log(u_0/u)/\log c \rceil_+$. Integrating over the selected action recovers the Taylor correction exactly, including derivative signs. Induction gives

$$|\mathbb{E}Z_u(s) - V^\star(s)| \le (Ld/2 + \alpha c)u \le u, \qquad \sup_{s,u} \mathbb{E}|Z_u(s)|^2 \le M_Z < \infty. \tag{13}$$

The bias statement needs only a mean-square proxy guarantee, not an unbiased proxy. The second moment follows from (11); M_Z is uniform over the bootstrap stages (Lemma 4).

4.2 ADJACENT ACCURACIES AND THE KEY CONDITIONAL ESTIMATE

The two members of a pair use independent fine proxies but share the coarse proposal, chosen action, and oracle observation. This permits the derivative weights to change smoothly without changing the sampled action law.

Let $D(u) = \sup_{M,s} \mathbb{E}|Z_u(s) - Z_{2u}(s)|^2$ under this coupling. When both members recurse, subtracting (12) gives $Z_u - Z_{2u} = \beta w_f D' + R_u$, where D' is the fresh child difference. Smoothness implies Lipschitz gradients; the fixed proposal floor and (13) then give $\mathbb{E}R_u^2 \le C'u/(1+\xi^{-1})$. The important conditioning is on *all* parent information: both proxies, anchor, proposal, action, reward, and successor, denoted $\mathcal{H}$. Fresh child randomness implies

$$\mathbb{E}[(D')^2 \mid \mathcal{H}] \le D(cu), \qquad \mathbb{E}[w_f^2 (D')^2] \le \mathbb{E}[w_f^2] D(cu) \le M D(cu). \tag{14}$$

Note that no independence between the importance weight and successor is needed. The uniform child estimate holds at every realized successor. Consequently

$$D(u) \le kD(cu) + C'u, \qquad kc < 1 \quad \Longrightarrow \quad D(u) \le C_D u. \tag{15}$$

At the mixed boundary, $D(u) \le 4M_Z$ is absorbed by a fixed constant. Lemma 5 gives the full calculation and numerical C_D. This argument uses second moments only; no fourth-moment proxy bound is hidden in the estimate.

Coupled routine $\mathrm{Pair}(s,u)$**, marginals** (Z_u, Z_{2u})**.**
if $u \geq u_0$: draw one $z = W(s)$; **return** (z, z).
if $u < u_0 \leq 2u$: **return** independent $(Z_u(s), W(s))$.
Draw one anchor and form π as in (10).
Independently draw proxies $q_f = \widehat{q}_u$, $q_c = \widehat{q}_{2u}$.
Draw one $A \sim \pi$; query one (r, S') at (s, A).
Set $w_f = \partial_A F_s(q_f)/\pi_A$, $w_c = \partial_A F_s(q_c)/\pi_A$.
Draw $(z_f, z_c) = \mathrm{Pair}(S', cu)$ with fresh child randomness.
return $(F_s(q_f) + w_f[r + \beta z_f - q_f(A)]$,
$F_s(q_c) + w_c[r + \beta z_c - q_c(A)])$.

Figure 1: Both-base and mixed-base cases are explicit. All unshared random draws are independent. No cross-action coupling of the transition oracle is required. The full single routine appears in Appendix B.2.

5 Allocation, exponent bootstrap, and confidence

We use the coupled estimators from the previous section to reduce the query-complexity exponent through multilevel sample allocation, then iterate this improvement to obtain the nearly quadratic bound and its high-probability guarantee.

5.1 An explicit multilevel transformation

At a root (s, a) write $Y_u = r + \beta Z_u(S')$ and $u_j = u_0 2^{-j}$. The base increment is $\Delta_0 = r + \beta W(S')$. For $j \geq 1$, one root observation and a call to $\mathrm{Pair}(S', u_j)$ give $\Delta_j = Y_{u_j} - Y_{u_{j-1}}$; the root reward cancels. The expectation telescopes to $\mathbb{E}Y_{u_J}$. Lemma 7 supplies known numerical bounds

$$\mathrm{Var}(\Delta_j) \leq v_j := V_0 u_j, \qquad \mathrm{Cost}(\Delta_j) \leq c_j := B_C(C_p + 1)u_j^{-p/2}\mathcal{L}(u_j)^b. \tag{16}$$

The constants V_0, B_C are the computable formulas in (44)–(45), rather than unspecified asymptotic factors. The cost already includes every recursive correction: $\sum_{\ell \geq 0} c^{-\ell p/2} \leq (1 - c^{-1})^{-1}$. There is no additional factor for recursion depth. The mechanism can be summarized by three estimates:

$$|\mathbb{E}Y_u - Q^\star| \lesssim u, \qquad \mathbb{E}|\Delta_u|^2 \lesssim u, \qquad \mathrm{Cost}(\Delta_u) \lesssim u^{-p/2}\mathcal{L}(u)^b. \tag{17}$$

The first places the finest resolution at $u_J \asymp t$. The other two make $\sqrt{v_j c_j}$ proportional, up to logarithms and input constants, to $u_j^{(2-p)/4}$. For $p > 2$, the sum is geometrically dominated by the finest levels. Its square contributes $t^{1-p/2}$; multiplying by the sampling factor t^{-2} gives $t^{-(1+p/2)}$. This is why a square-root-accuracy proxy and a stable recursive difference improve the exponent, rather than merely reduce a prefactor. For requested RMS tolerance $t \in (0, 1]$, put $t_* := \beta u_0$ and

$$e_t = \min(t, t_*), \quad J = \left\lceil \log_2 \frac{2\beta u_0}{e_t} \right\rceil, \quad S_J = \sum_{j=0}^{J} \sqrt{v_j c_j}, \quad n_j = \left\lceil 4e_t^{-2} S_J \sqrt{v_j / c_j} \right\rceil. \tag{18}$$

Average n_j independent copies of each increment, sum independently across levels, and clip to $[-B_Q, B_Q]$. Bias is at most $e_t/2$, variance at most $e_t^2/4$, and the deterministic cost, including sample ceilings, is

$$N \leq 4e_t^{-2}\left(\sum_{j=0}^{J} \sqrt{v_j c_j}\right)^2 + \sum_{j=0}^{J} c_j. \tag{19}$$

With $\omega = p - 2$, a geometric sum bounds the leading term by a constant times

$$(C_p + 1)(1 - 2^{-\omega/4})^{-2} e_t^{-(1+p/2)}\mathcal{L}(e_t)^b.$$

Since $(1 - 2^{-\omega/4})^{-2} = O(\omega^{-2})$ uniformly for $0 < \omega \leq 2$, and one fixed cutoff handles all $t > t_*$, Theorem 8 proves the transformation

$$(p, C_p) \longmapsto (1 + p/2,\ A(p-2)^{-2}(C_p + 1)). \tag{20}$$

The explicit A in (52) is independent of p and the stage.

5.2 CHOOSING FINITELY MANY STAGES

Proposition 2. *There are computable structural constants $C_0 \geq 1$, $b \geq 0$, and $A \geq 4$, and algorithms $\mathcal{A}_m$ such that*

$$p_m = 2 + d_m, \qquad d_m = 2^{1-m}, \tag{21}$$

$$C_{m+1} = A d_m^{-2}(C_m + 1), \tag{22}$$

$$\mathrm{RMS}(\mathcal{A}_m(s,a;t)) \leq t, \qquad \mathrm{Cost}(\mathcal{A}_m(s,a;t)) \leq C_m t^{-p_m} \mathcal{L}(t)^b \quad (0 < t \leq 1). \tag{23}$$

The same constants and cutoffs apply at every stage.

The signed SmoothCruiser baseline (Appendix F) supplies $p_0 = 4$; (20) proves the proposition inductively. The recurrence gives

$$C_m + 1 \leq (C_0 + 1)(A + 4)^m 2^{m^2 - 3m}, \qquad \log C_m = O(m^2 + 1). \tag{24}$$

Choose the finite, accuracy-dependent stage

$$m(t) = \lceil \log_2(2\mathcal{L}(t)) \rceil . \tag{25}$$

Then $t^{-(p_{m(t)}-2)} \leq e$, proving (8). Compute this finite sequence of cost bounds first. Stage $m+1$ uses stage m for fine proxies, stage zero for every anchor, and the same W_{u_0} for every base call. Appendix C.3 gives the explicit numerical budget $\mathcal{B}(t)$ and all rescaling details. For confidence, run the complete RMS algorithm at tolerance $\varepsilon/2$. Chebyshev's inequality bounds each failure probability by $1/4$. The median of an odd $r \geq 8\log(1/\delta)$ independent runs has failure probability at most $e^{-r/8} \leq \delta$. Every run includes fresh anchors, proxies, and base samples and has a deterministic budget. Taking $r = 2\lceil 4\log(1/\delta) \rceil + 1$ proves (9); no exponential tails for the intermediate recursive samples are required.

6 PROPOSAL CONTROLS AND END-TO-END ESTIMATION

We consider four three-state, two-action models-entropy, nearly tied entropy, signed sine, and signed linear-whose transition matrices, rewards, and backup functions are specified in Appendix E.1. The experiments address two questions: whether coupling reduces the second moment of differences between adjacent resolutions, and whether multilevel allocation reduces the query cost of accurate root estimation. Reference values are computed from the known finite-state models to numerical precision and used only for evaluation; the estimators access transitions through sampled observations. Figure 2 measures coupled differences, whereas Figure 3 measures estimation error relative to the reference solution. Our examples have only three states and two actions, with known transition matrices and rewards. Their Bellman equations are therefore a small deterministic system. Using the known matrices, one can solve this system by fixed-point iteration, without sampling transitions.

Fresh empirical-model proxies use $\lceil 1/u \rceil$ observations per state-action pair; anchors use 128 and the base routine 512. These practical proxies and constants differ from the worst-case construction.

Separating sharing, adaptation, and common randomness. Figure 2 compares five proposal and coupling variants: the shared anchored proposal; independent anchors with coupled action draws; a shared uniform proposal; proposals from the two fine proxies with coupled action draws; and a shared anchor with independent actions, transitions, and descendant trees. All share the initial root observation. The latter three controls separate gradient adaptation, accuracy-dependent proposals, and common randomness. We use 3,000 independent root pairs per configuration, and root-block bootstrap intervals; the shaded mixed-base point is excluded from slope fits.

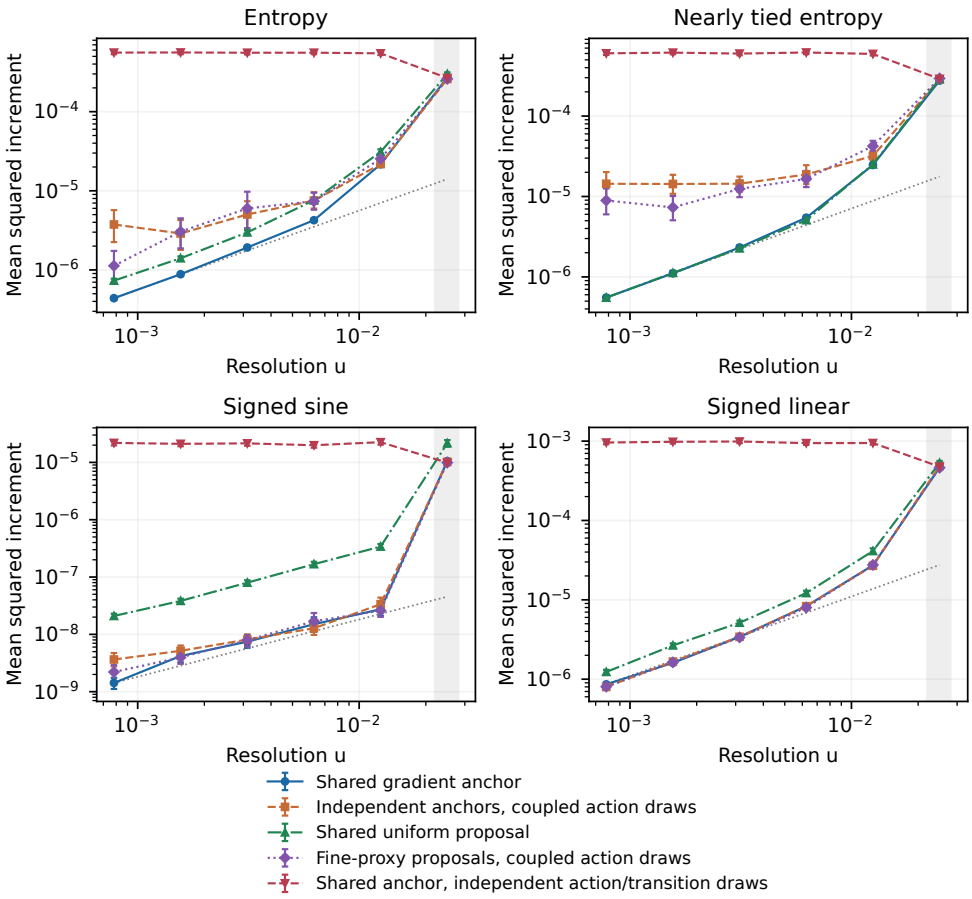

Figure 2: Proposal controls across six resolutions. Bars are 95% bootstrap intervals for increment second moments. The shaded coarsest point is the mixed-base case; gray dotted lines have slope one. Curves measure practical finite-state variants, not the conservative theorem constants.

Model	Shared-anchor slope	Moment ratio	Mismatch (%)
Entropy	1.35 [1.33, 1.37]	8.6 [5.1, 12.9]	0.37
Nearly tied entropy	1.33 [1.30, 1.35]	26.0 [17.1, 36.6]	1.30
Signed sine	1.04 [0.93, 1.15]	2.6 [1.8, 3.7]	0.70
Signed linear	1.23 [1.21, 1.26]	0.9 [0.9, 1.0]	0.00

Table 2: Descriptive slopes use levels 2–6; brackets give 95% bootstrap intervals. The finest-level moment ratio is independent anchors/shared anchor. Mismatch is the independent-anchor action disagreement at the first recursive correction, after the shared root observation. Appendix E.2 also reports pooled node frequencies and the fitting procedure.

The shared uniform control nearly matches the shared-anchor second moment for nearly tied entropy actions and costs fewer queries (70,295 versus 87,191 per finest-level pair). Sharing is beneficial here; gradient adaptation is not uniformly superior. For the standard entropy and signed sine examples, the anchored proposal reduces the measured second moment relative to uniform. Independent action/transition draws greatly increase the measured increment second moment. These are moment comparisons, not automatic efficiency gains.

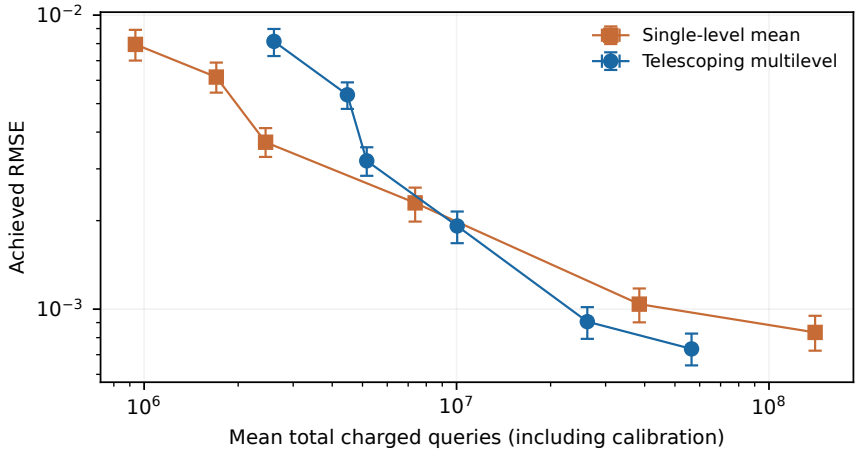

Figure 3: Achieved RMSE against all charged queries, including calibration. The six nominal target accuracies are listed in Appendix E.4; bars are 95% bootstrap intervals over 128 complete independent replications. Both estimators have the same finest marginal expectation. Their practical allocations use estimated variances.

Independent anchors have fixed accuracy, so persistent proposal disagreement is expected and is not evidence against every alternative shared proposal. Fine-proxy proposals instead refine with u; Appendix E.2 reports first-correction disagreement separately from pooled recursive rates. The linear proposal *distribution* is nonrandom, while its sampled action remains random; independent anchors should therefore leave its law unchanged.

Signed derivatives. In the signed sine model, 73% of evaluated active derivatives are negative and 33% have magnitude below .05, but no fine/coarse sign reversal was observed. A separate crossing stress model places a true derivative at zero; its paired derivatives reverse sign in approximately 16% of recursive comparisons. These frequencies distinguish negative weights from sensitivity at a zero derivative (Appendix E.3).

The complete estimation task. Figure 3 compares a single-level mean with a telescoping estimator at the same finest resolution on the entropy model. Both have the same finest-resolution expectation, $\mathbb{E}Y_{u_J}$, which may differ from $Q^\star$. Their overall accuracy is assessed against $Q^\star$, using empirical RMSE across 128 independent runs:

$$\widehat{\mathrm{RMSE}} = \sqrt{\frac{1}{128}\sum_{i=1}^{128}\left(\widehat{Q}_i - Q_{\mathrm{ref}}\right)^2}.$$

Each of 128 independent replications performs its own 48-sample variance calibration, then a fresh production run; *all calibration queries are charged*. Six nominal target accuracies control depth and allocation. Calibration makes the multilevel method more expensive at the coarser targets. At the finest target .001, its measured RMSE is .00073 at 56.5 million queries, compared with .00083 at 140.5 million for the single-level mean. This practical crossover supports the allocation mechanism without establishing a uniform speedup or the worst-case theorem's constants. The allocations here are data dependent and carry no certified high-probability guarantee.

7 Limits and Open Questions

A three-state linear subclass reduces root estimation to a Bernoulli mean. Testing two root values separated by 4ε and applying the adaptive transcript KL chain rule proves $N(\varepsilon,\delta) = \Omega(\varepsilon^{-2}\log(1/\delta))$ for small ε and $\delta \le 1/4$ (Propositions 9–10). This matches the confidence dependence, not the nonlinear accuracy overhead. For linear backups, a geometrically decaying path expansion admits exact $O(\varepsilon^{-2})$ cost (Appendix D.4).

At the nonlinear quadratic endpoint, the present bounds would give $v_j \asymp u_j$ and $c_j \asymp u_j^{-1}$ even with a quadratic-cost proxy. Every $\sqrt{v_j c_j}$ is then constant, producing $O(\varepsilon^{-2}\log^2(1/\varepsilon))$, rather than exact quadratic cost. Faster increment variance, such as $O(u_j^{1+\nu})$ at cost $O(u_j^{-1})$, would suffice; constructing it under the present assumptions remains open.

The result has fixed-parameter scope. Hard maximum is nonsmooth even for two actions. Approximating it to accuracy ε by entropy smoothing requires temperature decreasing with ε (for $K > 1$, the approximation error is at most $\tau \log K$), and hence an increasing smoothness bound. Likewise, growing K or $\beta G \uparrow 1$ changes the constants. The theorem does not give a uniform near-quadratic rate in these regimes. Its contribution is to separate how smoothness and recursive coupling approach the mean-estimation exponent, while leaving practical constant reduction and the exact quadratic nonlinear endpoint as concrete next questions.

AI USE STATEMENT

Generative AI (OpenAI ChatGPT/Codex) assisted with checking mathematical claims, identifying and comparing related literature, designing the diagnostic experiments, drafting, editing, and formatting this manuscript. The synthetic experiment results were generated by executing the supplied simulation code; AI assistance was used in designing and implementing that workflow. Verification performed during preparation includes primary-source reference checks, evaluation of the entropy decision identity and inequality on 10,000 numerical cases, deterministic-kernel recovery across the five proposal and coupling variants, query-accounting and embedded-table consistency checks, numerical checks of the two-action lower-bound cutoffs and layout checks. These computational checks were performed with AI assistance. The authors are responsible for the final submitted content, including AI-assisted material.

REPRODUCIBILITY STATEMENT

The model and oracle conventions are specified in Section 3. Appendices B–G provide full proofs, initialization, coupled pseudocode, and numerical bounds needed to specify all allocations. Appendix E and the accompanying code provide the finite models, random seeds, sample counts, raw measurements, and plotting scripts. The experiments test the coupling mechanism and do not implement the conservative full worst-case bootstrap algorithm.

REFERENCES

Christian Beck, Arnulf Jentzen, Konrad Kleinberg, and Thomas Kruse. Nonlinear Monte Carlo methods with polynomial runtime for Bellman equations of discrete time high-dimensional stochastic optimal control problems. *arXiv preprint arXiv:2303.03390*, 2023. URL https://arxiv.org/abs/2303.03390.

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, 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Michael B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015. doi: 10.1017/S096249291500001X. URL https://people.maths.ox.ac.uk/gilesm/files/acta15.pdf.

Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Ménard, Rémi Munos, and Michal Valko. Planning in entropy-regularized Markov decision processes and games. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/50982fb2f2cfa186d335310461dfa2be-Abstract.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine*

Learning Research, pp. 1861–1870, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 1324–1331, 1999. URL https://www.ijcai.org/Proceedings/99-2/Papers/093.pdf.

Yasa Syed and Guanyang Wang. Optimal randomized multilevel Monte Carlo for repeatedly nested expectations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33343–33364, 2023. URL https://proceedings.mlr.press/v202/syed23a.html.

SUPPLEMENTARY MATERIAL

This supplement gives complete proofs and algorithm specifications. The standing assumptions and notation are those of Section 3. Appendix A expands the theorem comparison; Appendices B–C prove the coupling and quantitative bootstrap; Appendix D treats confidence, lower bounds, and special cases. The remaining appendices document experiments, construct the signed SmoothCruiser baseline, fix all structural constants, and prove the regularized root-decision guarantee.

A THEOREM-LEVEL COMPARISON WITH EXISTING METHODS

SmoothCruiser. Grill et al. (2019), Theorems 1–2, establish a deterministic query count and high-probability pointwise accuracy for smooth regularized planning. The confidence parameter in Theorem 2 is chosen through a union bound involving the total query count. We inherit the local derivative correction. Our signed version is specified in Appendix F; the additions analyzed here are a common proposal, a coupled hierarchy, and the quantitative exponent bootstrap. Deterministic cost and high-probability estimation are not themselves new relative to that work.

READ. Syed & Wang (2023), Theorem 2.2 and Corollary 2.3, assume bounded second derivatives at the nonterminal levels and a finite 2^{D+1}-st terminal moment. Bounded terminal quantities satisfy this condition at every finite D; the issue here is the quantitative growth with D. For fixed nesting depth D, they obtain an unbiased finite-variance estimator and $O(\varepsilon^{-2})$ expected cost. One recursive sample has expected cost $\prod_{j=0}^{D-1} r_j/(2r_j-1)$, and its geometric branching is unbounded. Thus substituting $D \asymp \log(1/\varepsilon)$ requires controlling both depth-dependent moments and cost, followed by a separate deterministic-budget argument.

Multilevel fixed-point methods. In the bounded homogeneous specialization of Beck et al. (2023), equation (87), let $a < 1$ be the contraction, B the branching parameter, and K the vector size:

$$B > \frac{[1+a(2K-1)]^2}{(1-a)^2}, \qquad h_B = a + (1+aK)B^{-1/2},$$
$$\rho_B = \tfrac{1}{2}\left[h_B + \sqrt{h_B^2 + 4a(K-1)B^{-1/2}}\right], \qquad p_B = \frac{\log(3B)}{\log(1/\rho_B)} > 2. \tag{26}$$

Our comparison is an inference from these formulas: smoothness allows us a uniform exponent bootstrap beyond the parameter-dependent polynomial bound $O(\varepsilon^{-p_B})$ of Beck et al. (2023). Their primitive simulation/evaluation cost differs from our individual-query convention. With known deterministic rewards, our signed equation fits their framework via $v = Q - r$ and $f(s,z) = \beta F_s(r_s + z)$. Infinite horizon and recursive multilevel estimation are therefore shared features.

Multilevel allocation and initialization. The variance–cost allocation is the standard multilevel Monte Carlo principle reviewed by Giles (2015). We prove its finite-level bound with the ceiling terms retained. Sparse sampling (Kearns et al., 1999) provides the finite-tree precedent for our fixed-accuracy initializer. Our distinguishing ingredients are the quadratic Taylor residual and the stable coupling of signed derivative corrections, neither available under general Lipschitz assumptions alone.

B ESTIMATOR, SHARED PROPOSAL, AND CONDITIONAL COUPLING

Fix an input algorithm $\mathcal{A}_p$ with known numerical constants $C_p \geq 1, b \geq 0$ satisfying

$$\mathbb{E}|\mathcal{A}_p(s,a;t) - Q^\star(s,a)|^2 \leq t^2, \qquad \mathrm{Cost}(\mathcal{A}_p(s,a;t)) \leq C_p t^{-p}\mathcal{L}(t)^b,$$
$$0 < t \leq 1, \qquad 2 < p \leq 4. \tag{27}$$

For the iteration in Theorem 1, b is the fixed exponent of Appendix F.4. Set

$$c = \frac{1+\alpha}{2\alpha} > 1, \qquad d = \frac{1-\alpha c}{L} > 0. \tag{28}$$

Then $\alpha c < 1$ and $\alpha^2 c < 1$. Constants $\lambda, a_0, u_0, M, \xi, k$ are fixed once by the numerical recipe in Appendix G. In particular,

$$u_0 \le \min\{1/2, K/d\}, \qquad t_* := \beta u_0 > 0, \tag{29}$$

and neither cutoff changes with the bootstrap stage. First, (3) implies a uniform Lipschitz bound on the gradient:

$$\|\nabla F_s(x) - \nabla F_s(y)\|_2 \le H\|x - y\|_\infty, \qquad H = 3L \text{ is sufficient.} \tag{30}$$

Indeed, apply the Taylor-remainder bound at x to $y + tv$ and to y, and at y to $y + tv$. For $\|v\|_\infty = 1$ and $t = \|y - x\|_\infty > 0$, subtraction gives $|(\nabla F_s(y) - \nabla F_s(x)) \cdot v| \le 3L\|y - x\|_\infty$. Taking the supremum over v bounds the ℓ_1 norm of the gradient difference, and hence also its ℓ_2 norm.

At a requested state s, let $q = Q_s^\star$. Obtain a clipped coarse estimate $\widetilde{q} \in [-B_Q, B_Q]^K$ such that

$$\mathbb{E}\|\widetilde{q} - q\|_\infty^2 \le a_0^2. \tag{31}$$

Here $a_0 > 0$ is a fixed small constant, independent of the final accuracy. This is available at fixed finite deterministic cost: run the fixed algorithm $\mathcal{A}_0$ of Corollary 14 at accuracy $a_0/\sqrt{K}$ for each action and clip. This auxiliary routine is held fixed through all bootstrap iterations. Put $\widetilde{g} = \nabla F_s(\widetilde{q})$ and, for a fixed $\lambda > 0$, define

$$\pi_a = \frac{|\widetilde{g}_a| + \lambda/K}{\|\widetilde{g}\|_1 + \lambda}. \tag{32}$$

Thus every action has probability at least $\pi_{\min} = \lambda/[K(G + \lambda)] > 0$. Let $\widehat{q}_u$ be another clipped estimate satisfying

$$\mathbb{E}\|\widehat{q}_u - q\|_\infty^2 \le du. \tag{33}$$

It can be constructed by calling $\mathcal{A}_p$ at accuracy $\sqrt{du/K}$ for every action. Its numerical deterministic cost bound is calculated in Section C; it is part of the algorithm specification. If $A \sim \pi$, define the signed importance weight

$$w_u(A) = \frac{\partial_A F_s(\widehat{q}_u)}{\pi_A}. \tag{34}$$

The proposal and proxy are generated independently, followed by the action draw. The following second-moment property is decisive.

Lemma 3. *The fixed recipe in Appendix G gives $a_0, \lambda, u_0 > 0$, $M > G^2$, and $\xi > 0$ such that for all states, models, and $0 < u < u_0$,*

$$\mathbb{E}[w_u(A)^2] \le M, \qquad k := (1 + \xi)\beta^2 M, \qquad kc < 1. \tag{35}$$

All choices are uniform in the bootstrap iteration and in the particular RMS algorithm used for (33)*.*

Proof. The recipe places M strictly between G^2 and $1/(\beta^2 c)$ and gives $(1 + \xi)\beta^2 Mc < 1$. For its specified $\theta > 0$, write $g = \nabla F_s(\widehat{q}_u)$ and use

$$\begin{aligned}\sum_a \frac{g_a^2}{\pi_a} &\le (1 + \theta)\sum_a \frac{\widetilde{g}_a^2}{\pi_a} + (1 + \theta^{-1})\pi_{\min}^{-1}\|g - \widetilde{g}\|_2^2 \\ &\le (1 + \theta)G(G + \lambda) + (1 + \theta^{-1})\frac{K(G + \lambda)}{\lambda}\|g - \widetilde{g}\|_2^2.\end{aligned}$$

By (30), (31), and (33),

$$\mathbb{E}\|g - \widetilde{g}\|_2^2 \le 2H^2(du + a_0^2).$$

Appendix G bounds the first term by $G^2 + (M - G^2)/4$ and the expected remainder by $(M - G^2)/4$. This is less than M. The sum on the left is the second moment conditional on the anchor and proxy; averaging gives the asserted unconditional weight bound. □

B.1 A SAMPLE WITH SMALL BIAS AND BOUNDED SECOND MOMENT

Fix a bounded finite-cost base routine $W(s)$ with

$$|W(s)| \le B_V, \qquad \sup_{M,s} \mathbb{E}_M |W(s) - V_M^\star(s)| \le u_0. \tag{36}$$

Take $W = W_{u_0}$ from Lemma 12. The cutoff was already fixed in (29). Its known cost $C_W = W_{\mathrm{cost}}(u_0)$ is independent of u and of the stage. Define $Z_u(s)$ recursively. If $u \ge u_0$, return a fresh copy of $W(s)$. Otherwise, independently construct $\widetilde{q}$ and $\widehat{q}_u$, draw $A \sim \pi$, make one oracle query to obtain (r, S'), and return

$$Z_u(s) = F_s(\widehat{q}_u) + w_u(A)\big[r + \beta Z_{cu}(S') - \widehat{q}_u(A)\big]. \tag{37}$$

All randomness in the recursive child is fresh. Only the correction follows this recursion; construction of the proxy calls the already available algorithm $\mathcal{A}_p$. There is no random stopping issue: the correction tolerance increases by the fixed factor $c > 1$ at each step. Its depth is deterministically bounded by $\lceil \log(u_0/u)/\log c \rceil_+$. All calls to the proxy and coarse-proposal routines also have deterministic query budgets.

Lemma 4. *Uniformly over the model and state,*

$$|\mathbb{E} Z_u(s) - V^\star(s)| \le u, \tag{38}$$

$$\mathbb{E}|Z_u(s)|^2 \le M_Z < \infty. \tag{39}$$

The bound M_Z can be chosen independently of the bootstrap iteration.

Proof. The base cases follow from (36). Given the proxy and proposal, integration over A and its transition replaces the signed correction by $\nabla F_s(\widehat{q}_u) \cdot (q - \widehat{q}_u)$, up to a child-bias error at most $\beta G c u = \alpha c u$. Taylor's bound therefore gives

$$|\mathbb{E} Z_u - V^\star| \le \frac{L}{2} d u + \alpha c u = \frac{1 + \alpha c}{2} u \le u.$$

This is an induction over the finite correction depth and does not require an unbiased proxy.

For the second moment, set $b_u = F_s(\widehat{q}_u) + w_u(A)[r - \widehat{q}_u(A)]$. Since the proxy is clipped and $|w_u| \le G/\pi_{\min}$,

$$|b_u| \le D_0 := G B_Q + (G/\pi_{\min})(R + B_Q).$$

Let $\mathcal{H}$ contain the anchor, proxy, selected action, reward, and successor, but no child randomness. A fresh child at S' satisfies

$$\mathbb{E}[|Z_{cu}(S')|^2 \mid \mathcal{H}] \le \sup_x \mathbb{E}|Z_{cu}(x)|^2.$$

Since w_u is $\mathcal{H}$-measurable, the tower property gives

$$\mathbb{E}[w_u^2 |Z_{cu}(S')|^2] \le M \sup_x \mathbb{E}|Z_{cu}(x)|^2.$$

Using $|x + y|^2 \le (1 + \xi)|x|^2 + (1 + \xi^{-1})|y|^2$ therefore gives

$$\sup_{M,s} \mathbb{E}|Z_u(s)|^2 \le k \sup_{M,s} \mathbb{E}|Z_{cu}(s)|^2 + (1 + \xi^{-1}) D_0^2.$$

Because $k < 1$, this gives the uniform bound

$$M_Z = \max\{B_V^2, (1 + \xi^{-1}) D_0^2/(1 - k)\}.$$

The possible dependence of the successor state on A causes no problem: the child second-moment bound holds uniformly at every successor. The deterministic cost, with numerical constants, is proved in Lemma 7. □

B.2 EXPLICIT COUPLED ROUTINE

The following routines take $\mathcal{A}_p$ as a fixed input; their subscripts are suppressed. Every unspecified random draw is fresh. A proxy is built from K independent calls to $\mathcal{A}_p$ at $t(u) = \sqrt{du/K}$ and clipped coordinatewise.

Procedure $\mathrm{Single}(s, u)$**.**

if $u \geq u_0$: **return** a fresh $W(s)$.
Generate a fresh anchor $\widetilde{q}$ and, independently, a proxy q_u.
Form π by (32); draw $A \sim \pi$.
Query (r, S') at (s, A); set $w = \partial_A F_s(q_u)/\pi_A$.
Generate $z = \mathrm{Single}(S', cu)$ with fresh child randomness.
return $F_s(q_u) + w[r + \beta z - q_u(A)]$.

Procedure $\mathrm{Pair}(s, u)$**: marginals** (Z_u, Z_{2u})**.**

if $u \geq u_0$:
  Generate one fresh $w = W(s)$; **return** (w, w).
if $u < u_0 \leq 2u$:
  Generate $z_f = \mathrm{Single}(s, u)$ and $w = W(s)$ independently.
  return (z_f, w).
Generate one fresh anchor $\widetilde{q}$ and form the common proposal π.
Independently generate proxies $q_f = q_u$ and $q_c = q_{2u}$.
Draw one $A \sim \pi$ and query one common observation (r, S') at (s, A).
Set $w_f = \partial_A F_s(q_f)/\pi_A$ and $w_c = \partial_A F_s(q_c)/\pi_A$.
Generate $(z_f, z_c) = \mathrm{Pair}(S', cu)$ with fresh child randomness.
return $(F_s(q_f) + w_f[r + \beta z_f - q_f(A)]$,
  $F_s(q_c) + w_c[r + \beta z_c - q_c(A)])$.

The mixed base case specifies a concrete coupling rather than an unspecified boundary convention. In the recursive case the anchor, action, and oracle observation are shared; the two fine proxies are independent of each other and of the anchor. The child pair is fresh conditionally on the successor. These rules preserve both marginal laws, including at the boundary.

Lemma 5. *This coupling satisfies*

$$D(u) := \sup_{M,s} \mathbb{E}|Z_u(s) - Z_{2u}(s)|^2 \leq C_D u \qquad (0 < u \leq u_0), \tag{40}$$

with C_D independent of the bootstrap iteration. For $u \geq u_0$, define $D(u) = 0$, consistently with the shared base output. Its deterministic cost is at most the sum of the two marginal cost bounds.

Proof. For $u < u_0/2$, both members use (37). Write q_f, q_c for their proxies and w_f, w_c for their weights, and let D' be the coupled child difference. Subtracting their outputs gives

$$\begin{aligned} Z_u - Z_{2u} &= \beta w_f D' + R_u, \\ R_u &= F_s(q_f) - F_s(q_c) - w_f(q_f(A) - q_c(A)) \\ &\quad + (w_f - w_c)[r - q_c(A) + \beta Z_{2cu}(S')]. \end{aligned} \tag{41}$$

Because the proposal is common and has a fixed positive lower bound,

$$|w_f - w_c| \leq (H/\pi_{\min})\|q_f - q_c\|_\infty.$$

Put $W_{\max} = G/\pi_{\min}$ and define

$$C_R = 2(G + W_{\max})^2 + \frac{4H^2}{\pi_{\min}^2}\big[(R + B_Q)^2 + \beta^2 M_Z\big].$$

The first two terms of R_u have magnitude at most $(G + W_{\max})\|q_f - q_c\|_\infty$. Conditional on all parent randomness and S', the square of the bracket in its last term has expectation at most $2(R + B_Q)^2 + 2\beta^2 M_Z$. Thus

$$\mathbb{E}|R_u|^2 \leq C_R \mathbb{E}\|q_f - q_c\|_\infty^2 \leq 6C_R du.$$

For the leading term, let $\mathcal{H}$ contain both parent proxies, the anchor and proposal, action, reward, and successor. The child pair is generated with fresh randomness at this now fixed successor, so

$$\mathbb{E}[(D')^2 \mid \mathcal{H}] \leq D(cu), \qquad \mathbb{E}[w_f^2 (D')^2] = \mathbb{E}\big[w_f^2 \mathbb{E}[(D')^2 \mid \mathcal{H}]\big] \leq M D(cu). \tag{42}$$

Thus no independence between w_f and S' is asserted or needed. The uniform-in-state child bound is what makes their correlation harmless. Applying the squared-sum inequality to (41) yields

$$D(u) \leq kD(cu) + C'u, \qquad C' = 6d(1 + \xi^{-1})C_R. \tag{43}$$

For $u \geq u_0/2$, the uniform second moment gives $D(u) \leq 4M_Z$. Since $kc < 1$, induction in (43) proves (40), taking $C_D = \max\{8M_Z/u_0, C'/(1-kc)\}$.

Sharing the proposal, action, and transition can only save queries. Every recursive child pair has the correct marginal laws. Thus its cost is bounded by the sum of marginal worst-case budgets, including at the boundary between the recursive and base regimes. □

Remark 6. There is no assumption that the transition oracle can couple two different actions. Both members use exactly the same action, sampled from the common proposal, so reusing one transition sample is valid. The importance weights retain the signs of the derivatives. Positivity of F_s or its gradient is never used.

C Quantitative complexity and bootstrap constants

C.1 Numerical cost and variance bounds

Let C_0, b_{init} be the certified baseline constants in Appendix F.4. The incoming proxy has logarithmic exponent b from (27); the bootstrap iteration takes $b = b_{\mathrm{init}}$. Keep the anchor tied to its own certified exponent, even when considering another input algorithm. Define

$$\begin{aligned} t_a &= a_0/\sqrt{K}, \qquad C_W = W_{\mathrm{cost}}(u_0), \\ C_{\mathrm{an}} &= KC_0 t_a^{-4}\mathcal{L}(t_a)^{b_{\mathrm{init}}}, \qquad s_0 = \sqrt{d/K}, \\ B_P &= K\max\{s_0^{-2}, s_0^{-4}\}[1+\log_+(1/s_0)]^b, \qquad (44) \\ B_S &= \frac{B_P + C_{\mathrm{an}} + 1}{1-c^{-1}} + C_W, \qquad B_C = 1 + 2B_S, \\ V_0 &= \max\{B_Q^2/u_0, \beta^2 C_D\}. \qquad (45) \end{aligned}$$

Here C_{an} bounds all K anchor calls. These numerical constants may be very large. For a general input algorithm, B_P and the downstream constants B_S, B_C, A also depend on its logarithmic exponent b, held fixed during the exponent transformation. During the main bootstrap, $b = b_{\mathrm{init}}$ is itself a structural constant. The bounds do not depend on the input accuracy exponent $p \in (2, 4]$ or on the stage.

Lemma 7. *For $0 < u \leq u_0$, the routines above have the bounds*

$$\mathrm{Cost}(Z_u) \leq B_S(C_p+1)u^{-p/2}\mathcal{L}(u)^b, \qquad (46)$$

$$v_j := V_0 u_j, \qquad c_j := B_C(C_p+1)u_j^{-p/2}\mathcal{L}(u_j)^b, \qquad u_j = u_0 2^{-j}. \qquad (47)$$

The second line gives numerical upper bounds for the variance and cost of a root-level increment, including $j = 0$.

Proof. Put $f_p(u) = u^{-p/2}\mathcal{L}(u)^b$. Since $u \leq u_0 \leq 1$ and $t(u) = s_0\sqrt{u} \leq 1$, (27) gives a total proxy cost at most $B_P C_p f_p(u)$. Indeed $s_0^{-p} \leq \max\{s_0^{-2}, s_0^{-4}\}$ and $\mathcal{L}(s_0\sqrt{u}) \leq [1+\log_+(1/s_0)]\mathcal{L}(u)$. The remaining work at a recursive node is at most $C_{\mathrm{an}} + 1$ queries. Along the correction path,

$$f_p(c^\ell u) \leq c^{-\ell p/2} f_p(u), \qquad \sum_{\ell \geq 0} c^{-\ell p/2} \leq (1-c^{-1})^{-1}.$$

Also $f_p(u) \geq 1$. Summing node costs and adding the base cost C_W proves (46). The logarithmic number of correction steps is already included in this geometric sum; it does not multiply the cost again. Let $Y_u(s,a) = r + \beta Z_u(S')$. For $j \geq 1$, query one root observation and call $\mathrm{Pair}(S', u_j)$ to generate $\Delta_j = Y_{u_j} - Y_{u_{j-1}}$. Its second moment is at most $\beta^2 C_D u_j$. Its cost is bounded by one plus the two marginal costs, hence by c_j. At $j = 0$, use $\Delta_0 = r + \beta W(S')$, bounded in magnitude by B_Q and costing at most $1 + C_W$. These are also bounded by v_0, c_0 in (47). □

C.2 Fully specified allocation

For input tolerance $t \in (0, 1]$, set

$$e_t = \min\{t, t_*\}, \qquad J = \left\lceil \log_2 \frac{2\beta u_0}{e_t} \right\rceil \geq 1. \qquad (48)$$

Use exactly v_j, c_j from (47), and compute

$$S_J = \sum_{j=0}^{J} \sqrt{v_j c_j}, \qquad n_j = \left\lceil 4e_t^{-2} S_J \sqrt{v_j/c_j} \right\rceil. \tag{49}$$

Generate independent increments across levels and repetitions and return

$$\widehat{Q} = \mathrm{clip}_{[-B_Q, B_Q]} \left(\sum_{j=0}^{J} \frac{1}{n_j} \sum_{i=1}^{n_j} \Delta_j^{(i)} \right). \tag{50}$$

All allocation quantities are known before sampling. The cutoff e_t is necessary to make the same proof and constants cover coarse input tolerances at every stage.

Theorem 8. *Let* (27) *hold for known C_p, b, with $2 < p \le 4$. The estimator* (50) *has RMS error at most t for all $t \in (0, 1]$ and deterministic cost*

$$A(p-2)^{-2}(C_p+1)t^{-(1+p/2)}\mathcal{L}(t)^b, \tag{51}$$

where the same numerical A, depending on the structural parameters and the fixed input exponent b, works throughout $(2, 4]$. One valid choice is

$$\begin{aligned} H_\beta &= 1 + \log_+(4\beta), & D_\beta &= \max\{1, 4\beta\}, \quad E_\beta = \max\{1, (4\beta)^2\}, \\ B_g &= 32/(\log 2)^2, & B_* &= t_*^{-3}\mathcal{L}(t_*)^b, \\ A &= \max\{4,\ B_* B_C H_\beta^b [4V_0 D_\beta B_g + 8E_\beta]\}. \end{aligned} \tag{52}$$

Proof. The expectations of the increments telescope to $\mathbb{E}Y_{u_J}$. By construction, $\beta u_J \le e_t/2$, so the absolute bias is at most $e_t/2$. Independence and (49) give variance at most $e_t^2/4$. Clipping cannot increase MSE, so RMS error is at most $e_t \le t$. The sample ceilings give the exact upper bound

$$N \le 4e_t^{-2} S_J^2 + \sum_{j=0}^{J} c_j. \tag{53}$$

Write $\omega = p - 2 \in (0, 2]$. Monotonicity of $\mathcal{L}$ and a geometric sum give

$$S_J^2 \le V_0 B_C (C_p + 1) \frac{u_J^{-\omega/2}\mathcal{L}(u_J)^b}{(1 - 2^{-\omega/4})^2}, \tag{54}$$

$$\sum_{j=0}^{J} c_j \le 2B_C(C_p+1)u_J^{-p/2}\mathcal{L}(u_J)^b. \tag{55}$$

For $x = \omega \log 2/4 \le \log 2/2$, $1 - e^{-x} \ge xe^{-x} \ge x/\sqrt{2}$; consequently

$$(1 - 2^{-\omega/4})^{-2} \le B_g \omega^{-2}. \tag{56}$$

Minimality of J gives $e_t/(4\beta) < u_J \le e_t/(2\beta)$. Hence

$$u_J^{-\omega/2} \le D_\beta e_t^{-\omega/2}, \quad u_J^{-p/2} \le E_\beta e_t^{-p/2}, \quad \mathcal{L}(u_J)^b \le H_\beta^b \mathcal{L}(e_t)^b.$$

Since $e_t \le 1$ and $\omega \le 2$, substitution in (53) yields

$$N \le B_C H_\beta^b [4V_0 D_\beta B_g + 8E_\beta](C_p+1)\omega^{-2} e_t^{-(1+p/2)}\mathcal{L}(e_t)^b.$$

Finally $1 + p/2 \le 3$ and, for every $0 < t \le 1$,

$$e_t^{-(1+p/2)}\mathcal{L}(e_t)^b \le B_* t^{-(1+p/2)}\mathcal{L}(t)^b. \tag{57}$$

For $t \le t_*$ this is immediate; for $t > t_*$ use $(t/t_*)^{1+p/2} \le t_*^{-3}$ and $\mathcal{L}(t_*)/\mathcal{L}(t) \le \mathcal{L}(t_*)$. This proves (51). All constants and cutoffs are uniform in $p \in (2, 4]$. □

C.3 Accuracy-dependent number of stages

Define the computable numerical budget

$$\mathcal{B}(t) = e(C_0 + 1)(A + 4)^{m(t)} 2^{m(t)^2} t^{-2} \mathcal{L}(t)^b. \tag{58}$$

We prove $\mathrm{Cost}(\mathcal{A}_{m(t)}(s, a; t)) \le \mathcal{B}(t)$ for every $0 < t \le 1$. The baseline in Appendix F satisfies (27) for $p_0 = 4$ at every $t \in (0, 1]$. Theorem 8 therefore proves (21)–(23) by induction. To control the constants, put $D_m = C_m + 1$. Since $d_m \le 2$ and $D_m \ge 1$,

$$D_{m+1} = A d_m^{-2} D_m + 1 \le (A + 4) d_m^{-2} D_m.$$

Consequently

$$C_m \le D_m \le (C_0 + 1)(A + 4)^m 2^{m^2 - 3m} \le (C_0 + 1)(A + 4)^m 2^{m^2}. \tag{59}$$

Thus $\log C_m = O(m^2 + 1)$ with all constants fixed. The choice (25) ensures $d_{m(t)} \le 1/\mathcal{L}(t)$, so $t^{-d_{m(t)}} \le e$. Combining this with (59) proves (58) and (8).

Complete algorithm. Given t, compute $m(t)$, the sequence $p_0, \ldots, p_{m(t)}$, and the numerical bounds $C_0, \ldots, C_{m(t)}$. Stage zero calls the bounded baseline. Stage $m + 1$ invokes (48)–(50) with the preceding stage as its proxy algorithm. Every anchor always calls stage zero. Every correction base always calls W_{u_0}. The same t_*, u_0, λ, a_0 are used at every stage. The algorithm returns the stage-$m(t)$ result; it never takes an infinite limit of algorithms. To estimate $V^\star(s)$, call the root-action estimator for all K actions at $t = \min\{1, \varepsilon/(G\sqrt{K})\}$ and apply F_s. If the minimum selects 1, then $G\sqrt{K} \le \varepsilon$, so the same bound applies. Indeed

$$\mathbb{E}|F_s(\widehat{Q}_s) - V^\star(s)|^2 \le G^2 \sum_{a=1}^{K} \mathbb{E}|\widehat{Q}(s, a) - Q^\star(s, a)|^2 \le \varepsilon^2.$$

A fixed rescaling of t changes only structural constants in (8).

D Confidence, lower bounds, and the quadratic endpoint

D.1 High probability with a deterministic query budget

For a requested root, run the complete algorithm independently at RMS tolerance $\varepsilon/2$. Each result has probability at most $1/4$ of an error exceeding ε, by Chebyshev's inequality. Use an odd number r of independent complete runs and return their median. If the median fails, at least $r/2$ individual runs fail. Hoeffding's inequality therefore gives

$$\mathbb{P}(|\widehat{Q}_{\mathrm{med}} - Q^\star| > \varepsilon) \le e^{-r/8}.$$

The explicit choice $r = 2\lceil 4 \log(1/\delta) \rceil + 1 \le 8 \log(1/\delta) + 3$ gives the explicit budget

$$N(\varepsilon, \delta) \le [8 \log(1/\delta) + 3] \mathcal{B}(\varepsilon/2). \tag{60}$$

Fresh randomness includes every anchor, proxy, and base routine. Because every complete run has a deterministic budget, so does their median. No exponential-tail assumption on the intermediate Z_u samples is used. This completes Theorem 1.

The guarantee is pointwise at each prescribed root with uniform constants; it is not a simultaneous event over all states. Expected cost and deterministic cost have not been interchanged.

D.2 RMS and confidence-dependent lower bounds

Proposition 9. *For fixed nonzero R, G and $0 < \beta < 1$ with $\beta G < 1$, the full class in Section 3 has worst-case query complexity $\Omega(\varepsilon^{-2})$ for pointwise RMS error at most ε. The statement holds for either $Q^\star$ or $V^\star$.*

Proof. Take three states s_0, z_0, z_1, and use $F_s(q) = gq_1$ at every state, with a fixed $0 < g \le G$. All actions at z_i return reward iR and remain at z_i. All actions at s_0 return reward zero and move to z_1 with unknown probability θ, otherwise to z_0. This is an admissible linear, hence smooth, model. Its root action value is

$$Q^\star(s_0, a) = C_*\theta, \qquad C_* := \frac{\beta g R}{1 - \beta g} > 0.$$

All information about θ comes from Bernoulli observations at s_0, even if the algorithm chooses queries adaptively elsewhere. Compare $\theta_\pm = 1/2 \pm h$ with $h = 2\varepsilon/C_*$. For small ε, the root action values differ by 4ε. The Kullback–Leibler divergence of the two transcript laws for any algorithm with at most N queries is at most $C_1 N h^2$, by the chain rule and the Bernoulli divergence bound near $1/2$. An estimator with MSE at most ε^2 under both models gives a test, by thresholding at the midpoint, with error at most $1/4$ under each model. Therefore the total variation distance between transcript laws is at least $1/2$. Pinsker's inequality implies that their divergence is at least $1/2$, and thus $N \ge c\varepsilon^{-2}$. For the root value, replace C_* by $gC_* > 0$ in the same argument. □

The upper and lower bounds show that two is the infimum polynomial accuracy exponent for the full fixed-parameter class. They do not identify a matching logarithmic overhead or an exact quadratic upper bound.

Proposition 10. *For the same full class and fixed nondegenerate parameters, any algorithm with a deterministic query budget N and an ε-accuracy guarantee of probability at least $1 - \delta$ at a prescribed root under every model satisfies*

$$N \ge c\varepsilon^{-2}\log(1/\delta) \qquad (0 < \varepsilon \le \varepsilon_0, \quad 0 < \delta \le 1/4),$$

for positive structural constants c, ε_0.

Proof. Use the two models in Proposition 9 with $h = 2\varepsilon/C_*$ and $\varepsilon \le C_*/8$. Their root values are 4ε apart. Thresholding an ε-accurate estimate at the midpoint produces a test whose two error probabilities are at most δ. Data processing for relative entropy gives

$$\mathrm{KL}(P_+ \| P_-) \ge \mathrm{kl}(1 - \delta, \delta) = (1 - 2\delta)\log\frac{1 - \delta}{\delta} \ge \tfrac{1}{4}\log(1/\delta).$$

For $h \le 1/4$, the per-query Bernoulli divergence is at most $16h^2$. The adaptive transcript chain rule therefore gives $\mathrm{KL}(P_+ \| P_-) \le 64N\varepsilon^2/C_*^2$. Thus $N \ge C_*^2\varepsilon^{-2}\log(1/\delta)/256$. For a root-value target, replace C_* by gC_*. □

This matches the order of the confidence dependence. It does not match the nonconstant accuracy overhead in the nonlinear upper bound.

D.3 Why the present allocation has a logarithmic endpoint

Suppose an exact quadratic proxy algorithm were available, with $b = 0$ and cost Ct^{-2}. Our present bounds would give $v_j = V_0 u_j$ and $c_j = \overline{C} u_j^{-1}$. Each $\sqrt{v_j c_j}$ would then be constant, and (53) would give

$$O(\varepsilon^{-2}(J + 1)^2) + O(\varepsilon^{-1}) = O(\varepsilon^{-2}\mathcal{L}(\varepsilon)^2).$$

More generally an input logarithmic exponent b produces the bound $O(\varepsilon^{-2}\mathcal{L}(\varepsilon)^{b+2})$. This is a limitation of the available hierarchy and its certified allocation, not a lower bound for every possible estimator.

A sufficient alternative would be increments with $v_j = O(u_j^{1+\nu})$, $c_j = O(u_j^{-1})$, and $\nu > 0$. Then the variance–cost sum is bounded independently of J and (53) gives exact quadratic cost. The difficulty is to improve the increments without increasing proxy cost enough to remove that gain. Antithetic or higher-order corrections are possible research directions, not established solutions under the present assumptions.

D.4 AN EXACT QUADRATIC SPECIAL CASE: LINEAR BACKUPS

The lower-bound exponent is attained exactly when the backup is linear, including with signed coefficients. This provides a concrete sufficient condition stronger than general smoothness.

Proposition 11. *Suppose $F_s(q) = \sum_a f_s(a) q_a$, with known coefficients satisfying $\sum_a |f_s(a)| \le G$ and $\beta G < 1$. Then pointwise RMS error ε is achievable with deterministic cost $O(\varepsilon^{-2})$. Failure probability δ is achievable with deterministic cost $O(\varepsilon^{-2}[1 + \log(1/\delta)])$.*

Proof. Write $\rho(s) = \sum_a |f_s(a)|$. At s, sample A with probability $|f_s(A)|/\rho(s)$ and put $w = \rho(s)\,\mathrm{sgn}(f_s(A))$. If $\rho(s) = 0$, the continuation contributes zero and may stop. Along a simulated path (S_j, A_j, r_j) starting at s, define the single reward contribution

$$D_j = \beta^j \left(\prod_{i=0}^{j} w_i \right) r_j, \qquad j \ge 0.$$

It costs at most $j + 1$ queries to sample D_j and $|D_j| \le GR\alpha^j$. Iterating the linear Bellman equation shows

$$V^\star(s) = \sum_{j=0}^{\infty} \mathbb{E} D_j, \qquad \left| \sum_{j=H}^{\infty} \mathbb{E} D_j \right| \le \frac{GR\alpha^H}{1 - \alpha}.$$

Choose $H = O(\log(1/\varepsilon))$ to make this tail at most $\varepsilon/2$. Estimate each $\mathbb{E}D_j$, $0 \le j < H$, from independent samples, with variance bounds $v_j = (GR)^2 \alpha^{2j}$ and costs $c_j = j + 1$. The allocation used in Section C applies directly. Here

$$\sum_{j=0}^{H-1} \sqrt{v_j c_j} \le GR \sum_{j=0}^{\infty} \alpha^j \sqrt{j+1} < \infty.$$

The total deterministic cost is therefore $O(\varepsilon^{-2}) + O(H^2) = O(\varepsilon^{-2})$, with variance at most $\varepsilon^2/4$ and squared bias at most $\varepsilon^2/4$. For a prescribed root action, fix that initial action and use the same expansion from its successor; the root reward term has cost one and the remaining bounds change only by constants. Apply the median argument of Section D.1 for the confidence statement. □

This proof also shows why a logarithmic simulation horizon need not force a logarithmic factor multiplying ε^{-2}: increasingly small discounted reward contributions can be estimated with increasingly few samples. In the nonlinear case, producing equally effective contributions is exactly the difficulty addressed, but not fully eliminated, by the exponent-improvement construction.

D.5 THE ROLE OF LOCAL ACCURACY

The construction uses a proxy of RMS accuracy $O(\sqrt{u})$ to control the Taylor bias at scale u. Coupled increments then have second moment $O(u)$, so fine levels require fewer replicates. A local mean-estimation lower bound alone does not determine the cost of this allocation across recursive levels.

D.6 FINITE ACTIONS DO NOT MAKE MAXIMUM SMOOTH

For $K \ge 2$, $F(q) = \max_a q_a$ is not differentiable at ties and does not satisfy the quadratic remainder assumption. Even the restriction $q = (t, -t)$ gives $F(q) = |t|$, whose kink cannot have a quadratic Taylor remainder at zero. Finiteness of K removes neither that kink nor the statistical difficulty of locating a changing maximizer. Our construction therefore makes no quadratic or nearly quadratic claim for general unregularized Bellman optimality.

Approximating maximum by a soft maximum does not fix this distinction uniformly in accuracy. For the normalized entropy backup F_τ,

$$0 \le \max_a q_a - F_\tau(q) \le \tau \log K.$$

Both backups are 1-Lipschitz, so their bounded fixed points differ by at most $\tau \log K/(1 - \beta)$. To approximate an unregularized value to error $O(\varepsilon)$, this route requires $\tau \log K = O((1 - \beta)\varepsilon)$; its

smoothness parameter $L = 1/\tau$ then grows with $1/\varepsilon$. The constants in Theorem 1 are not uniform in this regime. Similarly, if $K = K(\varepsilon)$ increases, factors involving K, the probability floor, and the fixed-accuracy routines must be tracked. The fixed-K exponent alone does not answer that different question.

Remaining scope limitations. The intermediate Z_u samples have uniform second moments, not a uniform almost-sure envelope as $u \downarrow 0$; the final estimator can be clipped. All constants may deteriorate as $\beta G \uparrow 1$, $L \uparrow \infty$, or $K \uparrow \infty$. The explicit recurrence quantifies growth with the stage index while holding these parameters fixed. It does not assert favorable structural constants or numerical practicality for the worst-case construction.

E Experimental models, controls, and reproducibility

E.1 Models and practical proxies

We consider the following four models.

- **Entropy:** normalized entropy backup with $\tau = 0.2$.
- **Nearly tied entropy:** modified transitions and equalized rewards, with $\tau = 0.05$.
- **Signed sine:** $F(q) = 0.8 \sin(q_1)$, with a separate reward matrix.
- **Signed linear:** $F(q) = 0.55q_1 - 0.25q_2$.

We distinguish the 3,000-pair proposal-control runs from the earlier 6,000-pair bias and sensitivity diagnostics. The accompanying scripts regenerate both, and retain individual fine/coarse outputs for each sampled root pair. The transition matrices for actions 1 and 2 are

$$P_1 = \begin{pmatrix} .65 & .30 & .05 \\ .20 & .60 & .20 \\ .25 & .25 & .50 \end{pmatrix}, \qquad P_2 = \begin{pmatrix} .10 & .35 & .55 \\ .50 & .15 & .35 \\ .10 & .70 & .20 \end{pmatrix}, \qquad r = \begin{pmatrix} .15 & .12 \\ .05 & .10 \\ .20 & .00 \end{pmatrix}.$$

The default discount is $\beta = .7$. The entropy backup has $\tau = .2$. For the nearly tied case, write $\overline{P} = (P_1 + P_2)/2$ and replace P_1, P_2 by $\overline{P} \pm .01(P_1 - P_2)$, average the two rewards at each state, and use $\tau = .05$. The signed sine backup is $.8 \sin(q_1)$, with reward matrix $((1.05, .20), (1.50, -.20), (.65, .10))$. The linear control is $.55q_1 - .25q_2$ and uses the default rewards. These are fixed-point models; only the entropy examples have the planning interpretation of Section 3.2.

Each empirical-model proxy samples $n(u) = \lceil 1/u \rceil$ transitions at every state-action pair and solves the resulting finite Bellman equation. It uses fresh samples at every invocation. The anchor uses 128 observations per pair, and the base routine uses 512. Both also solve a fresh empirical model. The experimental cutoff is $u_0 = .05$ and the proposal parameter is $\lambda = .05$. The recursion uses the same $c = (1 + \beta G)/(2\beta G)$ as the theory. A deterministic iteration count makes the finite-model numerical truncation error at most 10^{-11}.

These proxies have a tractable finite-state interpretation. For S states, deterministic known rewards, and an exact empirical fixed point, a contraction argument gives

$$\mathbb{E}\|\widehat{Q} - Q^\star\|_\infty^2 \le \frac{\beta^2 B_V^2 SK}{(1-\alpha)^2 n}.$$

Indeed the empirical transition error applied to $V^\star$ has expected squared maximum at most $B_V^2 SK/n$, and the fixed-point perturbation contributes the factor $(1-\alpha)^{-1}$. Thus the proxies have the intended mean-square scaling in u. Their dimension-dependent construction is not the state-space-free bootstrap algorithm. The experimental constants are also not the conservative recipe of Appendix G; here u is a resolution parameter, not a certified bias tolerance. These runs test the mechanism and its sensitivity, not the full minimax complexity bound or a practical speedup over existing algorithms.

E.2 FIVE PROPOSAL CONTROLS AND UNCERTAINTY

There are four standard models, six resolutions $u = .05\,2^{-j}$, $j = 1, \ldots, 6$, and five modes: 120 configurations, each with 3,000 independent root pairs. At each recursive paired node:

1. **Shared anchor:** one empirical anchor determines a common proposal; one uniform action variate and one oracle response are shared.
2. **Independent anchors, coupled action draws:** two independent 128-observation anchors determine proposals; their actions use the same uniform variate, a maximal coupling for two actions.
3. **Shared uniform:** both proposals are $(1/2, 1/2)$, and the action and observation are shared; no anchor is sampled.
4. **Fine-proxy proposals:** each proposal is $(|\nabla F(q)| + \lambda/2)/(\|\nabla F(q)\|_1 + \lambda)$ using that member's own fine proxy; action draws use the same uniform variate. No anchor is sampled.
5. **Shared anchor, independent draws:** the parent proposal is shared, but actions and transitions are independent, including when the actions agree; the descendant single routines are also independent.

All modes use independent fine proxies and share the initial observation at the prescribed root action. For modes 1–4, equal actions share one transition and recurse as a pair; different actions use separate observations and independent single continuations with the appropriate marginal proposal rule. The mixed-base case always uses independent fine and base values. Both-base values are shared in modes 1–4 and independent in mode 5. Thus every marginal is well defined; no cross-action oracle coupling is assumed. The alternative proposal families need not satisfy the worst-case stability constants in our proof.

Marginal laws and common expectations. Variants 1, 2, and 5 use the same single-estimator anchored proposal rule; they preserve its marginal laws while changing the joint law of a pair. The uniform and fine-proxy rules (variants 3 and 4) generally change the individual estimator distributions. They are not merely different couplings of the original Z_u. Nevertheless, their expectations agree at each resolution when the proxy distributions and base routine are held fixed. To see this, for a positive-support proposal and a vector h fixed conditionally on $(\widehat{q}, \pi)$, importance sampling gives

$$\mathbb{E}_{A\sim\pi}\left[\frac{\partial_A F(\widehat{q})}{\pi_A} h_A \,\middle|\, \widehat{q}, \pi\right] = \sum_a \partial_a F(\widehat{q}) h_a.$$

At a recursive node take h_a to be the conditional mean of the fresh reward-plus-child correction minus $\widehat{q}_a$. Induction from the common base through the finite recursion makes this child mean independent of the proposal variant, so the proposal cancels at every step. This remains valid when π depends on $\widehat{q}$, as in variant 4. All five constructions have positive support. Equal expectations do not imply equal marginal laws or variances, and the common expectation may still be biased for the fixed point.

Every raw record stores the two root outputs, per-root query counts, and per-root sums of action/derivative diagnostics. The sample unit for uncertainty is a complete independent root pair, preserving dependence among its recursive nodes. We use 1,000 percentile bootstrap resamples within each configuration. Slope estimates are ordinary least squares of log empirical second moment against $\log u$, using levels 2–6 only; each bootstrap refits this same fixed range. Moment ratios divide bootstrap means from independently sampled configurations. These are descriptive finite-range intervals, not confidence statements about an asymptotic exponent. Node frequencies are ratios of total counts, with whole-root resampling for their intervals.

Fixed-accuracy independent anchors can disagree persistently as u decreases. This is a property of that control, not a general lower bound on proposals. Fine-proxy proposals become more accurate at the first correction as resolution improves; pooled node frequencies need not decrease at the same rate because coarser recursive nodes remain present. The raw CSV includes both definitions. Table 3 reports finest-level moments and costs for all five controls, so an improvement in moment alone is not presented as a cost advantage.

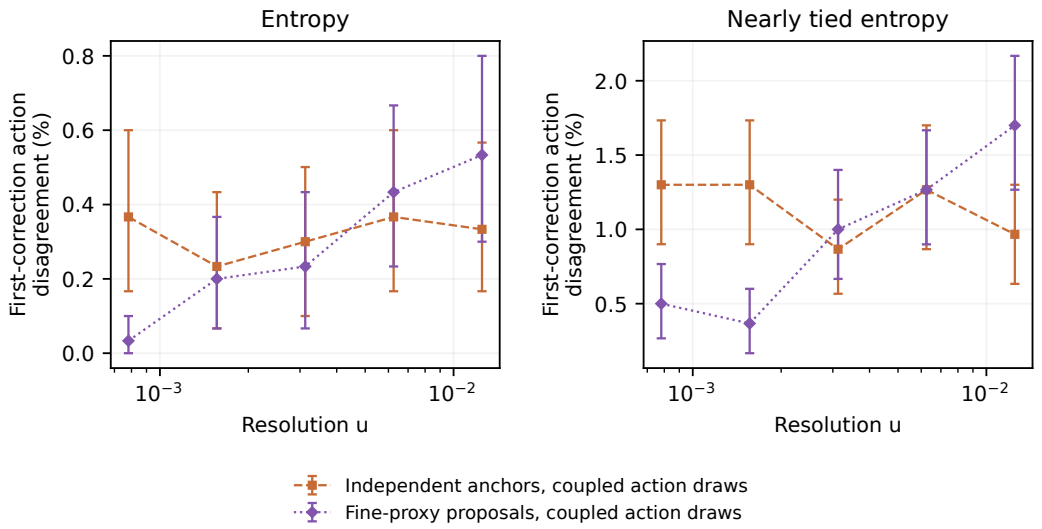

Figure 4: Disagreement at the first recursive correction for independent anchors and fine-proxy proposals. Shared anchored and shared uniform proposals have zero disagreement there. Colors and line styles match Figure 2; intervals use complete-root resampling. First-correction frequencies separate accuracy-dependent proposals from frequencies pooled over correction nodes at several resolutions. The independent- draw control has roughly 50% disagreement and is recorded in the raw results.

Model	Proposal / randomness	Second moment	Queries per pair
Entropy	Shared anchor	4.39e-07	87,191
Entropy	Independent anchors	3.76e-06	101,015
Entropy	Shared uniform	7.32e-07	70,295
Entropy	Fine-proxy proposals	1.13e-06	70,295
Entropy	Shared anchor, independent draws	5.58e-04	100,265
Nearly tied entropy	Shared anchor	5.53e-07	87,191
Nearly tied entropy	Independent anchors	1.44e-05	101,017
Nearly tied entropy	Shared uniform	5.52e-07	70,295
Nearly tied entropy	Fine-proxy proposals	8.92e-06	70,296
Nearly tied entropy	Shared anchor, independent draws	6.03e-04	100,265
Signed sine	Shared anchor	1.42e-09	56,330
Signed sine	Independent anchors	3.65e-09	64,778
Signed sine	Shared uniform	2.09e-08	46,346
Signed sine	Fine-proxy proposals	2.20e-09	46,346
Signed sine	Shared anchor, independent draws	2.19e-05	64,021
Signed linear	Shared anchor	8.61e-07	56,330
Signed linear	Independent anchors	7.98e-07	64,778
Signed linear	Shared uniform	1.23e-06	46,346
Signed linear	Fine-proxy proposals	8.04e-07	46,346
Signed linear	Shared anchor, independent draws	9.60e-04	64,021

Table 3: All controls at $u = .05/64$. Costs include proxy and anchor observations; counts can vary when coupled actions disagree. Mode 5 denotes a shared parent anchor with independent action/transition draws and descendants.

As a fixed-resolution efficiency diagnostic, the product of empirical increment variance (with denominator $n - 1$) and mean query cost is .0383 for the shared anchor versus .0514 for uniform on standard entropy. For nearly tied entropy, the corresponding products are .0482 and .0388, favoring uniform. These variances are computed from raw increments, not substituted second moments. The products describe independent averaging at this resolution, not complete estimator complexity; the supplementary CSV records all variants.

Metric (%)	Signed sine	Crossing stress
Negative active derivative	73.0 [72.4, 73.5]	44.7 [44.1, 45.3]
Active magnitude $< .05$	33.1 [32.5, 33.6]	31.4 [30.9, 31.9]
Opposite fine/coarse signs	0.0 [0.0, 0.0]	15.8 [15.4, 16.2]
Negative selected derivative	58.3 [57.7, 58.9]	32.8 [32.3, 33.3]
Zero selected derivative	16.6 [16.2, 16.9]	25.5 [25.1, 26.0]

Table 4: Frequencies over recursive nodes under the shared-anchor control at $u = .05/64$. Brackets are 95% bootstrap intervals. A zero observed reversal frequency is a finite-sample observation, not a proof that reversal is impossible.

E.3 Signed derivative diagnostics

For $F(q) = .8\sin(q_1)$, the active derivative is $.8\cos(q_1)$; the second coordinate is identically zero. We separately record negative active derivatives, active magnitudes below .05, opposite signs between fine/coarse active derivatives, and negative or zero *selected* derivatives. We do not count the structurally zero second coordinate as an active near-zero observation.

A crossing stress model uses the same transitions and discount, but sets $x = (\pi/2, \pi/2 - .15, \pi/2 + .15)$, $v_s = .8\sin(x_s)$, and

$$r(s,1) = x_s - \beta \sum_{s'} P_1(s,s')v_{s'}, \qquad (r(1,2), r(2,2), r(3,2)) = (.20, -.20, .10).$$

Its true active derivative vanishes at state 1; neighboring states have opposite derivative signs. This defines the synthetic model, not information given to the sampling routine. Five additional finest-resolution configurations use 3,000 independent pairs each. Table 4 reports the shared-anchor case for both signed models, with root-block bootstrap intervals. At the same resolution, the shared-anchor increment second moment is 5.75×10^{-12} for crossing stress versus 1.42×10^{-9} for the original signed model; both cost 56,330 queries per pair. Thus observed sign reversals coexist with small increments in this configuration; their presence alone does not imply poorer estimation. Other variants and raw-increment variance–cost products are available in the supplementary results.

E.4 Independent end-to-end estimator replications

On the standard entropy model, nominal targets are $.012, .008, .005, .003, .0015, .001$ and $J = \max\{1, \lceil \log_2(2\beta u_0/t) \rceil\}$. Both methods target the same marginal mean $\mathbb{E}Y_{u_J}$ and use the shared-anchor single or paired routine. For each target and method, 128 *complete* independent replications include their own calibration; no pilot is amortized across replications. Ground truth is used only for evaluation.

For the single-level estimator, 48 pilot samples of Y_{u_J} give an unbiased sample variance $\widehat{v}$; a fresh average uses $n = \max\{2, \lceil 2\max(\widehat{v}, 10^{-12})/t^2 \rceil\}$ samples. For the multilevel estimator, 48 pilot samples at every level yield $\widehat{v}_j$ and observed per-increment cost c_j. Set

$$v_j = \max(\widehat{v}_j, 10^{-12}), \quad S = \sum_{j=0}^{J} \sqrt{v_j c_j}, \quad n_j = \max\{2, \lceil 2t^{-2} S \sqrt{v_j/c_j} \rceil\}.$$

Fresh production samples are independent of all calibration draws. Outputs are the un-clipped single-level mean or sum of level means, respectively. All pilot and production observations are charged. Data-dependent allocations are practical, not certified replacements for the numerical bounds of Theorem 8; t is a nominal allocation parameter.

The first four targets were run before adding the two finest targets to examine the calibration-dominated regime more closely. All six are reported. This extension is descriptive; there was no selection of replications or removal of unfavorable target accuracies. RMSE and mean-query intervals resample complete estimator replications, including their independent calibrations.

Target	Single RMSE	Multilevel RMSE	Single queries (10^6)	Multilevel queries (10^6)
0.012	0.00795	0.00814	0.94	2.60
0.008	0.00616	0.00536	1.70	4.47
0.005	0.00370	0.00319	2.45	5.16
0.003	0.00230	0.00192	7.37	10.05
0.0015	0.00104	0.00091	38.45	26.21
0.001	0.00083	0.00073	140.49	56.51

Table 5: Complete estimator comparison; query counts are means and include fresh calibration in each replication. Figure 3 shows bootstrap intervals. Supplementary raw files retain every output, allocation, pilot variance, total cost, and calibration cost.

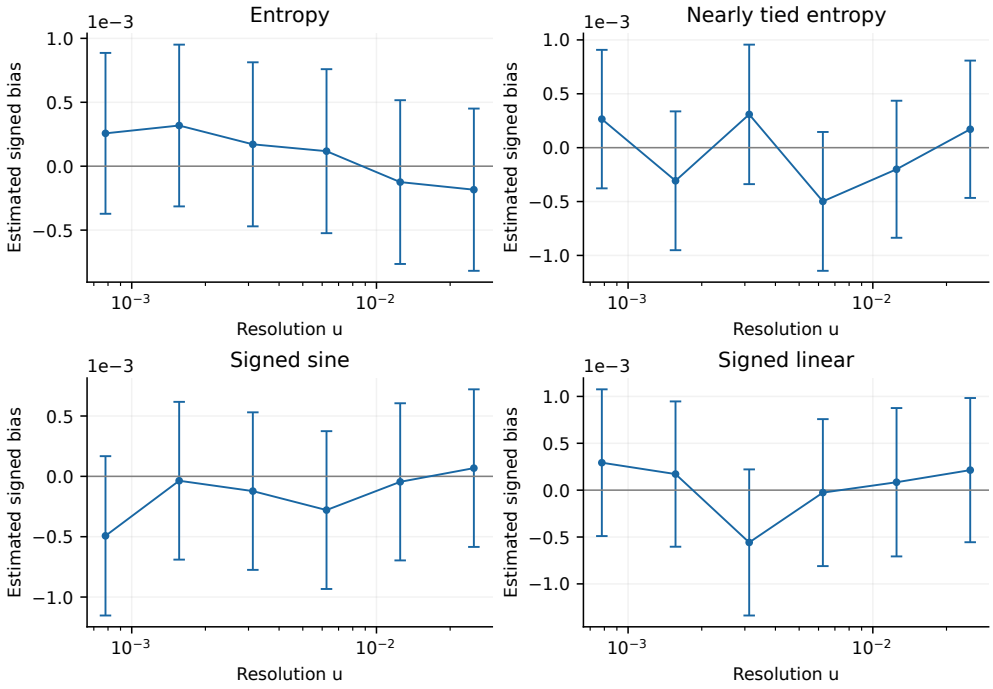

Figure 5: Signed bias of the fine output with shared anchors. Bars are 1.96 Monte Carlo standard errors over 6,000 independent pairs. The horizontal axis is resolution, not certified accuracy. Reference-solution error is below 5×10^{-12}.

E.5 Bias, sensitivity, and reproduction

The earlier 57-configuration diagnostic uses 6,000 pairs per configuration. Figure 5 retains its fine-output bias measurements; small biases are often unresolved at this replication count. Figure 6 varies discount, temperature, and proposal floor at $u = .05/8$. These runs measure sensitivities, not uniform constants or an optimal floor.

From the source-bundle root, `python experiments/reproduce.py` runs the simulations, numerical checks, bootstrap analysis, and figure/table generation. The `--from-raw` option regenerates figures and tables from the supplied individual outputs. The package records exact seeds and Python/NumPy/Matplotlib versions. Checks cover the entropy loss identity and bound, deterministic-kernel recovery for all five variants, reference residuals, and query-accounting consistency. They are numerical implementation checks, not a proof audit.

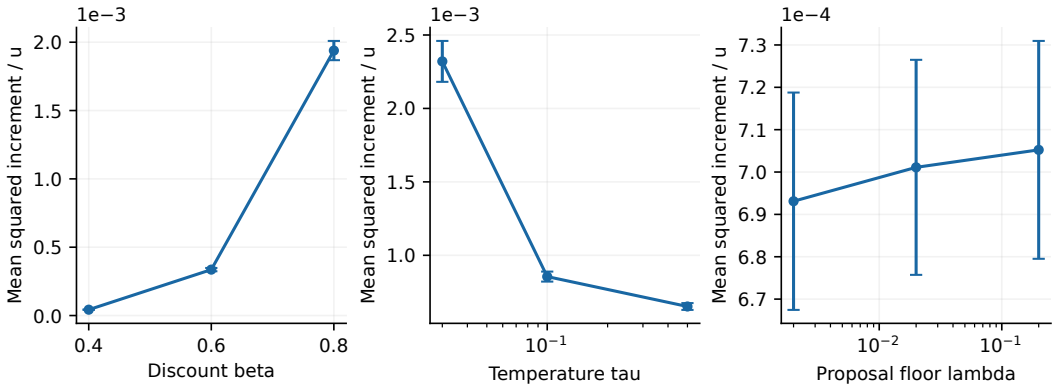

Figure 6: Shared-anchor entropy diagnostics at $u = .05/8$. Only the indicated parameter changes; bars are 1.96 Monte Carlo standard errors.

F A SELF-CONTAINED SIGNED SMOOTHCRUISER CONSTRUCTION

The baseline algorithm has two recursive calls with different roles. Many inexpensive child samples build an approximate action-value vector; a single, more accurate child supplies a linear correction. Taylor's theorem makes the residual bias quadratic in the vector's error. We specify the initialization and both calls explicitly.

F.1 A BOUNDED FINITE-DEPTH INITIALIZER

Lemma 12 (Initialization at any fixed accuracy)**.** *For every $h > 0$, there is a routine $W_h(s)$ with a finite deterministic query budget, depending only on h and the fixed parameters, such that*

$$|W_h(s)| \le B_V, \qquad \sup_{M,s} \|W_h(s) - V_M^\star(s)\|_{L^2} \le h.$$

Proof. Let $v_j = T^j 0$. For fixed integers $H, n \ge 1$, define random routines $W^{(0)} = 0$ and, for $j = 1, \dots, H$,

$$W^{(j)}(s) = \mathrm{clip}_{[-B_V, B_V]}\, F_s \left(\left(\frac{1}{n} \sum_{i=1}^{n} [r_{ai} + \beta W_{ai}^{(j-1)}(S'_{ai})] \right)_{a=1}^{K} \right). \tag{61}$$

Each summand uses a fresh query at (s, a) and an independent recursive copy. It is bounded in magnitude by $R + \beta B_V = B_Q$. For $D_j = \sup_{M,s} \|W^{(j)}(s) - v_j(s)\|_{L^2}$, the centered sampling error has expected squared maximum at most KB_Q^2/n. The vector of expectation errors has maximum norm at most βD_{j-1}. Lipschitz continuity and clipping therefore give

$$D_j \le GB_Q \sqrt{K/n} + \alpha D_{j-1}, \qquad D_0 = 0.$$

Also $\|v_H - V^\star\|_\infty \le \alpha^H B_V$. Thus

$$\sup_{M,s} \|W^{(H)}(s) - V^\star(s)\|_{L^2} \le \alpha^H B_V + \frac{GB_Q}{1-\alpha} \sqrt{K/n}.$$

Choose H with the first term at most $h/2$ and $n \ge 4G^2 B_Q^2 K/[(1-\alpha)^2 h^2]$. Then $W_h = W^{(H)}$ works and costs at most $\sum_{j=1}^{H} (Kn)^j$ queries. Its potentially large cost will only be used at fixed accuracies. For a numerical specification, choose

$$H(h) = \max\left\{1, \left\lceil \frac{\log_+(2B_V/h)}{\log(1/\alpha)} \right\rceil \right\}, \quad n(h) = \max\left\{1, \left\lceil \frac{4G^2 B_Q^2 K}{(1-\alpha)^2 h^2} \right\rceil \right\},$$

and write

$$W_{\mathrm{cost}}(h) = \sum_{j=1}^{H(h)} [Kn(h)]^j. \tag{62}$$

□

F.2 THE BOUNDED BASIC SAMPLE

Recall c, d from (28), so that $\alpha c < 1$ and $\alpha^2 c < 1$. Set

$$B_X = \frac{2GB_Q + GR}{1-\alpha}, \qquad \overline{B} = R + \beta B_X, \qquad \kappa^2 = \frac{d}{8\beta^2}. \tag{63}$$

Set $h_0 = \min\{1/2, \kappa^2/4\}$ and obtain W_{h_0} from Lemma 12.

Algorithm 1: bounded signed sample $X_u(s)$.

1. If $u \geq h_0$, return a fresh $W_{h_0}(s)$.
2. Otherwise let $n(u) = \lceil 8K\overline{B}^2/(du) \rceil$. For each action, average $n(u)$ independent observations $r + \beta X_{\kappa\sqrt{u}}(S')$, and clip the average to $[-B_Q, B_Q]$. Denote the resulting vector by $\widehat{q}$.
3. Compute $g = \nabla F_s(\widehat{q})$ and $\rho = \|g\|_1$. If $\rho = 0$, return $F_s(\widehat{q})$. Otherwise sample A with probability $|g_A|/\rho$, query (r, S') at (s, A) independently of the pilot, and return

$$X_u(s) = F_s(\widehat{q}) + \rho \operatorname{sgn}(g_A)[r + \beta X_{cu}(S') - \widehat{q}_A]. \tag{64}$$

   The correction child uses fresh randomness.

For $u < h_0$, the pilot tolerance obeys $\kappa\sqrt{u} \geq 2u$, and the correction tolerance is $cu > u$. Every recursive branch therefore reaches the base regime in a uniformly bounded number of steps for fixed input u. All sample counts are deterministic functions of the requested tolerance.

Proposition 13 (Bound, bias, and cost of a basic sample)**.** *Algorithm 1 satisfies*

$$|X_u(s)| \leq B_X, \tag{65}$$

$$|\mathbb{E}X_u(s) - V^\star(s)| \leq u, \tag{66}$$

$$S_X(u) := \sup_{M,s} \operatorname{Cost}(X_u(s)) \leq Cu^{-2}\mathcal{L}(u)^b \quad (0 < u < h_0), \tag{67}$$

for some finite constants C, b depending only on the fixed parameters.

Proof. Induct over the finite recursion. The base output has magnitude at most $B_V \leq B_X$ and absolute bias at most $h_0 \leq u$. Clipping gives $|F_s(\widehat{q})| \leq GB_Q$. If the children satisfy (65), then

$$|X_u| \leq GB_Q + G(R + \beta B_X + B_Q) = B_X,$$

including the case $\rho = 0$.

Write $q = Q_s^\star$. Each unaveraged pilot sample has magnitude at most $\overline{B}$ and expectation within $\beta\kappa\sqrt{u}$ of its target coordinate. A centered-variance bound and the squared-sum inequality give, also after clipping,

$$\mathbb{E}\|\widehat{q} - q\|_\infty^2 \leq \frac{2K\overline{B}^2}{n(u)} + 2\beta^2\kappa^2 u \leq \frac{du}{2} \leq du. \tag{68}$$

Conditional on the pilot, expectation of the signed action correction is $g \cdot (q - \widehat{q})$ up to absolute error at most $\beta\|g\|_1 cu \leq \alpha cu$. Consequently,

$$|\mathbb{E}X_u - V^\star| \leq \frac{L}{2}\mathbb{E}\|\widehat{q} - q\|_\infty^2 + \alpha cu \leq \left(\frac{1 - \alpha c}{2} + \alpha c\right) u \leq u.$$

When $\rho = 0$, the same Taylor estimate applies with no correction.

For the cost, some fixed $A_0 < \infty$ gives

$$S_X(u) \leq A_0 u^{-1}[1 + S_X(\kappa\sqrt{u})] + 1 + S_X(cu) \qquad (0 < u < h_0). \tag{69}$$

For sufficiently small u, one has $\mathcal{L}(\kappa\sqrt{u}) \leq \frac{3}{4}\mathcal{L}(u)$ and $\mathcal{L}(cu) \leq \mathcal{L}(u)$. Substitution of $Cu^{-2}\mathcal{L}(u)^b$ into the two recursive terms gives at most

$$Cu^{-2}\mathcal{L}(u)^b \left[A_0\kappa^{-2}(3/4)^b + c^{-2}\right].$$

Choose a finite $b \geq 0$ such that the bracket is strictly below one; this is possible because $c^{-2} < 1$. The additive terms $A_0 u^{-1} + 1$ are absorbed by the remaining slack for a sufficiently large C. Above a fixed small threshold the cost is uniformly finite: branching is bounded and every branch increases its tolerance by at least $\min\{2, c\} > 1$ until it reaches h_0. Increase C to cover that interval and the base routine. Induction over the finite recursion proves (67). □

F.3 From a basic sample to an RMS estimator

Corollary 14 (Fourth-power initialization)**.** *There is an algorithm $\mathcal{A}_0$ satisfying* (27) *with $p = 4$. A known constant C_0 and exponent b, valid for every $t \in (0, 1]$, are specified in Appendix F.4.*

Proof. At (s, a), let $Y = r + \beta X_{t/(2\beta)}(S')$ using a fresh root query. Its absolute bias for $Q^\star(s, a)$ is at most $t/2$ and $|Y| \le \overline{B}$. Average $n = \lceil 4\overline{B}^2/t^2 \rceil$ independent copies. Before clipping, its variance is at most $t^2/4$ and its MSE is at most $t^2/2 \le t^2$. Clip the result to $[-B_Q, B_Q]$, which can only reduce its MSE. Each copy costs $O(t^{-2}\mathcal{L}(t)^b)$; multiplying by n gives the stated cost. □

The fourth power has now been derived without an external initialization theorem. It comes from multiplying the number of outer samples by the cost of a basic sample. That multiplication is what the multilevel construction will improve.

F.4 Known baseline constants at every input tolerance

The following deliberately conservative recipe removes all hidden accuracy cutoffs from the initialization. Its exponent b is denoted b_{init} in Section C. Put

$$
\begin{aligned}
a_n &= 8K\overline{B}^2/d, \qquad a_X = K(a_n + 1), \qquad \gamma_X = (1 - c^{-2})/2,\\
b &= \left\lceil \frac{\log_+(a_X \kappa^{-2}/\gamma_X)}{\log(4/3)} \right\rceil, \qquad u_X = \min\{h_0/2, \exp[-1 - 4\log_+(1/\kappa)]\},\\
r_X &= \min\{2, c\}, \qquad D_X = \left\lceil \frac{\log(h_0/u_X)}{\log r_X} \right\rceil, \qquad n_X = \lceil a_n/u_X \rceil,\\
B_X^{\mathrm{tree}} &= Kn_X + 1, \qquad T_X = \max\{h_0, ch_0, \kappa\sqrt{h_0}\}.
\end{aligned}
$$

The use of B_X^{tree} here denotes a branching bound, not the sample envelope B_X in (63). Define

$$
\begin{aligned}
C_{\mathrm{strip}} &= W_{\mathrm{cost}}(h_0)(B_X^{\mathrm{tree}})^{D_X} + \sum_{j=1}^{D_X} (B_X^{\mathrm{tree}})^j,\\
C_X &= \max\{(a_X + 1)/\gamma_X, C_{\mathrm{strip}} \max\{1, T_X^2\}\},\\
C_0 &= (4\overline{B}^2 + 1)\{1 + C_X[1 + 4(1 + \log 2)^b]\}.
\end{aligned} \tag{70}
$$

These are finite numerical functions of the structural parameters. For $u \ge u_X$, every nonbase node has at most B_X^{tree} children and at most that many local queries; every branch has at most D_X nonbase nodes. This proves the C_{strip} bound. For $u < u_X$,

$$
\mathcal{L}(\kappa\sqrt{u}) \le \tfrac{3}{4}\mathcal{L}(u), \qquad a_X\kappa^{-2}(3/4)^b + c^{-2} \le 1 - \gamma_X.
$$

The additive node cost $a_X u^{-1} + 1$ is absorbed by $\gamma_X C_X u^{-2}\mathcal{L}(u)^b$. On $[u_X, T_X]$, the definition of C_X covers the strip, including children that first cross the base threshold. Induction therefore gives $S_X(u) \le C_X u^{-2}\mathcal{L}(u)^b$ for $0 < u \le T_X$; for every $u > 0$ it gives

$$
S_X(u) \le C_X[1 + u^{-2}\mathcal{L}(u)^b].
$$

At root tolerance $t \le 1$, put $u = t/(2\beta)$. Then $u^{-2} \le 4t^{-2}$ and $\mathcal{L}(u) \le (1 + \log 2)\mathcal{L}(t)$. The outer sample count is at most $(4\overline{B}^2 + 1)t^{-2}$. Including the root query yields precisely $C_0 t^{-4}\mathcal{L}(t)^b$. This completes the numerical specification of the baseline, with no stage-dependent qualification.

G ONE FIXED STRUCTURAL RECIPE

Use c, d from (28), and let $H = 3L$. Set

$$\begin{aligned} M &= \tfrac{1}{2}\left(G^2 + \frac{1}{\beta^2 c}\right), \qquad \Delta_M = M - G^2, \\ \xi &= \tfrac{1}{2}\left(\frac{1}{\beta^2 M c} - 1\right), \qquad k = (1+\xi)\beta^2 M, \\ \theta &= \min\{1, \Delta_M/(8G^2)\}, \\ \lambda &= \min\{G, \Delta_M/[8G(1+\theta)]\}, \\ E_g &= 2H^2(1+\theta^{-1})K(G+\lambda)/\lambda, \\ a_0^2 &= \min\{1, \Delta_M/(8E_g)\}, \\ u_0 &= \min\{1/2, K/d, \Delta_M/(8E_g d)\}. \end{aligned} \tag{71}$$

All quantities are strictly positive and $kc < 1$. Moreover,

$$(1+\theta)G(G+\lambda) \le G^2 + \Delta_M/4, \qquad E_g(du_0 + a_0^2) \le \Delta_M/4.$$

These verify Lemma 3. The anchor accuracy $a_0/\sqrt{K}$ is in $(0, 1]$, and every proxy accuracy for $u < u_0$ is in $(0, 1]$.

For completeness, the remaining second-moment constants are

$$\begin{aligned} \pi_{\min} &= \lambda/[K(G+\lambda)], \qquad W_{\max} = G/\pi_{\min}, \\ D_0 &= GB_Q + W_{\max}(R + B_Q), \\ M_Z &= \max\{B_V^2, (1+\xi^{-1})D_0^2/(1-k)\}, \\ C_R &= 2(G + W_{\max})^2 + 4H^2\pi_{\min}^{-2}[(R+B_Q)^2 + \beta^2 M_Z], \\ C_D &= \max\{8M_Z/u_0, 6d(1+\xi^{-1})C_R/(1-kc)\}. \end{aligned}$$

Together with (62), (70), (44)–(45), and (52), this is a deterministic recipe for every constant used in the allocation. It is intended to certify the algorithm, not to tune it efficiently.

H REGULARIZED ROOT-DECISION GUARANTEE

Fix a state and let $q = Q_s^\star$, $\pi^\star = \mathrm{softmax}(q/\tau)$, $\widehat{\pi} = \mathrm{softmax}(\widehat{q}/\tau)$. Write $J_q(\pi) = \pi \cdot q + \tau\mathcal{H}(\pi) - \tau \log K$. Because $\widehat{\pi}$ maximizes $J_{\widehat{q}}$,

$$0 \le J_q(\pi^\star) - J_q(\widehat{\pi}) \le (\pi^\star - \widehat{\pi}) \cdot (q - \widehat{q}) \le 2\|q - \widehat{q}\|_\infty.$$

Also $\log \pi_a^\star = q_a/\tau - \log \sum_b e^{q_b/\tau}$; substituting this identity into the relative entropy gives $\tau \mathrm{KL}(\widehat{\pi}\|\pi^\star) = F_\tau(q) - J_q(\widehat{\pi})$. The variational identity also gives

$$F_\tau(q) - J_q(\widehat{\pi}) = F_\tau(q) - F_\tau(\widehat{q}) - \nabla F_\tau(\widehat{q}) \cdot (q - \widehat{q}) \le \frac{\|q - \widehat{q}\|_\infty^2}{2\tau}.$$

Combining the two bounds proves (6). For $0 < \eta \le \min\{1, (2\tau)^{-1}\}$ set $\varepsilon_\eta = \sqrt{2\tau\eta} \le 1$. Apply (60) to each coordinate at accuracy ε_η and failure probability δ/K. The union bound gives loss at most η with probability at least $1 - \delta$, using at most

$$K[8\log(K/\delta) + 3]\mathcal{B}\Big(\sqrt{2\tau\eta}/2\Big)$$

queries. With fixed K, τ, substituting (58) yields (7). The guarantee concerns one decision with optimal continuation, not a policy learned uniformly over all states.

H.1 A MATCHING LOCAL DECISION LOWER BOUND

The leading accuracy exponent in (7) cannot be improved uniformly over the entropy subclass, even when the decision has only two actions and no nonzero continuation value.

Proposition 15 (Local entropy-decision lower bound)**.** *Fix $R, \tau > 0$ and $0 < \beta < 1$. For two-action normalized entropy backups, any algorithm that returns a decision $\widehat{\pi} \in \Delta_2$ with*

$$\Pr\{F_\tau(Q_s^\star) - J_{Q_s^\star}(\widehat{\pi}) \le \eta\} \ge 1 - \delta$$

at a prescribed root for every admissible model, using a deterministic query budget N, must satisfy

$$N \ge \frac{R^2}{512\tau\eta}\log(1/\delta), \qquad 0 < \eta \le \min\left\{\frac{\tau}{32}, \frac{R^2}{128\tau}\right\}, \quad 0 < \delta \le \frac{1}{4}. \tag{72}$$

The constant is not optimized. This is a fixed-temperature statement.

Proof. Use a root s and an absorbing state z. All actions at z give reward zero and remain at z, so $V^\star(z) = F_\tau(0,0) = 0$. At s, every query moves to z and gives reward in $\{-R, R\}$. In model $\theta \in \{-h, h\}$, the probability of reward R is $(1+\theta)/2$ for action 1 and $(1-\theta)/2$ for action 2. Thus

$$Q_s^\star = (R\theta, -R\theta), \qquad p_\theta := \pi_\theta^\star(1) = \frac{1}{1+\exp(-2R\theta/\tau)}.$$

Set $h = 4\sqrt{2\tau\eta}/R$. The cutoff in the proposition ensures $h \le 1/2$ and $Rh/\tau \le 1$. Since $\tanh x \ge x/2$ for $0 \le x \le 1$,

$$p_h - p_{-h} = \tanh(Rh/\tau) \ge \frac{Rh}{2\tau} = 4\sqrt{\frac{\eta}{2\tau}}.$$

For any returned decision, the entropy identity and Pinsker's inequality give

$$F_\tau(Q_s^\star) - J_{Q_s^\star}(\widehat{\pi}) = \tau\,\mathrm{kl}(\widehat{\pi}(1), p_\theta) \ge 2\tau|\widehat{\pi}(1) - p_\theta|^2.$$

Therefore loss at most η places $\widehat{\pi}(1)$ within $\sqrt{\eta/(2\tau)}$ of the correct p_θ. As $p_h + p_{-h} = 1$, thresholding the returned first-action probability at $1/2$ yields a test with both errors at most δ.

A root query, under either action, has relative entropy between the two models equal to

$$h\log\frac{1+h}{1-h} \le 4h^2.$$

A query at z has zero divergence. The adaptive transcript chain rule, including an algorithm's independent internal randomness, bounds total KL by $4Nh^2$. Data processing for the decision test bounds it below by $\mathrm{kl}(1-\delta, \delta) \ge \frac{1}{4}\log(1/\delta)$. Consequently $N \ge \log(1/\delta)/(16h^2)$; substituting $h^2 = 32\tau\eta/R^2$ proves (72). □

For fixed R, τ this matches the decision upper bound's η^{-1} power and confidence factor. It neither removes the upper bound's subpolynomial overhead nor establishes a lower bound for learning an entire policy. The construction is already a one-decision subproblem of the infinite-horizon class.