THE EXECUTABLE CONTENT OF GIBBS PATH TILTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gibbs trajectory reweighting can favor environmental outcomes that a controller cannot choose. Yet an infeasible path target need not induce suboptimal actions. We quantify this distinction in finite-horizon entropy-regularized control. Building on established information projections and residual-value correction, we express the discrepancy as an auxiliary control problem whose cost is local transition optimism. Its propagated value gives an exact criterion for lossless action extraction and an action-variance identity. As the cost scale λ tends to zero, relaxation optimism is $O(\lambda^2)$ and execution loss is $O(\lambda^4)$; either leading coefficient may vanish. We derive explicit coefficients over the full Markov decision process and characterize a sixth-order degenerate case. Under learned dynamics, small model error can reverse the benefit of correction near a decision tie. A confidence-certified selector protects the regularized objective; a paired transfer bound exploits cancellation between candidates. Supplied aggregate records cover $663{,}552$ random control problems and $12{,}779{,}520$ finite-sample trials. In the latter, unconditional correction is harmful in 38.46% of trials; the evaluated conservative selector switches in 7.06%, with no recorded harmful switch on a valid confidence event. These studies support the gap mechanisms and learned-model comparison, but do not test the asymptotic rates or paired refinement. The contribution is a quantitative account of decision-relevant path optimism, not a new control solver.

1 INTRODUCTION

A trajectory-level Gibbs law rewards low-cost paths regardless of whether their probability can be increased by choosing actions or only by changing the outcomes of environmental randomness. This distinction is central to stochastic control as inference (Rawlik et al., 2012; Levine, 2018). It does not, however, settle the decision question: *when do the action probabilities of an optimistic path target actually lose performance under the physical transition kernel?*

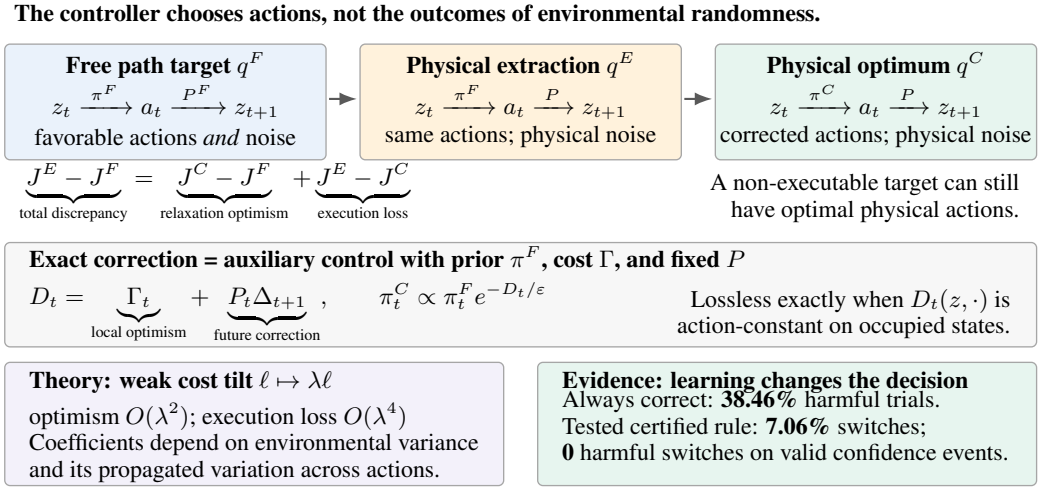
A one-action example separates the questions. Suppose the sole action leads equiprobably to terminal costs 0 and 1. At temperature $\varepsilon = 0.2$, the free Gibbs law puts approximately 0.993 probability on the zero-cost outcome and has regularized value 0.137, rather than the physical value 0.5. Its optimism is substantial, but extracting its action incurs no execution loss: there is no alternative action. More generally, common-mode transition optimism can leave every action probability unchanged. A measure of path non-executability alone is therefore not a measure of policy suboptimality.

This paper studies finite-horizon Markov decision processes (MDPs) and quantifies that separation. Let q^F be the free Gibbs law, q^E the law obtained by executing its action conditional under the physical kernel, and q^C the optimal physical law. Write J^X for the regularized objective of q^X, for $X \in \{F, C, E\}$, as defined in Section 2. Their *relaxation gap* $J^C - J^F$ and *execution gap* $J^E - J^C$ have different causes and need not have comparable magnitudes. The projection directions themselves are known: Lefebvre (2024) derives both forward and reverse Kullback–Leibler (KL) projections, and Arriojas et al. (2023) study optimistic policies in the physical environment and recover fixed-dynamics control through modified inference. We do not claim these constructions, stochastic optimism, or soft Bellman recursion as new.

Contribution. We characterize the *decision-relevant* content of optimism through a specific auxiliary problem, an exact propagated action criterion, and a finite-tilt variance identity. We derive a full-MDP second/fourth-order separation theorem with explicit coefficients and a sharp degenerate case. For learned dynamics, a local sign characterization and paired confidence bounds turn correction

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.

Figure 1: **Path-space optimism and execution loss are different quantities.** Here z_t, a_t are states and actions; P is the physical transition kernel and P^F its Gibbs tilt. The policies π^F, π^C give the free and optimal physical action probabilities. Local transition optimism is Γ; Δ is the executable-minus-free value difference, and $D_t = \Gamma_t + P_t\Delta_{t+1}$ includes its future correction. The temperature is ε, the trajectory cost is ℓ, and λ scales that cost; O denotes asymptotic order as λ tends to zero. Action-constant D leaves extraction lossless. The percentages summarize the finite-sample boundary study; they do not test the weak-tilt law (Theorem 6). Certification concerns the regularized objective.



Boundary evidence: 12,779,520 trials. The weak-tilt law and paired refinement are analytical results.

into a certified policy comparison. The existing experiments test null cases, propagation, and the help–harm boundary; they do not test the new asymptotic rates or paired selector. Small typical losses and conservative abstention are part of the evidence.

2 Three Laws and an Auxiliary Correction Problem

We assume a finite horizon T, finite state space $\mathcal{Z}$, nonempty finite action sets $\mathcal{A}_t(z)$, bounded stage costs c_t and terminal cost h, temperature $\varepsilon > 0$, and a reference policy with $\pi_{0,t}(a \mid z) > 0$. The initial distribution μ_0 and transition kernel P_t are fixed. A trajectory is $\tau = (z_0, a_0, \dots, z_T)$; capitals denote the corresponding random variables. Throughout, $t < T$ means $t \in \{0, \dots, T-1\}$. For finite distributions, $\mathrm{KL}(p\|q) = \sum_x p(x)\log[p(x)/q(x)]$, with zero-mass terms zero and divergence infinite if p assigns mass outside q's support. Write $\operatorname{supp} p = \{x : p(x) > 0\}$ and $p \ll q$ for this support inclusion. All logarithms are natural. We write $\mathbb{P}$ for probability, $\mathbb{E}$ for expectation, Var for variance, $\mathrm{TV}(p,q) = \frac{1}{2}\sum_x |p(x) - q(x)|$ for total variation, $\operatorname{osc}(f) = \sup f - \inf f$ for its range, and $[x]_+ = \max\{x, 0\}$. Subscripts indicate the law or variable involved. Write $P_t f(z,a) = \sum_{z'} P_t(z' \mid z, a) f(z')$ and $\ell(\tau) = \sum_{t<T} c_t(z_t, a_t) + h(z_T)$. The reference law and objective are

$$p_0(\tau) = \mu_0(z_0) \prod_{t<T} \pi_{0,t}(a_t \mid z_t) P_t(z_{t+1} \mid z_t, a_t), \tag{1}$$

$$\mathcal{F}(q) = \mathbb{E}_q \ell + \varepsilon\, \mathrm{KL}(q\|p_0). \tag{2}$$

A history-dependent policy $\alpha_t(\cdot \mid H_t)$, where $H_t = (Z_0, A_0, \dots, Z_t)$, induces a law q^α using μ_0 and P. Let $\mathcal{Q}_P^{\mathrm{hist}}$ be the class of all such laws, and let $\mathcal{Q}_\mu = \{q \ll p_0 : q(Z_0) = \mu_0\}$ also allow altered transition conditionals. Define

$$q^C = \arg\min_{q \in \mathcal{Q}_P^{\mathrm{hist}}} \mathcal{F}(q), \qquad q^F = \arg\min_{q \in \mathcal{Q}_\mu} \mathcal{F}(q), \qquad q^E = \mu_0 \prod_{t<T} \pi_t^F P_t. \tag{3}$$

Here π^F is the action conditional of q^F; $J^X = \mathcal{F}(q^X)$ for $X \in \{F, C, E\}$. We use $d_t^X(z) = q^X(Z_t = z)$ and $\rho_t^X(z,a) = q^X(Z_t = z, A_t = a)$. "Occupied" means positive probability under the specified law; "reachable" means positive probability under some physical law. All transition comparisons are on the physical row's support.

Established recursions and projections. Let V_t and Q_t^C denote the executable state and action values. The executable soft Bellman recursion is

$$V_T = h, \qquad Q_t^C = c_t + P_t V_{t+1}, \tag{4}$$

$$V_t(z) = -\varepsilon \log \sum_a \pi_{0,t}(a \mid z) e^{-Q_t^C(z,a)/\varepsilon},$$

$$\pi_t^C(a \mid z) = \pi_{0,t}(a \mid z) e^{-[Q_t^C(z,a) - V_t(z)]/\varepsilon}. \tag{5}$$

Let W_t denote the free state value and G_t its transition soft cost. The free recursion instead averages exponentiated continuation costs:

$$W_T = h, \qquad G_t = -\varepsilon \log P_t e^{-W_{t+1}/\varepsilon}, \tag{6}$$

$$W_t(z) = -\varepsilon \log \sum_a \pi_{0,t}(a \mid z) e^{-[c_t(z,a) + G_t(z,a)]/\varepsilon}, \tag{7}$$

$$\pi_t^F(a \mid z) = \pi_{0,t}(a \mid z) e^{-[c_t(z,a) + G_t(z,a) - W_t(z)]/\varepsilon}, \tag{8}$$

$$P_t^F(z' \mid z, a) = P_t(z' \mid z, a) e^{-[W_{t+1}(z') - G_t(z,a)]/\varepsilon}. \tag{9}$$

Thus $q^F = \mu_0 \prod_t \pi_t^F P_t^F$, $J^F = \mathbb{E}_{\mu_0} W_0$, and $J^C = \mathbb{E}_{\mu_0} V_0$. The Gibbs normalization is *conditional on* Z_0; a single global normalizer would generally change μ_0. These are standard regularized-control and probabilistic-control constructions (Kárný and Guy, 2006; Rawlik et al., 2012; Geist et al., 2019; Lefebvre, 2024). In particular, q^C minimizes $\mathrm{KL}(q\|q^F)$ and q^E minimizes $\mathrm{KL}(q^F\|q)$ over physical laws. Appendices A–B state and prove the underlying variational identities, and Appendix C.1 extends the reverse-projection statement to history-dependent policies.

From local optimism to a global decision problem. Define the transition optimism

$$\Gamma_t(z,a) = P_t W_{t+1}(z,a) - G_t(z,a) = \varepsilon\, \mathrm{KL}(P_t(\cdot \mid z,a) \| P_t^F(\cdot \mid z,a)) \geq 0. \tag{10}$$

Let $g_{\mathrm{relax}} = J^C - J^F$, $g_{\mathrm{exec}} = J^E - J^C$, and $g_{\mathrm{total}} = J^E - J^F$.

Theorem 1 (Auxiliary control and exact gap decomposition)**.** *For every physical law $q \in \mathcal{Q}_P^{\mathrm{hist}}$,*

$$\mathcal{F}(q) - J^F = \mathbb{E}_q \sum_{t<T} \Gamma_t(Z_t, A_t) + \varepsilon\, \mathrm{KL}(q\|q^E), \tag{11}$$

$$\mathrm{KL}(q\|q^F) = \mathrm{KL}(q\|q^C) + \mathrm{KL}(q^C\|q^F). \tag{12}$$

Consequently, correction is exactly an auxiliary entropy-regularized control problem with physical kernel P, stage cost Γ, prior π^F, and zero terminal cost. Moreover,

$$g_{\mathrm{relax}} = \varepsilon\, \mathrm{KL}(q^C\|q^F) = \sum_{t,z,a} \rho_t^C \Gamma_t + \varepsilon \sum_{t,z} d_t^C\, \mathrm{KL}(\pi_t^C\|\pi_t^F), \tag{13}$$

$$g_{\mathrm{exec}} = \varepsilon\, \mathrm{KL}(q^E\|q^C) = \varepsilon \sum_{t,z} d_t^E\, \mathrm{KL}(\pi_t^F\|\pi_t^C), \tag{14}$$

$$g_{\mathrm{total}} = \varepsilon\, \mathrm{KL}(q^E\|q^F) = \sum_{t,z,a} \rho_t^E \Gamma_t = g_{\mathrm{relax}} + g_{\mathrm{exec}}. \tag{15}$$

The relaxation gap is zero if and only if $\Gamma_t = 0$ on every q^C-occupied state–action cell, equivalently if and only if W_{t+1} is constant on each corresponding transition support; then $q^F = q^C = q^E$.

The Pythagorean equality is the familiar information-projection geometry (Csiszár, 1975). Its decision interpretation is useful: g_{total} is the cost of following the auxiliary prior, g_{relax} is the auxiliary optimum, and g_{exec} is the improvable part. The auxiliary-control form builds on residual soft-value correction (Adamczyk et al., 2023): with stage cost $c_t - \Gamma_t$ and terminal cost h, W is a physical soft value with policy π^F, since $c_t - \Gamma_t + P_t W_{t+1} = c_t + G_t$. Restoring the cost Γ_t gives exactly the residual problem above. Our comparison specializes this construction to the optimism induced by a Gibbs transition tilt. Occupancies cannot be interchanged: optimism is evaluated under q^C in (13), but under q^E in (15).

3 Only Propagated, Action-Differential Optimism Matters

Set the auxiliary value $\Delta_t = V_t - W_t$ and action value

$$\Delta_T = 0, \qquad D_t(z,a) = \Gamma_t(z,a) + P_t\Delta_{t+1}(z,a). \tag{16}$$

The continuation term is essential: an action can change which future transition corrections are encountered even when its immediate Γ_t is small. The following recursion identifies the exact correction, rather than prescribing an additional risk penalty.

Theorem 2 (Backward executable correction)**.** *For every t, z, a,*

$$\Delta_t(z) = -\varepsilon \log \mathbb{E}_{A\sim\pi_t^F(\cdot|z)} e^{-D_t(z,A)/\varepsilon}, \tag{17}$$

$$\pi_t^C(a \mid z) = \pi_t^F(a \mid z) e^{-[D_t(z,a)-\Delta_t(z)]/\varepsilon}, \tag{18}$$

$$\varepsilon\,\mathrm{KL}(\pi_t^F(\cdot \mid z)\|\pi_t^C(\cdot \mid z)) = \mathbb{E}_{\pi_t^F} D_t(z,A) - \Delta_t(z). \tag{19}$$

One backward pass recovers π^C from (W, Γ, π^F) exactly.

This is the Bellman solution of the auxiliary problem in Theorem 1. Its complexity is linear in the finite transition table, just as for direct soft dynamic programming; it is an interpretation and exact conversion, not a faster optimizer.

Corollary 3 (Exact lossless extraction)**.**

$$g_{\mathrm{exec}} = \sum_{t,z} d_t^E(z)\big[\mathbb{E}_{\pi_t^F} D_t(z,A) - \Delta_t(z)\big]. \tag{20}$$

It vanishes if and only if $D_t(z,\cdot)$ is action-constant at every q^E-occupied state. Then $q^E = q^C$, even when $g_{\mathrm{relax}} > 0$.

Corollary 4 (Common-mode invariance and a range bound)**.** *If $P_t(\cdot \mid z, a)$ is action-independent at every reachable state, then $\pi^F = \pi^C$ there and $g_{\mathrm{exec}} = 0$. In general,*

$$g_{\mathrm{exec}} \le \frac{1}{8\varepsilon} \sum_{t,z} d_t^E(z)\,\mathrm{osc}_a(D_t(z,\cdot))^2. \tag{21}$$

An exact action-variance representation. For $s \in [0,1]$, interpolate $\pi_{t,s}(a \mid z) \propto \pi_t^F(a \mid z) e^{-sD_t(z,a)/\varepsilon}$.

Theorem 5 (Finite-tilt execution loss)**.**

$$g_{\mathrm{exec}} = \frac{1}{\varepsilon} \sum_{t,z} d_t^E(z) \int_0^1 (1-s)\,\mathrm{Var}_{A\sim\pi_{t,s}(\cdot|z)}[D_t(z,A)]\,ds. \tag{22}$$

The identity uses standard log-partition differentiation (Wainwright and Jordan, 2008); a proof appears in Appendix C.2. Adding an action-independent constant to D_t changes neither the integrand nor the policy. Thus the relevant quantity is not the mean transition optimism, but variation of *total future* optimism across choices.

3.1 Second-order optimism, fourth-order execution loss

Fix $(P, \pi_0, \mu_0, \varepsilon)$ and scale all costs by λ: $c_t^\lambda = \lambda c_t$, $h^\lambda = \lambda h$. Define the unregularized value under the reference policy and an environmental variance-to-go by

$$U_T = h, \qquad U_t(z) = \mathbb{E}_{\pi_{0,t}}[c_t(z,A) + P_t U_{t+1}(z,A)], \tag{23}$$

$$B_T = 0, \quad M_t(z,a) = \mathrm{Var}_{P_t(\cdot|z,a)}(U_{t+1}) + P_t B_{t+1}(z,a), \quad B_t(z) = \mathbb{E}_{\pi_{0,t}} M_t(z,A). \tag{24}$$

We use $O(\lambda^k)$ for a remainder bounded by a constant times $|\lambda|^k$ near zero, with the finite model and ε fixed. Let d_t^0 be the occupancy under p_0. Unlike the variance of the whole return, B_t accumulates *environmental* continuation variance; action randomness enters the next theorem through variation of M_t across actions.

Theorem 6 (Full-MDP weak-tilt separation)**.** *Under the finite-state assumptions above, as* $\lambda \to 0$*,*

$$g^{\lambda}_{\mathrm{relax}} = \frac{\lambda^2}{2\varepsilon}\mathbb{E}_{\mu_0} B_0 + O(\lambda^3), \tag{25}$$

$$g^{\lambda}_{\mathrm{exec}} = \frac{\lambda^4}{8\varepsilon^3}\sum_{t,z} d^0_t(z)\,\mathrm{Var}_{\pi_{0,t}(\cdot|z)}[M_t(z,A)] + O(\lambda^5). \tag{26}$$

Uniformly over admissible state–action cells,

$$\pi^{C,\lambda}_t(a \mid z) - \pi^{F,\lambda}_t(a \mid z) = -\frac{\lambda^2}{2\varepsilon^2}\pi_{0,t}(a \mid z)[M_t(z,a) - B_t(z)] + O(\lambda^3). \tag{27}$$

Both coefficients are nonnegative and may vanish. If the coefficient in (26) vanishes, then $g^{\lambda}_{\mathrm{exec}} = O(\lambda^6)$*.*

Why the orders differ. Backward expansion gives $\Gamma^{\lambda}_t = \lambda^2\,\mathrm{Var}_{P_t}(U_{t+1})/(2\varepsilon) + O(\lambda^3)$, $D^{\lambda}_t = \lambda^2 M_t/(2\varepsilon) + O(\lambda^3)$, and $\Delta^{\lambda}_t = \lambda^2 B_t/(2\varepsilon) + O(\lambda^3)$. Transition optimism is therefore $O(\lambda^2)$, while execution loss is quadratic in its action-dependent part by (22). The proof tracks the dependence of values, policies, and occupancies on λ; it does not hold them fixed (Appendix C.3). If $\mathbb{E}B_0 > 0$, the ratio $g^{\lambda}_{\mathrm{exec}}/g^{\lambda}_{\mathrm{relax}}$ is $O(\lambda^2)$. The equivalent high-temperature orders are $O(\varepsilon^{-1})$ and $O(\varepsilon^{-3})$ (Corollary 14). Both the fourth-order and degenerate sixth-order powers are attained by explicit analytical examples (Appendix C.3). These are local asymptotics, not a claim that every finite-temperature gap is small or that the existing experiments establish either rate.

4 Learned Dynamics: Compare Before Correcting

A correction is optimal for its fitted kernel, not necessarily for the unknown physical kernel. We write $\widehat{P}$ for the fitted kernel and $\widehat{\pi}^F, \widehat{\pi}^C$ for its free and corrected policies. For a kernel K, $J_K(\alpha)$ is the objective of policy α with K used in both its physical and reference path laws. Thus $J^C = J_P(\pi^C)$ and $J^E = J_P(\pi^F)$, whereas J^F is the free path value. The benefit of correction can fail arbitrarily close to a decision tie. In the one-step, uniform-prior, two-action model with terminal costs $0, 1$, let both true bad-state probabilities be $\theta \in (0,1)$ and the fitted probabilities be $\theta - \xi, \theta + \xi$, where $|\xi| < \min\{\theta, 1-\theta\}$. With $G_\varepsilon(x) = -\varepsilon\log(1 - x + xe^{-1/\varepsilon})$, the Bernoulli instance of G_t, Proposition 11 gives

$$J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^C) = \frac{G'_\varepsilon(\theta)^2 - 1}{2\varepsilon}\xi^2 + O(\xi^4). \tag{28}$$

The sign changes at $\theta_* = (1 - e^{-1/\varepsilon})^{-1} - \varepsilon \in (0,1)$. Thus arbitrarily small error can favor either candidate. This is not a claim of large harm from small error: both policies have identical physical expected cost θ here, and the difference is an $O(\xi^2)$ action-regularization cost. The full finite-error formula and proof are in Appendix B.

A deterministic transfer identity yields a statistical selector. For a fixed Markov policy α, let $\rho^{P,\alpha}_t(z,a)$ be its state–action occupancy under P. Its fitted-model value satisfies

$$\widehat{U}^{\alpha}_T = h, \qquad \widehat{U}^{\alpha}_t(z) = \mathbb{E}_{a\sim\alpha_t}\left[c_t(z,a) + \varepsilon\log\frac{\alpha_t(a \mid z)}{\pi_{0,t}(a \mid z)} + \widehat{P}_t\widehat{U}^{\alpha}_{t+1}(z,a)\right]. \tag{29}$$

Use $0\log 0 = 0$. Suppose simultaneous row bounds $\sup_{z,a}\mathrm{TV}(P_t(\cdot \mid z,a), \widehat{P}_t(\cdot \mid z,a)) \le \delta_t$ hold; the radii δ_t need not equal the unknown errors.

Theorem 7 (Exact model transfer and a deployment interval)**.** *For every fixed Markov policy* α*,*

$$J_P(\alpha) - J_{\widehat{P}}(\alpha) = \sum_{t<T}\mathbb{E}_{\rho^{P,\alpha}_t}[(P_t - \widehat{P}_t)\widehat{U}^{\alpha}_{t+1}], \tag{30}$$

$$|J_P(\alpha) - J_{\widehat{P}}(\alpha)| \le R_\alpha := \sum_{t<T}\delta_t\,\mathrm{osc}(\widehat{U}^{\alpha}_{t+1}). \tag{31}$$

In particular, for $\widehat{g} = J_{\widehat{P}}(\widehat{\pi}^F) - J_{\widehat{P}}(\widehat{\pi}^C)$*,*

$$J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^C) \in [\widehat{g} - R_F - R_C, \widehat{g} + R_F + R_C], \quad R_X = R_{\widehat{\pi}^X}. \tag{32}$$

This is a policy-evaluation telescoping identity of the simulation-lemma type (Kearns and Singh, 2002), specialized to a comparison of the two fitted policies. Set $\widehat{L} = \widehat{g} - R_F - R_C$. Deploy $\widehat{\pi}^C$ only if $\widehat{L} > 0$; otherwise retain $\widehat{\pi}^F$. Denote the selected policy by $\widehat{\pi}^{\mathrm{sel}}$.

Corollary 8 (Certified no-harm correction)**.** *On the simultaneous row-confidence event,*

$$J_P(\widehat{\pi}^{\mathrm{sel}}) \le J_P(\widehat{\pi}^F), \qquad J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^{\mathrm{sel}}) \ge [\widehat{L}]_+. \tag{33}$$

The implication holds for all candidate policies on that same event, including policies fitted and selected using the data that define it; independence from the confidence set is unnecessary. A $1-\eta$ uniform confidence event thus gives, for $\eta \in (0,1)$, a $1-\eta$ no-harm guarantee for the *regularized objective*, not for task cost or safety metrics. Exact finite-sample radii appear in Proposition 22.

The sum of two policy radii is sufficient, not necessary. Because the policies share the same unknown kernel, their transfer errors can cancel. Proposition 15 gives a computable paired bound that vanishes for identical policies and can be intersected with (32). In the one-step Bernoulli setting, let $\theta_a, \widehat{\theta}_a$ be the true and fitted bad-state probabilities for action $a \in \{0,1\}$, let $u_X = \widehat{\pi}^X(0)$, and set $g_P = J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^C)$. A common radius $r \ge 0$ with $|\theta_a - \widehat{\theta}_a| \le r$ gives the exact contrast and a sharp interior-case interval:

$$g_P - \widehat{g} = (u_F - u_C)[(\theta_0 - \widehat{\theta}_0) - (\theta_1 - \widehat{\theta}_1)], \qquad |g_P - \widehat{g}| \le 2r|u_F - u_C|. \tag{34}$$

Here sharpness means both endpoints are attainable when each fitted probability lies in $[r, 1-r]$. For the suite's two-action, uniform-prior, zero-stage-cost, binary-loss candidates, G_ε is increasing, so u_F, u_C lie on the same side of $1/2$. Thus $|u_F - u_C| \le 1/2$ and $|g_P - \widehat{g}| \le r$ (Corollary 16). This certifies the evaluated half-radius rule: writing ϑ for the multiplier of $2r$, $\vartheta = 0.5$ is sufficient here, not in arbitrary MDPs. The policy-dependent paired rule remains unevaluated.

5 Evidence: Exact Mechanisms and a Statistical Boundary

The evidence supports a theory-of-control paper, not a large-model benchmark claim. Two source-backed studies contain executable rerun code, frozen aggregates, and verification records: a random finite-MDP suite and a finite-sample Bernoulli boundary suite. Archived cliff, inventory, learned-task, and rollout-estimation figures provide task-level illustrations, but lack their original code and row-level records. These are simulated tabular control studies, not real-world datasets or deployments. We preserve that distinction throughout; all recorded outcomes remain unchanged.

Does non-executability predict loss? Not by itself. The random suite crosses three transition families, three state counts, two action counts, four horizons, four temperatures, four Dirichlet concentrations, and three cost scales, with 256 instances per valid cell (Appendix G). Across 663,552 instances in 2,592 cells, the maximum three-law or correction residual is 6.47×10^{-14}. Deterministic $|g_{\mathrm{total}}|$ is at most 6.39×10^{-14}. Common-mode $|g_{\mathrm{exec}}|$ is at most 6.39×10^{-14} although its median relaxation gap is 0.350. These exact null cases distinguish optimism from action distortion. For generic MDPs, the median raw execution gap is 1.29×10^{-3}, the median normalized gap is 1.56×10^{-4}, and the normalized 95th percentile is 0.0110; 27.64% exceed the configured 10^{-3} threshold. Loss is therefore often small, not typically catastrophic.

Does the correction need its continuation term? Yes. The ablation uses $\pi^{\mathrm{imm}} \propto \pi^F e^{-\Gamma/\varepsilon}$, omitting $P\Delta_{t+1}$. Its excess objective exceeds the numerical tolerance 10^{-10} in 99.20% of generic instances. This is numerical detectability, not a large-effect threshold: the mean excess is 0.0185, with maximum 3.282. Execution loss and occupancy-weighted correction oscillation have Spearman correlation 0.977 over individual generic instances. These results support propagation and the finite-tilt mechanism. Neither this correlation nor the small median tests the weak-tilt expansion: the fixed grid was not a cost-scaling study.

Does learning preserve the benefit? No; certification trades power for protection. The boundary grid crosses three true base probabilities, eight action margins, five temperatures, thirteen sample sizes, and 8,192 repetitions per cell: 12,779,520 trials in 1,560 cells. Both beneficial and harmful rates

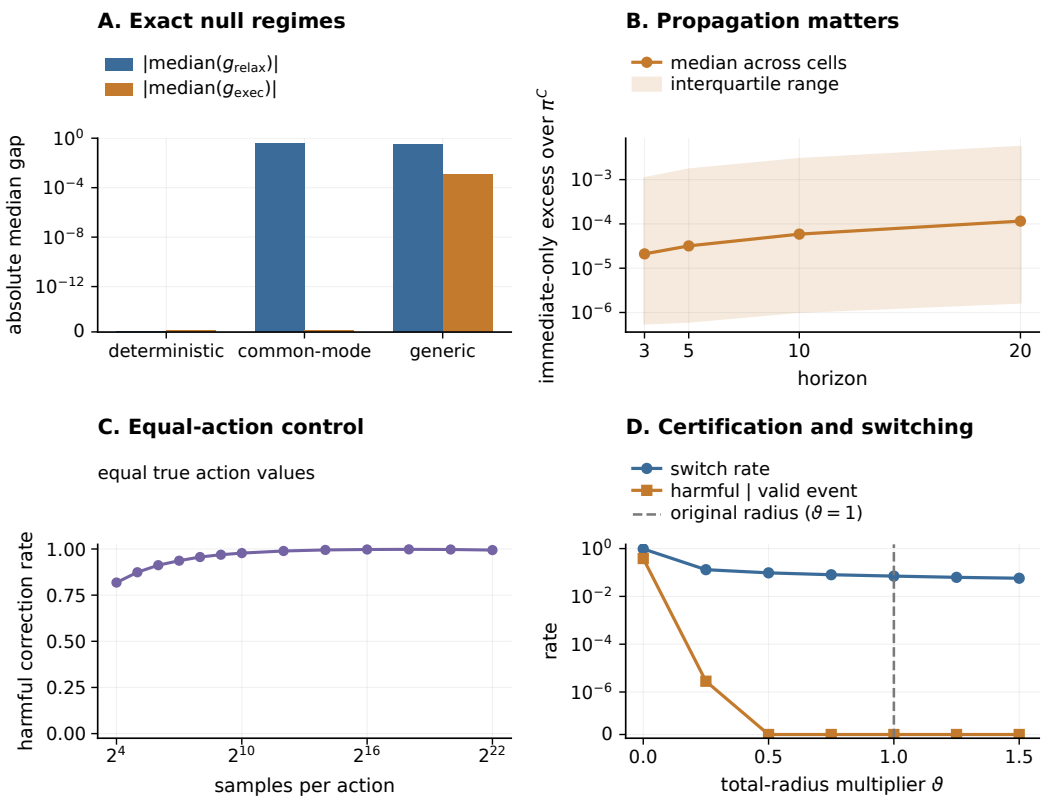


Figure 2: **Known-model exactness does not imply learned-model dominance. A:** deterministic dynamics have zero total gap; common-mode stochastic dynamics have zero execution gap despite optimism. **B:** immediate-only correction remains suboptimal across horizons; points are medians of generic cell-mean excesses, with interquartile bands across cells. **C:** the equal-action negative control frequently makes fitted correction worse in regularized objective, not physical task cost. **D:** increasing the confidence radius reduces switching. Total-radius multipliers $\vartheta = 1$ and $\vartheta = 0.5$ are certified in this Bernoulli setting and have no recorded harmful switch on a valid event (Corollary 16). Switching is unconditional; harm is conditioned on a valid confidence event.

exceed 1% in 712 cells. Always correcting is harmful in 38.46% overall, with maximum cellwise rate 99.98%. In the $1{,}597{,}440$ equal-action trials, the harm rate is 95.50%. These frequencies do not measure harm magnitude; in the equal-action control only the regularization cost changes. Across its frozen cells, mean always-correct regret relative to the better fitted candidate falls from 9.11×10^{-3} at 16 samples per action to 7.76×10^{-8} at $4{,}194{,}304$ samples per action. The equal-action subset also contains 814 beneficial trials. Its two estimates fluctuate independently, rather than symmetrically around the true probability as in Proposition 11. Harm and benefit use a fixed 10^{-12} objective tolerance.

The simultaneous confidence event holds in 99.70% of trials, and interval coverage is 100.00%. The evaluated selector switches in 7.06%, with zero harmful switches on a valid event. Its mean oracle regret is 0.002864, versus 0.001265 for always correcting and 0.004711 for always extracting: protection is not dominance in average regret. The plug-in rule switches in 97.01% and produces $4{,}914{,}851$ harmful deployments. A quarter-radius rule produces 36 harmful switches on valid events. The half-radius rule, certified here by Corollary 16, switches in 9.65% with mean oracle regret 0.002418 and zero recorded harmful switches on valid events. It is the same strategy as the "one-policy" rule, not an independent baseline. Appendix G.6 retains all unique strategies and descriptive paired comparisons. These existing outcomes do not evaluate the policy-dependent paired selector in (34).

Archived task-level illustrations. Appendix G retains the supplied cliff, inventory, learned-task, and rollout summaries separately from the source-backed studies. Their original experiment code, row-level records, and complete fit accounting are unavailable. They are not used to establish the quantitative claims above.

6 Positioning, Limitations, and Conclusion

What is established, and what is added. Path-integral and KL control exploit special relationships between control and noise (Kappen, 2005; Todorov, 2006); maximum causal entropy and regularized MDPs retain physical dynamics (Ziebart et al., 2010; Neu et al., 2017; Geist et al., 2019). Probabilistic control and inference already relate desired path densities to feasible policies (Kárný and Guy, 2006; Rawlik et al., 2012; Lefebvre, 2024). Arriojas et al. (2023) explicitly execute optimistic policies in the physical environment, while Bhole et al. (2025) separate policy and transition regularization in a broader risk-sensitive formulation. Action-constrained and trajectory-regularized formulations address other feasibility structures (Bušić and Meyn, 2018; Ito and Kashima, 2022; Kim and Sreenath, 2026). Residual value correction with a prior optimal policy is established by Adamczyk et al. (2023). We specialize it to a Gibbs target's transition optimism. Our additions concern propagated action variation, weak-tilt coefficients, and learned-policy comparison, not new projection or control solvers (Table 1).

In smooth, unconstrained settings, related second-order control errors and fourth-order value losses arise in nominal model predictive control under small process noise (Messerer et al., 2024) and in model predictive path integral control for deterministic systems under jointly shrinking exploration covariance and temperature (Homburger et al., 2025). Theorem 6 instead scales costs at fixed stochastic dynamics and temperature, identifying explicit propagated environmental-variance coefficients and a degenerate sixth-order case.

Scope and limitations. The exact theory assumes finite horizon, bounded costs, full prior support, and known tabular dynamics; the model-transfer results extend to standard-Borel spaces under Corollary 21. No tractable approximation guarantees are established for large neural policies. Learned-model guarantees require simultaneous uniform transition confidence sets, which can be expensive and conservative. Unlike robust planning (Iyengar, 2005; Nilim and El Ghaoui, 2005), the selector compares two fixed fitted policies rather than optimizing over the uncertainty set. Robust baseline-regret comparison already exploits shared model uncertainty to reduce conservatism (Petrik et al., 2016); here we derive an explicit finite-horizon transfer contrast for the learned Gibbs/executable pair. The paired refinement reduces avoidable cancellation loss, but has no empirical switching results here. Near a true tie, a small model error can reverse the sign without causing a large objective loss. Neither regularized no-harm nor the archived task outcomes prove deployment safety. The source-backed experiments are finite synthetic control studies; the task-level rerun evidence is incomplete.

Conclusion. An optimistic path distribution may be infeasible yet induce optimal actions. The auxiliary correction identifies the decision-relevant component: variation across actions in future optimism. Its variance representation explains the second- and fourth-order bounds under weak cost scaling. With learned dynamics, correction becomes a certified policy comparison rather than an unconditional improvement.

Reproducibility statement. The supplement contains proofs and archived illustrations. The code package supplies frozen aggregates and rerun source for the random-MDP and boundary suites (Appendix H), but lacks original task-level raw records and rerun code. New analytical results introduce no empirical outcomes. Archived verification reports are distinguished from checks reproducible directly from the supplied aggregates.

Artificial Intelligence (AI) Use Statement.

The authors proposed the research idea and developed the core mathematical claims and proof strategy. Generative AI tools were used to support research execution, including coding, debugging, and generating synthetic datasets; to suggest or refine intermediate proof steps; to support the literature

survey through summaries of author-supplied sources; and to polish or edit portions of the manuscript. All AI-assisted work was reviewed by the authors. In particular, all AI-assisted proof suggestions were independently checked by the authors and AI-assisted code and synthetic-data-generation procedures were reviewed and tested by the authors. The authors take responsibility for all aspects of the final work, including all AI-assisted content.

REFERENCES

Jacob Adamczyk, Argenis Arriojas, Stas Tiomkin, and Rahul V. Kulkarni. Utilizing prior solutions for reward shaping and composition in entropy-regularized reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6658–6665, 2023. doi:10.1609/aaai.v37i6.25817.

Argenis Arriojas, Jacob Adamczyk, Stas Tiomkin, and Rahul V. Kulkarni. Bayesian inference approach for entropy regularized reinforcement learning with stochastic dynamics. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of PMLR, pp. 99–109, 2023.

Ajinkya Bhole, Mohammad Mahmoudi Filabadi, Guillaume Crevecoeur, and Tom Lefebvre. Unifying entropy regularization in optimal control: From and back to classical objectives via iterated soft policies and path integral solutions. arXiv:2512.06109, 2025.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 9780199535255.

Ana Bušić and Sean Meyn. Action-constrained Markov decision processes with Kullback–Leibler cost. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of PMLR, pp. 1431–1444, 2018.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi:10.1214/aop/1176996454.

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 PMLR, pp. 2160–2169, 2019.

Hannes Homburger, Florian Messerer, Moritz Diehl, and Johannes Reuter. Optimality and suboptimality of MPPI control in stochastic and deterministic settings. *IEEE Control Systems Letters*, 9:763–768, 2025. doi:10.1109/LCSYS.2025.3574151.

Kaito Ito and Kenji Kashima. Kullback–Leibler control for discrete-time nonlinear systems on continuous spaces. *SICE Journal of Control, Measurement, and System Integration*, 15(2):119–129, 2022. doi:10.1080/18824889.2022.2095827.

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

Hilbert J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, 2005. doi:10.1088/1742-5468/2005/11/P11011.

Miroslav Kárný and Tatiana V. Guy. Fully probabilistic control design. *Systems & Control Letters*, 55(4):259–265, 2006. doi:10.1016/j.sysconle.2005.08.001.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002. doi:10.1023/A:1017984413808.

Mintae Kim and Koushil Sreenath. Trajectory-regularized stochastic optimal control via KL divergence. arXiv:2607.22201, 2026.

Tom Lefebvre. Probabilistic control and majorisation of optimal control. *Systems & Control Letters*, 190:105837, 2024. doi:10.1016/j.sysconle.2024.105837.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. arXiv:1805.00909, 2018.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Florian Messerer, Katrin Baumgärtner, Sergio Lucia, and Moritz Diehl. Fourth-order suboptimality of nominal model predictive control in the presence of uncertainty. *IEEE Control Systems Letters*, 8:508–513, 2024. doi:10.1109/LCSYS.2024.3396611.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. arXiv:1705.07798, 2017.

Gerhard Neumann. Variational inference for policy search in changing situations. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 817–824, 2011.

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

Marek Petrik, Mohammad Ghavamzadeh, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems 29*, pp. 2298–2306, 2016.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Robotics: Science and Systems VIII*, 2012. doi:10.15607/RSS.2012.VIII.045.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of PMLR, pp. 2380–2388, 2015.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems 19*, pp. 1369–1376, 2006.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi:10.1561/2200000001.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Hewlett–Packard Laboratories Technical Report HPL-2003-97(R.1), 2003.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1255–1262, 2010.

Paul H. Zipkin. *Foundations of Inventory Management*. McGraw–Hill, 2000. ISBN 9780256113792.

A Foundational Variational Identities

These standard identities are included with proofs to fix conditioning, support, and the admissible history-dependent policy class. The forward and reverse projection viewpoints build on prior work (Neumann, 2011; Rawlik et al., 2012; Lefebvre, 2024); we do not claim them as new. All logarithms are natural, with $0\log(0/q) = 0$ for $q > 0$. We use $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$ in the finite setting and $\mathrm{osc}(f) = \sup f - \inf f$; a subscript on osc specifies the variable. Conditional laws at zero-probability initial states can be assigned using the displayed backward recursions without changing any path law.

Notation guide. Superscripts F, C, E refer to the free path target, physical optimum, and physical extraction, respectively. The following symbols retain the same meaning in the main text and proofs.

Symbol	Meaning and defining location
p_0, π_0, P, μ_0	Reference path law, action prior, physical kernel, initial law; (1).
q^F, q^C, q^E	Free, optimal physical, and extracted path laws; (3).
$\mathcal{F}(q), J^X$	Regularized objective and its value at q^X; (2).
d_t^X, ρ_t^X	State and state–action occupancies; Section 2.
W_t, V_t, Γ_t	Free value, physical value, and transition optimism; (4)–(10).
Δ_t, D_t	Correction value and action value; (16).
$\lambda;\ U_t, B_t, M_t$	Weak-cost scale; prior value and variance-to-go quantities; (23)–(24).
$\widehat{P}, \widehat{\pi}^X$	Fitted kernel and its candidate policies; Section 4.
$\theta_a, \widehat{\theta}_a;\ \xi$	True/fitted bad-state probabilities; symmetric fit perturbation; Section 4.
$\delta_t,\ r;\ \eta$	Row total-variation radii; confidence failure probability; Section 4.
$R_\alpha, R^{\mathrm{pair}}_{\alpha,\beta}$	Single-policy and paired objective-transfer radii; (31), (55).
ϑ	Multiplier of the original total radius $2r$; Corollary 16.
$K, \Phi_t, \Lambda(s)$	Generic transition kernel, local free action value, log-partition function; Sections 4, B, C.2.
$m_{\mathrm{cat}}, \zeta, \omega$	Category count, smoothing pseudo-count, and mixture weight; Proposition 22.
$S_C, S_E;\ \overline{S}_C, \overline{S}_E$	Physical path scores and their upper bounds; Proposition 23.

Theorem 9 (Executable variational identity)**.** *For every history-dependent physical policy α,*

$$\mathcal{F}(q^\alpha) - \mathbb{E}_{\mu_0} V_0 = \varepsilon \mathbb{E}_{q^\alpha} \sum_{t=0}^{T-1} \mathrm{KL}\big(\alpha_t(\cdot \mid H_t) \| \pi_t^C(\cdot \mid Z_t)\big) = \varepsilon\, \mathrm{KL}(q^\alpha \| q^C). \tag{35}$$

Thus q^C is the unique optimal path law over $\mathcal{Q}_P^{\mathrm{hist}}$, $J^C := \mathcal{F}(q^C) = \mathbb{E}_{\mu_0} V_0$, and history dependence cannot improve on π^C.

Lemma 10 (Free law and local transition optimism)**.** *The unique minimizer over $\mathcal{Q}_\mu$ is*

$$q^F(\tau) = \mu_0(z_0) \frac{p_0(\tau \mid z_0) e^{-\ell(\tau)/\varepsilon}}{\mathbb{E}_{p_0(\cdot|z_0)} e^{-\ell/\varepsilon}} = \mu_0(z_0) \prod_t \pi_t^F(a_t \mid z_t) P_t^F(z_{t+1} \mid z_t, a_t), \tag{36}$$

with $J^F := \mathcal{F}(q^F) = \mathbb{E}_{\mu_0} W_0$. For every $q \in \mathcal{Q}_\mu$,

$$\mathcal{F}(q) - J^F = \varepsilon\, \mathrm{KL}(q \| q^F). \tag{37}$$

Moreover,

$$\Gamma_t(z, a) := \mathbb{E}_{P_t} W_{t+1} - G_t(z, a) = \varepsilon\, \mathrm{KL}(P_t(\cdot \mid z, a) \| P_t^F(\cdot \mid z, a)) \geq 0. \tag{38}$$

B Proofs of the Main Results

B.1 Proof of Theorem 9

Proof. Fix a history H_t ending at $Z_t = z$ and an arbitrary distribution ν on $\mathcal{A}_t(z)$. Substituting Equation (5) into the KL divergence gives the local identity

$$\varepsilon\, \mathrm{KL}(\nu \| \pi_t^C(\cdot \mid z)) = \varepsilon\, \mathrm{KL}(\nu \| \pi_{0,t}(\cdot \mid z)) + \sum_a \nu(a) Q_t^C(z, a) - V_t(z). \tag{39}$$

Apply this identity with $\nu = \alpha_t(\cdot \mid H_t)$, take expectation under q^α, and sum over t. Because q^α uses the physical kernel,

$$\mathbb{E}_{q^\alpha} \sum_{t=0}^{T-1} \left[c_t(Z_t, A_t) + \mathbb{E}_{P_t(\cdot|Z_t,A_t)} V_{t+1} - V_t(Z_t) \right] = \mathbb{E}_{q^\alpha} \left[\sum_{t=0}^{T-1} c_t(Z_t, A_t) + h(Z_T) \right] - \mathbb{E}_{\mu_0} V_0.$$

The KL chain rule for laws sharing μ_0 and P yields

$$\mathrm{KL}(q^\alpha \| p_0) = \mathbb{E}_{q^\alpha} \sum_{t=0}^{T-1} \mathrm{KL}(\alpha_t(\cdot \mid H_t) \| \pi_{0,t}(\cdot \mid Z_t)).$$

Combining these displays proves the first equality in Equation (35). Applying the same chain rule to q^α and q^C proves the second. The right-hand side is nonnegative and vanishes exactly when $\alpha_t(\cdot \mid H_t) = \pi_t^C(\cdot \mid Z_t)$ on every occupied history. This proves optimality and uniqueness of the induced path law. □

B.2 Proof of Lemma 10

Proof. Condition on $Z_0 = z_0$. The Gibbs variational formula gives the unique conditional minimizer

$$q^F(\tau \mid z_0) = \frac{p_0(\tau \mid z_0) e^{-\ell(\tau)/\varepsilon}}{\mathbb{E}_{p_0(\cdot|z_0)} e^{-\ell/\varepsilon}}.$$

The recursion in Equations (6)–(7) implies $\mathbb{E}_{p_0(\cdot|Z_t=z)} e^{-\ell_{t:T}/\varepsilon} = e^{-W_t(z)/\varepsilon}$, where $\ell_{t:T} = \sum_{s=t}^{T-1} c_s(Z_s, A_s) + h(Z_T)$. Moreover,

$$\pi_t^F(a \mid z) P_t^F(z' \mid z, a) = \pi_{0,t}(a \mid z) P_t(z' \mid z, a) \exp\{-[c_t(z, a) + W_{t+1}(z') - W_t(z)]/\varepsilon\}.$$

Multiplying over time telescopes the exponent and proves the factorization in Equation (36).

For any $q \in \mathcal{Q}_\mu$,

$$\begin{aligned} \varepsilon \,\mathrm{KL}(q \| q^F) &= \varepsilon \,\mathrm{KL}(q \| p_0) + \mathbb{E}_q \ell - \mathbb{E}_q W_0(Z_0) \\ &= \mathcal{F}(q) - \mathbb{E}_{\mu_0} W_0, \end{aligned}$$

which establishes Equation (37) and the value J^F. Finally, Equation (9) gives, on the support of P_t,

$$\log \frac{P_t(z' \mid z, a)}{P_t^F(z' \mid z, a)} = \frac{W_{t+1}(z') - G_t(z, a)}{\varepsilon}.$$

Taking expectation under $P_t(\cdot \mid z, a)$ proves Equation (10). □

B.3 Proof of Theorem 1

Proof. For $q = q^\alpha \in \mathcal{Q}_P^{\mathrm{hist}}$, the KL chain rule gives

$$\begin{aligned} \varepsilon \,\mathrm{KL}(q \| q^F) &= \varepsilon \mathbb{E}_q \sum_t \mathrm{KL}(\alpha_t(\cdot \mid H_t) \| \pi_t^F(\cdot \mid Z_t)) + \mathbb{E}_q \sum_t \Gamma_t(Z_t, A_t) \\ &= \varepsilon \,\mathrm{KL}(q \| q^E) + \mathbb{E}_q \sum_t \Gamma_t(Z_t, A_t). \end{aligned}$$

Combining with Lemma 10 proves Equation (11). The auxiliary problem therefore has the same minimizing path law as the original physical control problem. Its Bellman recursion has stage cost Γ, prior π^F, terminal cost zero, and value $\Delta = V - W$.

For any $q \in \mathcal{Q}_P^{\mathrm{hist}}$, Theorem 9 and Lemma 10 give

$$\begin{aligned} \varepsilon \,\mathrm{KL}(q \| q^F) &= \mathcal{F}(q) - J^F \\ &= \mathcal{F}(q) - J^C + J^C - J^F \\ &= \varepsilon \,\mathrm{KL}(q \| q^C) + \varepsilon \,\mathrm{KL}(q^C \| q^F), \end{aligned}$$

where the last equality uses Lemma 10 with $q = q^C$. This proves Equation (12). Setting $q = q^E$ gives $g_{\text{total}} = g_{\text{relax}} + g_{\text{exec}}$ and the three KL value identities.

The chain rule between the Markov factorizations of q^C and q^F yields

$$\varepsilon \,\mathrm{KL}(q^C \| q^F) = \varepsilon \sum_{t,z} d_t^C(z)\,\mathrm{KL}(\pi_t^C(\cdot \mid z) \| \pi_t^F(\cdot \mid z)) + \varepsilon \sum_{t,z,a} \rho_t^C(z,a)\,\mathrm{KL}(P_t(\cdot \mid z,a) \| P_t^F(\cdot \mid z,a)).$$

Substituting Equation (10) proves Equation (13). The laws q^E and q^C share P and differ only in their action conditionals, which proves Equation (14). The laws q^E and q^F share π^F and differ only in their transition conditionals, which proves Equation (15).

It remains to characterize the zero-relaxation case. By the KL identity, $g_{\text{relax}} = 0$ if and only if $q^C = q^F$. Equality of path laws implies $P_t^F = P_t$ on every q^C-occupied state–action cell, and Equation (10) then gives $\Gamma_t = 0$ there. Conversely, suppose $\Gamma_t = 0$ on every such cell. Because the reference policy has full support and all costs are finite, π^C has full support at every occupied state. We prove backward that $V_t = W_t$ and $\pi_t^C = \pi_t^F$ on every q^C-occupied state. The claim is immediate at $t = T$. If it holds at $t+1$, every successor in the support of an occupied (t, z, a) is occupied at $t+1$, so $V_{t+1} = W_{t+1}$ there. Since $\Gamma_t(z,a) = 0$,

$$G_t(z,a) = \mathbb{E}_{P_t} W_{t+1} = \mathbb{E}_{P_t} V_{t+1}$$

for every admissible action at the occupied state. The action logits in Equations (5) and (7) therefore agree, as do V_t and W_t. Induction gives $q^C = q^F$. Finally, Equation (10) shows that $\Gamma_t = 0$ is equivalent to $P_t^F = P_t$, and equality in Jensen's inequality for the strictly convex exponential is equivalent to W_{t+1} being constant on the corresponding transition support. □

B.4 Proof of Theorem 2

Proof. Let $\Phi_t(z,a) = c_t(z,a) + G_t(z,a)$. Equations (4), (10), and (16) imply

$$Q_t^C(z,a) = \Phi_t(z,a) + D_t(z,a).$$

Factoring the free normalizer out of the executable soft minimum gives

$$\begin{aligned} V_t(z) &= -\varepsilon \log \sum_a \pi_{0,t}(a \mid z) e^{-[\Phi_t(z,a) + D_t(z,a)]/\varepsilon} \\ &= W_t(z) - \varepsilon \log \sum_a \pi_t^F(a \mid z) e^{-D_t(z,a)/\varepsilon}, \end{aligned}$$

which is Equation (17). Dividing Equation (5) by Equation (8) and using $\Delta_t = V_t - W_t$ proves Equation (18). Taking the negative log ratio in expectation under π_t^F proves Equation (19).

Both policies have full support. Hence $\pi_t^C(\cdot \mid z) = \pi_t^F(\cdot \mid z)$ if and only if $D_t(z,a) - \Delta_t(z) = 0$ for every admissible action, equivalently if and only if $D_t(z,\cdot)$ is constant. For the constructive statement, use backward induction. The initialization $\Delta_T = 0 = V_T - W_T$ is exact. If $\Delta_{t+1} = V_{t+1} - W_{t+1}$, then Equation (16) gives $\Phi_t + D_t = Q_t^C$; the preceding soft-minimum identity returns $\Delta_t = V_t - W_t$, and Equation (18) returns π_t^C. □

B.5 Proof of Corollary 3

Proof. Substituting Equation (19) into Equation (14) proves Equation (20). Every summand is a nonnegative KL divergence. The sum therefore vanishes exactly when $\pi_t^F(\cdot \mid z) = \pi_t^C(\cdot \mid z)$ at every state with $d_t^E(z) > 0$. By Theorem 2, full action support makes this equivalent to $D_t(z,\cdot)$ being constant over the admissible actions at those states. Since the two laws then use the same action and transition conditionals on every history reached by q^E, induction over time gives $q^E = q^C$. The condition constrains only action-differential corrections, so it does not force $P_t^F = P_t$ and does not force $g_{\text{relax}} = 0$. □

B.6 Proof of Corollary 4

Proof. Suppose that at every reachable (t, z) the row $P_t(\cdot \mid z, a) = P_t(\cdot \mid z)$ is independent of a. Then $\mathbb{E}_{P_t} V_{t+1}$ is a common additive term in all executable action logits at (t, z), while $G_t(z, a) = -\varepsilon \log \mathbb{E}_{P_t(\cdot|z)} e^{-W_{t+1}/\varepsilon}$ is likewise independent of a in the free logits. Equations (5) and (7) therefore both reduce to the same normalization of $\pi_{0,t}(a \mid z) e^{-c_t(z,a)/\varepsilon}$. Hence $\pi_t^F(\cdot \mid z) = \pi_t^C(\cdot \mid z)$ on every reachable state. The laws q^E and q^C then share both action and transition conditionals, so $q^E = q^C$ and $g_{\text{exec}} = 0$. Positive g_{relax} remains possible because the free law can still tilt the common stochastic row; the one-action example in Corollary 20 is a special case.

For the general bound, Equation (19) is the Jensen gap of the bounded random variable $D_t(z, A)$ under $A \sim \pi_t^F(\cdot \mid z)$. Hoeffding's lemma applied to $-D_t(z, A)/\varepsilon$ gives

$$\varepsilon \, \mathrm{KL}(\pi_t^F(\cdot \mid z) \| \pi_t^C(\cdot \mid z)) \le \frac{\mathrm{osc}_a(D_t(z, \cdot))^2}{8\varepsilon}.$$

Multiply by $d_t^E(z)$, sum over (t, z), and use Equation (14). □

B.7 Exact finite-error learned correction boundary

Proposition 11 (Learned correction has no unconditional dominance)**.** *Consider one decision, two actions, a uniform action prior, no stage cost, and terminal loss* 1 *in a bad state and* 0 *otherwise. Under the true kernel both actions have bad-state probability* $\theta \in (0, 1)$*. Under the learned kernel their probabilities are* $\widehat{\theta}_0 = \theta - \xi$ *and* $\widehat{\theta}_1 = \theta + \xi$*, where* $0 < \xi < \min\{\theta, 1 - \theta\}$*. Let* $\mathrm{Ber}(u)$ *denote the Bernoulli distribution with success probability* u*, and let* $\mathrm{sigmoid}(x) = (1 + e^{-x})^{-1}$*. Define*

$$G_\varepsilon(x) = -\varepsilon \log\left(1 - x + x e^{-1/\varepsilon}\right), \qquad k(u) = \mathrm{KL}(\mathrm{Ber}(u) \| \, \mathrm{Ber}(1/2)). \tag{40}$$

Then

$$\widehat{\pi}^C(0) = \mathrm{sigmoid}(2\xi/\varepsilon), \qquad \widehat{\pi}^F(0) = \mathrm{sigmoid}\left(\frac{G_\varepsilon(\theta + \xi) - G_\varepsilon(\theta - \xi)}{\varepsilon}\right), \tag{41}$$

$$J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^C) = \varepsilon\{k(\widehat{\pi}^F(0)) - k(\widehat{\pi}^C(0))\}. \tag{42}$$

Thus correction harms if and only if $G_\varepsilon(\theta + \xi) - G_\varepsilon(\theta - \xi) < 2\xi$ *and helps if and only if the inequality is reversed. With* $\kappa_\varepsilon(\theta) = G'_\varepsilon(\theta) = \varepsilon(1 - e^{-1/\varepsilon})/[1 - \theta(1 - e^{-1/\varepsilon})]$*,*

$$J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^C) = \frac{\kappa_\varepsilon(\theta)^2 - 1}{2\varepsilon} \xi^2 + O(\xi^4). \tag{43}$$

In particular, if $\theta < (1 - e^{-1/\varepsilon})^{-1} - \varepsilon$*, correction is harmful for every sufficiently small nonzero* ξ*; when the inequality is reversed, it is locally beneficial. Both signs therefore occur under arbitrarily small model error.*

Proof. For a Bernoulli transition with bad-state probability x, the one-step free continuation value is $G_\varepsilon(x) = -\varepsilon \log[(1 - x) + x e^{-1/\varepsilon}]$. Under the learned kernel, the executable soft policy is proportional to $\frac{1}{2} e^{-\widehat{\theta}_a/\varepsilon}$, so its odds in favor of action 0 are $e^{2\xi/\varepsilon}$. The free action conditional is proportional to $\frac{1}{2} e^{-G_\varepsilon(\widehat{\theta}_a)/\varepsilon}$, so its odds are $e^{[G_\varepsilon(\theta+\xi) - G_\varepsilon(\theta-\xi)]/\varepsilon}$. This proves Equation (41).

Under the true kernel, both actions incur expected terminal loss θ. Since the reference action law is uniform and the physical and reference transition rows agree, the trajectory KL reduces to the action KL. Thus, for any action probability u on action 0,

$$J_P(u) = \theta + \varepsilon \, \mathrm{KL}(\mathrm{Ber}(u) \| \, \mathrm{Ber}(1/2)) = \theta + \varepsilon k(u),$$

which proves Equation (42). Both displayed policy probabilities are at least $1/2$. Because $k'(u) = \log[u/(1 - u)] > 0$ for $u > 1/2$ and the sigmoid is strictly increasing, the sign is determined exactly by comparing the two logits. This gives the stated finite-ξ help/harm criterion.

For the local expansion, analyticity on $[\theta - \xi, \theta + \xi] \subset (0, 1)$ gives

$$G_\varepsilon(\theta + \xi) - G_\varepsilon(\theta - \xi) = 2\kappa_\varepsilon(\theta)\xi + O(\xi^3).$$

Using $\mathrm{sigmoid}(x) = \frac{1}{2} + x/4 + O(x^3)$ and $k(\frac{1}{2} + x) = 2x^2 + O(x^4)$ yields

$$\widehat{\pi}^C(0) = \frac{1}{2} + \frac{\xi}{2\varepsilon} + O(\xi^3),$$
$$\widehat{\pi}^F(0) = \frac{1}{2} + \frac{\kappa_\varepsilon(\theta)\xi}{2\varepsilon} + O(\xi^3),$$

and hence Equation (43). Writing $\gamma = 1 - e^{-1/\varepsilon}$, the inequality $\kappa_\varepsilon(\theta) < 1$ is equivalent to $\varepsilon\gamma < 1 - \theta\gamma$, or $\theta < \gamma^{-1} - \varepsilon$. Therefore every sufficiently small nonzero ξ makes correction harmful below this threshold and beneficial above it. The threshold lies strictly in $(0, 1)$: with $x = 1/\varepsilon > 0$, $1 - e^{-x} < x$ gives $(1 - e^{-x})^{-1} - 1/x > 0$, while $e^x - 1 > x$ gives $(1 - e^{-x})^{-1} - 1/x < 1$. The row-wise total-variation error is exactly ξ, so either sign occurs under arbitrarily small model error. □

B.8 Proof of Theorem 7

Proof. Write $\widehat{U} = \widehat{U}^\alpha$. Subtract the fixed-policy recursions under P and $\widehat{P}$, then telescope the nominal Bellman residual along a rollout generated by P and α. The stage-cost and action-KL terms cancel, leaving

$$J_P(\alpha) - J_{\widehat{P}}(\alpha) = \sum_{t=0}^{T-1} \mathbb{E}_{\rho_t^{P,\alpha}}[(P_t - \widehat{P}_t)\widehat{U}_{t+1}],$$

which proves Equation (30). For any bounded measurable f, $|(P - \widehat{P})f| \le \mathrm{TV}(P, \widehat{P})\,\mathrm{osc}(f)$. Applying this inequality rowwise proves Equation (31). Applying that bound separately to $\widehat{\pi}^F$ and $\widehat{\pi}^C$, then using the triangle inequality, gives Equation (32). □

B.9 Proof of Corollary 8

Proof. If $\widehat{L} \le 0$, the selector returns $\widehat{\pi}^F$, so both claims in Equation (33) hold with equality. If $\widehat{L} > 0$, the lower endpoint of Equation (32) gives

$$J_P(\widehat{\pi}^F) - J_P(\widehat{\pi}^C) \ge \widehat{L} > 0.$$

The selector then returns $\widehat{\pi}^C$, proving both inequalities. The probability statement follows by restricting this deterministic implication to the simultaneous transition-confidence event. □

C Action-Differential Optimism: Further Identities and Proofs

C.1 Reverse projection over history-dependent physical laws

For completeness, we extend the established Markov-policy reverse projection (Lefebvre, 2024) to the admissible history-dependent class. This is background for the comparison, not a new projection principle.

Proposition 12 (Reverse projection)**.** *For every $q^\alpha \in \mathcal{Q}_P^{\mathrm{hist}}$,*

$$\mathrm{KL}(q^F \| q^\alpha) = \mathrm{KL}(q^F \| q^E) + \mathbb{E}_{q^F} \sum_{t<T} \mathrm{KL}(\pi_t^F(\cdot \mid Z_t) \| \alpha_t(\cdot \mid H_t)). \tag{44}$$

Thus q^E is the unique minimizing physical path law.

Proof. The fixed-μ_0 Gibbs law is Markov, so its action conditional given H_t is $\pi_t^F(\cdot \mid Z_t)$. The KL chain rule separates the action divergences from $\mathbb{E}_{q^F} \sum_t \mathrm{KL}(P_t^F \| P_t)$; the latter is exactly $\mathrm{KL}(q^F \| q^E)$. If an action in the support of π^F receives zero probability under α, both sides hold in the extended-real sense. Otherwise the chain rule is finite. Its action terms vanish if and only if $\alpha_t = \pi_t^F$ on every q^F-occupied history. Bounded costs and a full-support prior imply that q^F and p_0 have identical path support; q^E has this support too. Equality therefore fixes every conditional needed to determine the physical law q^E. This proves uniqueness of the path law, not uniqueness of policies at unreachable histories. □

The expectation in (44) is under q^F, not q^E. In particular, its remainder is not generally $\mathrm{KL}(q^E \| q^\alpha)$. The forward identity (12) and reverse projection therefore have different occupancy geometry.

C.2 Proof of Theorem 5

Proof. Fix (t, z), write $p = \pi_t^F(\cdot \mid z)$ and $D(a) = D_t(z, a)$, and let

$$\Lambda(s) = \log \mathbb{E}_p e^{-sD(A)/\varepsilon}, \qquad p_s(a) = p(a)e^{-sD(a)/\varepsilon - \Lambda(s)}.$$

Finite action support permits differentiation for every real s:

$$\Lambda(0) = 0, \qquad \Lambda'(0) = -\mathbb{E}_p D/\varepsilon, \qquad \Lambda''(s) = \mathrm{Var}_{p_s}(D)/\varepsilon^2.$$

Taylor's identity with integral remainder, which is exact here, gives

$$\mathbb{E}_p D + \varepsilon\Lambda(1) = \varepsilon[\Lambda(1) - \Lambda(0) - \Lambda'(0)] = \frac{1}{\varepsilon}\int_0^1 (1-s)\,\mathrm{Var}_{p_s}(D)\,ds.$$

By (17), $\varepsilon\Lambda(1) = -\Delta_t(z)$. Substitution into (20) proves the result. The original occupancy d_t^E remains fixed during this local interpolation; no interpolated trajectory occupancy is being substituted. □

Corollary 13 (Two-sided local variance bounds)**.** *Let $\chi = \mathrm{osc}(D)/\varepsilon$ and $v = \mathrm{Var}_p(D)$ in the notation of the preceding proof. Define $a_-(\chi) = (e^{-\chi} - 1 + \chi)/\chi^2$ and $a_+(\chi) = (e^{\chi} - 1 - \chi)/\chi^2$, with $a_-(0) = a_+(0) = 1/2$. Then*

$$\frac{v}{\varepsilon}a_-(\chi) \le \varepsilon\,\mathrm{KL}(p\|p_1) \le \frac{v}{\varepsilon}a_+(\chi). \tag{45}$$

Proof. The likelihood ratio p_s/p lies in $[e^{-s\chi}, e^{s\chi}]$. Using $\mathrm{Var}_q(D) = \min_b \mathbb{E}_q(D-b)^2$ gives $e^{-s\chi}\,\mathrm{Var}_p(D) \le \mathrm{Var}_{p_s}(D) \le e^{s\chi}\,\mathrm{Var}_p(D)$. Integrating against $(1-s)/\varepsilon$ gives (45). At $\chi = 0$, D is constant, $v = 0$, and both bounds are zero; the stated continuous extensions also follow from Taylor expansion of the exponentials. □

C.3 Proof of Theorem 6

All $O(\cdot)$ terms below are for a fixed finite MDP, horizon, and $\varepsilon > 0$. They are uniform over time, admissible state–action cells, and $s \in [0, 1]$ where relevant. They are not uniform over growing horizons, vanishing temperatures, or a changing family of models.

Proof. **Step 1: first-order values and baseline policies.** At $\lambda = 0$, all costs vanish, $V^0 = W^0 = 0$, and $\pi^{C,0} = \pi^{F,0} = \pi_0$, so their physical occupancy is d^0, the occupancy under p_0. Finite positive exponential sums make every backward value and policy analytic in a neighborhood of zero. Differentiating either value recursion at zero gives

$$\left.\frac{dV_t^\lambda}{d\lambda}\right|_0 = \left.\frac{dW_t^\lambda}{d\lambda}\right|_0 = \mathbb{E}_{\pi_{0,t}}[c_t + P_t U_{t+1}] = U_t, \qquad U_T = h.$$

Hence

$$V_t^\lambda = \lambda U_t + O(\lambda^2), \quad W_t^\lambda = \lambda U_t + O(\lambda^2), \quad \pi_t^{F,\lambda} = \pi_{0,t} + O(\lambda), \quad d_t^{E,\lambda} = d_t^0 + O(\lambda). \tag{46}$$

The last assertion follows either from finite products of analytic policies and fixed transition probabilities, or by induction in the forward occupancy recursion.

Step 2: local environmental optimism. For any fixed finite probability vector p and $w^\lambda = \lambda u + O(\lambda^2)$, centering $X^\lambda = w^\lambda - \mathbb{E}_p w^\lambda$ gives

$$\begin{aligned}\mathbb{E}_p w^\lambda + \varepsilon \log \mathbb{E}_p e^{-w^\lambda/\varepsilon} &= \varepsilon \log \mathbb{E}_p e^{-X^\lambda/\varepsilon} \\ &= \frac{1}{2\varepsilon}\mathbb{E}_p(X^\lambda)^2 + O(\lambda^3) = \frac{\lambda^2}{2\varepsilon}\,\mathrm{Var}_p(u) + O(\lambda^3).\end{aligned}$$

Apply this to each physical row with $w^\lambda = W_{t+1}^\lambda$:

$$\Gamma_t^\lambda(z, a) = \frac{\lambda^2}{2\varepsilon}\,\mathrm{Var}_{P_t(\cdot|z,a)}(U_{t+1}) + O(\lambda^3). \tag{47}$$

Step 3: propagation through the auxiliary problem. We prove backward from $\Delta_T^\lambda = B_T = 0$ that

$$D_t^\lambda = \frac{\lambda^2}{2\varepsilon} M_t + O(\lambda^3), \qquad \Delta_t^\lambda = \frac{\lambda^2}{2\varepsilon} B_t + O(\lambda^3). \tag{48}$$

If the claim holds at $t+1$, equations (16) and (47) give the assertion for D_t^λ. For a vector of order λ^2, its soft minimum differs from its mean by $O(\lambda^4)$. Therefore

$$\begin{aligned}\Delta_t^\lambda &= \mathbb{E}_{\pi_t^{F,\lambda}} D_t^\lambda + O(\lambda^4) \\ &= \frac{\lambda^2}{2\varepsilon} \mathbb{E}_{\pi_{0,t}} M_t + O(\lambda^3) = \frac{\lambda^2}{2\varepsilon} B_t + O(\lambda^3).\end{aligned}$$

This completes the induction. Since $g_{\mathrm{relax}}^\lambda = \mathbb{E}_{\mu_0} \Delta_0^\lambda$, (25) follows.

Step 4: action distortion and execution loss. By (18) and (48),

$$\begin{aligned}\pi_t^{C,\lambda} - \pi_t^{F,\lambda} &= \pi_t^{F,\lambda} \left[\exp\{-(D_t^\lambda - \Delta_t^\lambda)/\varepsilon\} - 1\right] \\ &= -\frac{\lambda^2}{2\varepsilon^2} \pi_{0,t}(M_t - B_t) + O(\lambda^3),\end{aligned}$$

which is (27). For the local interpolation, $\pi_{t,s}^\lambda = \pi_{0,t} + O(\lambda)$ uniformly in s, and

$$\mathrm{Var}_{\pi_{t,s}^\lambda}(D_t^\lambda) = \frac{\lambda^4}{4\varepsilon^2} \mathrm{Var}_{\pi_{0,t}}(M_t) + O(\lambda^5).$$

Insert this and $d_t^{E,\lambda} = d_t^0 + O(\lambda)$ into (22). The integral of $1 - s$ is $1/2$, which gives exactly the coefficient in (26).

Step 5: degenerate coefficients. Every variance is nonnegative. If their occupancy-weighted sum is zero, $M_t(z, \cdot)$ is action-constant at every state with $d_t^0(z) > 0$, because the prior has full support. These are precisely all physically reachable states, independently of λ: bounded exponential tilting never removes action support. Equation (48) then gives $\mathrm{osc}_a D_t^\lambda(z, \cdot) = O(\lambda^3)$ on every reachable state. The exact range bound (21) yields $g_{\mathrm{exec}}^\lambda = O(\lambda^6)$. This is a higher-order bound, not an assertion of exact losslessness for all λ. □

The variance-to-go has the explicit interpretation

$$B_t(z) = \mathbb{E}_{p_0}\left[\sum_{k=t}^{T-1} \mathrm{Var}_{P_k(\cdot|Z_k,A_k)}(U_{k+1}) \,\middle|\, Z_t = z\right]. \tag{49}$$

This follows from its backward recurrence. It isolates environmental innovation variance rather than variation caused by randomized action choice. If $\mathbb{E}_{\mu_0} B_0 > 0$, division of the two expansions further gives

$$\frac{g_{\mathrm{exec}}^\lambda}{g_{\mathrm{relax}}^\lambda} = \frac{\lambda^2}{4\varepsilon^2} \frac{\sum_{t,z} d_t^0(z) \mathrm{Var}_{\pi_{0,t}} M_t(z, A)}{\mathbb{E}_{\mu_0} B_0} + O(\lambda^3). \tag{50}$$

Strict positivity of a leading coefficient is required to replace an O statement by an exact asymptotic order. In particular, stochasticity alone does not imply a positive fourth-order coefficient.

The fourth- and sixth-order powers are attainable. These are analytical examples, not additional experimental outcomes. Take one initial state, $T = 1$, two actions with uniform prior, zero stage cost, and terminal cost 1 in a bad state and 0 otherwise. If the bad-state probabilities are $(1/4, 1/2)$, then $M_0 = (3/16, 1/4)$, $B_0 = 7/32$, and $\mathrm{Var}_{\pi_0} M_0 = 1/1024$. Theorem 6 gives

$$g_{\mathrm{relax}}^\lambda = \frac{7\lambda^2}{64\varepsilon} + O(\lambda^3), \qquad g_{\mathrm{exec}}^\lambda = \frac{\lambda^4}{8192\varepsilon^3} + O(\lambda^5). \tag{51}$$

For probabilities $(1/4, 3/4)$, $M_0 = (3/16, 3/16)$ is action-constant, but extraction is not exactly lossless. To see the first nonzero term, write $v(\theta) = \theta(1-\theta)$ and $\kappa_3(\theta) = \theta(1-\theta)(1-2\theta)$. The Bernoulli log-partition expansion gives

$$D_0^\lambda(\theta) = \Gamma_0^\lambda(\theta) = \frac{\lambda^2 v(\theta)}{2\varepsilon} - \frac{\lambda^3 \kappa_3(\theta)}{6\varepsilon^2} + O(\lambda^4).$$

Here $\kappa_3(1/4) = 3/32 = -\kappa_3(3/4)$, so the two action corrections differ by $\lambda^3/(32\varepsilon^2) + O(\lambda^4)$. Their variance under the uniform prior is $\lambda^6/(4096\varepsilon^4) + O(\lambda^7)$; the interpolated policies differ from that prior by $O(\lambda)$. Equation (22) therefore gives

$$g^{\lambda}_{\mathrm{relax}} = \frac{3\lambda^2}{32\varepsilon} + O(\lambda^3), \qquad g^{\lambda}_{\mathrm{exec}} = \frac{\lambda^6}{8192\varepsilon^5} + O(\lambda^7). \tag{52}$$

Thus the fourth-order coefficient can vanish without exact losslessness, and the sixth-order degenerate bound cannot generally be improved.

Corollary 14 (Fixed-cost high-temperature separation)**.** *Hold all costs and the MDP fixed and let $\varepsilon \to \infty$. With U, B, M, d^0 as in (23)–(24),*

$$g_{\mathrm{relax}}(\varepsilon) = \frac{\mathbb{E}_{\mu_0} B_0}{2\varepsilon} + O(\varepsilon^{-2}), \tag{53}$$

$$g_{\mathrm{exec}}(\varepsilon) = \frac{1}{8\varepsilon^3} \sum_{t,z} d^0_t(z)\, \mathrm{Var}_{\pi_{0,t}}\, M_t(z, A) + O(\varepsilon^{-4}). \tag{54}$$

If the displayed execution coefficient vanishes, then $g_{\mathrm{exec}}(\varepsilon) = O(\varepsilon^{-5})$.

Proof. For any fixed reference temperature $\varepsilon_0 > 0$ and $\lambda > 0$, $\mathcal{F}_{\lambda c, \lambda h, \varepsilon_0}(q) = \lambda \mathcal{F}_{c,h,\varepsilon_0/\lambda}(q)$. The minimizers and extracted policies on the two sides coincide, and every gap scales by λ. Substitute $\lambda = \varepsilon_0/\varepsilon$ in Theorem 6 and divide each gap by λ. □

C.4 Paired transfer: exploiting common model error

Proposition 15 (Paired policy-comparison radius)**.** *Let α, β be fixed Markov policies, and let the uniform row-TV bounds of Theorem 7 hold. Set*

$$b_t = \sup_z \mathrm{TV}(\alpha_t(\cdot \mid z), \beta_t(\cdot \mid z)), \qquad e_t = 1 - \prod_{k=0}^{t} (1 - b_k),$$

$$S^{\alpha}_t = \mathrm{osc}(\widehat{U}^{\alpha}_{t+1}), \qquad S^{\beta}_t = \mathrm{osc}(\widehat{U}^{\beta}_{t+1}).$$

For $g_K = J_K(\alpha) - J_K(\beta)$ define

$$R^{\mathrm{pair}}_{\alpha,\beta} = \sum_{t<T} \delta_t \left[\mathrm{osc}(\widehat{U}^{\alpha}_{t+1} - \widehat{U}^{\beta}_{t+1}) + 2e_t \min\{S^{\alpha}_t, S^{\beta}_t\} \right]. \tag{55}$$

Then

$$|g_P - g_{\widehat{P}}| \le \min\{R_\alpha + R_\beta,\ R^{\mathrm{pair}}_{\alpha,\beta}\}. \tag{56}$$

The paired radius is zero for identical policies. Replacing $R_F + R_C$ by the minimum in (56) preserves the selector guarantee on the same simultaneous event.

Proof. Couple the two physical rollouts with identical initial state. When their states agree, maximally couple their actions; the probability of different actions at time t, conditional on agreement so far, is at most b_t. When both state and action agree, use the same physical next-state draw. Thus the probability of any disagreement through action t is at most e_t, and the coupling bound gives

$$\mathrm{TV}(\rho^{P,\alpha}_t, \rho^{P,\beta}_t) \le e_t.$$

Write $\Xi_t = P_t - \widehat{P}_t$, suppressing arguments, and $\langle f, g \rangle = \sum_{z,a} f(z,a) g(z,a)$. Subtracting the exact transfer identities gives a sum whose tth term is

$$\begin{aligned} &\mathbb{E}_{\rho^{P,\alpha}_t}[\Xi_t \widehat{U}^{\alpha}_{t+1}] - \mathbb{E}_{\rho^{P,\beta}_t}[\Xi_t \widehat{U}^{\beta}_{t+1}] \\ &= \mathbb{E}_{\rho^{P,\alpha}_t}[\Xi_t (\widehat{U}^{\alpha}_{t+1} - \widehat{U}^{\beta}_{t+1})] + \langle \rho^{P,\alpha}_t - \rho^{P,\beta}_t, \Xi_t \widehat{U}^{\beta}_{t+1} \rangle. \end{aligned}$$

The first term is bounded in absolute value by $\delta_t\, \mathrm{osc}(\widehat{U}^{\alpha}_{t+1} - \widehat{U}^{\beta}_{t+1})$. Each value of $\Xi_t \widehat{U}^{\beta}_{t+1}$ lies between $-\delta_t S^{\beta}_t$ and $\delta_t S^{\beta}_t$. Its span is at most $2\delta_t S^{\beta}_t$, so the second term is at most $2e_t \delta_t S^{\beta}_t$ in absolute value. Reversing the decomposition uses $\rho^{P,\beta}_t$ for the value difference and gives the same first bound with S^{α}_t in the second. We may choose the smaller bound separately for every t and sum. This proves the paired radius. Intersecting with the original sum-radius interval proves (56). If $\alpha = \beta$, the fitted value difference and all e_t vanish. The selector conclusion follows from the same lower-endpoint argument as Corollary 8. □

The paired bound is not uniformly smaller than the sum-radius bound before taking their minimum; nor is it claimed minimax optimal for general horizons. Its useful structural property is cancellation when the candidates agree. Uniformity of the underlying row event again permits data-dependent candidates.

Corollary 16 (Bernoulli contrast and half-radius certification)**.** *In the one-step, unit terminal-loss model, fix two learned action probabilities u_F, u_C and suppose $|\theta_a - \widehat{\theta}_a| \le r$ for both actions. Then (34) holds. If $r \le \widehat{\theta}_a \le 1 - r$ for $a = 0, 1$, the two endpoints $\widehat{g} \pm 2r|u_F - u_C|$ are attainable within that rectangular confidence set.*

For the fitted free and corrected candidates with two actions, a uniform prior, zero stage costs, terminal loss in $\{0, 1\}$, and the same fitted probabilities $\widehat{\theta}_0, \widehat{\theta}_1$, one further has

$$|u_F - u_C| \le \tfrac{1}{2}, \qquad |g_P - \widehat{g}| \le 2r|u_F - u_C| \le r.$$

Consequently, switching only when $\widehat{g} > r$ preserves the no-harm conclusion of Corollary 8 on the same row-confidence event. Relative to the original total radius $2r$, every multiplier at least $1/2$ is sufficient for this candidate family.

Proof. For a fixed probability u of action zero, the objective is $J_P(u) = u\theta_0 + (1-u)\theta_1 + \varepsilon k(u)$. The regularization term is independent of the transition kernel. Subtract its fitted-model analogue, then subtract the same identity for the other candidate. This gives the exact formula. The two row errors each lie in $[-r, r]$, so their contrast lies in $[-2r, 2r]$. When both fitted probabilities are in $[r, 1-r]$, opposing endpoint errors are feasible and attain either endpoint after accounting for the sign of $u_F - u_C$.

For the specified fitted candidates, the free continuation cost $G_\varepsilon(\theta) = -\varepsilon \log(1 - \theta + \theta e^{-1/\varepsilon})$ is strictly increasing on $[0, 1]$. If $\widehat{\theta}_0 \le \widehat{\theta}_1$, both candidates assign action zero probability at least $1/2$; if $\widehat{\theta}_0 \ge \widehat{\theta}_1$, both probabilities are at most $1/2$. Hence u_F, u_C lie in the same interval of length $1/2$. The exact contrast bound therefore gives $|g_P - \widehat{g}| \le r$. On a switch, $g_P \ge \widehat{g} - r > 0$; otherwise the selected policy is the free candidate. Replacing r by $2\vartheta r$ for any $\vartheta \ge 1/2$ only widens the interval. This proves the special-case guarantee, including fitted probabilities at 0 or 1 and ties between actions. □

For $T = 1$, $\widehat{U}_1^\alpha = \widehat{U}_1^\beta = h$ and $e_0 = |u_F - u_C|$, so the general paired radius reduces to this exact result. The half-radius guarantee applies to the already-evaluated 0.5 multiplier; it is not a general half-radius theorem for arbitrary MDPs. The tighter policy-dependent radius $2r|u_F - u_C|$ remains a distinct, unevaluated selector. No frozen experimental outcome is changed.

C.5 Executable correction algorithm

Given a free backward solution, compute for $t = T - 1, \dots, 0$:

$$\Gamma_t = P_t W_{t+1} + \varepsilon \log P_t e^{-W_{t+1}/\varepsilon}, \qquad D_t = \Gamma_t + P_t \Delta_{t+1},$$
$$\Delta_t = -\varepsilon \log \mathbb{E}_{\pi_t^F} e^{-D_t/\varepsilon}, \qquad \pi_t^C \propto \pi_t^F e^{-D_t/\varepsilon}.$$

Initialize $\Delta_T = 0$. For numerical stability, evaluate normalizers by log-sum-exp on the support of each row and retain log policies when probabilities underflow. Exact zeros of P remain zeros. This costs $O(\sum_{t,z,a} |\operatorname{supp} P_t(\cdot \mid z, a)|)$ arithmetic operations, plus action normalization, once the free solution is available. It does not avoid storing or evaluating the physical kernel, and it is not an asymptotic speedup over solving the executable Bellman equations directly.

D Additional Theoretical Consequences

D.1 Continuation-range control

Corollary 17 (Continuation-range bound)**.** *Let*

$$\Omega_t(z, a) = \max_{z' \in \operatorname{supp} P_t(\cdot|z,a)} W_{t+1}(z') - \min_{z' \in \operatorname{supp} P_t(\cdot|z,a)} W_{t+1}(z'). \tag{57}$$

Then

$$0 \le \sum_{t,z,a} \rho_t^C(z,a)\Gamma_t(z,a) \le g_{\text{relax}} \le g_{\text{total}} \le \sum_{t,z,a} \rho_t^E(z,a)\frac{\Omega_t(z,a)^2}{8\varepsilon} \le \sum_{t=0}^{T-1} \max_{z,a} \frac{\Omega_t(z,a)^2}{8\varepsilon}. \tag{58}$$

Proof. Hoeffding's lemma (Boucheron et al., 2013) applied to $-W_{t+1}/\varepsilon$ under each transition row gives $\Gamma_t(z,a) \le \Omega_t(z,a)^2/(8\varepsilon)$. Substitute this bound into Equation (15); the remaining inequalities follow from nonnegativity and $g_{\text{total}} = g_{\text{relax}} + g_{\text{exec}}$. □

D.2 Correction magnitude and local metrics

Corollary 18 (Correction magnitude)**.** *For every t, z,*

$$\min_a D_t(z,a) \le \Delta_t(z) \le \mathbb{E}_{\pi_t^F} D_t(z,A) \le \max_a D_t(z,a), \tag{59}$$

$$0 \le \varepsilon\,\mathrm{KL}(\pi_t^F \| \pi_t^C) \le \frac{\mathrm{osc}_a(D_t(z,\cdot))^2}{8\varepsilon}, \qquad \mathrm{TV}(\pi_t^F, \pi_t^C) \le \frac{\mathrm{osc}_a(D_t(z,\cdot))}{4\varepsilon}. \tag{60}$$

Proof. The soft minimum in Equation (17) lies between the minimum and the mean, proving Equation (59). Hoeffding's lemma applied to $-D_t(z,A)/\varepsilon$ under π_t^F and Equation (19) gives the KL bound. Pinsker's inequality gives the TV bound. □

Corollary 19 (Local metric and fixed-vector cumulants)**.** *The local transition correction satisfies*

$$\Gamma_t(z,a) \ge 2\varepsilon\,\mathrm{TV}\big(P_t(\cdot \mid z,a), P_t^F(\cdot \mid z,a)\big)^2. \tag{61}$$

For a fixed finite distribution p and bounded vector w, define $\Gamma_\varepsilon(p,w) = \mathbb{E}_p w + \varepsilon \log \mathbb{E}_p e^{-w/\varepsilon}$, so $\Gamma_t(z,a) = \Gamma_\varepsilon(P_t(\cdot \mid z,a), W_{t+1})$. If $\sigma_p^2(w)$ and $\kappa_{3,p}(w)$ are the variance and third centered moment, then, as $\varepsilon \to \infty$,

$$\Gamma_\varepsilon(p,w) = \frac{\sigma_p^2(w)}{2\varepsilon} - \frac{\kappa_{3,p}(w)}{6\varepsilon^2} + O(\varepsilon^{-3}). \tag{62}$$

For a fixed action law p and correction vector D, the analogous expansion as $\varepsilon \to \infty$ is

$$\mathbb{E}_p D + \varepsilon \log \mathbb{E}_p e^{-D/\varepsilon} = \frac{\mathrm{Var}_p(D)}{2\varepsilon} - \frac{\kappa_{3,p}(D)}{6\varepsilon^2} + O(\varepsilon^{-3}). \tag{63}$$

Proof. Equation (61) is Pinsker's inequality applied to Equation (10). For the expansion, set $X = w - \mathbb{E}_p w$. The cumulant generating function of the bounded finite-support variable X is analytic near zero, and

$$\varepsilon \log \mathbb{E}_p e^{-X/\varepsilon} = \sigma_p^2(w)/(2\varepsilon) - \kappa_{3,p}(w)/(6\varepsilon^2) + O(\varepsilon^{-3}).$$

This proves Equation (62); replacing w by D proves Equation (63). These expansions hold with the displayed vectors fixed; the full-MDP expansion, which accounts for the temperature dependence of values and policies, is Corollary 14. These are high-temperature asymptotics; the experimental temperatures, all at most one, are not presented as tests of the asymptotic approximation. □

D.3 One-action systems

Corollary 20 (One action)**.** *If every reachable state has one admissible action, then $\pi^C = \pi^F$, $q^E = q^C$, and $g_{\text{exec}} = 0$, while*

$$g_{\text{relax}} = g_{\text{total}} = \sum_{t,z,a} \rho_t^C(z,a)\Gamma_t(z,a). \tag{64}$$

The free/executable gap is positive exactly when the free law changes a reachable physical transition row. Deterministic dynamics and constant continuation values are zero-gap special cases.

Proof. With one admissible action, both π^C and π^F are point masses, so $q^E = q^C$ and Equation (14) is zero. Equations (13) and (15) reduce to Equation (64). A deterministic transition row or a continuation value constant on its support has $\Gamma_t = 0$ by Theorem 1. □

D.4 SECONDARY MODEL-TRANSFER BOUNDS

Corollary 21 (Optimal, free-value, and learned-policy transfer)**.** *Let $J^C(K)$ and $J^F(K)$ denote the optimal physical and free values under kernel K. If $\widehat{V}$ is the executable optimum under $\widehat{P}$, then*

$$|J^C(P) - J^C(\widehat{P})| \le R_V := \sum_{t=0}^{T-1} \delta_t \operatorname{osc}(\widehat{V}_{t+1}). \tag{65}$$

For a row-TV bound $\delta \ge 0$ and a continuation range $\Omega \ge 0$, define

$$\Psi_\varepsilon(\delta, \Omega) = \varepsilon \log\Big[1 + \delta(e^{\Omega/\varepsilon} - 1)\Big]. \tag{66}$$

If $\widehat{W}$ is the free value under $\widehat{P}$, then

$$|J^F(P) - J^F(\widehat{P})| \le R_W := \sum_{t=0}^{T-1} \Psi_\varepsilon\Big(\delta_t, \operatorname{osc}(\widehat{W}_{t+1})\Big). \tag{67}$$

Moreover,

$$0 \le J_P(\widehat{\pi}^C) - J^C(P) \le R_{\widehat{\pi}^C} + R_V. \tag{68}$$

These statements and Theorem 7 extend to standard-Borel state and action spaces when regular conditional probabilities exist, $\alpha_t \ll \pi_{0,t}$, the displayed expectations and spans are bounded and measurable, and measurable optimizers exist for the optimal-value claims.

Proof. The soft minimum is monotone and translation equivariant, hence one-Lipschitz in the sup norm. Subtracting the executable Bellman recursions, applying the TV–span inequality to $\widehat{V}_{t+1}$, and inducting backward proves Equation (65).

For the free recursion, let $\mathcal{T}_P(f) = -\varepsilon \log \mathbb{E}_P e^{-f/\varepsilon}$. This map is one-Lipschitz in f. After subtracting $\inf f$, the random variable $e^{-f/\varepsilon}$ lies in $[e^{-\Omega/\varepsilon}, 1]$, where $\Omega = \operatorname{osc}(f)$. For $\mathrm{TV}(P, \widehat{P}) \le \delta$, the TV–span inequality and the lower bound on both normalizers give

$$|\mathcal{T}_P(f) - \mathcal{T}_{\widehat{P}}(f)| \le \varepsilon \log[1 + \delta(e^{\Omega/\varepsilon} - 1)].$$

Applying this rowwise with $f = \widehat{W}_{t+1}$ and inducting backward proves Equation (67). Finally,

$$\begin{aligned} J_P(\widehat{\pi}^C) - J^C(P) &\le J_{\widehat{P}}(\widehat{\pi}^C) + R_{\widehat{\pi}^C} - [J^C(\widehat{P}) - R_V] \\ &= R_{\widehat{\pi}^C} + R_V. \end{aligned}$$

The standard-Borel extension uses the same conditional-expectation argument under the stated measurability and existence assumptions. □

D.5 FINITE-SAMPLE TRANSITION RADII

Proposition 22 (Finite-sample radii used in the experiments)**.** *Fix $\eta \in (0, 1)$. Let $\mathbf{e}_a$ denote a point mass at a and* Unif *a uniform distribution. Suppose first that a cliff row is the pushforward of $(1 - p_{\mathrm{slip}})\mathbf{e}_a + p_{\mathrm{slip}}\,\mathrm{Unif}\{1, \dots, 4\}$ and $\widehat{p}_{\mathrm{slip}}$ is the mean of N independent Bernoulli(p_{slip}) slip indicators. Simultaneously over all rows sharing p_{slip}, with probability at least $1 - \eta$,*

$$\sup_{t,z,a} \mathrm{TV}\Big(P_t(\cdot \mid z, a), \widehat{P}_t(\cdot \mid z, a)\Big) \le \frac{3}{4}\sqrt{\frac{\log(2/\eta)}{2N}}. \tag{69}$$

Second, let p be a distribution on $m_{\mathrm{cat}} \ge 2$ categories, $\bar{p}_N$ its empirical law from N independent samples, and $\mathbf{1}$ the all-ones vector. Set $\widehat{p} = (N\bar{p}_N + \zeta\mathbf{1})/(N + m_{\mathrm{cat}}\zeta)$ for a smoothing pseudo-count $\zeta \ge 0$. Every transition row formed as a pushforward of p obeys, with probability at least $1 - \eta$,

$$\sup_{t,z,a} \mathrm{TV}(P_t(\cdot \mid z, a), \widehat{P}_t(\cdot \mid z, a)) \le \min\left\{1, \sqrt{\frac{\log(2^{m_{\mathrm{cat}}} - 2) + \log(1/\eta)}{2N}} + \frac{m_{\mathrm{cat}}\zeta}{N + m_{\mathrm{cat}}\zeta}\right\}. \tag{70}$$

Substituting either radius into Theorem 7 makes Corollary 8 a $1 - \eta$ finite-sample guarantee.

Proof. Hoeffding's inequality (Boucheron et al., 2013) gives $\mathbb{P}(|\widehat{p}_{\text{slip}} - p_{\text{slip}}| > r) \le 2e^{-2Nr^2}$. Before the state-transition pushforward, the TV distance between $(1-p_{\text{slip}})\mathbf{e}_a + p_{\text{slip}}\,\text{Unif}\{1,\dots,4\}$ and its $\widehat{p}_{\text{slip}}$ counterpart is $3|\widehat{p}_{\text{slip}} - p_{\text{slip}}|/4$. Total variation contracts under a measurable pushforward. Setting $r = \sqrt{\log(2/\eta)/(2N)}$ proves Equation (69) simultaneously for every row generated by the shared parameter.

For the categorical model, the finite-alphabet inequality of Weissman et al. (2003) implies

$$\mathbb{P}(\text{TV}(\bar{p}_N, p) > r) \le (2^{m_{\text{cat}}} - 2)e^{-2Nr^2}.$$

Choose $r = \sqrt{[\log(2^{m_{\text{cat}}} - 2) + \log(1/\eta)]/(2N)}$. Writing $\omega = m_{\text{cat}}\zeta/(N + m_{\text{cat}}\zeta)$ and Unif for the uniform law on these categories, $\widehat{p} = (1-\omega)\bar{p}_N + \omega\,\text{Unif}$, so

$$\text{TV}(p, \widehat{p}) \le \text{TV}(p, \bar{p}_N) + \omega\,\text{TV}(p, \text{Unif}) \le r + \omega.$$

Truncation at one and contraction under each task-transition pushforward give Equation (70). The final claim follows by applying Corollary 8 on the corresponding confidence event. □

E Physical-Rollout Estimation

Define the physical path scores

$$S_C(\tau) = \sum_{t=0}^{T-1} \left[\Gamma_t(Z_t, A_t) + \varepsilon\,\text{KL}(\pi_t^C(\cdot \mid Z_t) \| \pi_t^F(\cdot \mid Z_t))\right], \tag{71}$$

$$S_E(\tau) = \varepsilon \sum_{t=0}^{T-1} \text{KL}(\pi_t^F(\cdot \mid Z_t) \| \pi_t^C(\cdot \mid Z_t)). \tag{72}$$

Then $\mathbb{E}_{q^C} S_C = g_{\text{relax}}$ and $\mathbb{E}_{q^E} S_E = g_{\text{exec}}$. For the intervals below, write $g_C = g_{\text{relax}}$ and $g_E = g_{\text{exec}}$, let $\widehat{g}_C, \widehat{g}_E$ denote their score sample means, and fix a failure probability $\eta \in (0,1)$.

Proposition 23 (Simultaneous variance-adaptive rollout intervals)**.** *Let*

$$\overline{S}_C = \sum_{t=0}^{T-1} \max_{z,a} \left[\Gamma_t(z,a) + \varepsilon\,\text{KL}(\pi_t^C(\cdot \mid z) \| \pi_t^F(\cdot \mid z))\right],$$

$$\overline{S}_E = \varepsilon \sum_{t=0}^{T-1} \max_{z} \text{KL}(\pi_t^F(\cdot \mid z) \| \pi_t^C(\cdot \mid z)).$$

From $n_C, n_E \ge 2$ independent physical rollouts, let the two gap estimators be the score sample means with unbiased sample variances $\widehat{\sigma}_C^2, \widehat{\sigma}_E^2$. With probability at least $1-\eta$ simultaneously, for $j \in \{C, E\}$,

$$|\widehat{g}_j - g_j| \le \sqrt{\frac{2\widehat{\sigma}_j^2 \log(8/\eta)}{n_j}} + \frac{7\overline{S}_j \log(8/\eta)}{3(n_j - 1)}. \tag{73}$$

The Hoeffding half-width $\overline{S}_j\sqrt{\log(4/\eta)/(2n_j)}$ is a distribution-free reference.

Proof. Every summand in S_C and S_E is nonnegative. By construction, $0 \le S_C \le \overline{S}_C$ and $0 \le S_E \le \overline{S}_E$, and the expectations follow from Equations (13) and (14). Apply the two-sided empirical-Bernstein inequality derived from Maurer and Pontil (2009) to each scaled score, allocate $\eta/2$ to each score family, and take a union bound. Applying Hoeffding's inequality (Boucheron et al., 2013) in the same way gives the reference radii. □

F Relationship to Existing Control Formulations

The comparison below distinguishes established projection and control constructions from the quantitative results proved here. It does not infer absence of a method from differences in terminology.

Table 1: **Relationship to the closest prior constructions.** The first two columns acknowledge results already available; the last locates the additional question addressed here. No absence of a particular symbol is used as evidence that a prior work omitted the underlying concept.

Work	Established contribution	Relation to this paper
Kárný and Guy (2006); Neumann (2011); Rawlik et al. (2012)	Desired path distributions and stochastic control as KL-based inference with physical dynamics.	Executable variational identities are background, not a new control algorithm.
Levine (2018)	Explains optimistic inference under stochastic dynamics and the need to preserve the environment model.	We quantify when that optimism changes extracted actions and by how much.
Arriojas et al. (2023)	Studies optimistic policies in the physical environment; modified inference recovers fixed-dynamics entropy-regularized control.	We do not claim first physical evaluation or first correction. We isolate exact gap geometry and the propagated action criterion.
Lefebvre (2024)	Derives forward and reverse KL projections and their distinct recursions; relates them to stochastic and risk-sensitive control.	Both projection directions and the extracted policy are prior art. The focus here is their objective difference, full-MDP weak-tilt coefficients, and learned-policy contrast.
Adamczyk et al. (2023)	Residual soft-value correction uses a prior optimal policy to solve a modified reward or dynamics problem.	Our auxiliary problem is a finite-horizon instance with cost Γ. The added analysis concerns lossless extraction, weak-tilt coefficients, and fitted-policy contrasts.
Bhole et al. (2025)	Separates policy/transition regularization and unifies classical and risk-sensitive objectives through soft-policy iteration.	We retain the unmodified target and quantify decision-relevant optimism rather than proposing a broader regularizer family.
Bušić and Meyn (2018); Ito and Kashima (2022); Kim and Sreenath (2026)	Action constraints, continuous-space KL control, and trajectory regularization under their respective dynamics assumptions.	These feasibility formulations are complementary; our exact finite-horizon comparison concerns (q^F, q^E, q^C).

The log-partition Hessian equals the covariance of sufficient statistics, and exponential-family KL is a log-partition Bregman divergence (Wainwright and Jordan, 2008). Theorem 5 specializes this standard calculus to the propagated correction and physical occupancies; our weak-tilt coefficients track the additional dependence on the MDP recursion. Confidence-gated improvement over a baseline is also established (Thomas et al., 2015). Here the candidates are specifically learned Gibbs extraction and its executable correction, and the bound uses a model-transfer contrast, rather than the off-policy return-estimation procedure of that work.

G Additional Experimental Detail

The random-MDP and finite-sample boundary studies have bundled executable source, aggregate records, and verification outputs. The controlled, cliff, inventory, learned-task, and rollout-estimation studies are archived figures and tables without their original rerun code or row-level records. Their numerical summaries are retained as illustrations, not independently reproduced experiments. None of these studies is a real-world deployment or an evaluation on a learned large-scale model.

G.1 Controlled mechanism fixture

The controlled model has two states. For a probability $\varphi \in [0, 1]$, the one-action transition row is $(\varphi, 1 - \varphi)$. With two actions, the rows are $(\varphi, 1 - \varphi)$ and $(1 - \varphi, \varphi)$, and the second action has stage cost 0.25. The nonconstant terminal loss is $(0, 1)$; the constant control uses $(0.5, 0.5)$. This construction makes every zero-gap and positive-gap regime explicit.

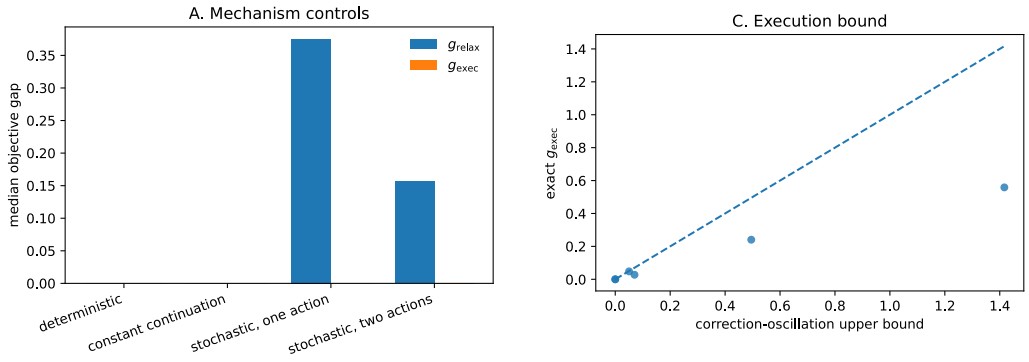


Figure 3: **Archived mechanism controls and execution bound.** Original panels A and C show null cases and the correction-oscillation bound. One-action stochastic systems isolate relaxation optimism without execution loss. These fixed mechanism checks do not test the cost-scaling or temperature asymptotics of Theorem 6 and Corollary 14.

Table 2: **Controlled gap mechanisms.** "Residual" is the largest numerical disagreement among the value, occupancy, and KL forms.

Regime	Rows	Median g_{relax}	Median g_{exec}	Maximum identity residual
Deterministic dynamics	24	0	0	6.7×10^{-16}
Constant continuation	96	0	0	1.0×10^{-15}
Stochastic, one action	36	0.375	0	5.6×10^{-16}
Stochastic, two actions	36	0.156	0.000107	1.1×10^{-15}

G.2 RANDOM-MDP PREVALENCE AND PROPAGATION SUITE

Every instance starts at state 0 with uniform reference actions and independent stage and terminal costs uniform on $[0, c_{\max}]$, where $c_{\max} > 0$ is the cost scale. Here Unif denotes a uniform distribution, and $|\mathcal{A}|$ below is the common number of actions. Transition rows are either random point masses (*deterministic*), one Dirichlet row shared across actions at each (t, z) (*common-mode*), or independent symmetric Dirichlet rows (*generic*). The grid crosses $|\mathcal{Z}| \in \{8, 16, 32\}$, $|\mathcal{A}| \in \{2, 4\}$, $T \in \{3, 5, 10, 20\}$, $\varepsilon \in \{.1, .2, .5, 1\}$, concentrations $\{.05, .2, 1, 5\}$, and cost scales $c_{\max} \in \{.25, 1, 4\}$, with 256 independent instances per cell. Omitting redundant concentration cells for deterministic rows leaves 663,552 instances in 2,592 cells; all contribute to the supplied aggregates.

Normalized gaps divide by $T \max_{t,z,a} c_t(z, a) + \max_z h(z)$; the configured nontrivial-gap threshold is 10^{-3}. The immediate-only ablation replaces D_t by Γ_t in action reweighting; the full correction uses Equation (16). The generic mean execution gap is 0.0717.

Table 3: **Random-MDP transition regimes.** All configured instances are included in the aggregates. P95 is the 95th percentile. Excess is the immediate-Γ objective minus J^C.

Regime	Instances	Median g_{relax}	Median g_{exec}	P95 normalized g_{exec}	Mean immediate-only excess
Deterministic	73,728	0	-2.22×10^{-16}	1.13×10^{-16}	-8.24×10^{-16}
Common-mode	294,912	0.350	-2.22×10^{-16}	1.49×10^{-16}	-1.25×10^{-15}
Generic	294,912	0.303	0.00129	0.0110	0.0185

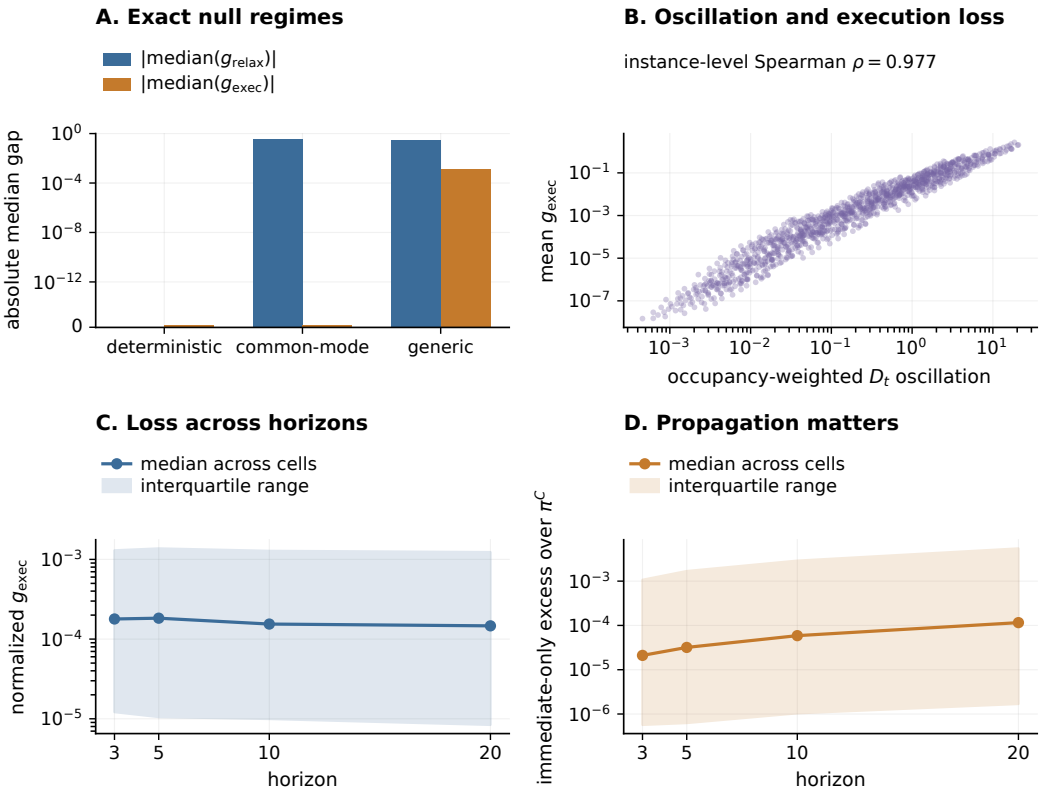


Figure 4: **Random-MDP mechanisms.** (A) Exact null cases. (B) Points are generic cell means; the Spearman statistic is computed over individual generic MDPs. (C) Medians of cell-median normalized execution gaps. (D) Medians of cell-mean immediate-only excesses. Bands in C–D are interquartile ranges across cells, not individual-instance uncertainty.

G.3 CLIFF CONTROL WITH KNOWN AND LEARNED KERNELS

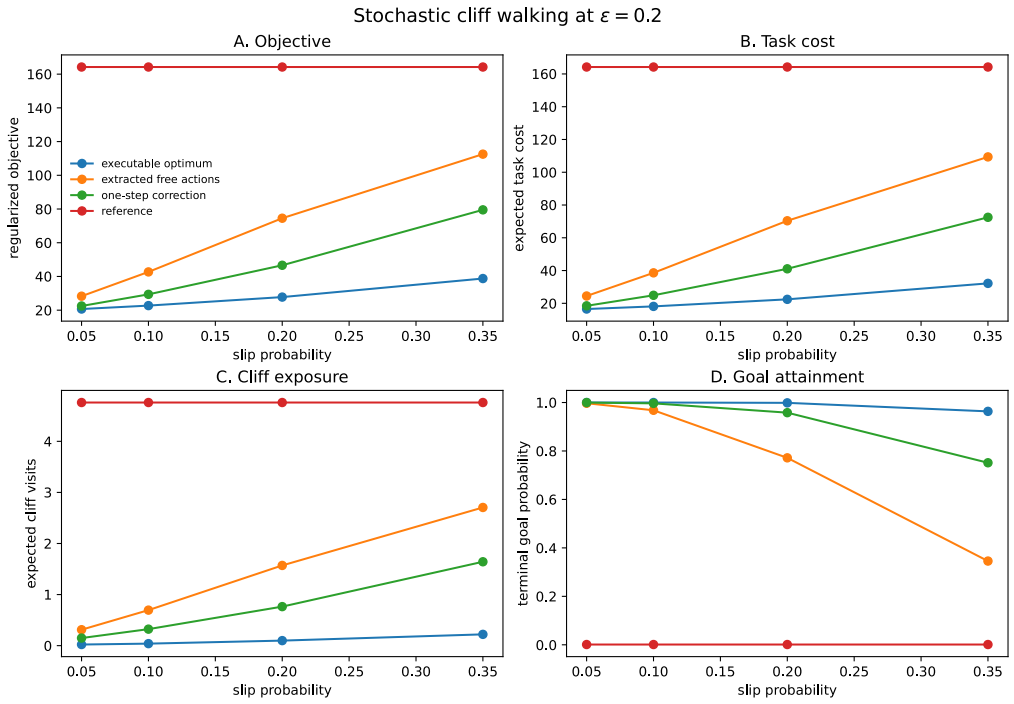


Figure 5: **Known-kernel stochastic cliff grid at $\varepsilon = .2$.** The propagated correction overlaps π^C. The legend's "one-step correction" is the immediate-only correction, which omits future correction-to-go.

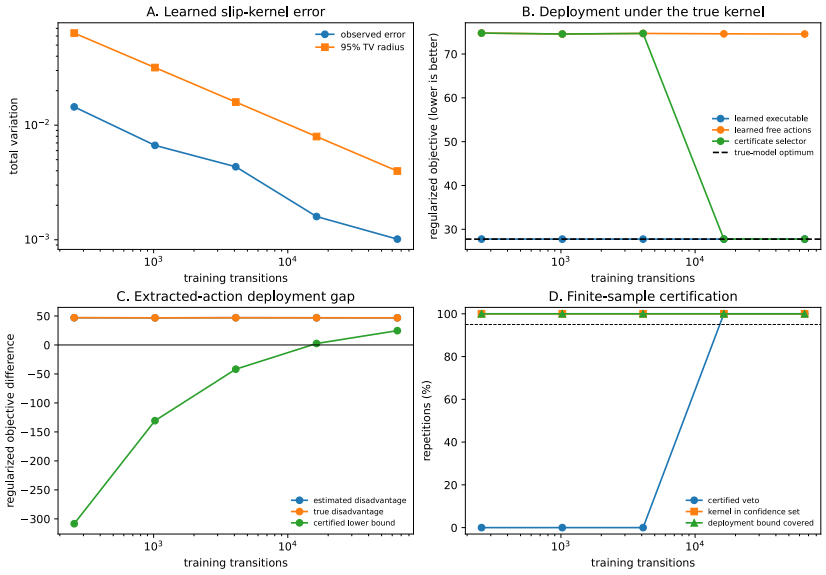


Figure 6: **Certified correction with a learned cliff kernel.** Curves aggregate 64 independent fits. Objective curves use post-hoc true-kernel evaluation; the confidence radius and deployment decision use only the learned model and sample size. "Certified veto" means rejecting the uncorrected candidate and switching to the corrected policy.

G.4 INVENTORY PARAMETERIZATION

Following the standard periodic-review inventory setting (Zipkin, 2000), the frozen task asset uses a 20-period, bounded-state, clipped approximation to a fully backlogged model with net inventory $x_t \in \{-20, \dots, 30\}$ and orders $u_t \in \{0, \dots, 15\}$. Let Y_t denote demand and define $\mathrm{clip}(v, l, u) = \min\{u, \max\{l, v\}\}$. The inventory evolves as

$$x_{t+1} = \mathrm{clip}(x_t + u_t - Y_t, -20, 30). \tag{74}$$

Demand Y_t has a negative-binomial mass function with mean $\mu > 0$ and dispersion $\nu_{\mathrm{NB}} > 0$, hence variance $\mu + \mu^2/\nu_{\mathrm{NB}}$. The tail beyond demand 30 is folded into the final demand bin. The stage cost is $0.08u_t + 0.08[x_t]_+ + 0.8[-x_t]_+$, with twice the inventory penalty at the terminal time. The reference policy is

$$\widetilde{\pi}_0(u \mid x) \propto \exp\left(-\frac{[u - \mathrm{clip}(8 - x, 0, 15)]^2}{2(2.5)^2}\right), \qquad \pi_0 = .98\widetilde{\pi}_0 + .02\,\mathrm{Unif}\{0, \dots, 15\}.$$

The uniform mixture guarantees the full-support assumption. The selected-case maximum identity residual is 1.1×10^{-14}.

Table 4: **Frozen task-domain summary under known kernels.** The supplied assets compare propagated correction with naively extracted free actions across every displayed slip/dispersion and temperature cell. Original rerun code and raw records were not supplied. Here pp denotes percentage points.

Domain	Cells	Objective better	Task cost better	Reported physical effect
Cliff	16	16/16	16/16	Task cost 23.8–70.9% lower.
Inventory	16	16/16	16/16	Task cost 6.8–53.5% lower; stockouts 4.3–42.7 pp lower.

Table 5: **Selected inventory horizon and grid-width checks.** Five of the thirteen documented sensitivity cases are supplied; the remaining eight case records are not bundled.

Case	Horizon	g_{relax}	g_{exec}	Executed boundary occupancy
Baseline	20	12.076	12.397	2.65×10^{-4}
Horizon 10	10	6.437	6.602	2.57×10^{-4}
Horizon 30	30	17.716	18.191	2.67×10^{-4}
Wide grid	20	12.080	12.415	2.42×10^{-6}
Wider grid	20	12.080	12.415	1.81×10^{-8}

Table 6: **Demand-model misspecification.** Policies are solved under the assumed law and evaluated under the true mean-6, dispersion-5 kernel. "Row TV" is the maximum transition-row total variation.

Assumed mean	Assumed dispersion	Maximum row TV	Corrected regret	Extracted regret
5.4	2.0	.161	.042	16.039
5.4	5.0	.067	.260	13.329
5.4	20	.131	.932	11.485
6.0	2.0	.150	.621	15.373
6.0	5.0	.000	0.000	12.397
6.0	20	.129	.268	10.280
6.6	2.0	.154	2.153	14.814
6.6	5.0	.062	.258	11.604
6.6	20	.156	.024	9.238

Learned nonparametric demand kernel. Figure 8 reports the archived unrestricted-demand fit and its finite-sample deployment certificate.

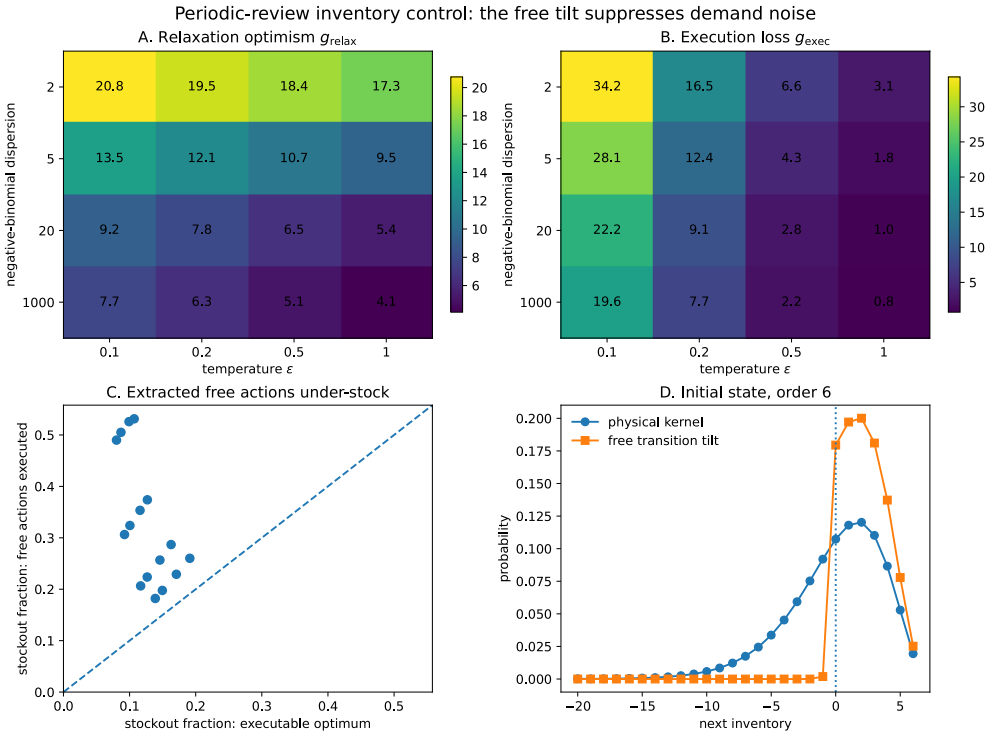


Figure 7: **Archived inventory mechanism.** Selected-case gaps are $g_{\text{relax}} = 12.076$ and $g_{\text{exec}} = 12.397$. The free law tilts demand outcomes, whereas execution uses P. The original vector figure is unchanged; task-level records and plotting code are not archived.

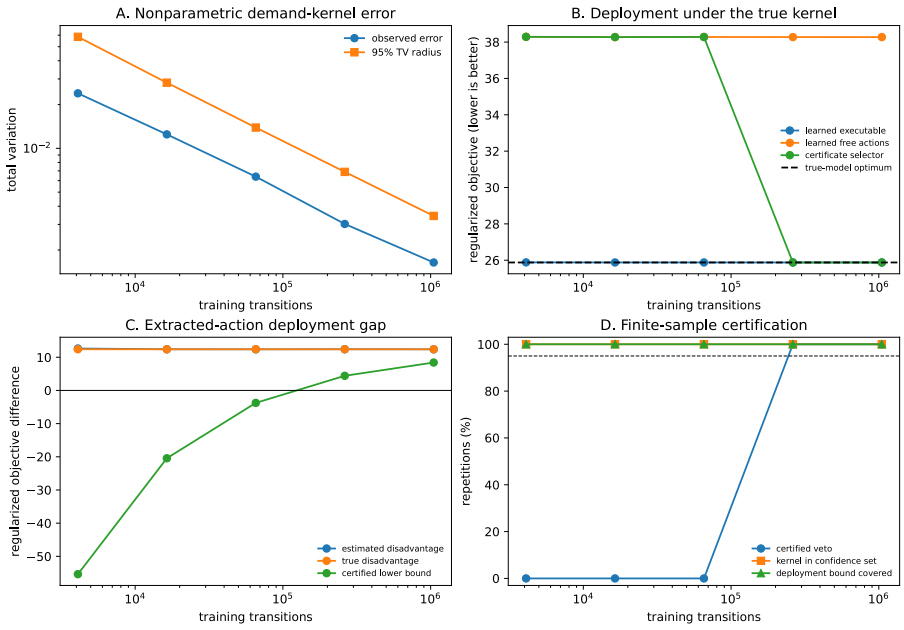


Figure 8: **Learning an unrestricted 31-bin demand distribution.** Objectives and deployment gaps use the hidden true kernel only for post-hoc evaluation; certificate quantities use the learned model and the finite-sample TV radius.

G.5 PHYSICAL-ROLLOUT ESTIMATION STUDY

For the selected inventory case, 64 independent repetitions use $n \in \{128, 512, 2048, 8192\}$ physical rollouts per estimator. Coverage is conditional on the fixed model and nominally simultaneous at 95%.

Table 7: **Repeated physical-rollout estimation.** EB denotes the empirical-Bernstein interval in Proposition 23.

Target	Rollouts	Median absolute error	Median EB half-width	Hoeffding half-width	EB coverage
g_{relax}	128	.120	13.005	17.658	100.0%
	512	.039	3.333	8.829	100.0%
	2,048	.018	.884	4.415	100.0%
	8,192	.014	.247	2.207	100.0%
g_{exec}	128	.036	1.601	1.910	100.0%
	512	.028	.459	.955	100.0%
	2,048	.014	.145	.478	100.0%
	8,192	.005	.051	.239	100.0%

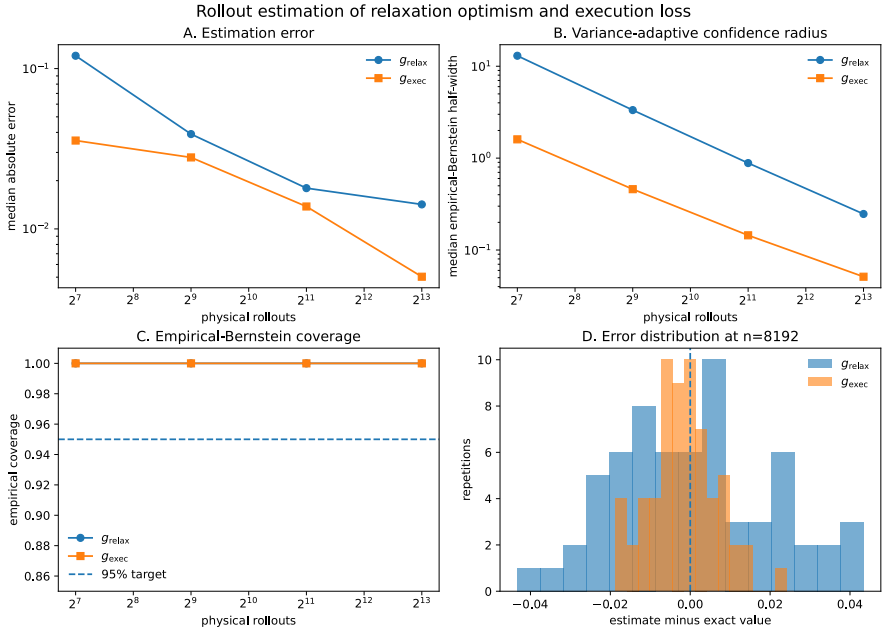


Figure 9: **Repeated rollout estimation.** The exact targets are $g_{\text{relax}} = 12.076$ and $g_{\text{exec}} = 12.397$; interval widths shrink with sample size and remain conservative in the frozen repetitions.

G.6 FINITE-SAMPLE HELP–HARM BOUNDARY

The source-backed boundary suite uses the one-step two-action Bernoulli model of Proposition 11, generalized to allow a true action margin. The true bad-state probabilities are $\theta_0 = b$ and $\theta_1 = b + m$, with $b \in \{.1, .3, .45\}$ and $m \in \{0, .005, .01, .02, .04, .08, .15, .25\}$. We cross $\varepsilon \in \{.05, .1, .2, .5, 1\}$ and

$$n \in \{16, 32, 64, 128, 256, 512, 1024, 4096, 16384, 65536, 262144, 1048576, 4194304\}$$

samples per action, using 8,192 independent binomial repetitions in every valid cell. The learned corrected and free policies are $\widehat{\pi}^C \propto e^{-\widehat{\theta}_a/\varepsilon}$ and $\widehat{\pi}^F \propto e^{-G_\varepsilon(\widehat{\theta}_a)/\varepsilon}$. The hidden true kernel is used only after both candidate policies and the deployment decision are fixed.

The zero-margin subset shares the proposition's equal true action values, but samples independent binomial counts $N_a \sim \mathrm{Bin}(n, \theta_a)$ and uses $\widehat{\theta}_a = N_a/n$, without smoothing. Thus $(\widehat{\theta}_0 + \widehat{\theta}_1)/2$ also fluctuates: the fitted rows are not constrained to the symmetric perturbation around the true θ in Proposition 11. The frozen cell counts contain 814 beneficial zero-margin trials, alongside the reported 95.50% harmful trials. The tested true probabilities lie on the proposition's locally harmful symmetric branch; the finite-sample study does not directly test its locally beneficial symmetric branch. Benefits are not restricted to nonzero true margins.

The fixed implementation labels improvement $g < -10^{-12}$ as harmful and $g > 10^{-12}$ as beneficial; the intervening values are numerical ties. Switching uses the strict test $\widehat{L} > 0$ without a sign tolerance. Recorded confidence-event membership allows 10^{-15} numerical slack, and interval coverage allows 10^{-12} endpoint slack. These are numerical reporting conventions, not modifications of the exact confidence theorem.

With $\eta = .05$, the simultaneous Hoeffding event is $\max_a |\widehat{\theta}_a - \theta_a| \le r := \sqrt{\log(4/\eta)/(2n)}$. Since the one-step continuation span is one, Theorem 7 gives the interval $\widehat{g} \pm 2r$. The configured total-radius multipliers are $\vartheta \in \{0, .25, .5, .75, 1, 1.25, 1.5\}$. The previously named "one-policy-radius" rule is exactly $\vartheta = .5$, because the theorem radius is the sum of two identical policy radii; it is therefore not a distinct strategy. Corollary 16 certifies every $\vartheta \ge .5$ for these candidates on the same confidence event. This conclusion depends on the two-action, uniform-prior, zero-stage-cost, binary-loss structure. All 12,779,520 trials in 1,560 cells are included in the supplied aggregates.

Both beneficial and harmful rates exceed 1% in 712 cells (the frozen aggregation criterion). Always correcting harms 38.46% of trials, with maximum cellwise rate 99.98%. For the $m = 0$ negative control, 1,597,440 trials have identical true actions, yet correction harms 95.50%. The plug-in rule switches in 97.01% of trials and produces 4,914,851 harmful deployments: a positive fitted-model gap alone is not a certificate.

The confidence event holds in 99.70% of trials; interval coverage is 100.00%. The theorem selector switches in 7.06%, never harmfully on a valid event. Its mean oracle regret is 0.001846 below always extracting but 0.001599 above always correcting, reflecting conservative abstention. At multiplier 0.25, 36 harmful switches occur on valid events. None occurs for multipliers at least 0.5, consistent with Corollary 16. The half-radius rule switches in 9.65% of trials with mean two-candidate oracle regret 0.002418, compared with 7.06% and 0.002864 for the original sum-radius rule. These are unchanged outcomes of the configured strategies, not results for the unevaluated policy-dependent paired selector.

Frequency and magnitude in the equal-action subset. Grouping the frozen zero-margin cells by sample size, with each cell weighted by its trial count, gives mean always-correct oracle regret 0.00910929 at $n = 16$ and 7.75851×10^{-8} at $n = 4{,}194{,}304$. The corresponding harmful-trial fractions are 81.82% and 99.42%. Regret is measured against the better fitted candidate, not the true optimal policy; it is averaged over all equal-action trials, not only harmful trials. This is a descriptive aggregation of existing records, not a new experiment or a population estimate.

Table 8: **All unique configured deployment strategies.** Oracle regret is relative to the better of learned extraction and learned correction under the hidden true kernel. The $0.5\times$ row is both the one-policy-radius rule and the half-radius rule certified by Corollary 16. The policy-dependent paired selector is not included. Column headers distinguish percentages from counts.

Strategy	Switch rate	Harmful deploy. (%)	Harmful on valid event (count)	Mean oracle regret
Always correct	100.00%	38.4588	4,893,761	0.001265
Always extract	0.00%	0	0	0.004711
Plug-in, $\vartheta = 0$	97.01%	38.4588	4,893,760	0.001265
$\vartheta = 0.25$	13.11%	0.000540	36	0.001808
$\vartheta = 0.50$ (certified half-radius)	9.65%	0	0	0.002418
$\vartheta = 0.75$	8.06%	0	0	0.002698
$\vartheta = 1.00$ (sum radius)	7.06%	0	0	0.002864
$\vartheta = 1.25$	6.27%	0	0	0.002942
$\vartheta = 1.50$	5.75%	0	0	0.003082

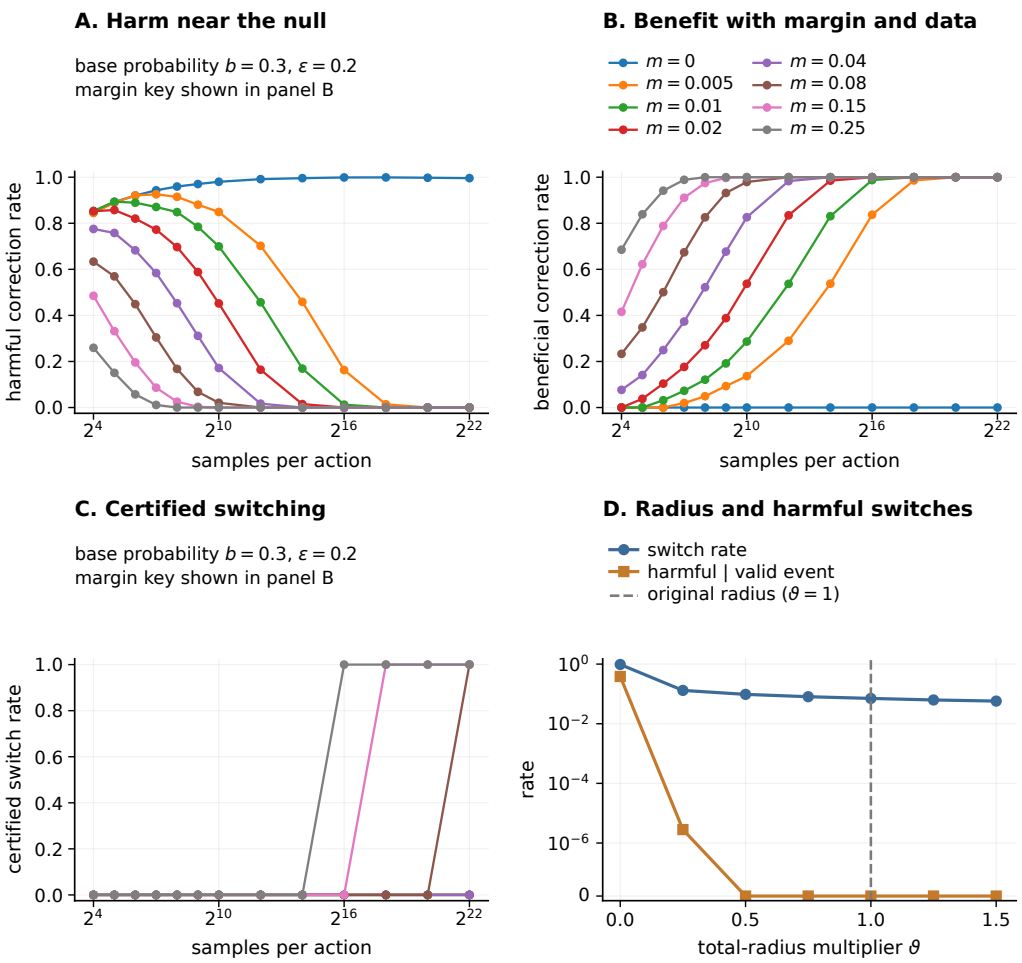


Figure 10: **Finite-sample correction boundary.** (A–B) Learned correction can be harmful or beneficial, with the transition controlled by true margin and sample size. (C) The valid selector switches only when the lower confidence bound is positive. (D) The vertical axis is symmetric-logarithmic so exact zeros remain zero. Reducing the radius increases switching; multipliers at least 0.5 are certified for these candidates, whereas 0.25 yields harmful switches on valid events. Switching is unconditional; the harmful-switch rate is conditional on a valid confidence event. The margin key in B applies to A–C.

Table 9: **Descriptive paired oracle-regret differences over the frozen design.** Rows are not treated as independent draws from a superpopulation; therefore no population-level p-values or confidence intervals are reported. d_z is the paired mean divided by the paired standard deviation (SD).

Paired oracle-regret difference	Mean	Paired SD	Cohen d_z
Selector − always correct	0.001599	0.011158	0.143
Selector − always extract	−0.001846	0.008694	−0.212
Selector − $0.5\times$ radius	0.000446	0.004266	0.105

H Reproducibility Protocol

Executable source and archived results. The supplementary package includes anonymous one-A100 source and aggregate results for the random-MDP and boundary suites. The source requires the NVIDIA CUDA computing platform, uses 64-bit floating-point arithmetic (float64), disables TensorFloat-32 (TF32), requests deterministic PyTorch algorithms, and provides input-hash-checked

atomic checkpoint/resume. The results records retain device-class and software metadata. From the maintained `code/` directory, launch

```
CUDA_VISIBLE_DEVICES=0 ./scripts/run_end_to_end.sh \
    results/full configs/full.yaml
```

and resume with `scripts/resume.sh`. The frozen configuration specifies master seed 20270825 and every cell, sample size, multiplier, and repetition reported above. Retained metadata records one NVIDIA A100-PCIE-40GB, CUDA 12.4, PyTorch 2.6.0+cu124, and Python 3.12.14. The maintained source and the original source have distinct hashes; the configuration and numerical results are unchanged. A public historical-source derivative records its documentation omissions and original hashes; fresh runs must use a new output directory rather than resume checkpoints made with a different source hash.

The archive contains aggregate comma-separated values (CSV) and JavaScript Object Notation (JSON) records, claim assessments, scientific input hashes, configuration and run summaries, and historical verification reports. Those reports record maximum random-MDP identity residual 6.47×10^{-14}, boundary recomputation error 4.44×10^{-16}, zero conditional interval failures, and zero harmful theorem-selector switches on valid events. Fresh executable runs additionally write compressed row-level outputs and checkpoints; these row-level outputs are not supplied here.

Figures, statistics, and evidence limits. `scripts/regenerate_extension_figures.py` reads the frozen aggregates and recreates the three extension figures with exact zeros on symmetric-log axes and legends outside the data regions. Individual vector panels and plotted-value records are exported alongside each composed figure. It accepts explicit input/output directories and defaults to bundled paths. Numeric macros and strategy tables use the same aggregates. The frozen claim assessment does not establish necessity of the sum-of-two-radii rule: Corollary 16 certifies the evaluated half-radius rule for this Bernoulli family. The policy-dependent paired selector was not tested. Historical narrative and claim labels are separately marked as superseded. The supplied configuration is not evidence of an independently time-stamped preregistration. The archived task-domain figures and tables lack original experiment code, raw records, and test suites. An archived task summary lists 480 learned fits without a per-fit ledger; these cannot be independently decomposed or verified. No original-task rerun or independent validation is claimed.