POST-EDIT REGRET FOR LIFTED QUOTIENTS: EXACT ACCOUNTING AND FEASIBLE CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

A state–action abstraction is deployed through a particular coarse model, optimizer, and policy lift; its realized regret reflects this full deployment, beyond the aggregation map. For finite-horizon Kullback–Leibler (KL)-regularized control, we derive an exact signed ledger for this solve–lift–deploy process, separating prior, transition, optimization, and evaluation-cost mismatch. We identify the local price of fixing the environment dynamics as an information projection, decompose it into successor tilting and action reweighting, and allocate the global price through counterfactual occupancies. A residual-control identity then combines this information cost with the original-to-surrogate cost difference. It yields a coupled, dynamics-feasible regret envelope contained in the separate-error envelope. Its raw width vanishes exactly when the combined residual is policy-invariant, equivalently a Bellman potential difference; cancellation can therefore give exact certificates even when separate error budgets remain positive. Finite-model benchmark results evaluate the separate envelope: among 370 strictly positive-width candidate pools, 54 uniquely certify the true optimum, 102 partially prune, and 214 abstain, with no false certifications. Taxi action refinement reduces exact regret by 37.1%; uniform model-error padding covers all 556 nonzero-error pools and remains informative in 494. Independent path enumeration and randomized algebraic audits support the underlying identities. Together, these results provide exact post-deployment accounting and an interpretable hierarchy of certificates for explaining abstraction effects and certifying decisions.

1 INTRODUCTION

The performance of an abstract solution depends on how it is deployed in the original environment. A state merge changes which distinctions the planner can use; an action merge changes the reference mass assigned to an abstract action; a declared coarse transition may predict a different continuation value from the transition realized after lifting. Evaluation may also use a cost that the abstract model does not represent. These mechanisms are different, can have opposite signs, and depend on the policy that actually reaches each state. Accounting for them explains realized regret beyond what an abstraction radius captures.

We study a recorded *state–action quotient declaration*: the maps, coarse prior, coarse dynamics, and coarse costs are all specified, the declared problem is solved, and its policy is lifted and executed under the unchanged fine dynamics. The question is *what determines the regret of this particular solve–lift–deploy operation*, rather than the worst-case quality of a representation class. The finite-horizon KL-regularized setting gives a tractable language for this question, but also exposes a subtle distinction: a Gibbs path law may improve its objective by tilting stochastic successors that a controller cannot choose (Todorov, 2009; Kappen et al., 2012; Levine, 2018; Geist et al., 2019).

From a signed ledger to an operational residual. Our first result expresses deployed regret as four exact, signed terms: reference-prior mismatch, continuation-value mismatch, coarse-versus-fine surrogate optimization, and transfer to the original cost. The ledger retains value-directed discrepancies and their cancellations. Our second result identifies the one-step cost h_t of enforcing the true transition as a reverse-KL projection. Its two nonnegative components distinguish within-action successor dispersion from between-action reweighting. A counterfactual occupancy integral allocates their total contribution across reachable states.

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.

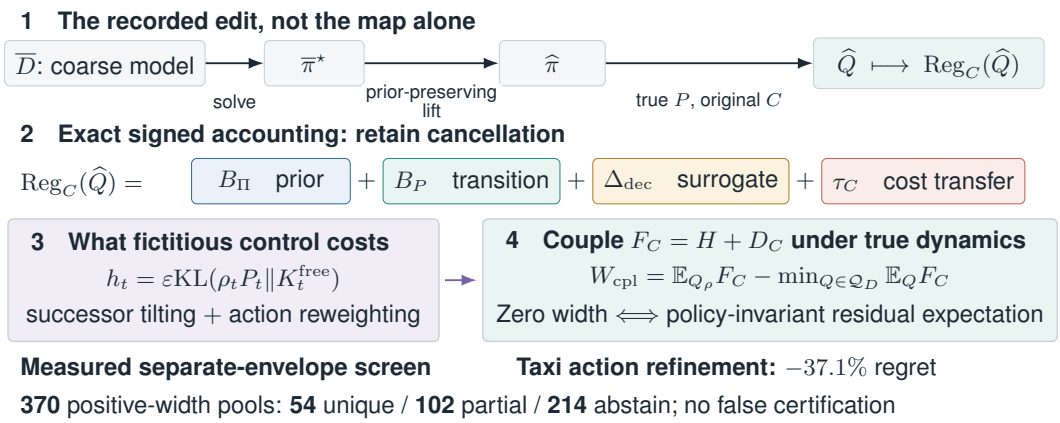


Figure 1. Accounting for a deployed edit without treating every mismatch as a loss. The declared coarse problem determines a lifted policy, but deployment keeps the fine transition P. The signed ledger explains its realized objective and regret. The information projection separates fictitious successor control from the induced action reweighting. Combining its local cost h with the evaluation-cost difference D_C produces a residual task and a coupled certificate: policy-invariant residuals give zero raw width even when separate budgets are positive. The empirical strip reports the *separate-envelope* screen, not an evaluation of the coupled refinement. All 370 pools have strictly positive candidate widths.

The key consequence is a residual-control representation. If ρ is the locally projected policy and $H = \sum_{t<T} h_t(X_t)$, then optimizing the original objective is exactly equivalent, up to a known constant, to controlling the *combined residual* $F_C = H + (C - \widetilde{C})$ with a KL penalty relative to ρ. This gives a feasible certificate that preserves both dynamic constraints and cancellations between the two residuals. We characterize its zero-width regime by a Bellman potential condition. In contrast, $\sum_t \mathrm{span}_{R_t}(h_t) + \Omega_f$ is the width of a particular *separate* envelope, not an information-theoretic obstruction to determining regret. Potential-based reward invariance and entropy-regularized task transformations are established ideas (Ng et al., 1999; Cao et al., 2021; Adamczyk et al., 2023); the contribution here is their connection to the quotient deployment ledger, the two-level information cost, and the resulting nested certificate hierarchy.

Evidence and scope. The experiments evaluate exact finite-model quantities. Randomized programs and independent exhaustive paths audit the algebra; Taxi and other finite control benchmarks test certificate decisions. The separate envelope uniquely resolves 54 of 370 positive-width pools, partially prunes 102, and abstains on 214. Uniform model-error padding is informative in 494 of 556 nonzero-error pools. These findings delineate useful certification and principled abstention. The coupled refinement is established by proof; the benchmark results concern the separate envelope.

A completely known fine model permits exact candidate ordering by policy evaluation: all regrets subtract the same optimal value. Our analysis complements this baseline by explaining what an edit changes, identifying when a certificate is exact, and transporting valid bounds when the fine model is uncertain, rather than reducing known-model selection cost.

2 A Recorded Solve–Lift–Deploy Operation

Consider $D = (\mathcal{X}, \mathcal{A}, T, \mu_0, \Pi_R, P, c, g)$ with finite state and action sets and horizon T, finite costs, normalized probability rows, and $\varepsilon > 0$. Policies may depend on history but use only actions with positive reference mass. Let R_D be the path law of (μ_0, Π_R, P), and let $\mathcal{Q}_D$ contain the laws with this initial distribution and these environment transitions and reference-action support constraints. For $C = \sum_{t<T} c_t(X_t, A_t) + g(X_T)$, define

$$J_C^D(Q) = \varepsilon \mathrm{KL}(Q \| R_D) + \mathbb{E}_Q C, \qquad V_C^D = \inf_{Q \in \mathcal{Q}_D} J_C^D(Q), \qquad \mathrm{Reg}_C^D(Q) = J_C^D(Q) - V_C^D. \quad (1)$$

We suppress D when fixed. The optimal path law is unique and has a Markov realization; policy choices on unreachable histories are immaterial. Write $\eta_t^Q(x, a)$ for state–action occupancy and

R_t for the states structurally reachable at time t from μ_0 through supported actions and transitions. Soft policies have positive occupancy on R_t. The policy-evaluation and optimality recursions are in Section A.

Surjections $\phi : \mathcal{X} \to \mathcal{Z}$ and $\psi : \mathcal{A} \to \mathcal{B}$ specify the state and action classes. The *declaration* $\overline{D} = (\mathcal{Z}, \mathcal{B}, T, \overline{\mu}_0, \overline{\Pi}^{\mathrm{dec}}, \overline{P}^{\mathrm{dec}}, \overline{c}, \overline{g})$ specifies the model actually solved, with $\overline{\mu}_0 = \phi_{\#}\mu_0$ and the same ε. Every coarse state encountered under fine deployment must have a specified coarse Bellman row, including states not reached by the declared model. The coarse policy is the Bellman-optimal policy at all such rows, not an arbitrary extension from its own initial support. Each declared positive action class must have positive fine reference mass at every fine state in that class; unsupported branches are assigned zero lift probability. These conditions do not require lumpability or equality of declared and realized rows.

Let $\overline{\pi}^\star$ and $\overline{V}_t$ be the declared soft-optimal policy and value. Pull back its costs to $\widetilde{c}_t(x,a) = \overline{c}_t(\phi(x), \psi(a))$ and $\widetilde{g}(x) = \overline{g}(\phi(x))$, defining $\widetilde{C}$. With $r_t(b|x) = \sum_{a:\psi(a)=b} \Pi_R(a|x,t)$, the deployed policy is

$$\widehat{\pi}_t(a|x) = \overline{\pi}^\star_t(\psi(a)|\phi(x)) \frac{\Pi_R(a|x,t)}{r_t(\psi(a)|x)}. \tag{2}$$

The displayed ratio applies when $r_t(\psi(a)|x) > 0$; otherwise the lift is zero. It preserves the fine reference proportions within each selected action class. Executing $\widehat{\pi}$ with (μ_0, P) produces $\widehat{Q}$. Its evaluation cost is C, which need not equal $\widetilde{C}$. The maps alone therefore do not specify the audited object.

3 An Exact Signed Regret Ledger

For $z = \phi(x)$ and $b = \psi(a)$, the two deployment witnesses are

$$\beta^{\Pi}_t(x,a) = \varepsilon \log \frac{\overline{\Pi}^{\mathrm{dec}}_t(b|z)}{r_t(b|x)}, \tag{3}$$

$$\beta^{P}_t(x,a) = \mathbb{E}_{P_t(\cdot|x,a)} \overline{V}_{t+1}(\phi(X')) - \mathbb{E}_{\overline{P}^{\mathrm{dec}}_t(\cdot|z,b)} \overline{V}_{t+1}(Z'). \tag{4}$$

Set the witness to zero on zero-deployment branches and define $B_j(Q) = \sum_{t,x,a} \eta^Q_t(x,a) \beta^j_t(x,a)$ for $j \in \{\Pi, P\}$.

Theorem 3.1 (Declaration-to-deployment bridge)**.** *For the lift* (2) *and* $\overline{V}^{\mathrm{dec}} = \mathbb{E}_{\overline{\mu}_0} \overline{V}_0$*,*

$$J_{\widetilde{C}}(\widehat{Q}) - \overline{V}^{\mathrm{dec}} = B_\Pi(\widehat{Q}) + B_P(\widehat{Q}). \tag{5}$$

Any partition of deployed time–state–action rows gives an exact additive allocation.

Proof idea. The lift gives $\log(\widehat{\pi}/\Pi_R) = \log(\overline{\pi}^\star/\overline{\Pi}^{\mathrm{dec}}) + \log(\overline{\Pi}^{\mathrm{dec}}/r)$. Insert the coarse soft Bellman equality, subtract its continuation expectation from the one realized by P, and telescope under $\widehat{Q}$. The initial term cancels because $\overline{\mu}_0 = \phi_{\#}\mu_0$.

The transition witness is value-directed: different kernels can have the same expectation of $\overline{V}_{t+1}$. Neither witness is a distance, and zero total mismatch can reflect cancellation across visited rows. Both are signed and unaffected by additive shifts of the continuation value. They explain this declaration and this deployment, not counterfactual causal effects of independently repairing one component.

Write $d_t = c_t - \widetilde{c}_t$, $d_T = g - \widetilde{g}$, $D_C = C - \widetilde{C}$, and $\widetilde{V} = \inf_{Q \in \mathcal{Q}_D} J_{\widetilde{C}}(Q)$. For $\lambda \in [0,1]$, let $V(\lambda)$ and $Q^\star_\lambda$ be the soft value and optimal law for $\widetilde{C} + \lambda D_C$. Define

$$\Delta_{\mathrm{dec}} = \overline{V}^{\mathrm{dec}} - \widetilde{V}, \qquad \tau_C(Q) = \mathbb{E}_Q D_C - \int_0^1 \mathbb{E}_{Q^\star_\lambda} D_C \, d\lambda. \tag{6}$$

Theorem 3.2 (Exact post-edit ledger)**.**

$$\boxed{\mathrm{Reg}_C(\widehat{Q}) = B_\Pi(\widehat{Q}) + B_P(\widehat{Q}) + \Delta_{\mathrm{dec}} + \tau_C(\widehat{Q}).} \tag{7}$$

The four terms separate reference-prior mismatch, continuation-value mismatch, surrogate optimization, and evaluation-cost transfer. They may cancel.

The proof adds and subtracts $\overline{V}^{\mathrm{dec}}$ and $\widetilde{V}$, then uses $V'(\lambda) = \mathbb{E}_{Q^\star_\lambda} D_C$. Totals are computed by finite dynamic programs: $\tau_C(Q) = \mathbb{E}_Q D_C - [V_C - \widetilde{V}]$. The homotopy additionally allocates transfer by differences between deployed occupancies and integrated optimal occupancies; evaluating that allocation requires numerical quadrature, not merely a fixed number of exact Bellman sweeps. Sections B and D give complete proofs and a conditional-KL path decomposition. This exact accounting retains the computational cost of obtaining the coarse and fine optima.

4 The Information Price of Fixed Dynamics

The Gibbs relaxation can choose a joint action–successor law; a controller can choose only the action marginal. For a continuation value v, the corresponding backups are

$$(\mathcal{T}_t^{\mathrm{free}} v)(x) = -\varepsilon \log \sum_{a,x'} \Pi_R(a|x,t) P_t(x'|x,a) e^{-[\widetilde{c}_t(x,a)+v(x')]/\varepsilon}, \tag{8}$$

$$(\mathcal{T}_t^{\mathrm{fix}} v)(x) = -\varepsilon \log \sum_{a} \Pi_R(a|x,t) e^{-[\widetilde{c}_t(x,a)+\mathbb{E}_{P_t(\cdot|x,a)} v]/\varepsilon}. \tag{9}$$

Set $U_T = \widetilde{g}$, $U_t = \mathcal{T}_t^{\mathrm{free}} U_{t+1}$, and

$$h_t = \mathcal{T}_t^{\mathrm{fix}} U_{t+1} - U_t \geq 0, \qquad V_{\widetilde{C}}^{\mathrm{init}} = \mathbb{E}_{\mu_0} U_0. \tag{10}$$

This relaxation retains the initial distribution. It differs from a fully tilted path partition function, which may also change X_0 (Section B).

4.1 A local projection with two distinct costs

For each supported row let $m = \mathbb{E}_P U_{t+1}$ and

$$\delta_t(x,a) = \varepsilon \log \mathbb{E}_P e^{-[U_{t+1}(X')-m]/\varepsilon}, \tag{11}$$

$$P_t^\dagger(x'|x,a) \propto P_t(x'|x,a) e^{-[U_{t+1}(x')-m]/\varepsilon}. \tag{12}$$

Write $\rho_t = \pi_t^{\mathrm{fix}}$ for the Gibbs action law in (9) evaluated at U_{t+1}, $\pi_t^{\mathrm{free}} \propto \rho_t e^{\delta_t/\varepsilon}$, $K_t^{\mathrm{fix}} = \rho_t P_t$, and $K_t^{\mathrm{free}} = \pi_t^{\mathrm{free}} P_t^\dagger$. Crucially, ρ is a *one-step projection of the relaxed solution*, not in general the fine soft-optimal policy.

Theorem 4.1 (Fixed-transition information projection)**.** *For every reachable state and supported action,*

$$\delta_t(x,a) = \varepsilon \mathrm{KL}(P_t(\cdot|x,a) \| P_t^\dagger(\cdot|x,a)), \tag{13}$$

$$h_t(x) = \varepsilon \mathrm{KL}(K_t^{\mathrm{fix}}(\cdot,\cdot|x) \| K_t^{\mathrm{free}}(\cdot,\cdot|x)) \tag{14}$$

$$= \underbrace{\mathbb{E}_{a\sim\rho_t} \delta_t(x,a)}_{h_t^{\mathrm{succ}}(x)} + \underbrace{\varepsilon \mathrm{KL}(\rho_t \| \pi_t^{\mathrm{free}})}_{h_t^{\mathrm{act}}(x)}. \tag{15}$$

More strongly, for every supported action distribution κ,

$$\varepsilon \mathrm{KL}(\kappa P_t \| K_t^{\mathrm{free}}) = h_t(x) + \varepsilon \mathrm{KL}(\kappa \| \rho_t). \tag{16}$$

Thus $\rho_t P_t$ is the unique reverse-KL projection onto laws retaining P_t.

The successor component vanishes exactly when U_{t+1} is constant on each supported transition row; the action component vanishes exactly when $\delta_t(x,a)$ is constant across supported actions. Stochasticity alone implies neither action reweighting nor a positive successor cost. Two matching variance integrals measure within-row continuation dispersion and between-action dispersion of δ_t, respectively (Corollary C.1 and Proposition C.2). The identities follow directly from Gibbs likelihood ratios and the KL chain rule; (16) also proves the minimization assertion, not only its value.

4.2 From a local premium to an exact residual control problem

Let Q_ρ be the law of (μ_0, ρ, P), $H = \sum_{t<T} h_t(X_t)$, and $F_C = H + D_C$. Both Q_ρ and the audited law are feasible for the true dynamics.

Theorem 4.2 (Residual-control representation)**.** *For every $Q \in \mathcal{Q}_D$,*

$$J_C(Q) - V_{\widetilde{C}}^{\mathrm{init}} = \mathbb{E}_Q F_C + \varepsilon \mathrm{KL}(Q\|Q_\rho). \tag{17}$$

Consequently, with $v_F = \inf_{Q\in\mathcal{Q}_D}\{\mathbb{E}_Q F_C + \varepsilon \mathrm{KL}(Q\|Q_\rho)\}$,

$$V_C = V_{\widetilde{C}}^{\mathrm{init}} + v_F, \qquad \mathrm{Reg}_C(Q) = \mathbb{E}_Q F_C + \varepsilon \mathrm{KL}(Q\|Q_\rho) - v_F. \tag{18}$$

The original problem and this residual problem have the same optimal path law.

Proof idea. The projected Gibbs law gives $\widetilde{c}_t + \varepsilon \log(\pi_t/\Pi_R) + P_t U_{t+1} - U_t = h_t + \varepsilon \log(\pi_t/\rho_t)$. Take expectations, telescope U, and add D_C. This identity is valid for history-dependent policies as well. Within the soft Bellman framework, it places the dynamic infeasibility premium and evaluation mismatch in *one feasible residual task*.

For the surrogate ($D_C = 0$), the price $\alpha = \widetilde{V} - V_{\widetilde{C}}^{\mathrm{init}}$ also has an exact component allocation: $\alpha = \sum_{k<T} \int_0^1 \langle \nu_{k,s}, h_k \rangle \, ds$, where $\nu_{k,s}$ is the occupancy of the true-dynamics optimal prefix with terminal value $U_k + s h_k$. Every such occupancy is positive on R_k. This yields nonnegative global successor and action components with the same reachable-row zero conditions as above (Theorem C.3). These are counterfactual prefix occupancies, not the deployment occupancy of a single policy. The full projection, allocation, and residual proofs are in Sections C and E.

5 Separate and Coupled Regret Certificates

The intervals below are computational envelopes: positive width measures envelope slack rather than statistical non-identifiability of regret. Define the feasible cost-transfer extrema $d_f^- = \min_{Q\in\mathcal{Q}_D} \mathbb{E}_Q D_C$, $d_f^+ = \max_{Q\in\mathcal{Q}_D} \mathbb{E}_Q D_C$, and $\Omega_f = d_f^+ - d_f^-$. They are obtained by ordinary additive-cost dynamic programs. Write $m_h = \sum_{t<T} \min_{R_t} h_t$, $M_h = \sum_{t<T} \max_{R_t} h_t$, and $G_C(Q) = J_C(Q) - V_{\widetilde{C}}^{\mathrm{init}}$. The separate envelope used in the reported experiments is

$$L_C(Q) = \max\{0, G_C(Q) - M_h - d_f^+\}, \tag{19}$$

$$U_C(Q) = G_C(Q) - m_h - d_f^-. \tag{20}$$

Theorem 5.1 (Width of the separate envelope)**.** *The interval (19)–(20) contains $\mathrm{Reg}_C(Q)$, and its width* before lower-endpoint clipping *is*

$$W_{\mathrm{sep}} = U_C - L_C^{\mathrm{raw}} = \sum_{t<T} \mathrm{span}_{R_t}(h_t) + \Omega_f. \tag{21}$$

It is zero exactly when each h_t is constant on R_t and $\Omega_f = 0$. In that case the common endpoint equals regret and clipping is inactive.

One proof uses the counterfactual bounds $m_h \le \alpha \le M_h$ and the cost homotopy. Another follows immediately from Theorem 4.2: the combined residual expectation is at least $m_h + d_f^-$, while evaluating ρ bounds v_F by $M_h + d_f^+$. The magnitude of α and this envelope's width are different: positive layerwise constant information premiums contribute to α but not W_{sep}.

5.1 A stronger envelope that retains feasibility and cancellation

Set

$$a_- = \min_{Q\in\mathcal{Q}_D} \mathbb{E}_Q F_C, \qquad a_\rho = \mathbb{E}_{Q_\rho} F_C. \tag{22}$$

The first quantity is an unregularized control value for the residual cost; the second is one policy evaluation. Neither uses the fine soft-optimal value.

Theorem 5.2 (Coupled feasible certificate and its exactness criterion)**.** *For every feasible Q, define*

$$L_C^{\mathrm{cpl}}(Q) = \max\{0, G_C(Q) - a_\rho\}, \qquad U_C^{\mathrm{cpl}}(Q) = G_C(Q) - a_-. \tag{23}$$

Then

$$[L_C^{\mathrm{cpl}}, U_C^{\mathrm{cpl}}] \subseteq [L_C, U_C], \qquad \mathrm{Reg}_C(Q) \in [L_C^{\mathrm{cpl}}, U_C^{\mathrm{cpl}}]. \tag{24}$$

Its raw width is $W_{\mathrm{cpl}} = a_\rho - a_-$: the unregularized residual-task regret of ρ. It vanishes iff $\mathbb{E}_Q F_C$ is the same for every feasible Q. Equivalently, writing $f_t = h_t + d_t$ and $f_T = d_T$, there are potentials v_t such that

$$v_T = d_T, \qquad f_t(x,a) + \mathbb{E}_{P_t(\cdot|x,a)} v_{t+1} = v_t(x) \quad (x \in R_t,\ \Pi_R(a|x,t) > 0). \tag{25}$$

In this regime the common endpoint is the exact regret.

Proof idea. Nonnegativity of KL and feasibility of Q_ρ give $a_- \le v_F \le a_\rho$; also $m_h + d_f^- \le a_-$ and $a_\rho \le M_h + d_f^+$. For exactness, let v^- solve the unregularized residual Bellman equation. Then $a_\rho - a_-$ is the ρ-occupancy sum of the nonnegative Bellman disadvantages $f_t + P_t v_{t+1}^- - v_t^-$. Since ρ reaches every supported row, zero width forces every disadvantage to vanish, which is (25). The converse is telescoping.

Why a positive separate width need not be irreducible. Take the admissible evaluation-cost difference $D_C = -H$. Then $F_C = 0$, the coupled envelope is exact, and $\mathrm{Reg}_C(Q) = \varepsilon \mathrm{KL}(Q\|Q_\rho)$; nevertheless $W_{\mathrm{sep}} > 0$ whenever some h_t has nonzero reachable span. This is an analytical cancellation example, not an experimental result. More generally, W_{cpl} is invariant to adding a Bellman potential difference to F_C. Applied to D_C alone, the same potential condition characterizes $\Omega_f = 0$ and equality of the original and surrogate soft-optimal policies (Theorem E.1).

Decision rule and computational boundary. For valid intervals $[L_e, U_e]$ on a finite candidate pool, retain

$$S_{\mathrm{cert}} = \{e : L_e \le u^\star\}, \qquad u^\star = \min_e U_e. \tag{26}$$

Corollary 5.3 (Safe screening)**.** *Every true-regret minimizer survives. A singleton certifies the unique pool minimizer; multiple survivors do not certify a unique choice. If all raw widths vanish, the survivor set is exactly the argmin, including ties.*

The coupled envelope can only remove survivors relative to the separate envelope when applied to the same candidates. This is an analytic guarantee, not a measured improvement in the reported runs. Both constructions use the fine model and can cost more than direct policy evaluation. Exact known-model ordering remains $\arg\min_e \mathrm{Reg}_C(Q_e) = \arg\min_e J_C(Q_e)$ (Proposition A.1).

5.2 Certificate computation

For one complete declaration:

1. Solve the coarse soft dynamic program and form the prior-preserving lift.
2. Forward-evaluate the deployed fine policy to obtain $\eta^{\widehat{Q}}$ and $J_{\widetilde{C}}(\widehat{Q})$, $J_C(\widehat{Q})$.
3. Evaluate (3)–(4) on deployed rows and sum B_Π, B_P.
4. Run the free recursion (8) to obtain U; evaluate δ_t, h_t, their reachable minima/maxima, and the surrogate interval.
5. Run the dynamics-feasible additive extrema recursions in Section G to obtain $d_f^\pm$ and form (19)–(20).
6. For the coupled certificate, form $f_t = h_t + d_t$, $f_T = d_T$ and the projected policy ρ. Evaluate a_ρ under P and compute a_- by (65); substitute into (23).
7. When a local cost-transfer allocation is required, solve the one-dimensional homotopy. At each quadrature node λ, run the ordinary soft dynamic program, then forward-evaluate $Q_\lambda^\star$.

Table 1. Benchmark decisions and the direct-evaluation boundary. The first three rows use the separate envelope; the last uses direct objectives. All true pool optima survive in the primary screening studies. "Informative" means at least one candidate is removed; unique decisions are a subset. Positive-width and model-error counts are pool evaluations, not independent training runs.

Study	Scale	Observed decision and boundary
Original edit pools	81 pools / 1,927 candidates	41 unique, all correct; 40 passenger-cost pools abstain
All-positive-width pools	370 pools	54 unique / 102 partial / 214 abstain; zero false certifications
Nonzero model error	556 pools	100% padded coverage; 396 unique; 494 informative
Known-model direct evaluation	32 pools	32/32 select a true optimum; four pools have numerical ties

Each dynamic-program sweep uses finite sums. Homotopy integration is approximated by numerical quadrature; the endpoint value difference supplies an independent check. For dense transition matrices, one Bellman sweep is $O(T|\mathcal{X}|^2|\mathcal{A}|)$; sparse transitions replace $|\mathcal{X}|$ by the number of nonzeros per row. The homotopy multiplies this cost by the number of quadrature nodes. The certificate is therefore not generally cheaper than direct policy evaluation, consistent with Proposition A.1. The source-audited known-model check in Section 7 confirms the direct ordering rule; no speed advantage is claimed.

6 Transferring the Audit Across Model Error

Suppose D and $\widehat{D}$ share $\mathcal{X}, \mathcal{A}, T, \mu_0, \varepsilon, \Pi_R$ but differ in costs and transitions. In the half-ℓ_1 total-variation convention assume

$$\|c_t - \widehat{c}_t\|_\infty \le \kappa_t, \quad \|g - \widehat{g}\|_\infty \le \kappa_T, \quad \sup_{x,a} \mathrm{TV}(P_t(\cdot|x,a), \widehat{P}_t(\cdot|x,a)) \le \delta_t. \tag{27}$$

With p_t the smallest positive reference-action mass, set

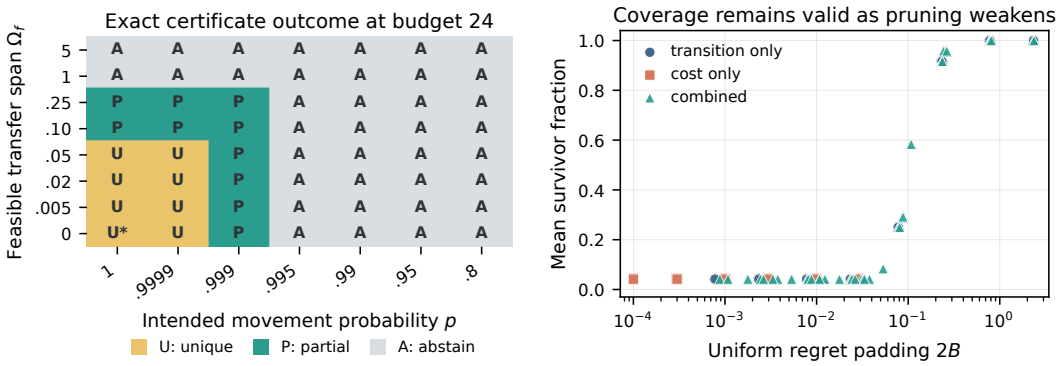
$$\widehat{L}_t = \mathrm{span}(\widehat{c}_t) + \varepsilon \log(1/p_t), \quad \widehat{S}_{t+1} = \mathrm{span}(\widehat{g}) + \sum_{s=t+1}^{T-1} \widehat{L}_s, \quad B = \kappa_T + \sum_{t<T} (\kappa_t + \delta_t \widehat{S}_{t+1}). \tag{28}$$

Theorem 6.1 (Uniform model-error padding)**.** *Every supported Markov policy satisfies* $|J_D(\pi) - J_{\widehat{D}}(\pi)| \le B$ *and* $|\mathrm{Reg}_D(\pi) - \mathrm{Reg}_{\widehat{D}}(\pi)| \le 2B$. *Thus an estimated-model regret interval* $[L, U]$ *transfers to* $[\max\{0, L - 2B\}, U + 2B]$ *in the true model.*

A finite-horizon simulation argument transports either certificate above. The error bounds must cover all rows, including states reachable only under D. Section F proves a tighter candidate-specific *asymmetric* expansion and a simultaneous-event wrapper for data-dependent policies and candidate pools. That wrapper is conditional on a valid simultaneous row-confidence event; it does not itself construct trajectory confidence sets. The experiments evaluate uniform padding under controlled deterministic radii.

7 Finite-Model Evaluation

Protocol and scope. The primary campaign uses float64 dynamic programming on one NVIDIA A100-PCIE-40GB. It evaluates exact finite control benchmarks, with Taxi using 501 planning states, six actions, horizon 50, and $\varepsilon = 1$. Candidate pools, perturbations, seeds, bootstrap counts, and numerical gates are fixed before post-hoc access to exact regret. Every screening rate below concerns the *separate envelope*, not the new coupled refinement. Random programs and exhaustive small path cases are algebraic verification, distinct from benchmark decision studies. Figure 2 shows both decision boundaries. Protocols and full results appear in Section H. A source-audited supplementary run provides the known-model boundary check and a path replication (Section I).



(a) Separate-envelope decision frontier

(b) Uniform-padding nonvacuity

Figure 2. Certificate decisions across stochasticity and model error. (a) For geometric-transition pools at budget 24, greater transition stochasticity (right) or cost-transfer span Ω_f (up) moves the frozen pool from unique certification (U), through partial pruning (P), to abstention (A); U* is the zero-width $p = 1, \Omega_f = 0$ reference. **(b)** In the geometric-transition nonzero-error sweep, uniform $2B$ padding preserves coverage, while larger radii generally weaken pruning. These are controlled model-error radii, not estimated trajectory-confidence radii.

Algebraic checks and zero-width regimes. Across 4,096 randomized programs, the largest ledger/transfer residual is 4.48×10^{-10} (rounded) and the raw-width residual is at most 7.11×10^{-15}. All prespecified gates pass. The separate-width condition predicts all 4,096 numerical singleton outcomes, including 77 stochastic cost-matched cases. A separate implementation enumerates every positive-probability path for 96 small programs and reproduces the conditional-KL identity and path ledger to 3.55×10^{-15} (rounded). These checks complement the proofs with independent implementation verification. Adaptive-order diagnostics and source/configuration hashes are in the supplement.

Useful decisions and principled abstention. The complete positive-width grid varies intended movement probability $p \in \{1, .9999, .999, .995, .99, .95, .8\}$ and feasible transfer span $\Omega_f \in \{0, .005, .02, .05, .1, .25, 1, 5\}$ while holding candidate structures fixed. Of 392 evaluations, 370 have positive raw width for every candidate: 156 are informative, including 54 unique certifications, and no optimum is lost. An exhaustive seven-candidate Taxi movement-action refinement reduces exact regret from 55.991 to 35.242, a reduction of 20.749 (37.1%). All 40 passenger-cost pools abstain despite 53–73% of their candidates being harmful. Here abstention signals unresolved ordering rather than endorsing every candidate.

Uncertainty and the known-model boundary. Across the 556 nonzero-error pools, unpadded estimated intervals cover true regret in 0.472% on average. Uniform $2B$ expansion covers every candidate, retains every optimum, and remains informative in 494 pools, with 396 unique decisions and no false unique. The studied perturbations are small, registered deterministic error radii; these findings do not establish nonvacuity for data-derived radii. In the source-audited known-model check, direct policy evaluation selects a computed pool optimum in all 32 pools; 28 direct screens are singletons and four return numerical ties (tolerances in Section H).

Mechanism localization and ranking behavior. The original objective-unit witness bank recovers 7/7 held-out single faults and 24/24 mixed active sets (Figure 3); these are not measurements of the nested-KL components. Rain activates the fixed-dynamics span, and passenger blindness the transfer span. Ranking comparisons remain inconclusive: the paired Taxi certificate-minus-uniform Spearman interval is $[-0.0365, 0.4957]$ and the stratified cross-domain interval is $[-0.0948, 0.6895]$; both include zero. Navigation correlation is -0.2381.

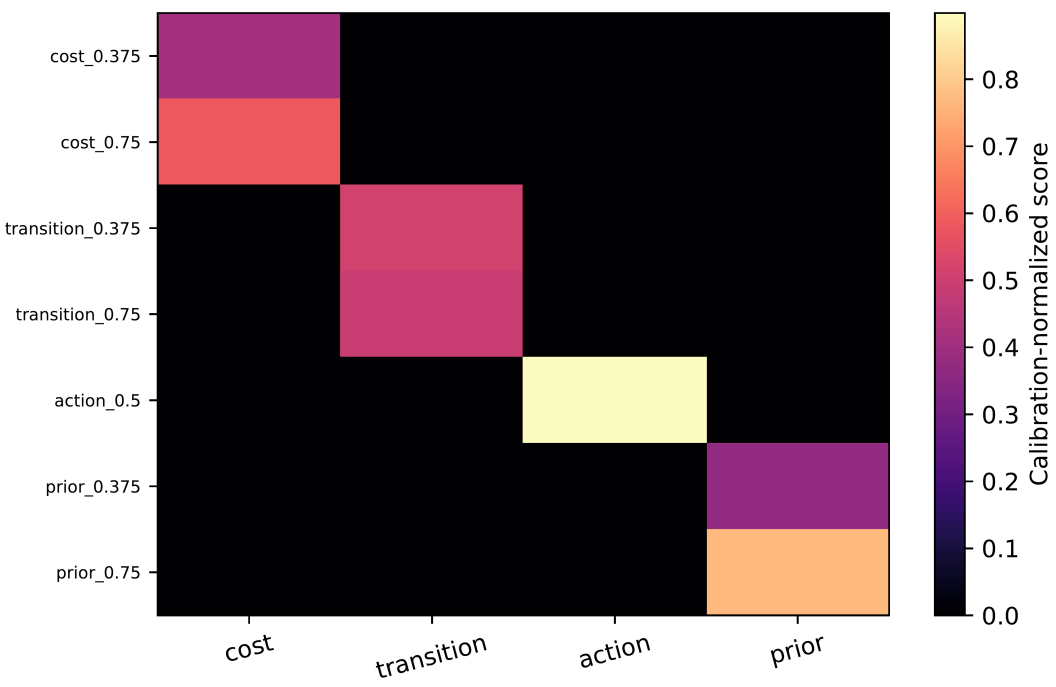


Figure 3. Held-out mechanism localization. Calibration-normalized objective-unit witnesses recover the injected ledger channel for all seven held-out single-fault interventions. Black off-channel entries indicate no spurious activation at the plotted precision.

8 Relation to Prior Work

Regularized control and residual tasks. We build on soft Bellman and Gibbs identities, stochastic-dynamics cautions, the KL chain rule, and the envelope theorem (Todorov, 2009; Kappen et al., 2012; Geist et al., 2019; Levine, 2018; Csiszár, 1975; Milgrom and Segal, 2002). We derive the fixed-transition projection and nested allocation and, building on corrective soft tasks (Adamczyk et al., 2023, Theorem 1) and reward invariance (Cao et al., 2021; Ng et al., 1999; Skalse et al., 2023), a coupled certificate contained in the separate envelope, with zero width characterized by feasible policy invariance.

Abstraction and robustness. Classical abstraction and bisimulation study behavior or value preservation (Givan et al., 2003; Ferns et al., 2004; Li et al., 2006; Abel et al., 2016); value equivalence and compositional semantics motivate continuation-based structure (Grimm et al., 2020; Zhang et al., 2026). We give exact accounting for a declared lift, including model misspecification. Simulation lemmas transport fixed-policy error (Kearns and Singh, 2002; Lobel and Parr, 2024), whereas robust Markov decision processes optimize over uncertainty sets (Iyengar, 2005; Nilim and El Ghaoui, 2005; Mai and Jaillet, 2021; Chen and Huang, 2024).

9 Limitations and Conclusion

The theory assumes finite horizons and spaces, additive costs, positive temperature, common reference support, and a recorded declaration; full support is needed for reachable-row necessity. Ledger terms are not causal effects, certificates may be expensive or vacuous, and no speed advantage is claimed. Tabular benchmarks enable exact evaluation; validation of learned policies, continuous control, real-world systems, and the coupled and uncertainty extensions remains future work.

The ledger accounts for the edit, the projection isolates relaxed changes, and the residual identifies what remains decision-relevant. Feasible cancellation, not separate error magnitudes, determines exactness. Together, these results support an exact, interpretable certify-or-abstain audit.

Reproducibility statement. We provide anonymized verification and result-reconstruction code, later-run source code, and experiment records in the supplementary material.

AI use statement. The author developed the research idea and proofs. AI tools provided coding, debugging, writing, and editing support, including polishing the exposition and presentation of the proofs. The author reviewed and verified the resulting material and retains responsibility for the research questions, methods, proofs, experiments, interpretation, and final content.

REFERENCES

David Abel, David Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2915–2923, 2016.

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.

Haoyang Cao, Samuel N. Cohen, and Łukasz Szpruch. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 12362–12373, 2021.

Ziyi Chen and Heng Huang. Accelerated policy gradient for s-rectangular robust MDPs with large state spaces. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6847–6880, 2024.

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

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 162–169, 2004.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2160–2169, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5541–5552, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

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

Hilbert J. Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, 2012. doi: 10.1007/s10994-012-5278-7.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002.

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

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006.

Sam Lobel and Ronald Parr. An optimal tightness bound for the simulation lemma. *Reinforcement Learning Journal*, 2:785–797, 2024. arXiv:2406.16249.

Tien Mai and Patrick Jaillet. Robust entropy-regularized Markov decision processes. *arXiv preprint arXiv:2112.15364*, 2021.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287, 1999.

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

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32033–32058, 2023.

Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009. doi: 10.1073/pnas.0710743106.

Yivan Zhang, Ziyan Luo, and Manuel Baltieri. Compositional behavioral semantics for state abstraction in reinforcement learning. *arXiv preprint arXiv:2606.25357*, 2026.

A Notation, Support Semantics, and Soft Dynamic Programs

For a finite set $\mathcal{S}$, $\Delta(\mathcal{S})$ is its probability simplex. For $f : \mathcal{S} \to \mathbb{R}$, $\operatorname{span}_{\mathcal{S}}(f) = \max_{\mathcal{S}} f - \min_{\mathcal{S}} f$. Total variation is $\mathrm{TV}(p, q) = \frac{1}{2} \sum_s |p(s) - q(s)|$. The elementary inequality

$$|p^\top f - q^\top f| \le \mathrm{TV}(p, q) \operatorname{span}(f) \tag{29}$$

will be used repeatedly.

For a supported Markov policy π, the trajectory KL in (1) reduces by the chain rule to

$$\mathrm{KL}(Q^\pi \| R_D) = \sum_{t<T} \mathbb{E}_{X_t \sim \eta_t^\pi} \mathrm{KL}(\pi_t(\cdot|X_t) \| \Pi_R(\cdot|X_t, t)).$$

Its regularized cost-to-go obeys

$$u_T^\pi(x) = g(x), \tag{30}$$

$$u_t^\pi(x) = \sum_a \pi_t(a|x) \left[c_t(x, a) + \varepsilon \log \frac{\pi_t(a|x)}{\Pi_R(a|x, t)} + \mathbb{E}_{P_t(\cdot|x,a)} u_{t+1}^\pi(X') \right]. \tag{31}$$

The optimal value recursion is

$$V_T(x) = g(x), \qquad V_t(x) = -\varepsilon \log \sum_a \Pi_R(a|x, t) \exp\left[-\frac{c_t(x, a) + \mathbb{E}_{P_t(\cdot|x,a)} V_{t+1}}{\varepsilon} \right], \tag{32}$$

with optimizer

$$\pi_t^\star(a|x) = \frac{\Pi_R(a|x, t) \exp[-(c_t(x, a) + \mathbb{E}_{P_t} V_{t+1})/\varepsilon]}{\sum_{a'} \Pi_R(a'|x, t) \exp[-(c_t(x, a') + \mathbb{E}_{P_t(\cdot|x,a')} V_{t+1})/\varepsilon]}. \tag{33}$$

All recursions are restricted to positive reference actions. Strict convexity of the conditional KL gives uniqueness on this face. Forward occupancies satisfy

$$\eta_0^\pi(x) = \mu_0(x), \quad \eta_t^\pi(x, a) = \eta_t^\pi(x) \pi_t(a|x), \quad \eta_{t+1}^\pi(x') = \sum_{x,a} \eta_t^\pi(x, a) P_t(x'|x, a).$$

The quotient occupancy-Markovization used for bookkeeping is

$$\overline{\mu}_0(z) = \Pr_{R_D}(\phi(X_0) = z), \tag{34}$$

$$\overline{\Pi}_t^{\mathrm{occ}}(b|z) = \Pr_{R_D}(\psi(A_t) = b \mid \phi(X_t) = z), \tag{35}$$

$$\overline{P}_t^{\mathrm{occ}}(z'|z, b) = \Pr_{R_D}(\phi(X_{t+1}) = z' \mid \phi(X_t) = z, \psi(A_t) = b). \tag{36}$$

The declaration may replace the conditional prior and transition rows, but not the initial pushforward $\overline{\mu}_0 = \phi_\# \mu_0$. Every replacement and every row needed by fine deployment must be recorded. Where the occupancy conditionals above are undefined, an explicit declaration is required; no arbitrary version is silently used. This distinction is essential: the bridge witnesses compare the recorded declaration with the fine rows realized after lifting.

Proposition A.1 (Known-model ordering is direct)**.** *For a finite pool of Markov policy laws $\{Q_e : e \in \mathcal{E}\} \subset \mathcal{Q}_D$ in a completely known fine model,*

$$\arg\min_e \mathrm{Reg}_C^D(Q_e) = \arg\min_e J_C^D(Q_e).$$

Each objective is obtained by finite-horizon policy evaluation; V_C^D is unnecessary for ordering.

Proof. V_C^D is common to all candidates. Forward occupancies or the equivalent backward policy-evaluation recursion compute every candidate objective exactly. □

A.1 Regret as a path-law divergence

For completeness, a standard soft performance identity also fixes the meaning of regret. Let $Q_C^\star$ be the fine optimal law. Then every $Q \in \mathcal{Q}_D$ satisfies

$$\mathrm{Reg}_C(Q) = \varepsilon \mathrm{KL}(Q \| Q_C^\star). \tag{37}$$

Indeed, for a supported action the Gibbs equation gives $c_t(x,a) + P_t V_{t+1}(x,a) - V_t(x) = \varepsilon \log[\Pi_R(a|x,t)/\pi_t^\star(a|x)]$. Adding the local policy/reference log ratio and taking conditional expectations produces $\varepsilon \mathrm{KL}(\pi_t(\cdot|\mathcal{H}_t) \| \pi_t^\star(\cdot|X_t))$, where $\mathcal{H}_t$ is the observed history. Telescoping V and applying the path KL chain rule proves (37). All sums are finite and supported, so this argument also covers history-dependent policies. Equality forces equality of path laws, not uniqueness of policy values on unreachable histories. This identity is a specialization of the usual regularized Bellman variational principle (Geist et al., 2019).

B Proofs for the Exact Ledger

B.1 Proof of Theorem 3.1

Let $z = \phi(x)$ and $b = \psi(a)$. The lift (2) gives the exact conditional log-ratio identity

$$\begin{aligned} \log \frac{\widehat{\pi}_t(a|x)}{\Pi_R(a|x,t)} &= \log \frac{\overline{\pi}_t^\star(b|z)}{r_t(b|x)} \\ &= \log \frac{\overline{\pi}_t^\star(b|z)}{\overline{\Pi}_t^{\mathrm{dec}}(b|z)} + \log \frac{\overline{\Pi}_t^{\mathrm{dec}}(b|z)}{r_t(b|x)}. \end{aligned} \tag{38}$$

Moreover, conditional on (x,b), the lift distributes fine actions according to the fine reference prior restricted to the class:

$$\widehat{\pi}_t(a|x,b) = \frac{\Pi_R(a|x,t)}{r_t(b|x)}.$$

Hence grouping fine actions by b turns any expectation of a function depending only on b into an expectation under $\overline{\pi}_t^\star(\cdot|z)$.

The coarse soft Bellman equality under its optimal policy is

$$\overline{V}_t(z) = \sum_b \overline{\pi}_t^\star(b|z) \left[\overline{c}_t(z,b) + \varepsilon \log \frac{\overline{\pi}_t^\star(b|z)}{\overline{\Pi}_t^{\mathrm{dec}}(b|z)} + \mathbb{E}_{\overline{P}_t^{\mathrm{dec}}(\cdot|z,b)} \overline{V}_{t+1}(Z') \right]. \tag{39}$$

For a deployed fine row, add and subtract the declared continuation expectation in the one-step quantity

$$\widetilde{c}_t(x,a) + \varepsilon \log \frac{\widehat{\pi}_t(a|x)}{\Pi_R(a|x,t)} + \mathbb{E}_{P_t(\cdot|x,a)} \overline{V}_{t+1}(\phi(X')) - \overline{V}_t(\phi(x)).$$

Using (38), (39), and $\widetilde{c}_t(x,a) = \overline{c}_t(z,b)$, its conditional expectation under $a \sim \widehat{\pi}_t(\cdot|x)$ equals

$$\mathbb{E}_{a \sim \widehat{\pi}_t(\cdot|x)} \left[\beta_t^\Pi(x,a) + \beta_t^P(x,a) \right].$$

Multiply by the deployed state occupancy and sum over t, x. The terms involving $\overline{V}_{t+1}(\phi(X_{t+1})) - \overline{V}_t(\phi(X_t))$ telescope under the true deployed transition. Since $\overline{V}_T(\phi(x)) = \overline{g}(\phi(x)) = \widetilde{g}(x)$ and $\overline{\mu}_0 = \phi_\# \mu_0$, the remaining left-hand side is $J_{\widetilde{C}}(\widehat{Q}) - \overline{V}^{\mathrm{dec}}$. The right-hand side is $B_\Pi(\widehat{Q}) + B_P(\widehat{Q})$, proving (5).

For a fixed partition of time–state–action rows, both B_Π and B_P are finite sums; regrouping summands by cells preserves the equality exactly. □

B.2 Conditional path laws and the initial-distribution distinction

For a fine trajectory $\omega = (x_0, a_0, \ldots, a_{T-1}, x_T)$, let $F(\omega) = (\phi(x_0), \psi(a_0), \ldots, \psi(a_{T-1}), \phi(x_T))$. Write $\overline{C}$ for the additive coarse cost on this path, so $\widetilde{C} = \overline{C} \circ F$. Put $p_F = F_\# R_D$, $q = F_\# Q$, and $B_q = q \otimes R_D(\cdot \mid F)$. Conditionals on zero-$q$ fibers are immaterial. Neither p_F nor q need be Markov, and B_q need not retain the fine dynamics. Let $q^\star \propto p_F e^{-\overline{C}/\varepsilon}$ and $V_{\widetilde{C}}^{\mathrm{all}} = -\varepsilon \log \mathbb{E}_{R_D} e^{-\widetilde{C}/\varepsilon}$.

Proposition B.1 (Conditional-KL decomposition). *Every feasible Q satisfies*

$$J_{\widetilde{C}}(Q) - V_{\widetilde{C}}^{\mathrm{all}} = \varepsilon \mathrm{KL}(Q\|B_q) + \varepsilon \mathrm{KL}(q\|q^\star). \tag{40}$$

The fully tilted partition value here permits a change of the initial distribution. In contrast, $V_{\widetilde{C}}^{\mathrm{init}} = \mathbb{E}_{\mu_0} U_0$ fixes it. Specifically, with $\mu^G(x) \propto \mu_0(x) e^{-U_0(x)/\varepsilon}$,

$$V_{\widetilde{C}}^{\mathrm{init}} - V_{\widetilde{C}}^{\mathrm{all}} = \varepsilon \mathrm{KL}(\mu_0\|\mu^G). \tag{41}$$

To see this, backward expansion of the free recursion gives $\mathbb{E}_{R_D}[e^{-\widetilde{C}/\varepsilon} \mid X_0 = x] = e^{-U_0(x)/\varepsilon}$. Substituting μ^G in the KL expression yields (41). Thus the local h_t premiums do not inadvertently include an initial-distribution tilt.

B.3 Proof of Proposition B.1

Let $q = F_\# Q$ and $p_F = F_\# R_D$. The conditional chain rule for relative entropy gives

$$\mathrm{KL}(Q\|R_D) = \mathrm{KL}(q\|p_F) + \mathbb{E}_{\overline{\omega}\sim q} \mathrm{KL}(Q(\cdot|\overline{\omega})\|R_D(\cdot|\overline{\omega}))\,. \tag{42}$$

By construction, $B_q(\omega) = q(F(\omega)) R_D(\omega|F(\omega))$, so the second term in (42) is exactly $\mathrm{KL}(Q\|B_q)$. Because $\widetilde{C}$ is F-measurable, $\mathbb{E}_Q \widetilde{C} = \mathbb{E}_q \overline{C}$. Therefore

$$J_{\widetilde{C}}(Q) = \varepsilon \mathrm{KL}(Q\|B_q) + \left[\varepsilon \mathrm{KL}(q\|p_F) + \mathbb{E}_q \overline{C}\right].$$

For $q^\star(\overline{\omega}) \propto p_F(\overline{\omega}) e^{-\overline{C}(\overline{\omega})/\varepsilon}$, direct substitution yields the finite Gibbs identity

$$\varepsilon \mathrm{KL}(q\|p_F) + \mathbb{E}_q \overline{C} = -\varepsilon \log \mathbb{E}_{p_F} e^{-\overline{C}/\varepsilon} + \varepsilon \mathrm{KL}(q\|q^\star).$$

The log-partition term equals $-\varepsilon \log \mathbb{E}_{R_D} e^{-\widetilde{C}/\varepsilon} = V_{\widetilde{C}}^{\mathrm{all}}$. Combining the two displays proves (40). □

B.4 Proof of Theorem 3.2

Add and subtract the declared and fine surrogate values:

$$\begin{aligned} \mathrm{Reg}_C(\widehat{Q}) &= J_C(\widehat{Q}) - V_C \\ &= \left[J_{\widetilde{C}}(\widehat{Q}) - \overline{V}^{\mathrm{dec}}\right] + \left[\overline{V}^{\mathrm{dec}} - \widetilde{V}\right] \\ &\quad + \left[\mathbb{E}_{\widehat{Q}} D_C - (V_C - \widetilde{V})\right]. \end{aligned} \tag{43}$$

The first bracket is $B_\Pi(\widehat{Q}) + B_P(\widehat{Q})$ by Theorem 3.1; the second is Δ_{dec}. By (55) and the fundamental theorem of calculus,

$$V_C - \widetilde{V} = V(1) - V(0) = \int_0^1 \mathbb{E}_{Q_\lambda^\star} D_C \, d\lambda.$$

Expanding both expectations in additive occupancy form and interchanging finite sums with the integral shows that the last bracket in (43) is precisely $\tau_C(\widehat{Q})$ as defined in (6). This proves (7).

Explicitly, define $\overline{\eta}_t = \int_0^1 \eta_t^{Q_\lambda^\star} \, d\lambda$ for state–action occupancies when $t < T$ and terminal-state occupancies when $t = T$. Then the complete local allocation is

$$\tau_C(Q) = \sum_{t<T} \langle \eta_t^Q - \overline{\eta}_t, d_t \rangle + \langle \eta_T^Q - \overline{\eta}_T, d_T \rangle. \tag{44}$$

Both kinds of occupancy have unit mass; the formula follows by linearity and integration of finite sums.

The forward recursions compute deployed occupancies and the backward recursions compute $\overline{V}$, $\widetilde{V}$, and V^λ. Additive constants in $\overline{V}_{t+1}$ cancel in (4); stagewise constants in d_t cancel because both η_t^Q and $\overline{\eta}_t$ have unit mass. □

C Proofs for the Fixed-Dynamics Relaxation

C.1 Componentwise and global characterizations

Corollary C.1 (Componentwise zero conditions)**.** *Both h_t^{succ} and h_t^{act} are nonnegative. At a reachable state, $h_t^{\mathrm{succ}} = 0$ iff every supported transition row is continuation-value homogeneous, whereas $h_t^{\mathrm{act}} = 0$ iff $\delta_t(x, a)$ is constant over supported actions. Hence the action component measures heterogeneity of rowwise relaxation costs, not transition stochasticity by itself.*

The two levels admit matching variance representations. Let $\overline{\delta}_t(x) = \mathbb{E}_{\pi_t^{\mathrm{fix}}} \delta_t(x, A)$, $P_{t,s} \propto P_t e^{-s(U_{t+1}-m)/\varepsilon}$, and $\pi_{t,s}^{\delta}(a|x) \propto \pi_t^{\mathrm{fix}}(a|x) e^{s(\delta_t(x,a)-\overline{\delta}_t(x))/\varepsilon}$.

Proposition C.2 (Two-level entropic variance)**.** *For every reachable state and supported action,*

$$\delta_t(x, a) = \frac{1}{\varepsilon} \int_0^1 (1 - s) \operatorname{Var}_{P_{t,s}}[U_{t+1}(X')]\, ds, \tag{45}$$

$$h_t^{\mathrm{succ}}(x) = \mathbb{E}_{\pi_t^{\mathrm{fix}}} \delta_t(x, A), \tag{46}$$

$$h_t^{\mathrm{act}}(x) = \frac{1}{\varepsilon} \int_0^1 (1 - s) \operatorname{Var}_{\pi_{t,s}^{\delta}}[\delta_t(x, A)]\, ds. \tag{47}$$

Thus successor dispersion is measured in continuation-value units within transition rows, whereas action reweighting is measured by dispersion of the rowwise relaxation costs.

Let $V_{\widetilde{C}}^{\mathrm{init}} = \mathbb{E}_{\mu_0} U_0$ and $\alpha = \widetilde{V} - V_{\widetilde{C}}^{\mathrm{init}}$. If $G_k(v)$ is the optimal fixed-dynamics prefix value with terminal potential v, then $\nabla G_k(v)$ is its time-k occupancy.

Theorem C.3 (Counterfactual occupancy identity)**.** *For $\nu_{k,s} = \nabla G_k(U_k + s h_k)$,*

$$\alpha = \sum_{k<T} \int_0^1 \langle \nu_{k,s}, h_k \rangle ds, \tag{48}$$

$$\sum_{k<T} \min_{R_k} h_k \le \alpha \le \sum_{k<T} \max_{R_k} h_k. \tag{49}$$

Moreover, with $\alpha^j = \sum_{k<T} \int_0^1 \langle \nu_{k,s}, h_k^j \rangle ds$ for $j \in \{\mathrm{succ}, \mathrm{act}\}$,

$$\alpha = \alpha^{\mathrm{succ}} + \alpha^{\mathrm{act}}, \qquad \alpha^{\mathrm{succ}}, \alpha^{\mathrm{act}} \ge 0. \tag{50}$$

Every $\nu_{k,s}$ is positive on R_k. Hence the successor component vanishes globally iff every reachable transition row is continuation-value homogeneous; the action component vanishes globally iff $\delta_k(x, a)$ is action-constant at every reachable state.

C.2 Proof of the nested information identity

Fix (t, x) and abbreviate $P_a = P_t(\cdot|x, a)$, $Y_a = U_{t+1}(X')$, $m_a = \mathbb{E}_{P_a} Y_a$, and

$$Z_a = \mathbb{E}_{P_a} \exp[-(Y_a - m_a)/\varepsilon] = \exp[\delta_t(x, a)/\varepsilon].$$

The tilted row in (12) satisfies

$$\log \frac{P_a(x')}{P_a^{\dagger}(x')} = \log Z_a + \frac{U_{t+1}(x') - m_a}{\varepsilon}.$$

Taking expectation under P_a cancels the centered second term and gives

$$\mathrm{KL}(P_a \| P_a^{\dagger}) = \log Z_a = \delta_t(x, a)/\varepsilon,$$

which proves (13).

Let

$$w_a = \Pi_R(a|x, t) \exp[-(\widetilde{c}_t(x, a) + m_a)/\varepsilon], \qquad W = \sum_a w_a.$$

The fixed backup induces $\pi_t^{\mathrm{fix}}(a|x) = w_a/W$. In the free backup,

$$\Pi_R(a|x,t)\sum_{x'} P_a(x')\exp[-(\widetilde{c}_t(x,a)+U_{t+1}(x'))/\varepsilon]$$
$$= w_a \mathbb{E}_{P_a}\exp[-(Y_a - m_a)/\varepsilon] = w_a Z_a.$$

Consequently

$$\pi_t^{\mathrm{free}}(a|x) = \frac{\pi_t^{\mathrm{fix}}(a|x)e^{\delta_t(x,a)/\varepsilon}}{\mathbb{E}_{\pi_t^{\mathrm{fix}}} e^{\delta_t(x,A)/\varepsilon}}$$

and

$$h_t(x) = \varepsilon \log \mathbb{E}_{\pi_t^{\mathrm{fix}}} e^{\delta_t(x,A)/\varepsilon}. \tag{51}$$

Directly,

$$\varepsilon \mathrm{KL}(\pi_t^{\mathrm{fix}} \| \pi_t^{\mathrm{free}}) = h_t(x) - \mathbb{E}_{\pi_t^{\mathrm{fix}}}\delta_t(x,A). \tag{52}$$

The KL chain rule for the joint laws $K_t^{\mathrm{fix}}(a,x'|x) = \pi_t^{\mathrm{fix}}(a|x)P_a(x')$ and $K_t^{\mathrm{free}}(a,x'|x) = \pi_t^{\mathrm{free}}(a|x)P_a^\dagger(x')$ gives

$$\begin{aligned}\varepsilon \mathrm{KL}(K_t^{\mathrm{fix}} \| K_t^{\mathrm{free}}) &= \varepsilon \mathrm{KL}(\pi_t^{\mathrm{fix}} \| \pi_t^{\mathrm{free}}) + \varepsilon \mathbb{E}_{\pi_t^{\mathrm{fix}}} \mathrm{KL}(P_a \| P_a^\dagger) \\ &= h_t(x) - \mathbb{E}_{\pi_t^{\mathrm{fix}}}\delta_t(x,A) + \mathbb{E}_{\pi_t^{\mathrm{fix}}}\delta_t(x,A) \\ &= h_t(x).\end{aligned}$$

This proves (14)–(15).

For Corollary C.1, define $h_t^{\mathrm{succ}} = \mathbb{E}_{\pi_t^{\mathrm{fix}}}\delta_t$ and $h_t^{\mathrm{act}} = \varepsilon \mathrm{KL}(\pi_t^{\mathrm{fix}} \| \pi_t^{\mathrm{free}})$. Both are nonnegative. Because every supported action has positive π_t^{fix} and every $\delta_t(x,a) \geq 0$, $h_t^{\mathrm{succ}} = 0$ iff $\delta_t(x,a) = 0$ for every supported action. By (13) and (12), this is equivalent to continuation-value homogeneity on every supported transition row. Moreover, $h_t^{\mathrm{act}} = 0$ iff $\pi_t^{\mathrm{fix}} = \pi_t^{\mathrm{free}}$. The exponential-tilting formula for π_t^{free} shows that this holds iff $e^{\delta_t(x,a)/\varepsilon}$, and hence $\delta_t(x,a)$, is constant over supported actions. This proves the corollary.

All supported actions have strictly positive π_t^{fix} because costs are finite and the reference mass is positive. Thus $h_t(x) = 0$ if and only if the joint KL is zero, which forces $P_a = P_a^\dagger$ for every supported action. From (12), this occurs if and only if $U_{t+1}(X') - m_a$ is constant P_a-almost surely; because it is centered, the constant is zero. The converse is immediate. □

The full projection equality. The free Gibbs law satisfies

$$K_t^{\mathrm{free}}(a,x'|x) = \Pi_R(a|x,t)P_t(x'|x,a)\exp[-(\widetilde{c}_t(x,a)+U_{t+1}(x') - U_t(x))/\varepsilon].$$

Therefore, for any supported action law κ,

$$\begin{aligned}\varepsilon \mathrm{KL}(\kappa P_t \| K_t^{\mathrm{free}}) &= \varepsilon \mathrm{KL}(\kappa \| \Pi_R) + \mathbb{E}_{a\sim\kappa}[\widetilde{c}_t(x,a) + P_t U_{t+1}(x,a)] - U_t(x) \\ &= \varepsilon \mathrm{KL}(\kappa \| \rho_t) + (\mathcal{T}_t^{\mathrm{fix}} U_{t+1})(x) - U_t(x).\end{aligned}$$

The second equality is the finite Gibbs variational identity at the same state and continuation. This is (16). Since ρ_t is positive on all supported actions, the remaining KL is zero exactly for $\kappa = \rho_t$, establishing uniqueness of the projection, including faces with zero reference mass.

C.3 Proof of Proposition C.2

For one transition row write $Z = Y - \mathbb{E}Y$ and

$$f(s) = \log \mathbb{E}_P e^{-sZ/\varepsilon}, \qquad s \in [0,1].$$

The exponentially tilted row is P_s, and differentiation under the finite sum yields

$$f'(s) = -\frac{1}{\varepsilon}\mathbb{E}_{P_s} Z, \qquad f''(s) = \frac{1}{\varepsilon^2}\mathrm{Var}_{P_s}(Z).$$

Since $f(0) = f'(0) = 0$, Taylor's formula with integral remainder gives $\varepsilon f(1) = \varepsilon^{-1}\int_0^1 (1-s)\,\mathrm{Var}_{P_s}(Y)ds$, proving (45). Averaging over π_t^{fix} gives (46).

For the outer action level, abbreviate $\overline{\delta} = \mathbb{E}_{\pi^{\mathrm{fix}}}\delta(A)$ and define

$$g(s) = \log \mathbb{E}_{\pi^{\mathrm{fix}}} e^{s(\delta(A)-\overline{\delta})/\varepsilon}.$$

Its exponential tilt is π_s^δ, and

$$g(0) = g'(0) = 0, \qquad g''(s) = \frac{1}{\varepsilon^2}\operatorname{Var}_{\pi_s^\delta}[\delta(A)].$$

By (51) and (52), $h_t^{\mathrm{act}} = \varepsilon g(1)$. The same integral-remainder identity proves (47).

For completeness, Jensen's inequality gives $\varepsilon \log \mathbb{E}e^{\delta/\varepsilon} \geq \mathbb{E}\delta$; log-sum-exp is at most the maximum, so $h_t(x) \leq \max_a \delta_t(x,a)$. Let $r_t^U(x,a) = \operatorname{span}_{\operatorname{supp} P_t(\cdot|x,a)}(U_{t+1})$ denote the continuation range on a row, distinct from the policy ρ_t. Hoeffding's lemma (Hoeffding, 1963) gives $\delta_t(x,a) \leq r_t^U(x,a)^2/(8\varepsilon)$. □

C.4 Proof of Theorem C.3

For $k \in \{0,\dots,T\}$ and terminal potential v at time k, define

$$G_k(v) = \inf_{\pi_0,\dots,\pi_{k-1}} \mathbb{E}\left[\sum_{t<k}\left(\widetilde{c}_t(X_t,A_t) + \varepsilon \log \frac{\pi_t(A_t|X_t)}{\Pi_R(A_t|X_t,t)}\right) + v(X_k)\right],$$

where the expectation uses μ_0, the policy, and the fixed fine dynamics. Finite compositions of strictly positive sums, exponentials, and logarithms show that G_k is real analytic in the vector v.

Let $\pi^{k,v}$ be the Bellman-optimal prefix policy. Differentiating the finite objective at its unique optimal path law (or applying the finite-dimensional envelope theorem) gives, for each x,

$$\frac{\partial G_k(v)}{\partial v(x)} = \Pr_{\pi^{k,v}}(X_k = x).$$

Hence $\nabla G_k(v)$ is the time-k occupancy. Positive reference actions and finite costs make every action on a structurally reachable path receive positive soft probability; therefore this occupancy is strictly positive on R_k.

The fixed-dynamics Bellman recursion for the prefix implies

$$G_{k+1}(U_{k+1}) = G_k(\mathcal{T}_k^{\mathrm{fix}} U_{k+1}) = G_k(U_k + h_k).$$

Also $G_T(U_T) = \widetilde{V}$ because $U_T = \widetilde{g}$, and $G_0(U_0) = \mathbb{E}_{\mu_0}U_0 = V_{\widetilde{C}}^{\mathrm{init}}$. Thus

$$\begin{aligned}\alpha &= G_T(U_T) - G_0(U_0)\\ &= \sum_{k<T}[G_k(U_k+h_k) - G_k(U_k)]\\ &= \sum_{k<T}\int_0^1 \langle \nabla G_k(U_k + sh_k), h_k\rangle\, ds,\end{aligned}$$

which is (48).

Each gradient is a probability vector supported on R_k, so its inner product with h_k lies between the reachable minimum and maximum. Integration and summation prove (49). Substituting $h_k = h_k^{\mathrm{succ}} + h_k^{\mathrm{act}}$ into (48) proves (50). Both component allocations are nonnegative. Since the occupancies are strictly positive on R_k, a component allocation is zero exactly when its local component vanishes at every reachable state. Corollary C.1 gives the stated row-homogeneity and action-constancy conditions. In particular, the total gap is zero exactly when every reachable transition row is continuation-value homogeneous; Dirac rows satisfy this condition for every continuation.

Finally, if $\{C_j\}_j$ partitions R_k, then

$$\int_0^1 \langle \nu_{k,s}, h_k\rangle ds = \sum_j \int_0^1 \sum_{x\in C_j} \nu_{k,s}(x) h_k(x)\, ds,$$

a nonnegative exact cellwise allocation. □

D Proofs for Cost Transfer, Exact Width, and Screening

D.1 The separate envelope and cost homotopy

Let $\widetilde{\Gamma}^{\mathrm{init}}(Q) = J_{\widetilde{C}}(Q) - V^{\mathrm{init}}_{\widetilde{C}}$. The occupancy bounds in (48) give

$$\widetilde{L}(Q) = \widetilde{\Gamma}^{\mathrm{init}}(Q) - \sum_{t<T} \max_{R_t} h_t, \tag{53}$$

$$\widetilde{U}(Q) = \widetilde{\Gamma}^{\mathrm{init}}(Q) - \sum_{t<T} \min_{R_t} h_t, \tag{54}$$

so $\widetilde{L}(Q) \le \mathrm{Reg}_{\widetilde{C}}(Q) \le \widetilde{U}(Q)$.

For $D_C = C - \widetilde{C}$, define $d_f^- = \inf_{Q\in\mathcal{Q}_D} \mathbb{E}_Q D_C$, $d_f^+ = \sup_{Q\in\mathcal{Q}_D} \mathbb{E}_Q D_C$, and $\Omega_f = d_f^+ - d_f^-$. Additive dynamic programs compute these extrema. The cost homotopy obeys the following identities. Writing $D_C = \sum_{t<T} d_t(X_t, A_t) + d_T(X_T)$, let $m_t^\lambda = \partial_\lambda V_t^\lambda$ and $\xi_t^\lambda(x,a) = d_t(x,a) + \mathbb{E}_{P_t(\cdot|x,a)} m_{t+1}^\lambda(X')$. Then

$$V'(\lambda) = \mathbb{E}_{Q^\star_\lambda} D_C, \tag{55}$$

$$V''(\lambda) = -\frac{1}{\varepsilon} \sum_{t<T} \mathbb{E}\,\mathrm{Var}_{\pi_t^\lambda}[\xi_t^\lambda] \le 0, \tag{56}$$

and therefore

$$\mathrm{Reg}_C(Q) - \mathrm{Reg}_{\widetilde{C}}(Q) = \mathbb{E}_Q D_C - \int_0^1 \mathbb{E}_{Q^\star_\lambda} D_C \, d\lambda. \tag{57}$$

D.2 Proof of the cost-homotopy identities

For each λ, the soft value recursion is

$$V_T^\lambda(x) = \widetilde{g}(x) + \lambda d_T(x), \tag{58}$$

$$V_t^\lambda(x) = -\varepsilon \log \sum_a \Pi_R(a|x,t) \exp\left[-\frac{\widetilde{c}_t(x,a) + \lambda d_t(x,a) + \mathbb{E}_{P_t(\cdot|x,a)} V_{t+1}^\lambda}{\varepsilon} \right]. \tag{59}$$

Finite log-sum-exp compositions are real analytic, so V_t^λ and $V(\lambda) = \mathbb{E}_{\mu_0} V_0^\lambda$ are real analytic.

Let π_t^λ be the Gibbs policy induced by (59) and $m_t^\lambda = \partial_\lambda V_t^\lambda$. Differentiation gives

$$m_T^\lambda = d_T, \qquad m_t^\lambda(x) = \mathbb{E}_{A\sim\pi_t^\lambda(\cdot|x)} \left[d_t(x,A) + \mathbb{E}_{P_t(\cdot|x,A)} m_{t+1}^\lambda(X') \right] = \mathbb{E}_{\pi_t^\lambda} \xi_t^\lambda(x,A).$$

Unrolling this policy-evaluation recursion under μ_0 proves

$$V'(\lambda) = \mathbb{E}_{\mu_0} m_0^\lambda = \mathbb{E}_{Q^\star_\lambda} D_C.$$

For the second derivative write $n_t^\lambda = \partial_\lambda^2 V_t^\lambda$. If

$$q_t^\lambda(x,a) = \widetilde{c}_t(x,a) + \lambda d_t(x,a) + \mathbb{E}_{P_t(\cdot|x,a)} V_{t+1}^\lambda,$$

then $\partial_\lambda q_t^\lambda = \xi_t^\lambda$ and the Gibbs derivative is

$$\partial_\lambda \pi_t^\lambda(a|x) = -\frac{1}{\varepsilon} \pi_t^\lambda(a|x) \left[\xi_t^\lambda(x,a) - m_t^\lambda(x) \right].$$

Differentiating $m_t^\lambda = \mathbb{E}_{\pi_t^\lambda} \xi_t^\lambda$ therefore yields

$$n_T^\lambda = 0, \qquad n_t^\lambda(x) = \mathbb{E}_{A\sim\pi_t^\lambda} \mathbb{E}_{P_t(\cdot|x,A)} n_{t+1}^\lambda(X') - \frac{1}{\varepsilon} \mathrm{Var}_{A\sim\pi_t^\lambda}[\xi_t^\lambda(x,A)].$$

Unrolling under the occupancy of $Q^\star_\lambda$ gives

$$V''(\lambda) = -\frac{1}{\varepsilon} \sum_{t<T} \mathbb{E}_{X_t\sim\eta_t^{Q^\star_\lambda}} \mathrm{Var}_{A_t\sim\pi_t^\lambda(\cdot|X_t)}[\xi_t^\lambda(X_t,A_t)] \le 0.$$

This proves (55)–(56). Integration over $\lambda \in [0,1]$ yields

$$V_C - \widetilde{V} = V(1) - V(0) = \int_0^1 \mathbb{E}_{Q_\lambda^\star} D_C \, d\lambda.$$

Since $J_C(Q) - J_{\widetilde{C}}(Q) = \mathbb{E}_Q D_C$, subtracting the value identity proves (57). Expanding both expectations in occupancies gives (6). Finally, $\mathrm{Var}(Z) \le \mathrm{span}(Z)^2/4$ gives the optional curvature bound

$$0 \le -V''(\lambda) \le \frac{1}{4\varepsilon} \sum_{t<T} \mathbb{E}_{\eta_t^{Q_\lambda^\star}} \mathrm{span}_{a:\Pi_R(a|X_t,t)>0} \big(\xi_t^\lambda(X_t,a)\big)^2.$$

□

D.3 Proof of the feasible-policy transfer interval

Every homotopy optimizer $Q_\lambda^\star$ is dynamics-feasible, hence

$$d_f^- \le \mathbb{E}_{Q_\lambda^\star} D_C \le d_f^+.$$

Average over λ and insert the result into (57) to obtain

$$\mathrm{Reg}_{\widetilde{C}}(Q) + \mathbb{E}_Q D_C - d_f^+ \le \mathrm{Reg}_C(Q) \le \mathrm{Reg}_{\widetilde{C}}(Q) + \mathbb{E}_Q D_C - d_f^-.$$

Substituting the surrogate endpoints (53)–(54) gives (19)–(20); clipping the lower endpoint at zero preserves validity. Both $\mathbb{E}_Q D_C$ and the homotopy average lie in $[d_f^-, d_f^+]$, proving $|\mathrm{Reg}_C(Q) - \mathrm{Reg}_{\widetilde{C}}(Q)| \le \Omega_f$. Let d_s^- and d_s^+ be the minimum and maximum of $D_C(\omega)$ over $\omega \in \mathrm{supp}\, R_D$, and set $\Omega_s = d_s^+ - d_s^-$. Every dynamics-feasible path law is supported on R_D, so $[d_f^-, d_f^+] \subseteq [d_s^-, d_s^+]$ and $\Omega_f \le \Omega_s$. □

D.4 Proof of Theorem 5.1

Subtract the untruncated lower endpoint from the upper endpoint:

$$\begin{aligned} U_C(Q) - L_C^{\mathrm{raw}}(Q) &= \widetilde{U}(Q) - \widetilde{L}(Q) + d_f^+ - d_f^- \\ &= \sum_{t<T} \left[\max_{x \in R_t} h_t(x) - \min_{x \in R_t} h_t(x) \right] + \Omega_f \\ &= \sum_{t<T} \mathrm{span}_{R_t}(h_t) + \Omega_f. \end{aligned}$$

All summands are nonnegative, so the raw width is zero if and only if every reachable-layer span and Ω_f vanish. Both raw endpoints are valid; if they coincide, their common value is the nonnegative regret, hence clipping is inactive and the interval equals the true regret. Note that layerwise constancy of h_t, not vanishing, is what is required: continuation-value homogeneity is the stronger special case that makes every h_t zero by Theorem 4.1, whereas if all supported state–action rows in a layer share one successor law and U_{t+1} is nonconstant on its support, then their common $\delta_t > 0$ gives $\mathrm{span}_{R_t}(h_t) = 0$ with $h_t > 0$. Policy-invariant expectation of D_C makes $\Omega_f = 0$. □

D.5 Proof of Corollary 5.3

Let $e^\star$ minimize true regret and let $\bar{e}$ attain $u^\star = \min_e U_e$. Interval validity gives

$$L_{e^\star} \le \mathrm{Reg}_C(Q_{e^\star}) \le \mathrm{Reg}_C(Q_{\bar{e}}) \le U_{\bar{e}} = u^\star,$$

so $e^\star \in S_{\mathrm{cert}}$. Therefore every true minimizer survives, and a singleton survivor is the unique minimizer. If all raw intervals are singletons, then $L_e = U_e = \mathrm{Reg}_C(Q_e)$ candidatewise, so the survivor set equals the exact argmin, including ties. □

E Residual Control, Coupled Certificates, and Policy Invariance

This section gives the complete proofs of the residual representation and the coupled certificate. The potential-invariance equivalences below specialize established reward shaping and entropy-regularized identifiability results (Ng et al., 1999; Cao et al., 2021; Adamczyk et al., 2023; Skalse et al., 2023) to the finite-horizon, fixed-support setting. They are used to characterize this certificate, not claimed as a new general theory of reward invariance.

Connection to corrective soft tasks. We use the reward-change construction of Adamczyk et al. (2023, Theorem 1), in finite-horizon cost notation. By (10), the auxiliary stage cost $\widetilde{c}_t - h_t$ with terminal cost $\widetilde{g}$ has soft value U_t and optimizer ρ_t under the original reference and dynamics. Changing that cost to C leaves the residual $F_C = H + D_C$ and changes the reference to ρ. The following proof establishes the corresponding identity for every feasible path law, including history-dependent policies; no new general reward-transformation principle is being asserted.

E.1 Proof of Theorem 4.2

For every supported (t, x, a), the projected Gibbs policy is

$$\rho_t(a|x) = \Pi_R(a|x,t) \exp\left[-\frac{\widetilde{c}_t(x,a) + P_t U_{t+1}(x,a) - U_t(x) - h_t(x)}{\varepsilon}\right].$$

Here $P_t v(x,a) = \sum_{x'} P_t(x'|x,a) v(x')$. Rearranging yields the pointwise identity

$$\widetilde{c}_t(x,a) + \varepsilon \log \frac{\pi_t(a|\mathcal{H}_t)}{\Pi_R(a|x,t)} + P_t U_{t+1}(x,a) - U_t(x) = h_t(x) + \varepsilon \log \frac{\pi_t(a|\mathcal{H}_t)}{\rho_t(a|x)}, \tag{60}$$

whenever the action has positive probability under the policy history $\mathcal{H}_t$ ending at x. Zero-policy terms are interpreted through $0 \log 0 = 0$ in expectation. Both sides use the same environment kernel P_t. Take expectation under Q, sum over time, and use $U_T = \widetilde{g}$ to telescope the continuation terms. The left side becomes $J_{\widetilde{C}}(Q) - \mathbb{E}_{\mu_0} U_0$; the right side becomes $\mathbb{E}_Q H + \varepsilon \mathrm{KL}(Q\|Q_\rho)$ by the conditional KL chain rule. All supported ρ_t are positive, so this KL is finite even when Q is deterministic. Adding $\mathbb{E}_Q D_C$ proves (17). Taking the infimum over the same feasible set proves both value and regret identities in (18). The objectives differ by the single constant $V^{\mathrm{init}}_{\widetilde{C}}$, so their sets of minimizing path laws coincide. Uniqueness follows from the soft Bellman recursion.

Equivalent dynamic program. Let $f_t(x,a) = h_t(x) + d_t(x,a)$ and $f_T(x) = d_T(x)$, where $D_C = \sum_{t<T} d_t(X_t, A_t) + d_T(X_T)$. The residual soft value satisfies

$$W_T(x) = d_T(x), \tag{61}$$

$$W_t(x) = -\varepsilon \log \sum_{a:\Pi_R(a|x,t)>0} \rho_t(a|x) \exp\left[-\frac{f_t(x,a) + P_t W_{t+1}(x,a)}{\varepsilon}\right]. \tag{62}$$

Then $v_F = \mathbb{E}_{\mu_0} W_0$ and $V_{C,t} = U_t + W_t$. The latter equality holds at T and propagates backward after substituting the displayed formula for ρ_t. The residual Gibbs policy is therefore the original fine optimal policy. In particular, this reformulation has the same asymptotic Bellman complexity; it is not a computational shortcut for obtaining V_C.

E.2 Proof of interval validity and nesting in Theorem 5.2

Because KL is nonnegative,

$$v_F = \inf_Q \{\mathbb{E}_Q F_C + \varepsilon \mathrm{KL}(Q\|Q_\rho)\} \;\geq\; \min_Q \mathbb{E}_Q F_C = a_-.$$

Evaluating the feasible law Q_ρ gives $v_F \leq a_\rho$. For every feasible Q, the reachable-layer bounds imply $m_h \leq \mathbb{E}_Q H \leq M_h$ and the cost extrema imply $d_f^- \leq \mathbb{E}_Q D_C \leq d_f^+$. Hence

$$m_h + d_f^- \;\leq\; a_- \;\leq\; v_F \;\leq\; a_\rho \;\leq\; M_h + d_f^+. \tag{63}$$

The residual identity gives $\mathrm{Reg}_C(Q) = G_C(Q) - v_F$. Subtracting (63) from $G_C(Q)$ proves validity and containment of the raw intervals. Since regret is nonnegative, clipping either lower endpoint at zero preserves validity; monotonicity of $x \mapsto \max\{0, x\}$ preserves containment as well. Subtracting the coupled raw endpoints gives $W_{\mathrm{cpl}} = a_\rho - a_-$. This quantity does not depend on the evaluated candidate Q when its surrogate is held fixed; different declarations may induce different surrogates and hence different widths.

E.3 Complete zero-width characterization

Define the hard residual value by backward induction:

$$v_T^-(x) = d_T(x), \tag{64}$$

$$v_t^-(x) = \min_{a:\Pi_R(a|x,t)>0} \{f_t(x,a) + P_t v_{t+1}^-(x,a)\}. \tag{65}$$

It attains $a_- = \mathbb{E}_{\mu_0} v_0^-$, including among history-dependent feasible policies. The Bellman disadvantage

$$A_t^-(x,a) = f_t(x,a) + P_t v_{t+1}^-(x,a) - v_t^-(x)$$

is nonnegative at every supported row. Telescoping along an arbitrary feasible law gives

$$\mathbb{E}_Q F_C - a_- = \sum_{t<T,x,a} \eta_t^Q(x,a) A_t^-(x,a). \tag{66}$$

For $Q = Q_\rho$, the left side is the raw coupled width. Every structurally reachable state–action row has positive occupancy under Q_ρ: a positive-reference path exists to the state, and all its action factors remain positive under ρ. Thus $W_{\mathrm{cpl}} = 0$ if and only if all supported reachable disadvantages vanish. Taking $v = v^-$ gives exactly (25).

Conversely, suppose potentials satisfy that condition. Summing their defining equality under any feasible law and using $f_T = v_T$ gives $\mathbb{E}_Q F_C = \mathbb{E}_{\mu_0} v_0$, independently of Q. It follows that $a_\rho = a_-$ and v_F equals their common value, since the KL minimum is attained by Q_ρ. The interval is then the singleton $\mathrm{Reg}_C(Q)$, a nonnegative number, so lower clipping is inactive. If instead all feasible residual expectations are equal from the outset, they include both a_- and a_ρ, so zero width follows and the previous argument constructs the potentials. This proves all equivalences.

Policy invariance does not require a deterministic residual. Under the potential condition, the random total residual admits the pathwise decomposition

$$F_C = v_0(X_0) + \sum_{t<T} \big[v_{t+1}(X_{t+1}) - P_t v_{t+1}(X_t, A_t)\big]. \tag{67}$$

Each bracket has zero conditional expectation given the state, action, and past. Thus F_C may remain stochastic, but its mean cannot be changed by any feasible controller. This distinguishes feasible-policy invariance from constancy on every trajectory. The necessity above would fail for a projected policy lacking positive mass on some structurally reachable action; finite costs and positive temperature rule that out here.

Endpoint slack and the meaning of width. Let $Q_C^\star$ be the common original/residual optimizer. The standard soft performance identity applied with reference Q_ρ yields

$$\mathrm{Reg}_C(Q) - \big[G_C(Q) - a_\rho\big] = a_\rho - v_F = \varepsilon \mathrm{KL}(Q_\rho \| Q_C^\star), \tag{68}$$

$$\big[G_C(Q) - a_-\big] - \mathrm{Reg}_C(Q) = v_F - a_-. \tag{69}$$

Hence width is the sum of the projected policy's soft suboptimality and the residual soft-over-hard optimal-value gap. These formulas concern the untruncated lower endpoint. They do not assert that either endpoint is an optimal bound given the entire model. Indeed, solving (62) determines the value exactly at comparable Bellman complexity.

E.4 Potential shifts and an analytical cancellation family

Fix the surrogate, and hence U, h, ρ. For any finite potentials w_t, define an additive shift by $s_t(x,a) = w_t(x) - P_t w_{t+1}(x,a)$ and $s_T(x) = w_T(x)$. Its feasible expectation is $\mathbb{E}_{\mu_0} w_0$ under every policy. Adding it to F_C changes a_-, a_ρ, v_F by that same constant and leaves W_{cpl} unchanged. If this shift is implemented through the evaluation cost C, then $G_C(Q)$ shifts by the same constant, so both certificate endpoints and actual regret remain unchanged.

For the particular admissible choice $d_t(x,a) = -h_t(x)$ and $d_T = 0$, the residual is identically zero. Thus $v_F = a_- = a_\rho = 0$, $Q_C^\star = Q_\rho$, and $\mathrm{Reg}_C(Q) = \varepsilon \mathrm{KL}(Q \| Q_\rho)$ for every feasible law.

Whenever some h_t is nonconstant on R_t, the separate raw width is nevertheless positive. This does not require altering the fixed surrogate to define h; one first fixes the surrogate and then chooses the original evaluation cost $c_t = \widetilde{c}_t - h_t$, $g = \widetilde{g}$. All such costs remain finite.

For an explicit controllable instance, take $T = 2$, $\varepsilon = 1$, a single initial state s, intermediate states x, y, and terminal values $\widetilde{g}(u) = 0$, $\widetilde{g}(v) = 2$. Two uniformly referenced actions at s deterministically choose x or y. At time 1, both actions at x lead deterministically to u, while both actions at y lead equally to u, v. Let $\widetilde{c}_t = 0$ and set $b = -\log[(1 + e^{-2})/2]$, $k = 1 - b = \log\cosh(1) > 0$. Then $U_1(x) = 0$, $U_1(y) = b$, $h_0(s) = 0$, $h_1(x) = 0$, and $h_1(y) = k$. With $d_1 = -h_1$ and $d_0 = d_2 = 0$, the residual is zero, but $\Omega_f = k$ and $\sum_t \mathrm{span}_{R_t}(h_t) = k$, giving $W_{\mathrm{sep}} = 2k > 0 = W_{\mathrm{cpl}}$. The projected initial action law is proportional to $(1, e^{-b})$, so policies with different initial actions can have positive regret despite zero coupled width. This is a mathematical example, not a benchmark or empirical observation.

E.5 Evaluation-cost invariance and homotopy curvature

Theorem E.1 (Equivalent forms of zero feasible transfer span)**.** *Under the finite-support assumptions of Section 2, the following are equivalent:*

1. $\Omega_f = 0$*, that is,* $\mathbb{E}_Q D_C$ *is constant over* $Q \in \mathcal{Q}_D$*.*
2. *There are finite potentials* $w_T = d_T$ *and* $d_t(x, a) + P_t w_{t+1}(x, a) = w_t(x)$ *at every supported reachable row.*
3. *The original and surrogate soft-optimal policies agree at every reachable state.*
4. $\mathrm{Reg}_C(Q) = \mathrm{Reg}_{\widetilde{C}}(Q)$ *for every feasible* Q*.*
5. $V''(\lambda_0) = 0$ *at at least one* $\lambda_0 \in [0, 1]$ *along the cost homotopy.*

When these hold, the homotopy policies are identical for all λ*,* $V(\lambda)$ *is affine, and* $V_C - \widetilde{V} = \mathbb{E}_{\mu_0} w_0$*.*

Proof. The equivalence of the first two conditions is the zero-width argument above applied to D_C rather than F_C, using any policy positive on every reference-supported action. Indeed, subtract the minimum-expectation Bellman value, sum its nonnegative disadvantages under that full-support policy, and use constancy of expectation to force all of them to zero. Conversely the potential identity telescopes to a policy-independent expectation.

If the second condition holds, substituting $V_t^\lambda = V_t^0 + \lambda w_t$ in the soft Bellman equation proves this equality by backward induction. The action-wise shift in the soft Q values is $\lambda w_t(x)$, so it cancels in normalization of the policy. This proves the third condition and the last claims of the theorem. Conversely, if the original and surrogate optimal policies coincide, subtract their supported Gibbs log-policy identities. With $w_t = V_{C,t} - V_{\widetilde{C},t}$, the result is $d_t + P_t w_{t+1} - w_t = 0$ on every supported reachable row, and $w_T = d_T$.

The exact objective identity gives $\mathrm{Reg}_C(Q) - \mathrm{Reg}_{\widetilde{C}}(Q) = \mathbb{E}_Q D_C - (V_C - \widetilde{V})$. It proves that the first condition implies the fourth (the optimum shifts by the same constant), and that the fourth implies the first.

Finally, the second condition makes $V(\lambda)$ affine and hence gives the fifth. Conversely, the curvature formula (56) is the negative sum of nonnegative conditional action variances, weighted by the strictly positive occupancies of $Q^\star_{\lambda_0}$. If it vanishes, every such variance is zero. Positivity of all supported action probabilities then gives $\xi_t^{\lambda_0}(x, a) = m_t^{\lambda_0}(x)$ for every supported reachable action. Since $\xi_t^{\lambda_0} = d_t + P_t m_{t+1}^{\lambda_0}$ and $m_T^{\lambda_0} = d_T$, taking $w_t = m_t^{\lambda_0}$ establishes the second condition. This also handles endpoints, since the finite soft recursion is analytic on a neighborhood of $[0, 1]$. □

E.6 SCREEN CONTAINMENT

For a common pool, let S_{sep} and S_{cpl} be the survivor sets formed from their respective intervals. For each candidate, nesting gives $L_e^{\mathrm{cpl}} \geq L_e$ and $U_e^{\mathrm{cpl}} \leq U_e$. Consequently, if $e \in S_{\mathrm{cpl}}$, then

$$L_e \leq L_e^{\mathrm{cpl}} \leq \min_j U_j^{\mathrm{cpl}} \leq \min_j U_j,$$

so $e \in S_{\mathrm{sep}}$. Thus $S_{\mathrm{cpl}} \subseteq S_{\mathrm{sep}}$ while both retain every true optimum. This statement is deterministic and makes no claim about how many additional candidates would be removed in the supplied empirical pools.

F PROOFS FOR MODEL-ERROR EXTENSIONS

F.1 SHARPER RADII AND SIMULTANEOUS VALIDITY

A policy-specific radius uses the actual estimated cost-to-go rather than a global range. Let $\widehat{u}_t^\pi$ be the regularized cost-to-go of π in $\widehat{D}$ and

$$B_\pi = \kappa_T + \sum_{t<T} \kappa_t + \sum_{t<T} \delta_t \operatorname{span}(\widehat{u}_{t+1}^\pi). \tag{70}$$

Theorem F.1 (Candidate-specific asymmetric padding)**.** *For every supported Markov policy π,*

$$|J_D(\pi) - J_{\widehat{D}}(\pi)| \leq B_\pi \leq B, \tag{71}$$

$$-B \leq V_D - V_{\widehat{D}} \leq B_\star, \tag{72}$$

where $\widehat{\pi}^\star$ is optimal in $\widehat{D}$ and $B_\star = B_{\widehat{\pi}^\star}$. Thus any estimated-model regret interval $[L, U]$ remains valid in the true model after expansion to

$$[\max\{0, L - B_\pi - B_\star\},\ U + B_\pi + B], \tag{73}$$

which is no wider on either side than uniform $2B$ padding.

Corollary F.2 (Direct objective screen)**.** *Intervals $[J_{\widehat{D}}(\pi_e) - B_{\pi_e}, J_{\widehat{D}}(\pi_e) + B_{\pi_e}]$ combined with (26) retain every true objective—and hence regret—minimizer.*

Corollary F.3 (Data-dependent pools)**.** *If the simultaneous row event (27) holds with probability $1 - \zeta$, all preceding intervals hold simultaneously with probability $1 - \zeta$ for arbitrary candidate pools of reference-supported Markov policies selected from the same data. No candidate-count union bound is needed.*

F.2 UNIFORM PADDING

Fix a reference-supported Markov policy π. Its local regularized costs in D and $\widehat{D}$ are

$$\ell_t^\pi(x) = \sum_a \pi_t(a|x) c_t(x,a) + \varepsilon \mathrm{KL}(\pi_t(\cdot|x) \| \Pi_R(\cdot|x,t)), \tag{74}$$

$$\widehat{\ell}_t^\pi(x) = \sum_a \pi_t(a|x) \widehat{c}_t(x,a) + \varepsilon \mathrm{KL}(\pi_t(\cdot|x) \| \Pi_R(\cdot|x,t)). \tag{75}$$

The common regularizer cancels, so $\|\ell_t^\pi - \widehat{\ell}_t^\pi\|_\infty \leq \kappa_t$. Also

$$0 \leq \mathrm{KL}(\pi_t(\cdot|x) \| \Pi_R(\cdot|x,t)) \leq \log(1/p_t),$$

and therefore

$$\operatorname{span}(\widehat{\ell}_t^\pi) \leq \operatorname{span}(\widehat{c}_t) + \varepsilon \log(1/p_t) = \widehat{L}_t.$$

The estimated policy-evaluation recursion and the elementary span inequality $\operatorname{span}(\widehat{P} f) \leq \operatorname{span}(f)$ imply by backward induction that

$$\operatorname{span}(\widehat{u}_{t+1}^\pi) \leq \widehat{S}_{t+1}. \tag{76}$$

Subtract the two policy-evaluation recursions:

$$u_t^\pi - \widehat{u}_t^\pi = (\ell_t^\pi - \widehat{\ell}_t^\pi) + P_t^\pi (u_{t+1}^\pi - \widehat{u}_{t+1}^\pi) + (P_t^\pi - \widehat{P}_t^\pi) \widehat{u}_{t+1}^\pi, \tag{77}$$

where $P_t^\pi(\cdot|x) = \sum_a \pi_t(a|x) P_t(\cdot|x,a)$. Convexity of TV gives $\mathrm{TV}(P_t^\pi(\cdot|x), \widehat{P}_t^\pi(\cdot|x)) \le \delta_t$. Using (29), (76), and backward induction,

$$\|u_t^\pi - \widehat{u}_t^\pi\|_\infty \le \kappa_T + \sum_{s=t}^{T-1} \kappa_s + \sum_{s=t}^{T-1} \delta_s \widehat{S}_{s+1}.$$

Taking the common initial expectation proves $|J_D(\pi) - J_{\widehat{D}}(\pi)| \le B$.

Finite-horizon Markov policies attain both soft optima. Since the objective difference is bounded by B uniformly over policies,

$$|V_D - V_{\widehat{D}}| = \left|\inf_\pi J_D(\pi) - \inf_\pi J_{\widehat{D}}(\pi)\right| \le B.$$

The triangle inequality then gives the regret bound $2B$, and expanding an estimated-model interval by $2B$ preserves coverage. □

F.3 Candidate-specific padding

Returning to (77), use the exact estimated span rather than (76). Backward induction gives

$$\|u_0^\pi - \widehat{u}_0^\pi\|_\infty \le \kappa_T + \sum_{t<T} \kappa_t + \sum_{t<T} \delta_t \,\mathrm{span}(\widehat{u}_{t+1}^\pi) = B_\pi,$$

so (71) follows. Equation (76) immediately implies $B_\pi \le B$ for every supported π, including $\widehat{\pi}^\star$.

Let $\pi_D^\star$ be optimal in the true model. Evaluating the estimated optimum in the true model yields

$$V_D \le J_D(\widehat{\pi}^\star) \le J_{\widehat{D}}(\widehat{\pi}^\star) + B_\star = V_{\widehat{D}} + B_\star.$$

Conversely, the uniform objective bound applied to $\pi_D^\star$ gives

$$V_{\widehat{D}} \le J_{\widehat{D}}(\pi_D^\star) \le J_D(\pi_D^\star) + B = V_D + B.$$

Thus $-B \le V_D - V_{\widehat{D}} \le B_\star$, proving (72).

For a candidate π, write

$$\Delta_J = J_D(\pi) - J_{\widehat{D}}(\pi) \in [-B_\pi, B_\pi], \qquad \Delta_V = V_D - V_{\widehat{D}} \in [-B, B_\star].$$

Then

$$\mathrm{Reg}_D(\pi) - \mathrm{Reg}_{\widehat{D}}(\pi) = \Delta_J - \Delta_V \in [-B_\pi - B_\star,\ B_\pi + B].$$

Adding this interval to any valid estimated-model regret interval gives (73). Since $B_\pi, B_\star \le B$, neither side expands by more than $2B$. □

F.4 Direct objective screening

By (71), $J_D(\pi_e) \in [L_e^J, U_e^J]$ for every candidate. The same interval-elimination argument as in Corollary 5.3 shows that every true objective minimizer survives: if $e^\star$ minimizes J_D and $\bar{e}$ minimizes U_e^J, then

$$L_{e^\star}^J \le J_D(\pi_{e^\star}) \le J_D(\pi_{\bar{e}}) \le U_{\bar{e}}^J.$$

Because V_D is a single constant shared by every candidate, objective and regret minimizers coincide; this is the same observation underlying Proposition A.1, and it does not require the model to be known. □

F.5 High-probability wrapper

Let $\mathcal{E}_{\mathrm{row}}$ be the simultaneous event (27). Conditional on any realized dataset in $\mathcal{E}_{\mathrm{row}}$, the preceding theorems are deterministic and hold for *every* reference-supported Markov policy. This includes policies and finite pools chosen as arbitrary measurable functions of that same dataset. Therefore all candidate intervals and both safe screens are simultaneously valid on $\mathcal{E}_{\mathrm{row}}$. Since $\Pr(\mathcal{E}_{\mathrm{row}}) \ge 1 - \zeta$, the claimed joint probability follows. No union bound over the random candidate pool is needed. □

G Dynamic Programs and Computational Accounting

G.1 Cost-difference extrema

Write $D_C = \sum_{t<T} d_t(X_t, A_t) + d_T(X_T)$. Path-support suffix extrema satisfy

$$\ell_T(x) = u_T(x) = d_T(x), \tag{78}$$

$$\ell_t(x) = \min_{a:\Pi_R(a|x,t)>0} \left\{ d_t(x,a) + \min_{x':P_t(x'|x,a)>0} \ell_{t+1}(x') \right\}, \tag{79}$$

$$u_t(x) = \max_{a:\Pi_R(a|x,t)>0} \left\{ d_t(x,a) + \max_{x':P_t(x'|x,a)>0} u_{t+1}(x') \right\}. \tag{80}$$

Then $d_s^- = \min_{x\in\operatorname{supp}\mu_0} \ell_0(x)$ and $d_s^+ = \max_{x\in\operatorname{supp}\mu_0} u_0(x)$.

Dynamics-feasible expectation extrema replace successor extrema by the fixed transition:

$$m_T(x) = M_T(x) = d_T(x), \tag{81}$$

$$m_t(x) = \min_{a:\Pi_R(a|x,t)>0} \left\{ d_t(x,a) + \sum_{x'} P_t(x'|x,a) m_{t+1}(x') \right\}, \tag{82}$$

$$M_t(x) = \max_{a:\Pi_R(a|x,t)>0} \left\{ d_t(x,a) + \sum_{x'} P_t(x'|x,a) M_{t+1}(x') \right\}. \tag{83}$$

Then $d_f^- = \mu_0^\top m_0$ and $d_f^+ = \mu_0^\top M_0$. Deterministic Markov actions attain these extrema even though $\mathcal{Q}_D$ allows history dependence, by standard finite-horizon backward induction.

G.2 Certificate computation

The computation sequence and complexity are given in Section 5.2. The recursions above supply its feasible cost-transfer extrema and the additional path-support extrema.

G.3 Adaptive homotopy quadrature

The primary report documents Gauss–Legendre quadrature with a configured base order. For a candidate order q, it computes: (i) the occupancy integral in (6); (ii) the direct value difference $V(1) - V(0)$; and (iii) the curvature integral implied by (56). If the independent identities disagree by more than 10^{-10}, the order is doubled up to 512. An unresolved case remains a numerical failure rather than being silently accepted. The requested order, final order, and convergence status are stored per case.

H Experimental Protocol and Detailed Results

H.1 Experimental settings

Table 2. Prespecified finite-model settings and perturbation grids.

Field	Value
Taxi	501 planning states; six actions; horizon 50; $\varepsilon = 1$
Original budgets / offsets	$\{8, 16, 24, 48\}$ / $\{0, \ldots, 9\}$
Positive-width p	$\{1, .9999, .999, .995, .99, .95, .8\}$
Positive-width Ω_f	$\{0, .005, .02, .05, .1, .25, 1, 5\}$
Model-error TV	$\{0, 10^{-8}, 3\times10^{-8}, \ldots, 3\times10^{-5}\}$
Model-error cost error	$\{0, 10^{-6}, 3\times10^{-6}, \ldots, 3\times10^{-4}\}$
Supplementary known-model check	
Taxi	horizon 40; intended probability .95; $\varepsilon = 0.5$
Grid	14×14; horizon 45; intended probability .90; $\varepsilon = 0.4$

Candidate structures, perturbations, seeds, bootstrap counts, and numerical gates were fixed before evaluation. Exact regret was used only for post-hoc assessment, not pool construction or certificate decisions. Screening and model-error results use the primary campaign; the supplementary run contributes the known-model check and path replication. The full registry, software settings, and numerical diagnostics are in the supplementary protocol.

H.2 Algebraic and path-law audits

Table 3. Maximum residuals in the validated theorem audit. All entries satisfy their prespecified gates; the loosest recorded gate is 5×10^{-8}.

Identity	Maximum absolute residual
Declaration bridge	1.886×10^{-13}
Four-term ledger / transfer	4.480×10^{-10}
Raw width / interval containment	7.105×10^{-15}
Homotopy curvature identity	1.960×10^{-8}
Occupancy quadrature (α)	2.445×10^{-9}
Local entropic reconstruction	1.174×10^{-14}
Counterfactual occupancy integral	2.109×10^{-15}
Conditional-KL split	3.220×10^{-15}
Independent path ledger	3.553×10^{-15}
Path versus dynamic-program regret	2.665×10^{-15}

The observed and predicted singleton counts coincide in all random programs:

Table 4. Random-instance exactness phase diagram: observed singletons / total.

	$\Omega_f = 0$	$\Omega_f > 0$
Deterministic dynamics	1,128 / 1,128	0 / 920
Stochastic dynamics	77 / 1,096	0 / 952

Numerical singletons use the registered 10^{-9} tolerance. The 77 stochastic singleton cases agree with the theorem that determinism is not necessary. The path audit enumerates every positive-probability fine and coarse path and does not call the dynamic-program ledger implementation.

H.3 Original pools and positive-width frontier

Table 5. Original no-leakage pool decisions. Gains are post-hoc exact-regret reductions.

Pool family	Pools	Size	Decision	Post-hoc outcome
Geometric transition	40	8–48	unique 40/40	gain 0.060–0.197
Movement actions	1	7	unique 1/1	gain 20.749 (37.1%)
Passenger cost	40	8–48	abstain 40/40	53–73% harmful

Of 392 frontier evaluations, 370 have raw width exceeding the registered 5×10^{-8} endpoint tolerance for every candidate. Among them, 54 uniquely certify, 102 partially eliminate, and 214 abstain. All true pool optima survive and there are no false certifications. The phase diagrams in Figure 2 hold candidate structure fixed while changing stochasticity and cost-transfer variation; thus they directly visualize the two terms in (21).

Table 6. Ten-seed geometric robustness at budget 24. "Unique" and "Informative" are fractions over the ten pools; "Survivor frac." is the mean fraction of surviving *candidates*. The $p = 1, \Omega_f = 0$ row is the zero-width regime and is therefore excluded from the 370 all-positive-width evaluations.

p	Ω_f	Unique	Informative	Survivor frac.	Min-U gain
.8	0	0	0	1.000	0.237489
.8	.1	0	0	1.000	0.237718
.99	0	0	0	1.000	0.164036
.99	.1	0	0	1.000	0.164505
.999	0	0	1	0.917	0.193339
.999	.1	0	1	0.958	0.193826
1	0	1	1	0.042	0.196570
1	.1	0	1	0.083	0.197059

H.4 Uniform model-error sweep

Across all 556 nonzero-error pool evaluations, mean unpadded coverage is 0.004721 and mean uniformly padded coverage is 1. All exact optima remain in the survivor set; 396 pools uniquely certify, 494 are informative, with no false certification or padded-interval violation. This sweep evaluates Theorem 6.1 with prescribed radii; candidate-specific and data-derived extensions remain theoretical.

Table 7. Selected nonzero-error conditions. Coverage is the fraction of true regrets inside the uniformly padded interval.

Condition	TV	Cost err.	Mean $2B$	Coverage / survivor
Geometric—transition	3×10^{-5}	0	2.336694	1 / 1.000
Geometric—cost	0	3×10^{-4}	0.030000	1 / 0.042
Geometric—combined	10^{-7}	3×10^{-4}	0.037789	1 / 0.042
Move-action—transition	3×10^{-5}	0	2.336694	1 / 0.286
Move-action—cost	0	3×10^{-4}	0.030000	1 / 0.143
Move-action—combined	10^{-5}	3×10^{-4}	0.808898	1 / 0.143

H.5 Mechanism, ranking, and known-model boundary

The held-out intervention heatmap is shown in Figure 3.

The objective-unit witness bank obtains held-out single-fault accuracy 1 and mixed active-set recovery 1 over 24 cases. Taxi upper-endpoint Spearman correlation is 0.633911 versus 0.404079 for the uniform score, but the paired 95% bootstrap difference interval $[-0.036505, 0.495652]$ includes zero. Macro correlations across four benchmark families are 0.583333, 0.298395, and 0.565476 for certificate, uniform, and fine-value oracle; the stratified certificate-minus-uniform interval $[-0.094782, 0.689494]$ also includes zero. Navigation certificate correlation is -0.238095. The intervals and navigation result support a setting-dependent interpretation of ranking performance.

The source-audited supplementary run also checks the known-model boundary. On 16 Taxi and 16 grid pools with a known fine model, exact policy evaluation selects a computed pool optimum in all 32. Direct zero-width objective screens return singletons in all 16 Taxi pools and 12 of 16 grid pools; the other four grid pools return numerical ties. The supplied screen keeps candidates within 10^{-12} of the minimum upper endpoint; its post-hoc optimum mask uses 10^{-10}. These outcomes are consistent with Proposition A.1; ties are numerical at the stated tolerances.

I Scope of Supplementary Evidence

The supplementary execution contributes the known-model direct-evaluation check and an independent conditional-KL path replication. The primary campaign supports the separate-envelope screening, uniform-padding, and objective-unit mechanism results in Section H.

Auxiliary certificate, nested-KL diagnostic, and uncertainty outputs do not implement the corresponding mathematical estimands and are excluded from those empirical claims. In particular, they are not used as evidence for candidate-specific padding, comparison with a rectangular robust-control solver, or trajectory-derived confidence radii. These exclusions leave the primary separate-envelope and uniform-padding results unchanged. The supplementary source audit records the implementation-level findings, inspected source hashes, and full execution inventory.

J Reproducibility and Supporting Materials

The supplementary archive contains anonymized source code for the supplementary execution, primary candidate-level records, frozen configurations, manifests, and verification and result-reconstruction code. The evidence map links the reported screening, coverage, mechanism, and known-model results to these records. Full registry counts, software settings, compute accounting, quadrature diagnostics, and execution details are provided in the supplementary protocol and source audit.

A separate NumPy reference implementation checks the soft, free, hard-residual, and policy-evaluation recursions, interval nesting, and potential-invariance cases. These algebraic fixtures complement the proofs in Sections B to F; they are not additional benchmark experiments. Empirical certificate claims concern the separate envelope and uniform padding; the coupled refinement and candidate-specific or data-dependent extensions are supported by their stated proofs.