WHAT ACTION GEOMETRY BUYS: A DESIGN SPACE FOR OPTIMAL-TRANSPORT BELLMAN BACKUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many Reinforcement Learning (RL) action spaces carry geometry that maximum-entropy backups ignore. Laplacian optimal transport (LOT) regularization admits a closed form—an average of cost-shifted local softmax—governed by one anchor-normalized action kernel; we turn this closed form into a design space, proving which kernel property buys which guarantee. Concentration buys planning: exact pruning error and a localized planner replace branching over K actions by an effective count d_{ker}, still targeting the unpruned regularized value. Uniform cost stability buys learning: reusable reward-free geometry and same-stream tabular Q-learning. Poisson truncation buys local computation: a graph-local kernel targets the *exact* heat-kernel value with no explicit K dependence under polynomial growth. Strictly positive couplings buy interpretation: inverse policy curvature is effective resistance on a value-dependent transport graph, reducing to Fisher–Rao at zero cost. Matched obstructions show these properties are not free; choosing a geometry is choosing a row of our map.

1 INTRODUCTION

In RL, many large discrete action spaces have useful structure: discretized controls have nearby torques, motion libraries contain related behaviors, and graph actions inherit local neighborhoods. The Bellman maximum ignores these relations. MaxEnt RL smooths action values, but even with a nonuniform prior it does not encode pairwise action similarity. A geometry-aware backup could instead focus planning on nearby alternatives, reuse transition-derived similarity across reward tasks, and reveal which policy changes the regularizer considers small.

Regularized MDPs provide the general Bellman interface (Geist et al., 2019), while policy mirror descent allows broad convex regularizers (Lan, 2023; Zhan et al., 2023). Optimal transport (OT) is attractive because moving probability between similar actions can be made cheaper, but entropic OT usually requires Sinkhorn iterations (Cuturi, 2013; Peyré & Cuturi, 2019). For the semi-relaxed OT Bellman backup, Dam (2026) showed that optimizing the policy frees one coupling marginal and yields a closed-form mixture of local softmax backups. We take this decomposition as our starting point and make its action kernel operational at the MDP level. We show how properties of the anchor-normalized kernel $w^{(s,j)}$ in (14) yield distinct guarantees for planning, learning, local computation, and policy geometry.

Main contributions: a map of the design space. Each guarantee below is purchased by an identifiable kernel property, and matched obstructions show the properties are not free.

- *Concentration buys error-controlled local planning.* For deterministic action sets fixed from kernel weights before sampling, we derive the exact worst-case soft-backup error and prove that top-weight sets minimize the worst-anchor count d_{ker}. We then give LOCALLOTPLAN, an unclipped recursion with asymmetric child accuracies whose bias, moments, termination, and oracle cost are proved jointly. It samples only retained actions while estimating the scalar *unpruned regularized* root value; sub-K arithmetic additionally requires local kernel access.
- *Dense obstructions versus constructive heat locality.* Uniform MaxEnt and bounded diffusion-distance Gibbs kernels generally remain dense. Under explicit heat-tail and volume-growth assumptions, exact heat columns admit small high-mass neighborhoods but remain entrywise dense. We construct a finite Poisson random-walk kernel with exact graph-ball support and propagate its tail error through the

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.

fixed point and planner. Under polynomial growth and before ball saturation, the resulting estimate of the exact-heat root value has no explicit K dependence.

- *Uniform cost stability buys geometry learning.* We construct reward-free transition geometry and prove linear propagation of cost error to backups and fixed points. With n transitions per action, the batch error is $\widetilde{O}(\kappa_C[t/\sqrt{Kn} + 1/\sqrt{n}])$, with matching $n^{-1/2}$ order on an explicit subclass and a single-trajectory extension. Bounded, predictable, uniformly consistent online costs yield almost-sure tabular Q-learning convergence without summable geometry bias or timescale separation. The target fixed points use P; learned-model error is excluded.
- *Strictly positive couplings buy the dual policy metric.* Building on the Laplacian Hessian, we show that inverse policy curvature is the pseudoinverse of a value-dependent transport co-occurrence Laplacian. The second-order cost of transferring mass from k to i is $\tau\lambda R_Q(i,k)$; at zero cost this becomes the $\tau\lambda$-scaled Fisher–Rao metric. This pointwise identity is not an optimization algorithm.

Section 7 gives controlled numerical illustrations of kernel concentration, local planning, and reward-free geometry learning.

Related work. Our regularizer acts on policy mass, unlike manifold regularization of the value function (Yan et al., 2017) or W-MCTS transport of return distributions (Dam et al., 2025). For planning, randomized multilevel integration can obtain expected action-space-independent complexity under its own stability assumptions (Meunier et al., 2025); our guarantee is a deterministic worst-case oracle bound for geometry-fixed neighborhoods and does not lower-bound randomized methods. For learning, we retain the classical tabular Q-learning setting (Watkins & Dayan, 1992; Jaakkola et al., 1994; Tsitsiklis, 1994); our task differs from plug-in model-based planning (Agarwal et al., 2020) and complements generative-model bounds for regularized RL (Kozuno et al., 2022): we estimate geometry, not a planning model.

Graph spectra have long supplied RL representations and value bases (Mahadevan & Maggioni, 2007; Machado et al., 2017; Wu et al., 2019); related work develops Laplacian and diffusion representations (Wang et al., 2021; Shribak et al., 2024). Here spectral geometry parameterizes the policy regularizer rather than the value representation. Our heat analysis builds on Gaussian and Davies-type decay (Folz, 2011; Horn et al., 2019; Keller & Rose, 2024; Davies, 1993), local heat approximation (Chung & Simpson, 2015; Bai et al., 2024), uniformization (Jensen, 1953), and heat-based transport kernels (Solomon et al., 2015); the new step is their composition with a Bellman fixed point and localized planner. Finally, Fisher–Rao geometry underlies natural policy gradients (Amari, 1998; Kakade, 2002); LOT replaces it locally by effective resistance on a response-dependent graph, distinct from the Laplacian Hessian studied for quadratically regularized OT (Essid & Solomon, 2018).

2 PRELIMINARIES

MDPs and regularized Bellman backups. Consider a discounted Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with $K = |\mathcal{A}|$ actions, transition law P, mean reward r, and discount $\gamma \in (0,1)$. Write $\Delta_{\mathcal{A}} = \{\pi \in \mathbb{R}^K_{\geq 0} : \mathbf{1}^\top \pi = 1\}$ for the action-probability simplex and $Q = Q(s,\cdot) \in \mathbb{R}^K$ for the action values at state s. For a convex penalty Ω_s, including its regularization scale,

$$F_s^\Omega(Q) = \max_{\pi \in \Delta_{\mathcal{A}}} \{\langle \pi,\, Q\rangle - \Omega_s(\pi)\}, \tag{1}$$
$$(\mathcal{T}^\Omega Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[F_{s'}^\Omega(Q(s',\cdot))].$$

Entropy-regularized backup. For $\mathsf{H}(\pi) = -\sum_i \pi_i \log \pi_i$ and $T_0 > 0$,

$$F_{T_0}^{\mathrm{ent}}(Q) = \max_{\pi \in \Delta_{\mathcal{A}}} \{\langle \pi,\, Q\rangle + T_0 \mathsf{H}(\pi)\} = T_0 \log \sum_i e^{Q_i/T_0}. \tag{2}$$

Optimal-transport Bellman smoothing. Following Dam (2026), fix $\tau, \lambda > 0$, and let $\mu_s \in \Delta_{\mathcal{A}}$ be a reference distribution and $C_s \in \mathbb{R}^{K\times K}_{\geq 0}$ an action cost. For the coupling set $\Pi(\pi, \mu_s) := \{\Gamma \in \mathbb{R}^{K\times K}_{\geq 0} : \Gamma\mathbf{1} = \pi,\ \Gamma^\top \mathbf{1} = \mu_s\}$, define

$$W_\lambda(\pi, \mu_s) = \min_{\Gamma \in \Pi(\pi,\mu_s)} \left\{ \langle \Gamma,\, C_s\rangle + \lambda \sum_{ij} \Gamma_{ij}(\log \Gamma_{ij} - 1) \right\}. \tag{3}$$

The OT aggregator and its value-normalized form, with $T_0 = \tau\lambda$, are

$$F_s^{\mathrm{OT}}(Q) = \max_{\pi \in \Delta_{\mathcal{A}}} \{\langle \pi,\, Q\rangle - \tau W_\lambda(\pi, \mu_s)\}, \tag{4}$$

$$\overline{F}_s^{\mathrm{OT}}(Q) := F_s^{\mathrm{OT}}(Q) - T_0(1 + \mathsf{H}(\mu_s)) = T_0 \sum_j \mu_s(j) \log \sum_i \exp\left(\frac{Q_i - \tau C_{s,ij}}{T_0}\right). \tag{5}$$

Thus $\overline{F}_s^{\mathrm{OT}}\Big|_{C_s=0} = F_{T_0}^{\mathrm{ent}}$.

Graph Laplacian and spectral distances. For an undirected action graph with symmetric nonnegative edge-weight matrix W, let $D = \mathrm{diag}(W\mathbf{1})$ and $L = D - W = \sum_k \lambda_k \phi_k \phi_k^\top$, with orthonormal eigenvectors ϕ_k and $0 = \lambda_1 \le \lambda_2 \le \cdots$. We use e_i for a coordinate vector and † for the Moore–Penrose pseudoinverse. A nonnegative spectral filter g defines

$$C_{ij}^L = \sum_k g(\lambda_k)(\phi_k(i) - \phi_k(j))^2. \tag{6}$$

For connected G, $g_{\mathrm{ER}}(0) = 0$ and $g_{\mathrm{ER}}(\lambda) = 1/\lambda$ yield the effective resistance $(e_i - e_j)^\top L^\dagger (e_i - e_j)$ Klein & Randić (1993); on disconnected graphs this is only a spectral pseudodistance across components. The filter $g_t(\lambda) = e^{-2t\lambda}$ yields squared diffusion distance from $H_t = e^{-tL}$.

This fixed spectral cost, used by LOT in Section 3, differs from the later effective-resistance metric induced by a Q-dependent transport graph. Moreover, Gibbs similarities $w_i^{(s,j)} \propto \rho_s(i) e^{-C_{s,ij}/\lambda}$ may remain dense even when direct heat columns $w^{(s,j)} = H_t(:, j)$ concentrate locally.

3 The LOT Bellman Backup

The Laplacian optimal-transport (LOT) backup extends the OT-smoothed backup by replacing ordinary coupling entropy with relative entropy with respect to $\rho_s \otimes \mu_s$ and using a graph-derived cost C_s. In a coupling $\Gamma \in \Pi(\pi, \mu_s)$, row i represents candidate action i and column j reference action j; ρ_s is the candidate-action prior and μ_s weights the reference actions. The kernel-relative transport penalty is

$$W_{\lambda,s}^{\rho}(\pi, \mu_s) := \min_{\Gamma \in \Pi(\pi, \mu_s)} \left\{ \langle \Gamma,\, C_s\rangle + \lambda \sum_{ij} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho_s(i)\mu_s(j)} \right\}, \tag{7}$$

with $0 \log 0 = 0$; the second term is $\lambda D_{\mathrm{KL}}(\Gamma \| \rho_s \otimes \mu_s)$. With $T_0 = \tau\lambda$, define

$$F_s^{\mathrm{LOT}}(Q) = \max_{\pi \in \Delta_{\mathcal{A}}} \left\{ \langle \pi,\, Q\rangle - \tau W_{\lambda,s}^{\rho}(\pi, \mu_s) \right\}. \tag{8}$$

Costs may come from an action graph or dynamics-derived features; direct heat instead supplies normalized similarity columns.

Assumption 1 (Full-support reference and action prior)**.** For every s, $\mu_s, \rho_s \in \Delta_{\mathcal{A}}$ satisfy $\min_a \mu_s(a) \ge \mu_{\min} > 0$ and $\min_a \rho_s(a) \ge \rho_{\min} > 0$.

Closed form. Optimizing the row marginal $\pi = \Gamma\mathbf{1}$ leaves only $\Gamma^\top \mathbf{1} = \mu_s$ fixed, so the problem separates over columns (Appendix F).

Lemma 1 (Closed-form LOT backup)**.** *Fix s, let $\tau, \lambda > 0$, let ρ_s, μ_s have full support, and let $C = C_s$ be entrywise finite. With $\mathsf{K}_{C,ij} = e^{-C_{ij}/\lambda}$, $u_i = \rho_s(i) e^{Q_i/(\tau\lambda)}$, and $v_j = \mu_s(j)/(\mathsf{K}_C^\top u)_j$, the optimal plan and maximizer are $\Gamma^\star_{ij} = u_i \mathsf{K}_{C,ij} v_j$ and $\pi^{\mathrm{LOT}} = \Gamma^\star\mathbf{1}$, and*

$$F_s^{\mathrm{LOT}}(Q) = \tau\lambda \sum_j \mu_s(j) \log Z_j(Q), \qquad Z_j(Q) = \sum_i \rho_s(i) \exp\left(\frac{Q_i - \tau C_{ij}}{\tau\lambda}\right). \tag{9}$$

Thus no iterative transport balancing is required.

Theorem 1 (Mixture of local softmax backups)**.** *Fix s, write $C = C_s$, and use $Z_j(Q)$ from* (9)*. Then $\nabla_Q F_s^{\mathrm{LOT}}(Q) = \pi^{\mathrm{LOT}}(Q)$, where $p^{(j)}$ is the local softmax of anchor j and*

$$p_i^{(j)}(Q) = \rho_s(i) \exp\big((Q_i - \tau C_{ij})/(\tau\lambda)\big)/Z_j(Q), \tag{10}$$

$$\Gamma^\star_{ij} = \mu_s(j) p_i^{(j)}(Q), \qquad \pi^{\mathrm{LOT}}(Q) = \sum\nolimits_j \mu_s(j) p^{(j)}(Q), \tag{11}$$

$$F_s^{\mathrm{LOT}}(Q) = \sum\nolimits_j \mu_s(j) f_j(Q), \qquad f_j(Q) = \tau\lambda \log Z_j(Q), \tag{12}$$

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \tfrac{1}{\tau\lambda} \sum\nolimits_j \mu_s(j) \big[\mathrm{diag}(p^{(j)}) - p^{(j)} p^{(j)\top}\big], \tag{13}$$

so F_s^{LOT} is $1/(2\tau\lambda)$-smooth in Euclidean norm.

The covariance mixture (13) does more than certify smoothness: it also encodes the metric that the optimized regularizer imposes on policy space. We return to this dual geometry in Section 6, after the planning and learning results that use the kernel as an input.

MaxEnt limit and ordering. At $C_s = 0$, LOT becomes prior-normalized MaxEnt at temperature T_0; for uniform ρ_s, its value equals the usual log-sum-exp minus the essential offset $T_0 \log K$. For brevity we write $F_s^C(Q) = F_s^{\mathrm{LOT}}(Q)$. For bounded nonnegative costs, ordinary uncentered LOT has values no greater than matched prior-normalized MaxEnt (Proposition 4, Appendix D.2). Here "matched" means using the same action prior ρ_s and temperature $T_0 = \tau\lambda$. The corresponding prior-normalized MaxEnt backup is $F_s^\rho(Q) = T_0 \log \sum_i \rho_s(i) e^{Q_i/T_0}$. This is value ordering, not policy-return ordering.

Under Assumption 1 and bounded costs, F_s^{LOT} is monotone and 1-Lipschitz in sup norm, so its Bellman operator $\mathcal{T}^{\mathrm{LOT}}$ ((1) with $\Omega_s = \tau W^\rho_{\lambda,s}$) is a γ-contraction with unique fixed point $Q^\star_{\mathrm{LOT}}$; Proposition 3 gives the hard-max bias bound.

Implementation. All constructions use the normalized-kernel form

$$\mathcal{F}_s(Q) = \sum_j \mu_s(j) \left[b_{s,j} + T_0 \log \sum_i w_i^{(s,j)} e^{Q_i/T_0} \right], \qquad w^{(s,j)} \in \Delta_{\mathcal{A}}, \tag{14}$$

Here i indexes candidate actions and j anchors. For cost-based LOT with $C_s \geq 0$ and $C_{s,jj} = 0$, set $\kappa_{s,j} = \sum_i \rho_s(i) e^{-C_{s,ij}/\lambda}$, $w_i^{(s,j)} = \rho_s(i) e^{-C_{s,ij}/\lambda}/\kappa_{s,j}$, and $b_{s,j} = T_0 \log \kappa_{s,j}$. Thus $w^{(s,j)}$ is the normalized prior-weighted similarity distribution, and $b_{s,j}$ restores the normalization so that $\mathcal{F}_s$ equals ordinary LOT exactly. Centered direct heat instead uses $w^{(s,j)} = H_t(:, j)$ and $b_{s,j} = 0$, since each heat column is already a probability distribution.

Algorithm 1 gives the numerically stable implementation.

4 Local Planning from Action Geometry

We ask two questions. First, when can a geometry-fixed action neighborhood replace branching over all K actions with controlled error? Second, when can that neighborhood and its weights be computed locally? We answer the first with an exact pruning rule and LocalLOTPlan, and the second with a finite-walk heat kernel. Detailed schedules, pseudocode, and proofs are in Appendices G.8 and C.

4.1 Exact pruning and localized planning

The second-order planner of Dam (2026) may build a baseline for all K actions. We instead retain, before transition sampling, the largest kernel weights around each anchor. The total error separates as

$$|\widehat{V} - V^\star_{\mathcal{F}}| \leq \underbrace{|\widehat{V} - V^\star_{\mathcal{I}}|}_{\text{estimation}} + \underbrace{|V^\star_{\mathcal{I}} - V^\star_{\mathcal{F}}|}_{\text{pruning}},$$

and we assign $\varepsilon/2$ to each term. All values are evaluated at the queried root state s_0. Here $\widehat{V}$ is the estimate, $V^\star_{\mathcal{F}}$ is the exact unpruned regularized value, and $V^\star_{\mathcal{I}}$ is the exact value for fixed retained

sets $\mathcal{I} = \{I_{s,j}\}_{s,j}$. The two terms are finite-sample estimation error and deterministic pruning bias. These values satisfy $Q^\star_{\mathcal{F}}(s,a) = r(s,a) + \gamma \mathbb{E}_{S' \sim P(\cdot|s,a)}[V^\star_{\mathcal{F}}(S')], V^\star_{\mathcal{F}}(s) = \mathcal{F}_s(Q^\star_{\mathcal{F}}(s,\cdot))$. The pruned values are defined analogously using $F^{\mathcal{I}}_s$. The target is the scalar *unpruned regularized* root value, not a best action or policy return.

For $w \in \Delta_{\mathcal{A}}$ and $I \subseteq \mathcal{A}$, write $w(I) = \sum_{i \in I} w_i$ and $w(I^c) = 1 - w(I)$. For retained sets $\mathcal{I} = \{I_{s,j}\}_{s,j}$, define

$$f^{\mathcal{I}}_{s,j}(Q) = b_{s,j} + T_0 \log \sum_{i \in I_{s,j}} w_i^{(s,j)} e^{Q_i/T_0}, \qquad F^{\mathcal{I}}_s(Q) = \sum_j \mu_s(j) f^{\mathcal{I}}_{s,j}(Q). \tag{15}$$

The weights are not renormalized here, so $F^{\mathcal{I}}_s \le \mathcal{F}_s$ and the pruning error has a known sign.

Definition 1 (Kernel effective action count)**.** For $\alpha \in (0,1)$, let

$$d_{\mathrm{ker}}(\alpha) = \sup_{s,j} \min\{|I| : w^{(s,j)}(I) \ge 1 - \alpha\}. \tag{16}$$

It is the largest, over anchors, of the smallest action counts retaining mass $1 - \alpha$; it is not a union of the anchor neighborhoods.

The next lemma converts discarded kernel mass into pruning error, uniformly over the span $\mathrm{osc}(Q) := \max_i Q_i - \min_i Q_i$.

Lemma 2 (Sharp worst-case error from pruning a fixed action set)**.** *Fix $w \in \Delta_{\mathcal{A}}$, $T_0 > 0$, and a nonempty retained set $I \subseteq \mathcal{A}$. Let $f_w(Q) = T_0 \log \sum_i w_i e^{Q_i/T_0}$ and $f^I_w(Q) = T_0 \log \sum_{i \in I} w_i e^{Q_i/T_0}$, where the retained weights are not renormalized, and let $\alpha := w(I^c) = \sum_{i \notin I} w_i < 1$. Then the worst-case pruning error over action-value vectors with span at most S is*

$$\sup_{\mathrm{osc}(Q) \le S} \big[f_w(Q) - f^I_w(Q) \big] = T_0 \log\big(1 + e^{S/T_0} \tfrac{\alpha}{1-\alpha}\big), \tag{17}$$

attained when all retained actions share one value and every discarded action exceeds it by S. Consequently, for a one-backup error tolerance $\xi > 0$, the fixed set I guarantees

$$0 \le f_w(Q) - f^I_w(Q) \le \xi \qquad \text{for all } Q \text{ with } \mathrm{osc}(Q) \le S \tag{18}$$

if and only if its discarded kernel weight satisfies

$$w(I^c) \le \alpha^\star(S,\xi) := \frac{e^{\xi/T_0} - 1}{e^{S/T_0} + e^{\xi/T_0} - 1}. \tag{19}$$

For each state–anchor kernel, top-weight sets meet this requirement with the smallest cardinality among deterministic, Q-independent choices. Hence the smallest possible worst-anchor action count is

$$d_{\mathrm{fixed}}(S,\xi) = d_{\mathrm{ker}}\big(\alpha^\star(S,\xi)\big). \tag{20}$$

Averaging over anchors and using Bellman contraction gives

$$0 \le V^\star_{\mathcal{F}}(s) - V^\star_{\mathcal{I}}(s) \le \frac{\xi}{1-\gamma}. \tag{21}$$

Hence $\xi = (1-\gamma)\varepsilon/2$ limits pruning error to $\varepsilon/2$.

Assumption 2 (Local-kernel access)**.** Given (s, j, α), a geometry routine returns a set with omitted mass at most α, its retained weights, and sampling access to the restricted Gibbs law (the local softmax conditioned on the retained set) in $\widetilde{O}(|I_{s,j}|)$ time, without scanning all K actions; it also samples $J \sim \mu_s$ in $\widetilde{O}(1)$ time after preprocessing. This is needed only for sub-K arithmetic, not for the transition-sample guarantee.

The recursive estimator uses a second-order bias correction and median-of-means confidence amplification; details and pseudocode are in Appendix B.

Theorem 2 (LocalLOTPlan guarantee)**.** *Assume $T_0 > 0$, $\gamma < 1$, $|R| \le R_{\max}$, an independent generative model (each call at (s,a) returns a fresh reward R and successor $S' \sim P(\cdot \mid s,a)$, with $\mathbb{E}[R \mid s,a] = r(s,a)$), and $B_F = \sup_s |\mathcal{F}_s(0)| < \infty$. Put $\bar{B}_Q = \frac{R_{\max} + \gamma(B_F + T_0 \log 2)}{1-\gamma}$, $S_Q = 2\bar{B}_Q$.*

There are $\varepsilon_0, C, c > 0$, depending only on the displayed structural parameters and the fixed recursion choices, such that, for $0 < \varepsilon \le \varepsilon_0$ and $\delta \in (0,1)$, define $\alpha_\varepsilon = \min\left\{\frac{1}{2}, \frac{e^{(1-\gamma)\varepsilon/(2T_0)}-1}{e^{S_Q/T_0}+e^{(1-\gamma)\varepsilon/(2T_0)}-1}\right\}$. Algorithm 2, with the recursive sample of Algorithm 3, satisfies

$$\Pr\left\{|\widehat{V}_{\mathrm{LLP}}(s_0) - V^\star_{\mathcal{F}}(s_0)| \le \varepsilon\right\} \ge 1-\delta \tag{22}$$

using at most N_{plan} generative-model calls, where

$$N_{\mathrm{plan}} \le \frac{C}{\varepsilon^3}[C d_{\mathrm{ker}}(\alpha_\varepsilon)]^{c[1+\log\log(e+\bar{B}_Q/\varepsilon)]}\log\frac{2}{\delta}. \tag{23}$$

No transition is sampled outside the retained set of the sampled anchor; sub-K arithmetic additionally requires Assumption 2.

Geometry replaces dense dependence on K by $d_{\mathrm{ker}}(\alpha_\varepsilon)$. The guarantee concerns the regularized value; hard-max bias is controlled separately by Propositions 4 and 3.

4.2 Constructive heat-kernel planning

To turn transition-query savings into arithmetic savings, the planner must find the relevant actions and their weights without first scanning all K actions. We obtain such weights from a sparse action-similarity graph $G = (\mathcal{A}, E, W)$, which is distinct from the MDP transition model $P(s' \mid s, a)$. Assume that G is connected and undirected, let $L = D - W$ be its Laplacian, and fix $t > 0$. Starting one unit of mass at anchor action j and diffusing it for time t gives

$$H_t = e^{-tL}, \qquad H_t(i,j) = \text{diffusion mass from anchor } j \text{ to candidate } i. \tag{24}$$

Each column $H_t(:,j)$ is a probability distribution over candidate actions. Small t concentrates most mass near j, whereas larger t spreads it farther. Unlike Laplacian representations over states (Mahadevan & Maggioni, 2007; Machado et al., 2017; Wu et al., 2019), here the graph relates actions and directly defines their regularization.

At state s, the distribution μ_s selects reference anchors. For anchor j, the factor $H_t(i,j)$ favors candidates near j, while e^{Q_i/T_0} favors candidates with high value. Averaging these anchor-specific soft decisions gives the centered heat backup

$$F^{\mathrm{HK}}_{s,t}(Q) = T_0 \sum_j \mu_s(j) \log \sum_i H_t(i,j) e^{Q_i/T_0}. \tag{25}$$

Thus t controls geometric reach, whereas T_0 controls sensitivity to value differences. With a uniform action prior and $C^{\mathrm{HK}}_{ij} = -\lambda \log H_t(i,j)$, $F^{\mathrm{HK}}_{s,t} = F^{\mathrm{LOT}}_s + T_0 \log K$. This centering makes $F^{\mathrm{HK}}_{s,t}(0) = 0$ and leaves the policy and derivatives unchanged (Solomon et al., 2015).

Although G is sparse, exact heat is dense: it includes graph walks of every length and therefore assigns positive mass to every action pair. To obtain finite local support, choose $\nu_{\mathrm{u}} > 0$ with $\nu_{\mathrm{u}} \ge d_{\max} := \max_i D_{ii}$, and set $P_G = I - L/\nu_{\mathrm{u}}, \theta = \nu_{\mathrm{u}} t$. Here P_G is a random walk on the action graph, not an MDP transition kernel. Classical uniformization (Jensen, 1953) gives $H_t = \sum_{k\ge 0} a_k P_G^k, a_k = e^{-\theta}\theta^k/k!$. Equivalently, $H_t(:,j)$ is the endpoint distribution obtained by drawing $N \sim \mathrm{Pois}(\theta)$ and taking N graph-walk steps from anchor j. Conditioning on $N \le r$, for $r \in \mathbb{N}_0$, gives

$$H_{t,r} = \frac{1}{s_r}\sum_{k=0}^{r} a_k P_G^k, \qquad s_r = \sum_{k=0}^{r} a_k, \qquad \beta_r = \Pr(N > r). \tag{26}$$

Let d_G be shortest-path distance in hops along positive-weight edges. For $B_G(j,r) = \{i : d_G(i,j) \le r\}$, $\mathrm{supp}(H_{t,r}(:,j)) = B_G(j,r)$. Hence $H_{t,r}$ is computable from the r-hop neighborhood alone. It is not exact heat cropped to that ball: it discards all walks longer than r, including long walks that return to the ball. Let $F^{\mathrm{HK}}_{s,t,r}$ denote (25) with H_t replaced by $H_{t,r}$.

For each action, define $\mathcal{N}_G(i) := \{\ell \ne i : W_{\ell i} > 0\}, \deg_G(i) := |\mathcal{N}_G(i)|, \Delta_G := \max_i \deg_G(i)$. Thus Δ_G is the maximum number of graph neighbors, whereas $d_{\max} = \max_i D_{ii}$ is the maximum weighted degree.

Theorem 3 (Local heat approximation)**.** *Assume local weighted-adjacency access: querying a vertex returns its weighted degree and weighted neighbor list in time proportional to the list size. Let Δ_G be the maximum number of positive-weight neighbors of any action, as defined above, and let $d_r = \max_j |B_G(j, r)|$ be the largest number of actions in an r-hop ball. A column of $H_{t,r}$ is constructed within its r-hop ball using $O((r+1)(\Delta_G+1)d_r)$ graph arithmetic, and*

$$\max_j \|H_t(:, j) - H_{t,r}(:, j)\|_1 \le 2\beta_r. \tag{27}$$

For $\delta_r(S) = T_0 \log(1 + \beta_r(e^{S/T_0} - 1))$,

$$\sup_{\mathrm{osc}(Q)\le S} |F^{\mathrm{HK}}_{s,t}(Q) - F^{\mathrm{HK}}_{s,t,r}(Q)| \le \delta_r(S). \tag{28}$$

Write $(Q^\star_t, V^\star_t)$ and $(Q^\star_{t,r}, V^\star_{t,r})$ for the action/state fixed-point values under exact and truncated heat, respectively. If $\sup_s \mathrm{osc}(Q^\star_t(s, \cdot)) \le S$, then

$$\|Q^\star_{t,r} - Q^\star_t\|_\infty \le \frac{\gamma\delta_r(S)}{1-\gamma}, \qquad \|V^\star_{t,r} - V^\star_t\|_\infty \le \frac{\delta_r(S)}{1-\gamma}. \tag{29}$$

Thus β_r is the probability of omitted long walks, $\delta_r(S)$ is the resulting one-backup error, and Bellman recursion amplifies it to at most $\delta_r(S)/(1-\gamma)$ in state value.

Corollary 1 (Local planning for the exact heat value)**.** *Under the assumptions of Theorem 2, let $S_{\mathrm{h}} = 2(R_{\max} + \gamma T_0 \log 2)/(1-\gamma)$,*

$$b_\varepsilon = \min\Big\{\tfrac{1}{2}, \tfrac{\exp((1-\gamma)\varepsilon/(2T_0))-1}{\exp(S_{\mathrm{h}}/T_0)-1}\Big\}, \qquad r_\varepsilon = \big\lceil \max\{1, 2e\nu_{\mathrm{u}}t, \log_2(1/b_\varepsilon)\}\big\rceil - 1, \tag{30}$$

and $d_\varepsilon = \max_j |B_G(j, r_\varepsilon)|$. The localized estimator on H_{t,r_ε} returns $|\widehat{V}(s_0) - V^\star_t(s_0)| \le \varepsilon$ with probability $1-\delta$ (statistical error and heat bias each at most $\varepsilon/2$) using at most N_{heat} generative-model calls:

$$|\widehat{V} - V^\star_t| \le |\widehat{V} - V^\star_{t,r_\varepsilon}| + |V^\star_{t,r_\varepsilon} - V^\star_t| \le \varepsilon/2 + \varepsilon/2, \tag{31}$$

$$N_{\mathrm{heat}} \le C\varepsilon^{-3}(Cd_\varepsilon)^{c[1+\log\log(e+1/\varepsilon)]} \log(2/\delta). \tag{32}$$

Under polynomial volume growth $|B_G(j, r)| \le C_{\mathrm{vol}}(r+1)^{d_g}$ (fixed $C_{\mathrm{vol}} \ge 1$, $d_g \ge 0$) and $\nu_{\mathrm{u}}t = O(1)$, before ball saturation $N_{\mathrm{heat}} = (1/\varepsilon)^{3+o(1)} \log(2/\delta)$, with no explicit dependence on K.

When does heat help? Under the Gaussian heat-tail and polynomial volume-growth assumptions of Theorem 6, $d^{\mathrm{HK}}_{\mathrm{ker}}(\alpha) = O((\log(1/\alpha))^{d_g/2})$, provided the selected radius remains within the Gaussian-tail range and the graph balls have not saturated. By contrast, uniform MaxEnt and uniform-prior Gibbs kernels with $0 \le C_{s,ij} \le C_{\max}$ satisfy

$$d^{\mathrm{MaxEnt}}_{\mathrm{ker}}(\alpha) = \lceil (1-\alpha)K \rceil, \qquad d^{\mathrm{Gibbs}}_{\mathrm{ker}}(\alpha) \ge K(1 - \alpha e^{C_{\max}/\lambda})_+.$$

Here $(x)_+ = \max\{x, 0\}$ and α is an exact-kernel tail budget, not the Poisson tail β_r. Heat can therefore replace global branching by local branching when graph balls grow slowly and remain unsaturated. The gain disappears when t or the required radius r makes a ball cover all actions. These fixed-neighborhood comparisons are not lower bounds for adaptive planners, and the result estimates a scalar regularized root value rather than proving best-action identification or policy return. The complete local-column construction, heat-specialized recursion, cache interface, and outer median-of-means procedure are given in Appendix C.4.

5 Learning Action Geometry from Transition Samples

The preceding planner treated the action cost C as known. We now use transition samples only to learn *which actions have similar effects*, through $\widehat{C}$. We never replace the true dynamics P inside a Bellman backup by an estimated model, so this is geometry learning, not model learning. For a cost matrix A write $\|A\|_{\max} := \max_{a,b} |A_{ab}|$, and let $\eta := \sup_s \|\widehat{C}_s - C_s\|_{\max}$ be the largest cost-estimation error over all states and action pairs. Let $\mathcal{T}^C$ denote the LOT Bellman operator with cost family $(C_s)_s$, and let $Q^\star_C$ be its fixed point. The resulting planning error satisfies

$$\sup_{s,Q} \big|F^{\widehat{C}}_s(Q) - F^C_s(Q)\big| \le \tau\eta, \quad \sup_Q \big\|\mathcal{T}^{\widehat{C}}Q - \mathcal{T}^C Q\big\|_\infty \le \gamma\tau\eta, \quad \big\|Q^\star_{\widehat{C}} - Q^\star_C\big\|_\infty \le \tfrac{\gamma\tau}{1-\gamma}\eta. \tag{33}$$

These bounds follow from unit-mass couplings and Bellman contraction; Theorem 18 also controls the induced policy error.

Reward-free batch geometry. Fix a reward-independent distribution ν over starting states and assume resettable generative access. For every action a, draw n transitions $S_{a\ell} \sim \nu$ and $S'_{a\ell} \sim P(\cdot \mid S_{a\ell}, a)$, independently across (a, ℓ). With e_s denoting the one-hot vector of state s, define $x_{a\ell} := e_{S'_{a\ell}} - e_{S_{a\ell}}$, $z_a := \mathbb{E}[x_{a1}]$, $\widehat{z}_a := \frac{1}{n}\sum_{\ell=1}^{n} x_{a\ell}$, $L_\nu := \frac{1}{K}\sum_a \mathbb{E}\big[x_{a1}x_{a1}^\top/2\big]$ and $\widehat{L}_\nu := \frac{1}{Kn}\sum_{a,\ell} x_{a\ell}x_{a\ell}^\top/2$. Here z_a is the average occupancy change produced by action a, and L_ν pools undirected transition edges into a state-space diffusion geometry (Appendix D.1).

For diffusion time $t \geq 0$ and cost scale $\kappa_C > 0$, define

$$C_{ab}^t := \kappa_C \big\| e^{-tL_\nu}(z_a - z_b) \big\|_2^2, \qquad \widehat{C}_{ab}^t := \kappa_C \big\| e^{-t\widehat{L}_\nu}(\widehat{z}_a - \widehat{z}_b) \big\|_2^2. \tag{34}$$

These costs are shared across states: $C_s = C^t$ and $\widehat{C}_s = \widehat{C}^t$ for every s. A small C_{ab}^t means similar smoothed, ν-averaged effects, not identical state-conditional transition kernels. This reward-free geometry can be reused across reward tasks sharing the dynamics, ν, and geometry settings.

Theorem 4 (Two guarantees for learning action geometry)**.** *Assume finite state and action sets and fixed full-support distributions ρ_s, μ_s for every s.* (i) Frozen reward-free geometry. *For deterministic $t \geq 0$, $\kappa_C > 0$, and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\big\| \widehat{C}^t - C^t \big\|_{\max} \leq \varepsilon_{n,\delta} = \widetilde{O}\Big(\kappa_C \big[\tfrac{t}{\sqrt{Kn}} + \tfrac{1}{\sqrt{n}}\big]\Big), \qquad \big\| Q^\star_{\widehat{C}^t} - Q^\star_{C^t} \big\|_\infty \leq \tfrac{\gamma\tau}{1-\gamma}\varepsilon_{n,\delta}, \tag{35}$$

where $\widetilde{O}$ hides log factors; the $t/\sqrt{Kn}$ and $1/\sqrt{n}$ terms come from estimating the shared Laplacian L_ν from all Kn transitions and each action signature z_a from its own n samples, respectively. The event involves only transitions, so the same $\widehat{C}^t$ and bound hold simultaneously for every bounded-reward task sharing P, ν, the priors, and the regularization parameters (both fixed points use the true P).

(ii) Predictable same-stream geometry. *Here t indexes online updates rather than diffusion time. Under Assumption 3 (bounded rewards and nonnegative costs, every (s, a) visited infinitely often, Robbins–Monro step sizes α_t), let Q_t be the Q-learning table of Algorithm 8, updated with $\widehat{C}_t = (\widehat{C}_{s,t})_s$ computed before the current transition. From any finite initialization, geometry consistency* (36) *implies $\|Q_t - Q^\star_C\|_\infty \to 0$ a.s. and, for every s, convergence* (37) *of the induced LOT policy, i.e. the smooth backup's action distribution $\nabla_Q F_s$:*

$$\sup_s \big\| \widehat{C}_{s,t} - C_s \big\|_{\max} \xrightarrow{\text{a.s.}} 0, \tag{36}$$

$$\big\| \nabla_Q F_s^{\widehat{C}_{s,t}}\big(Q_t(s,\cdot)\big) - \nabla_Q F_s^{C_s}\big(Q^\star_C(s,\cdot)\big) \big\|_1 \xrightarrow{\text{a.s.}} 0. \tag{37}$$

No stronger summability condition ($\sum_t \alpha_t \sup_s \|\widehat{C}_{s,t} - C_s\|_{\max} < \infty$) or separate geometry/value timescale is needed.

Sharpness and a single trajectory. Part (i) gives a finite-sample guarantee for frozen geometry; part (ii) gives asymptotic convergence for simultaneous updates. The $n^{-1/2}$ order is minimax-sharp on an explicit subclass (Theorem 11); Theorem 19 extends geometry learning to one trajectory under mixing and coverage. Further details are in Appendix D.1.

Predictable online implementation. At iteration t, the cost estimate uses only transitions observed before t; the current transition enters the geometry estimate only after forming the temporal-difference target. This past-measurability is the meaning of *predictable*. The construction and OnlineLOTQ loop are in Appendix D.1.

6 The Dual Policy Geometry: Effective Resistance

Sections 4 and 5 used the kernel as an *input*: geometry to be exploited by the planner or estimated from data. We now identify the *output* geometry—the metric the optimized regularizer imposes on policy space—and it is not the input graph.

The Hessian in (13) also determines the local geometry of the LOT regularizer. Let $\Omega_s(\pi) := \tau W^\rho_{\lambda,s}(\pi, \mu_s)$, $\pi_Q := \nabla_Q F_s^{\mathrm{LOT}}(Q)$, and let Γ_Q be the optimal coupling at Q. Define the *transport co-occurrence graph*

$$A_Q = \Gamma_Q \operatorname{diag}(\mu_s)^{-1}\Gamma_Q^\top, \qquad L_Q = \operatorname{diag}(A_Q\mathbf{1}) - A_Q. \tag{38}$$

Thus $(A_Q)_{ik}$ is large when actions i and k receive transport mass through the same anchors; this graph is induced by the current optimal coupling and generally differs from the input graph used to construct C_s.

Theorem 5 (LOT induces an effective-resistance policy metric)**.** *Under Assumption 1 and entrywise-finite C_s, let $T = \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\}$ be the simplex tangent space, with orthogonal projector $P_T = I - \mathbf{1}\mathbf{1}^\top/K$; derivatives ∇_T, ∇_T^2 are taken on T. Then π_Q is an interior policy and*

$$F_s^{\mathrm{LOT}} = (\Omega_s + \iota_{\Delta_\mathcal{A}})^\star, \qquad P_T Q = \nabla_T \Omega_s(\pi_Q), \tag{39}$$

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = L_Q/T_0, \qquad \nabla_T^2 \Omega_s(\pi_Q) = T_0 L_Q^\dagger \text{ on } T, \tag{40}$$

where $g^\star(Q) := \sup_\pi\{\langle Q, \pi\rangle - g(\pi)\}$ is the Fenchel conjugate and $\iota_{\Delta_\mathcal{A}}$ is 0 on $\Delta_\mathcal{A}$ and $+\infty$ outside. Consequently, the local quadratic cost of transferring policy mass from action k to action i is T_0 times their effective resistance in the transport co-occurrence graph:

$$(e_i - e_k)^\top \nabla_T^2 \Omega_s(\pi_Q)(e_i - e_k) = T_0 R_Q(i,k), \qquad R_Q(i,k) := (e_i - e_k)^\top L_Q^\dagger (e_i - e_k). \tag{41}$$

Actions connected through many shared anchors have low resistance and are locally easy to substitute; actions separated by a transport bottleneck have high resistance, so moving policy mass between them is penalized more strongly. Because Γ_Q depends on Q, this is a pointwise, value-dependent metric rather than a fixed global policy distance.

At zero cost, $\Omega_s(\pi) = T_0 \mathrm{KL}(\pi\|\rho_s)$ and the metric becomes Fisher–Rao scaled by T_0, with squared tangent norm $T_0 \sum_i v_i^2/\pi_{Q,i}$ for $v \in T$ (Corollary 4, Appendix D.5).

This is a local geometric identity, not a natural-policy-gradient algorithm or a convergence guarantee; kernels with exact zeros, such as truncated heat kernels, may introduce additional null directions and require a separate rank check.

7 Numerical Results

We use controlled experiments to examine kernel locality, approximation error, planning work, reward-free geometry learning, and policy curvature. FullExactHeat denotes the full exact-heat Bellman fixed-point reference. On a path with $K = 4096$ actions in the kernel-concentration experiment (Appendix H.1), exact heat retains 99% of a kernel column's mass in nine actions (Figure 1). On a 25-action cycle in the pruning and truncation experiment (Appendix H.2), radius-six truncated heat uses at most 13 actions and has fixed-point error below 10^{-3} against FullExactHeat (Figure 2). In the $K = 512$ ring-control experiment (Appendix H.3), dense planning uses 4.84 times as many exact expected-Q evaluations as local finite-walk heat, with median root-value errors 0.02755 and 0.028, respectively, against FullExactHeat (Figure 3); additional randomized action-selection comparisons appear in Figure 4. In the reward-free geometry-learning experiment (Appendix H.4), 2048 transitions per action give a median of 0.0092 across five goal-specific median errors $\|Q^\star_{\widehat{C}^t} - Q^\star_{C^t}\|_\infty$ (Figure 5). In the policy-curvature experiment (Appendix H.5), finite differences agree with the effective-resistance identity (Figure 6). Each planning evaluation computes $r(s,a) + \gamma \mathbb{E}_P[V(S')]$ exactly; the reported cost counts these evaluations, not environment samples.

8 Scope and Conclusion

LOT turns pairwise action geometry into a tractable local-softmax Bellman interface. Table 2 maps the kernel properties needed for pruning, local heat planning, geometry learning, and policy curvature. No single construction provides all four benefits. Matched MaxEnt and bounded diffusion-distance Gibbs similarities are provably dense; exact heat can concentrate while remaining entrywise dense; and zeros in $H_{t,r}$ violate the inverse-curvature assumptions and require separate support and rank analysis. For matched ordinary uncentered LOT, nonnegative cost does not improve hard-max bias over MaxEnt. Two scope limits recur: the planning guarantees estimate scalar regularized root values rather than best actions or policy returns, and effective resistance implies neither input-graph policy smoothness nor an optimization rate. Complete assumptions, expanded statements, algorithms, constants, and proofs are in the appendices.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the experimental evaluation, including coding the experiments reported in the paper. We also used generative AI tools to polish writing and to identify related work. We did not use generative AI tools to develop the theoretical model or conceptual framework, formulate mathematical claims, provide ingredients for or write mathematical proofs, propose or refine hypotheses, design the research methodology or experiments, generate or analyze datasets, or interpret experimental results.

All AI-generated experimental code was reviewed, verified, and tested for correctness by the authors. AI-identified related work was independently reviewed and validated by the authors. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 67–83, 2020.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Jiahe Bai, Baojian Zhou, Deqing Yang, and Yanghua Xiao. Faster local solvers for graph diffusion equations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Fan Chung and Olivia Simpson. Computing heat kernel PageRank and a local clustering algorithm. *arXiv preprint arXiv:1503.03155*, 2015.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Tuan Dam. Second-order smooth planning with optimal-transport bellman smoothing. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Tuan Dam, Pascal Stenger, Lukas Schneider, Joni Pajarinen, Carlo D'Eramo, and Odalric-Ambrym Maillard. Monte-carlo tree search with uncertainty propagation via optimal transport. In *International Conference on Machine Learning (ICML)*, 2025.

Edward Brian Davies. Large deviations for heat kernels on graphs. *Journal of the London Mathematical Society*, 47(1):65–72, 1993. doi: 10.1112/jlms/s2-47.1.65.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Stephan Eckstein and Marcel Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm. *SIAM Journal on Mathematical Analysis*, 54(6):5922–5948, 2022. arXiv:2110.06798.

Montacer Essid and Justin Solomon. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

Matthew Folz. Gaussian upper bounds for heat kernels of continuous time simple random walks. *Electronic Journal of Probability*, 16(62):1693–1722, 2011. doi: 10.1214/EJP.v16-926.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning (ICML)*, 2019.

Paul Horn, Yong Lin, Shuang Liu, and Shing-Tung Yau. Volume doubling, Poincaré inequality and gaussian heat kernel estimate for nonnegative curvature graphs. *Journal für die reine und angewandte Mathematik*, 757:89–130, 2019.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

Arne Jensen. Markoff chains as an aid in the study of markoff processes. *Skandinavisk Aktuarietidskrift*, 36:87–91, 1953. doi: 10.1080/03461238.1953.10419459.

Sham Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2002.

Matthias Keller and Christian Rose. Gaussian upper bounds, volume doubling and sobolev inequalities on graphs. *arXiv preprint arXiv:2406.19879*, 2024.

Douglas J. Klein and Milan Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12: 81–95, 1993.

Tadashi Kozuno, Wenhao Yang, Nino Vieillard, Toshinori Kitamura, Yunhao Tang, Jincheng Mei, Pierre Ménard, Mohammad Gheshlaghi Azar, Michal Valko, Rémi Munos, Olivier Pietquin, Matthieu Geist, and Csaba Szepesvári. KL-entropy-regularized reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:2205.14211*, 2022.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 198:1059–1106, 2023.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231, 2007.

Matthieu Meunier, Christoph Reisinger, and Yufei Zhang. Efficient learning for entropy-regularized Markov decision processes via multilevel Monte Carlo. *arXiv preprint arXiv:2503.21224*, 2025.

Joe Neeman, Bobby Shi, and Rachel Ward. Concentration inequalities for sums of Markov-dependent random matrices. *Information and Inference: A Journal of the IMA*, 13(4):iaae032, 2024. doi: 10.1093/imaiai/iaae032.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1–32, 2015. doi: 10.1214/EJP.v20-4039. Corrected version: arXiv:1212.2015v5.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*, volume 11. Now Publishers, 2019.

Yury Polyanskiy and Yihong Wu. Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*, 2021.

Dmitry Shribak, Chen-Xiao Gao, Yitong Li, Chenjun Xiao, and Bo Dai. Diffusion spectral representation for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.16121.

Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (SIGGRAPH)*, 34(4), 2015.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16(25):262–270, 2011. doi: 10.1214/ECP.v16-1624.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16 (3):185–202, 1994.

Kaixin Wang, Kuangqi Zhou, Qixin Zhang, Jie Shao, Bryan Hooi, and Jiashi Feng. Towards better Laplacian representation in reinforcement learning with generalized graph drawing. In *International Conference on Machine Learning (ICML)*, 2021.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

Yifan Wu, George Tucker, and Ofir Nachum. The Laplacian in RL: Learning representations with efficient approximations. In *International Conference on Learning Representations (ICLR)*, 2019.

Xinyan Yan, Krzysztof Choromanski, Byron Boots, and Vikas Sindhwani. Manifold regularization for kernelized LSTD. *arXiv preprint arXiv:1710.05387*, 2017.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023. arXiv:2105.11066.

Appendix of "What Action Geometry Buys: A Design Space for Optimal-Transport Bellman Backups"

Table of Contents

A Tables of Notations and Symbols **15**

B Implementation details for LOCALLOTPLAN **16**

C Constructive Heat-Kernel Planning: Details **18**

C.1 High-mass neighborhoods of exact heat 21

C.2 A locally computable finite-walk heat kernel 23

C.3 Choosing the radius for the exact heat target 28

C.4 Complete end-to-end heat-kernel planning algorithm 29

D Additional Assumptions and Full Theorem Statements **33**

D.1 Additional tabular and finite-sample learning results 33

D.2 Smoothness and second-order identities used by the planner 36

D.3 Second-order smooth planning implications 37

D.4 Error-controlled action pruning and the recursive root-value estimator 38

D.5 Additional results for the local policy metric 39

E Extended Discussion and Limitations **47**

F Proof of the Closed-Form LOT Backup **49**

F.1 Proof of Lemma 1 49

G Proofs of the Remaining Results **52**

G.1 How to navigate the proofs 52

G.2 Bellman backup and contraction 52

G.2.1 Proof of Theorem 1 52

G.2.2 Proof of Proposition 3 58

G.3 Online and batch learning of action geometry 61

G.3.1 Proof of Theorem 9 61

G.3.2 Proof of Theorem 10 64

G.3.3 Proof of Theorem 11 68

G.3.4 Proof of Theorem 4 72

G.4 Smoothness, curvature, and second-order identities 72

G.4.1 Proof of Proposition 5 72

G.4.2 Proof of Theorem 12 74

G.4.3 Proof of Theorem 13 . . . 76
G.4.4 Proof of Lemma 3 . . . 80
G.5 Fixed deterministic action pruning . . . 82
G.5.1 Proof of Lemma 4 . . . 82
G.5.2 Proof of Lemma 2 . . . 86
G.5.3 Proof of Theorem 14 . . . 90
G.6 Exact heat support and locally computable truncated heat . . . 92
G.6.1 Proof of Theorem 6 . . . 92
G.6.2 Proof of Proposition 1 . . . 96
G.6.3 Proof of Proposition 2 . . . 97
G.6.4 Proof of Theorem 7 . . . 98
G.6.5 Proof of Theorem 3 . . . 98
G.6.6 Proof of Theorem 8 . . . 103
G.6.7 Proof of Corollaries 1 and 2 . . . 103
G.7 Full recursive estimator: value bounds, bias, moments, and cost . . . 107
G.7.1 Proof of Theorem 15 . . . 108
G.7.2 Proof of Theorem 2 . . . 110
G.8 Complete specification of LOCALLOTPLAN . . . 113
G.9 Local policy metric, MaxEnt limit, and transport locality . . . 113
G.9.1 Proof of Corollary 3 . . . 113
G.9.2 Proof of Theorem 5 . . . 116
G.9.3 Proof of Corollary 4 . . . 119
G.9.4 Proof of Proposition 6 . . . 122
G.9.5 Proof of Theorem 16 . . . 124
G.9.6 Proof of Proposition 4 . . . 127
G.9.7 Proof of Theorem 17 . . . 130
G.10 Learned geometry and single-trajectory estimates . . . 133
G.10.1 Proof of Theorem 18 . . . 133
G.10.2 Proof of Corollary 5 . . . 137
G.10.3 Proof of Corollary 6 . . . 139
G.10.4 Proof of Theorem 19 . . . 139
G.10.5 Proof of Corollary 7 . . . 140
G.11 Spectral approximation and moving references . . . 140
G.11.1 Proof of Theorem 20 . . . 140
G.11.2 Proof of Theorem 21 . . . 141
G.11.3 Proof of Theorem 22 . . . 141
G.11.4 Proof of Corollary 8 . . . 142
G.12 Continuous actions, finite grids, and the Gaussian example . . . 142

G.12.1 Proof of Theorem 23 142
G.12.2 Proof of Corollary 9 143
G.12.3 Proof of Proposition 7 143

H Experiment Details **143**

H.1 Kernel concentration 144
H.2 Pruning and heat-truncation error 145
H.3 Ring-control planning 146
H.4 Reward-free geometry learning 148
H.5 Effective-resistance curvature 150

A Tables of Notations and Symbols

Table 1: Recurring notation. Symbols are defined at first use; this table is a reading aid only.

Symbol	Meaning	Symbol	Meaning
$\mathcal{S}, \mathcal{A}, K, \gamma$	states, actions, $\lvert\mathcal{A}\rvert$, discount	$w^{(s,j)}, b_{s,j}$	action-similarity distribution, value offset
$\Delta_{\mathcal{A}}$	simplex over $\mathcal{A}$	$p^{(s,j)}(q)$	local softmax at anchor j
$Q,\ \pi$	action values, policy	$F_s^{\mathrm{LOT}},\ \mathcal{F}_s$	LOT aggregator, normalized form
$\tau, \lambda,\ T_0 = \tau\lambda$	temperatures; effective temperature	$\mathcal{T}^{\mathrm{LOT}}$	LOT Bellman operator
ρ_s	action prior (base distribution)	$\Gamma^\star,\ \pi^{\mathrm{LOT}}$	optimal plan, LOT policy
μ_s	reference-action distribution	$A,\ L_A$ (also A_Q, L_Q)	transport co-occurrence graph, Laplacian
$C_s,\ \mathsf{K}_C = e^{-C/\lambda}$	action cost, Gibbs similarity kernel	$d_{\mathrm{ker}}(\alpha)$	largest retained neighborhood size
$L,\ (\lambda_k, \phi_k)$	graph Laplacian, eigenpairs	$\alpha^\star(S, \xi)$	exact omitted-mass threshold
$g,\ C^L$	spectral filter, induced squared cost	$\mathrm{osc}(q)$	$\max_i q_i - \min_i q_i$
$H_t = e^{-tL},\ H_{t,r}$	heat kernel, normalized Poisson head	$B_G(j, r),\ d_r,\ \Delta_G$	r-ball, max ball size, max degree
$\nu,\ \nu_{\mathrm{u}}$	learning distribution, uniformization rate	$\theta = \nu_{\mathrm{u}} t,\ \beta_r$	Poisson mean, omitted Poisson mass

Table 2: **What each geometry buys.** Rows are kernel constructions; columns are the guarantees of this paper; each cell names the governing result. The properties do not compose for free: dense kernels forfeit pruning, and truncation forfeits the strict positivity behind the metric.

Construction	Concentration $d_{\mathrm{ker}}(\alpha)$	Local computation	Planning target	Policy metric
MaxEnt ($C = 0$)	$\lceil(1-\alpha)K\rceil$ (Prop. 1)	none needed	Thm. 2 with $d_{\mathrm{ker}} \approx K$; ordering: Prop. 4	Fisher–Rao (Cor. 4)
Diffusion Gibbs	$\geq K(1 - \alpha e^{C_{\max}/\lambda})_+$ (Prop. 2)	dense spectral columns	Thm. 2	eff. resistance (Thm. 5)
Exact heat H_t	$O((\log(1/\alpha))^{d_g/2})$ (Thm. 6)	entrywise dense	Thm. 2	eff. resistance (Thm. 5)
Truncated heat $H_{t,r}$	support $B_G(j, r)$, size $\leq d_r$ (Thm. 3)	$O(r(\Delta_G{+}1)d_r)$, graph-local	*exact-heat* value (Cor. 1)	rank check required
Learned $\widehat{C}^t$	not analyzed	—	fixed-point error $\frac{\gamma\tau}{1-\gamma}\varepsilon_{n,\delta}$ (Thm. 4)	Thm. 5 at $\widehat{C}$

Table 3: **Three graph constructions used in the paper.**

Construction	Retained actions	Computation
Diffusion-distance Gibbs	$d_{\ker}(\alpha) \geq K(1-\alpha e^{C_{\max}/\lambda})_+$ for uniform ρ_s	Dense spectral cost and similarity columns
Exact heat $H_t(:,j)$	$O((\log(1/\alpha))^{d_g/2})$ under Gaussian-tail and growth assumptions	Dense; retained weights still require local access
Truncated heat $H_{t,r}(:,j)$	Exactly $B_G(j,r)$, of size at most d_r	$O(r(\Delta_G+1)d_r)$ arithmetic; Poisson-tail error β_r

B IMPLEMENTATION DETAILS FOR LOCALLOTPLAN

The factor ε^{-3} is ε^{-2} outer averaging times roughly ε^{-1} work per recursive sample. Geometry replaces dense dependence on K by $d_{\ker}(\alpha_\varepsilon)$. The guarantee concerns the regularized value; the hard-max comparison remains

$$|\widehat{V} - V^\star| \leq |\widehat{V} - V^\star_{\mathcal{F}}| + |V^\star_{\mathcal{F}} - V^\star|. \tag{42}$$

The second term is controlled separately by Propositions 4 and 3.

Numerically stable LOT backup. In (14), i ranges over candidate actions and j over anchors. For cost-based LOT, given a state-dependent nonnegative action-cost matrix $C_s \in \mathbb{R}^{K\times K}_{\geq 0}$ whose entry $C_{s,ij}$ measures the dissimilarity of candidate action i from anchor action j (with $C_{s,jj}=0$), set $\kappa_{s,j} = \sum_i \rho_s(i) e^{-C_{s,ij}/\lambda}$, $w_i^{(s,j)} = \rho_s(i)e^{-C_{s,ij}/\lambda}/\kappa_{s,j}$, and $b_{s,j} = T_0 \log \kappa_{s,j}$. Here $\kappa_{s,j}$ is the total prior-weighted similarity mass around anchor j (large when many probable actions lie near j, small when j is isolated), $w^{(s,j)} \in \Delta_{\mathcal{A}}$ is that mass normalized into the action-similarity distribution the backup uses in place of a global prior, and $b_{s,j}$ is the known value offset restoring the normalization, so that $\mathcal{F}_s$ with (w,b) equals the ordinary LOT backup exactly rather than up to an anchor-dependent constant. Centered direct heat instead uses $w^{(s,j)} = H_t(:,j)$ and $b_{s,j}=0$: the heat column is already a probability distribution, so no normalization, and hence no offset, is needed.

Algorithm 1 LOTBACKUP

Input: Q, $\{(w^{(s,j)}, b_{s,j})\}_j$, μ_s, T_0

For each anchor j: $m_j \leftarrow \max\{Q_i : w_i^{(s,j)} > 0\}$; $z_j \leftarrow \sum_i w_i^{(s,j)} e^{(Q_i - m_j)/T_0}$; $p_i^{(j)} \leftarrow w_i^{(s,j)} e^{(Q_i-m_j)/T_0}/z_j$

return (F_s, π_s) with $F_s = \sum_j \mu_s(j)[b_{s,j} + m_j + T_0 \log z_j]$ and $\pi_s(i) = \sum_j \mu_s(j) p_i^{(j)}$

Algorithm 2 LOCALLOTPLAN (overview)

Input: root s_0, accuracy ε, confidence δ, span bound S_Q, geometry (μ, w, b), and a generative model

Set $\xi \leftarrow (1-\gamma)\varepsilon/2$ and $\alpha_\varepsilon \leftarrow \min\{1/2, \alpha^\star(S_Q, \xi)\}$

For each encountered (s,j), retain the shortest top-weight prefix $I_{s,j}$ with mass at least $1-\alpha_\varepsilon$

Set $u \leftarrow \varepsilon/2$ and the proof-calibrated (η, M, B) from Appendix G.8

For $b = 1, \ldots, B$, average M fresh samples $Y_\eta^{\mathcal{I}}(s_0)$ from Algorithm 3 to obtain $\overline{Z}_b$

return $\mathrm{median}\{\overline{Z}_1, \ldots, \overline{Z}_B\}$

Normalized retained weights. The pruned backup in (15) uses unnormalized weights so that its error has a known sign. To sample actions, define

$$m_{s,j} := w^{(s,j)}(I_{s,j}), \qquad w_i^{\mathcal{I},(s,j)} := \frac{\mathbf{1}\{i \in I_{s,j}\} w_i^{(s,j)}}{m_{s,j}}. \tag{43}$$

The removed mass reappears in the known offset

$$F_s^{\mathcal{I}}(0) = \sum_j \mu_s(j)\big[b_{s,j} + T_0 \log m_{s,j}\big], \qquad B_Q^{\mathcal{I}} = \frac{R_{\max} + \gamma \sup_s |F_s^{\mathcal{I}}(0)|}{1-\gamma}.$$

For

$$\bar{f}_{s,j}^{\mathcal{I}}(Q) := F_s^{\mathcal{I}}(0) + T_0 \log \sum_{i \in I_{s,j}} w_i^{\mathcal{I},(s,j)} e^{Q_i/T_0}, \tag{44}$$

we have $\mathbb{E}_{J\sim\mu_s}[\bar{f}_{s,J}^{\mathcal{I}}(Q)] = F_s^{\mathcal{I}}(Q)$.

Why the recursive estimator and the median? Applying log-sum-exp to noisy action values creates plug-in bias. For a coarse baseline $\widehat{Q}$ and $\Delta = Q - \widehat{Q}$, a local backup is, up to a cubic remainder,

$$\bar{f}(\widehat{Q}) + p^\top \Delta + \frac{1}{4T_0}\mathbb{E}_{A,A' \overset{\text{iid}}{\sim} p}[(\Delta_A - \Delta_{A'})^2].$$

The recursive sample estimates all three terms; two independent residuals make the quadratic product unbiased. Its bias and moments are controlled, but it need not be sub-Gaussian, so block averaging reduces variance and a median over $O(\log(1/\delta))$ blocks gives high confidence. Retained-weight normalization, child accuracies, and the full recursion are in Appendix G.8.

Accuracy schedule. For a requested per-sample bias scale η, set

$$\begin{aligned} &r_\eta = \eta^{1/3}, \quad N_\eta = \lceil c_N r_\eta^{-2} \rceil, \quad \eta_b = c_b r_\eta, \\ &\eta_\ell = a\eta, \qquad \eta_q = c_q \eta^{2/3}, \qquad 1 < a < 1/\gamma. \end{aligned} \tag{45}$$

The baseline accuracy r_η makes the cubic Taylor remainder $O(r_\eta^3) = O(\eta)$. The linear child uses the coarser tolerance $a\eta$, while $\gamma a < 1$ preserves contraction and makes the recursion terminate. The quadratic child uses the intermediate tolerance $O(r_\eta^2)$ because its dominant cross-error is $O(r_\eta \eta_q) = O(\eta)$. With fixed constants chosen as in the proof, the resulting sample satisfies

$$\sup_s \big|\mathbb{E}Y_\eta^{\mathcal{I}}(s) - V_{\mathcal{I}}^\star(s)\big| \le C_{\mathrm{b}}\eta, \qquad \sup_{s,\eta} \mathbb{E}|Y_\eta^{\mathcal{I}}(s)|^3 \le C_3.$$

Algorithm 3 One unclipped localized second-order sample $Y_\eta^{\mathcal{I}}(s)$

Input: state s, tolerance η, retained sets $\mathcal{I}$
If $\eta \ge \eta_0$: **return** $\mathsf{Coarse}_{\eta_0}^{\mathcal{I}}(s)$
Set $(r_\eta, N_\eta, \eta_b, \eta_\ell, \eta_q)$ by (45); draw $J \sim \mu_s$; set $I \leftarrow I_{s,J}$
For a fresh oracle call at (s,i), write $X_\theta(s,i) = R + \gamma Y_\theta^{\mathcal{I}}(S')$
For each $i \in I$, set $\widehat{Q}_i \leftarrow \Pi_{[-B_Q^{\mathcal{I}}, B_Q^{\mathcal{I}}]}\big(N_\eta^{-1}\sum_{n=1}^{N_\eta} X_{\eta_b}^{(n)}(s,i)\big)$
Set $p_i \leftarrow w_i^{\mathcal{I},(s,J)} e^{\widehat{Q}_i/T_0} / \sum_{k\in I} w_k^{\mathcal{I},(s,J)} e^{\widehat{Q}_k/T_0}$
Draw $A_0 \sim p$ and $A, A' \overset{\text{iid}}{\sim} p$; $G_{\mathrm{lin}} \leftarrow X_{\eta_\ell}(s, A_0) - \widehat{Q}_{A_0}$
For $\ell = 1, 2$, independently set

$$\overline{D}_\ell \leftarrow \frac{1}{m_2}\sum_{h=1}^{m_2}\big[X_{\eta_q}^{(\ell h)}(s,A) - \widehat{Q}_A - X_{\eta_q}^{\prime(\ell h)}(s,A') + \widehat{Q}_{A'}\big]$$

return $\bar{f}_{s,J}^{\mathcal{I}}(\widehat{Q}) + G_{\mathrm{lin}} + \overline{D}_1\overline{D}_2/(4T_0)$

All calls marked fresh, recursive subtrees, and the two quadratic batches are mutually independent. The two copies in the last term are necessary: squaring one noisy residual would add its variance and bias the quadratic correction. Clipping is used only for the baseline $\widehat{Q}$; the final sample is left unclipped to avoid an additional clipping bias.

Outer confidence amplification. In Algorithm 2, let $u = \varepsilon/2$ and use

$$\eta = \frac{u}{2C_{\mathrm{b}}}, \qquad M = \left\lceil \frac{16C_3^{2/3}}{u^2} \right\rceil, \qquad B = 2\left\lceil 4\log\frac{2}{\delta} \right\rceil + 1.$$

Then recursive bias is at most $u/2$. The third-moment bound implies $\mathrm{Var}(Y_\eta^{\mathcal{I}}) \le C_3^{2/3}$, so each block mean is within $u/2$ of its expectation with probability at least $3/4$. The median of the B independent block means fails with probability at most δ. Together with the pruning allocation,

$$\underbrace{\varepsilon/2}_{\text{pruning}} + \underbrace{u/2}_{\text{recursive bias}} + \underbrace{u/2}_{\text{sampling}} = \varepsilon.$$

The constants $c_b, c_q, c_N, a, m_2, \eta_0, C_{\mathrm{b}}, C_3$ are fixed proof constants depending only on the structural quantities in Theorem 2, not on K, ε, or δ.

C Constructive Heat-Kernel Planning: Details

Purpose and RL setup. This appendix explains how an action graph produces a locally computable LOT backup. Let $\mathcal{A} = \{1, \dots, K\}$ be the finite action set, let $T_0 = \tau\lambda > 0$ be the LOT temperature, and let s_0 denote the root state whose scalar value is estimated. At an MDP state s, the regularized Bellman equations have the form

$$Q(s,i) = r(s,i) + \gamma\mathbb{E}[V(S') \mid s,i], \qquad V(s) = \mathcal{F}_s(Q(s,\cdot)),$$

where $P(\cdot \mid s,i)$ is the MDP next-state distribution and $\gamma \in (0,1)$ is the discount factor. The LOT aggregator uses an anchor distribution $\mu_s \in \Delta_{\mathcal{A}}$, samples $J \sim \mu_s$, and, conditional on $J = j$, evaluates

$$f_{s,j}(Q) = T_0 \log \sum_{i=1}^{K} w_i^{(s,j)} e^{Q_i/T_0}. \tag{46}$$

Here i is a *candidate action*, j is an *anchor action*, and $w^{(s,j)} \in \Delta_{\mathcal{A}}$ satisfies $w_i^{(s,j)} \ge 0$ and $\sum_i w_i^{(s,j)} = 1$. T_0 is the soft-backup temperature. We specialize to centered normalized kernels, so the usual anchor offset is $b_{s,j} = 0$; adding a known offset does not change the local action distribution. Directly evaluating (46) inspects all K actions.

The pruning result in the main text shows that a high-mass subset is enough for accurate planning. It does not by itself solve the computational problem: finding the large entries of a dense kernel may already require a scan over all K actions. This appendix supplies the missing construction. We place the actions on a sparse weighted graph, interpret graph heat as the anchor-conditioned weights, replace dense exact heat by a finite-walk kernel supported on a graph ball, and propagate the approximation error to the regularized root value.

For readability we use one action graph G for all states. A state-dependent family G_s is covered by applying every construction statewise and taking uniform bounds over s.

Three different transition objects. Table 4 distinguishes the three matrices: In particular, P_G is not an estimate or approximation of the MDP kernel P.

Table 4: Three transition objects in the heat construction

Object	Space on which it acts	Meaning
$P(\cdot \mid s,i)$	MDP states	Environment dynamics queried by the planning generative model.
P_G	Actions	A discrete random walk used only to compute graph diffusion.
$H_t = e^{-tL}$	Actions	The exact continuous-time heat kernel defining the target LOT backup.

Objects used below. The symbols $e_j \in \mathbb{R}^K$ and $\mathbf{1} \in \mathbb{R}^K$ denote the jth one-hot vector and the all-ones vector. For a vector x, $\mathrm{diag}(x)$ is the diagonal matrix with diagonal x; $M(:, j)$ is column j of a matrix; $\mathrm{supp}(x) = \{i : x_i > 0\}$; and $(x)_+ = \max\{x, 0\}$. We write

$$\mathrm{osc}(Q) = \max_i Q_i - \min_i Q_i, \qquad w(I) = \sum_{i \in I} w_i,$$
$$I^c = \mathcal{A} \setminus I, \qquad w(I^c) = 1 - w(I).$$

The notation $\mathcal{L}(X)$ means the probability law of a random variable X. A call to an independent generative model at (s, i) returns an independent reward and next state with the correct MDP law; one such call is counted as one transition sample. Constants denoted by C, c, or decorated versions thereof, are positive proof constants independent of K, ε, δ, and the radius r unless stated otherwise.

From an action graph to heat. Let $G = (\mathcal{A}, E, W)$ be a finite connected undirected weighted action graph. We take $W_{ii} = 0$, $W_{ik} = W_{ki} > 0$ on an edge, and $W_{ik} = 0$ otherwise; a larger edge weight means that two actions are considered more similar. Define

$$d_i = \sum_k W_{ik}, \qquad D = \mathrm{diag}(d_1, \ldots, d_K), \qquad L = D - W.$$

The matrix L is the combinatorial graph Laplacian. It is symmetric, positive semidefinite, and satisfies $L\mathbf{1} = 0$. Write its eigendecomposition as

$$L = \sum_{\ell=1}^{K} \lambda_\ell \phi_\ell \phi_\ell^\top, \qquad 0 = \lambda_1 < \lambda_2 \le \cdots \le \lambda_K, \qquad \phi_1 = \frac{\mathbf{1}}{\sqrt{K}}.$$

Connectedness gives the strict spectral gap $\lambda_2 > 0$.

Place one unit of mass at anchor j and let it flow along graph edges. The graph heat equation and its solution are

$$\frac{d}{du} h_j(u) = -L h_j(u), \qquad h_j(0) = e_j, \qquad h_j(t) = e^{-tL} e_j =: H_t(:, j). \tag{47}$$

In coordinates, this differential equation reads

$$\frac{d}{du} h_{j,i}(u) = \sum_k W_{ik} \big(h_{j,k}(u) - h_{j,i}(u)\big). \tag{48}$$

Each edge therefore transfers mass from an action with more heat to a neighbor with less heat. Moreover, $\frac{d}{du} \mathbf{1}^\top h_j(u) = -\mathbf{1}^\top L h_j(u) = 0$, so total mass remains one. Equivalently,

$$H_t = e^{-tL} = \sum_{\ell=1}^{K} e^{-t\lambda_\ell} \phi_\ell \phi_\ell^\top. \tag{49}$$

Equation (49) explains the smoothing effect: graph modes that vary rapidly across neighboring actions have large λ_ℓ and are suppressed quickly by $e^{-t\lambda_\ell}$, whereas slowly varying modes persist. Thus $t > 0$ is a diffusion-scale parameter, not an MDP time index.

Because $-L$ is the generator of a continuous-time random walk, every entry of H_t is nonnegative and every column sums to one. Hence $H_t(i, j)$ is both the amount of heat and the diffusion probability at candidate i when the process starts at anchor j. Symmetry makes H_t doubly stochastic. The limiting cases are

$$H_0 = I, \qquad H_t \longrightarrow \frac{1}{K} \mathbf{1}\mathbf{1}^\top \quad \text{as } t \to \infty. \tag{50}$$

Small t keeps most mass near the anchor; very large t approaches the uniform kernel and removes locality. For connected G and every $t > 0$, H_t is entrywise positive: there is a positive-probability continuous-time walk along a graph path from any anchor to any candidate. The Poisson-walk representation below makes this argument explicit. Exact heat is therefore dense even when most of its mass lies nearby.

Graph balls. Let $d_G(i, j)$ be the minimum number of positive-weight edges on a path from j to i, and define the r-hop ball

$$B_G(j, r) = \{i \in \mathcal{A} : d_G(i, j) \le r\}. \tag{51}$$

For the path $1 - 2 - 3 - 4 - 5 - 6 - 7$ and anchor $j = 4$,

$$B_G(4, 0) = \{4\}, \qquad B_G(4, 1) = \{3, 4, 5\}, \qquad B_G(4, 2) = \{2, 3, 4, 5, 6\}.$$

Hop distance determines which actions a bounded-length walk can reach; edge weights determine how probability is distributed among those actions.

There are now two different notions of locality. The dense exact column $H_t(:, j)$ may place almost all its mass in a small ball; this is *mass locality*. An algorithm has *computational locality* only if it can construct the required weights by accessing that ball rather than scanning all K actions. Table 5 summarizes the distinction. The argument follows the chain

Table 5: Mass and computational locality in the heat construction

Object	Local property	Role in the analysis
H_t	Dense, but its mass may concentrate in a small graph ball	Exact centered heat-kernel target.
Oracle exact-heat set	A small set can contain most of an H_t column	Reduces transition branching if exact retained weights are available; sub-K arithmetic needs Assumption 2.
$H_{t,r}$	Exactly supported on an r-hop ball and locally computable	Finite-support surrogate whose bias to H_t is controlled by a Poisson tail.

$$\boxed{\text{action graph } G} \longrightarrow \boxed{H_t} \longrightarrow \boxed{H_{t,r}} \longrightarrow \boxed{F^{\mathrm{HK}}_{s,t,r}} \longrightarrow \boxed{\widehat{V}(s_0)}.$$

We first bound how many actions contain most of an exact heat column. We then construct $H_{t,r}$ using only local weighted-adjacency lists, quantify its error to H_t, and propagate that error through the Bellman fixed point and the localized planner. Heat diffusion, Laplacian representations, and continuous-time-chain uniformization are classical (Mahadevan & Maggioni, 2007; Machado et al., 2017; Wu et al., 2019; Jensen, 1953). The contribution here is their decision-theoretic composition: a locally computable action kernel, an explicit LOT backup error, and an end-to-end regularized root-value guarantee. Detailed proofs appear in Section G.

The exact kernel has full support, so "effective support" below always means a high-mass set rather than literal matrix support.

Two locality routes and their error quantities. Table 6 distinguishes the two uses of a graph ball and their error quantities. The first route keeps exact heat weights and assumes an oracle can return

Table 6: Error quantities for exact and truncated heat

Quantity	What it measures	Where it is used
α	allowed exact-heat mass outside a retained set	Oracle pruning of the dense column $H_t(:, j)$.
$\Psi_t(r)$	assumed exact-heat mass outside $B_G(j, r)$	Converts an oracle mass budget α into a spatial radius.
β_r	probability that a Poisson graph walk uses more than r steps	Error from replacing H_t by the computable kernel $H_{t,r}$.
$\delta_r(S)$	worst one-backup value error caused by β_r	Bellman and root-value approximation.

them locally. The second route changes the kernel to $H_{t,r}$ so that the weights are actually computable from adjacency lists. Both use an r-hop ball, but their radii are selected from different error criteria.

C.1 HIGH-MASS NEIGHBORHOODS OF EXACT HEAT

For $0 \le \alpha < 1$, define the smallest worst-anchor set retaining at least $1-\alpha$ exact heat mass by

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) = \max_j \min\left\{ |I| : \sum_{i \notin I} H_t(i,j) \le \alpha \right\}. \tag{52}$$

This quantity controls how many candidate actions a fixed, geometry-only rule must keep after an anchor is selected. Read the definition from the inside out. For a fixed anchor j, the inner minimum chooses the smallest set I whose omitted heat mass is at most α. Because all weights are nonnegative, this minimum is attained by retaining the largest entries of $H_t(:,j)$. The outer maximum then records the most difficult anchor. Thus $d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha)$ is a worst-case local action count, not a dimension of the ambient action space. The tolerance α concerns pruning the *exact* heat column. It should not be confused with the Poisson tail β_r introduced later, which measures the error of the constructive finite-walk surrogate.

The next result uses two geometric assumptions. Let $\Psi_t : \mathbb{N}_0 \to [0,1]$ be a heat-tail envelope, let $v_0 \ge 1$ be a volume constant, and let $d_g \ge 0$ be a graph-volume exponent. The tail bound in (53) says that little heat escapes an r-hop ball; the volume bound says that such a ball contains few actions. Either condition alone is insufficient: concentrated heat is not useful if the ball already contains all K actions, and a small ball is not useful if most heat lies outside it. *Uniformly over a graph family* means that these envelopes and constants do not depend on the graph instance, its size K, or the anchor j.

Theorem 6 (High-mass neighborhoods of exact heat)**.** *Suppose, uniformly over all anchors and a graph family, that*

$$\sum_{i : d_G(i,j) > r} H_t(i,j) \le \Psi_t(r), \qquad |B_G(j,r)| \le v_0(1+r)^{d_g}, \tag{53}$$

where $\Psi_t(r) \to 0$ as $r \to \infty$. For $0 < \alpha < 1$, let $r_\alpha = \min\{m \in \mathbb{N}_0 : \Psi_t(m) \le \alpha\}$. Then

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) \le v_0(1+r_\alpha)^{d_g}. \tag{54}$$

At the endpoint, strict positivity of exact heat gives $d_{\mathrm{ker}}^{\mathrm{HK}}(0) = K$.

Suppose additionally that there are constants $c_H, c_t > 0$ and a largest radius $r_{\mathrm{diff}} \in \mathbb{N}_0$ on which the Gaussian estimate $\Psi_t(r) \le c_H e^{-r^2/(c_t t)}$ holds. If $0 < \alpha \le \min\{1, c_H\}$ and

$$\left\lceil \sqrt{c_t t \log(c_H/\alpha)} \right\rceil \le r_{\mathrm{diff}},$$

then

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) \le v_0 \left(1 + \left\lceil \sqrt{c_t t \log(c_H/\alpha)} \right\rceil\right)^{d_g}. \tag{55}$$

Thus $d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) = O((\log(1/\alpha))^{d_g/2})$ while the selected radius remains in the Gaussian range and the graph balls have not saturated at all K actions.

Reading Theorem 6. Given a permitted omitted mass $0 < \alpha < 1$, choose the first radius r_α for which the heat tail is below α. The ball $B_G(j, r_\alpha)$ is then an admissible retained set for every anchor, and the volume bound turns that radius into an action count. Under a Gaussian tail, $r_\alpha = O(\sqrt{t \log(1/\alpha)})$ and hence the retained count is only polylogarithmic in $1/\alpha$ for fixed t and d_g. This statement is valid only before the graph ball saturates at all K actions.

The theorem is a kernel-mass statement. It does not claim that H_t is sparse or that exact values $H_t(i,j)$ inside the ball can be computed from local adjacency lists alone. For comparison, two common alternatives can remain dense even under the same fixed-set pruning criterion.

For anchor-conditioned distributions $w^{(j)}$, we use the effective action count from Definition 1:

$$d_{\mathrm{ker}}(\alpha) = \max_j \min\{|I| : w^{(j)}(I^c) \le \alpha\}.$$

Proposition 1 (Uniform-prior MaxEnt remains dense)**.** *For $w_i^{(j)} = 1/K$, $d_{\mathrm{ker}}(\alpha) = \lceil (1-\alpha)K \rceil$.*

Indeed, every action carries mass $1/K$, so retaining mass $1-\alpha$ requires retaining at least $(1-\alpha)K$ actions. There is no graph concentration to exploit.

Proposition 2 (Bounded diffusion-distance Gibbs kernels remain dense)**.** *Let*

$$w_i^{(j)} = \frac{e^{-C_{ij}/\lambda}}{\sum_k e^{-C_{kj}/\lambda}}$$

with uniform action prior and $0 \le C_{ij} \le C_{\max}$*. Then*

$$d_{\ker}(\alpha) \ge K(1 - \alpha e^{C_{\max}/\lambda})_+. \tag{56}$$

For squared diffusion distance $C_{ij} = (e_i - e_j)^\top e^{-2tL}(e_i - e_j)$ *on a connected graph,* $C_{\max} \le 2e^{-2t\lambda_2}$*, where* λ_2 *is the spectral gap of* L*.*

The bounded-cost obstruction has the same interpretation. Since no Gibbs weight can be smaller than $e^{-C_{\max}/\lambda}/K$, every omitted action contributes a nonnegligible amount of mass. Unless the permitted tail α is large enough to absorb many such terms, the retained set must again contain a linear fraction of all actions. These comparisons explain why smoothness alone does not imply locality: locality requires a kernel with concentrated tails.

Although both constructions use the word "diffusion," they are different. Direct heat uses $H_t(i,j)$ itself as the anchor-conditioned weight. The Gibbs alternative first converts the heat embedding into the bounded squared distance $(e_i - e_j)^\top e^{-2tL}(e_i - e_j)$ and then exponentiates that distance. Bounded distances keep all Gibbs weights within a fixed ratio, whereas a direct heat column can decay strongly with graph distance.

These propositions concern deterministic, value-independent sets selected by retained kernel mass. They do not lower-bound adaptive, randomized, or value-dependent planners, nor do they say that the final LOT policy is dense for every Q.

Why heat is a LOT kernel. The direct-heat backup below uses $H_t(:,j)$ as the candidate-action prior conditioned on anchor j. This is not a separate regularizer. For LOT parameters $\tau, \lambda > 0$, recall that $T_0 = \tau\lambda$. With uniform base prior $\rho_i = 1/K$ and logarithmic cost $C_{ij}^{\mathrm{HK}} = -\lambda \log H_t(i,j)$,

$$\rho_i e^{-C_{ij}^{\mathrm{HK}}/\lambda} = \frac{H_t(i,j)}{K}.$$

Because each heat column sums to one, its total unnormalized mass is $1/K$; normalization therefore recovers exactly $H_t(:,j)$. The factor $1/K$ produces the known constant $-T_0 \log K$ in the uncentered LOT value. Adding $T_0 \log K$ removes this offset. Centering changes values by a deterministic translation but does not change the local policies or derivatives. We retain the index t in $F_{s,t}^{\mathrm{HK}}$ to distinguish the exact heat target from its finite-walk approximation.

Conditional on anchor j, the corresponding action response is

$$p_i^{(j)}(Q) = \frac{H_t(i,j)e^{Q_i/T_0}}{\sum_k H_t(k,j)e^{Q_k/T_0}}, \qquad \pi(i \mid s) = \sum_j \mu_s(j) p_i^{(j)}(Q(s,\cdot)).$$

The heat factor favors candidates geometrically close to the anchor, while e^{Q_i/T_0} favors candidates with high action value. At $t = 0$, the centered backup is $\sum_j \mu_s(j) Q_j$; as $t \to \infty$, it becomes the centered uniform-prior MaxEnt backup $T_0 \log(K^{-1} \sum_i e^{Q_i/T_0})$. The finite logarithmic-cost representation requires $t > 0$: at exactly $t = 0$, the off-diagonal entries of $H_0 = I$ vanish and their logarithmic costs are $+\infty$, although the centered backup has the stated $t \downarrow 0$ limit.

For clarity, the uncentered uniform-prior backup mentioned below is

$$F_{s,t}^{\mathrm{logH}}(Q) = T_0 \sum_j \mu_s(j) \log \left[\frac{1}{K} \sum_i H_t(i,j) e^{Q_i/T_0} \right].$$

Theorem 7 (Exact heat as a centered LOT kernel)**.** *Let* G *be connected and* $H_t = e^{-tL}$ *with* $t > 0$*. The centered direct-heat backup is*

$$F_{s,t}^{\mathrm{HK}}(Q) = T_0 \sum_j \mu_s(j) \log \sum_i H_t(i,j) e^{Q_i/T_0}. \tag{57}$$

It satisfies $F^{\mathrm{HK}}_{s,t}(0) = 0$. With the uniform prior and $C^{\mathrm{HK}}_{ij} = -\lambda \log H_t(i,j)$, the normalized LOT columns are $w^{(j)} = H_t(:,j)$. Let $c = T_0 \log K$, and let $(Q^\star_{\mathrm{HK}}, V^\star_{\mathrm{HK}})$ and $(Q^\star_{\mathrm{logH}}, V^\star_{\mathrm{logH}})$ denote the Bellman solutions of the centered and uncentered backups. Then

$$F^{\mathrm{HK}}_{s,t}(Q) = F^{\mathrm{logH}}_{s,t}(Q) + c, \qquad \begin{aligned} Q^\star_{\mathrm{HK}} &= Q^\star_{\mathrm{logH}} + \frac{\gamma c}{1-\gamma}\mathbf{1}, \\ V^\star_{\mathrm{HK}}(s) &= V^\star_{\mathrm{logH}}(s) + \frac{c}{1-\gamma}. \end{aligned} \tag{58}$$

The policies and derivatives agree at the corresponding translated values. The logarithmic heat cost need not have zero diagonal; LOT itself does not require $C_{jj} = 0$.

Under (53)*, the bounds* (54)*–* (55) *hold. If the exact retained heat weights are supplied by Assumption 2, then for the span envelope S_Q in Theorem 2, set*

$$\alpha_\varepsilon = \min\left\{\frac{1}{2},\ \frac{\exp((1-\gamma)\varepsilon/(2T_0)) - 1}{\exp(S_Q/T_0) + \exp((1-\gamma)\varepsilon/(2T_0)) - 1}\right\}.$$

This is precisely the pruning budget (2)*; for fixed structural parameters, $\alpha_\varepsilon = \Theta(\varepsilon)$. Substituting the effective action count into Algorithm 2 and Theorem 2 gives the following number of MDP transition samples the localized planner gives the following number of MDP transition samples for estimating the scalar root value to accuracy ε and confidence $1-\delta$:*

$$N^{\mathrm{oracle}}_{\mathrm{heat}}(\varepsilon,\delta) \le \varepsilon^{-3} \exp\left(O([\log\log(1/\varepsilon)]^2)\right) \log\frac{2}{\delta} = (1/\varepsilon)^{3+o(1)} \log\frac{2}{\delta}, \tag{59}$$

provided the Gaussian range and graph-ball nonsaturation remain valid. Without exact local access, the same result describes transition branching but not sub-K graph arithmetic. The dense comparisons are Propositions 1 and 2.

The last part of the theorem describes an *oracle local-access route*: Assumption 2 means here that, given (j,α), an oracle returns retained indices and their exact values $H_t(i,j)$ without scanning all K actions. If such entries can be returned directly, the high-mass ball reduces transition branching. The construction below gives a different, fully local route. It replaces exact heat by a finite-walk kernel whose weights can be generated from adjacency lists and then pays an explicit approximation error.

Why a second construction is needed. The preceding theorem explains when exact heat has a small high-mass set, but exact in-ball heat includes contributions from arbitrarily long walks. Local adjacency access therefore does not generally recover its exact retained weights. We next replace H_t by a normalized finite-walk kernel that is exactly computable in a graph ball, and then bound the resulting bias to the exact heat target.

C.2 A LOCALLY COMPUTABLE FINITE-WALK HEAT KERNEL

Directly forming the matrix exponential e^{-tL} is generally a global operation. Uniformization rewrites the same continuous diffusion as a random number of ordinary graph-walk steps. Because each such step uses only one weighted adjacency list, restricting the number of steps creates a locally computable surrogate.

Local weighted-adjacency access means that querying action i returns its weighted degree D_{ii} and the list $\{(\ell, W_{\ell i}) : W_{\ell i} > 0\}$ in time proportional to the number of returned neighbors. The arithmetic bounds below exclude storage of the input graph and any one-time preprocessing used to obtain ν_{u} or sample anchors from μ_s.

Let

$$d_{\max} = \max_i D_{ii}, \qquad \Delta_G = \max_i |\{k : W_{ik} > 0\}|,$$

where $d_{\max}$ is a maximum *weighted* degree and Δ_G is a maximum neighbor count. Fix $t > 0$ and choose a known uniformization rate $\nu_{\mathrm{u}} \in (0,\infty)$ with $\nu_{\mathrm{u}} \ge d_{\max}$, then define

$$P_G = I - \frac{L}{\nu_{\mathrm{u}}}, \qquad \theta = \nu_{\mathrm{u}} t. \tag{60}$$

The inequality $\nu_{\mathrm{u}} \geq d_{\max}$ makes every diagonal entry $1 - D_{ii}/\nu_{\mathrm{u}}$ nonnegative, so P_G is a symmetric stochastic *action-walk* kernel. It is used only to compute graph diffusion and must not be confused with the MDP transition kernel $P(\cdot \mid s, a)$. Its nonzero entries are

$$P_G(i,i) = 1 - \frac{D_{ii}}{\nu_{\mathrm{u}}}, \qquad P_G(\ell, i) = \frac{W_{\ell i}}{\nu_{\mathrm{u}}} \quad (\ell \neq i),$$

so multiplying by P_G needs only the weighted adjacency list of each currently active action. We use the column convention $P_G(\ell, i) = \Pr(X_{k+1} = \ell \mid X_k = i)$; symmetry makes P_G both row- and column-stochastic. Since $L = \nu_{\mathrm{u}}(I - P_G)$ and $\theta = \nu_{\mathrm{u}} t$, classical uniformization (Jensen, 1953) gives

$$\begin{aligned} H_t = e^{-tL} &= e^{-\theta(I - P_G)} = e^{-\theta} e^{\theta P_G} \\ &= \sum_{k=0}^{\infty} a_k P_G^k, \qquad\qquad a_k = e^{-\theta} \frac{\theta^k}{k!}. \end{aligned}$$

Equivalently, if $N \sim \mathrm{Pois}(\theta)$ is independent of the discrete chain and $X_{k+1} \sim P_G(\cdot \mid X_k)$, then

$$H_t(:, j) = \mathcal{L}(X_N \mid X_0 = j).$$

In words, draw a Poisson number N of graph-walk steps and take exactly N steps from anchor j. The endpoint distribution is the exact heat column. Since $\mathbb{E}[N] = \theta = \nu_{\mathrm{u}} t$, larger diffusion time generally requires more walk steps. A step may be a self-loop because $P_G(i, i)$ can be positive. The choice of ν_{u} changes this representation, but not the exact matrix $H_t = e^{-tL}$.

For $r \in \mathbb{N}_0$, put

$$s_r = \sum_{k=0}^{r} a_k, \qquad \beta_r = 1 - s_r = \Pr\{N \geq r + 1\}, \qquad H_{t,r} = \frac{1}{s_r} \sum_{k=0}^{r} a_k P_G^k. \tag{61}$$

Thus $H_{t,r}(:, j) = \mathcal{L}(X_N \mid X_0 = j, N \leq r)$: it renormalizes the finite Poisson head and is a probability distribution, rather than merely dropping the tail mass. We write $F^{\mathrm{HK}}_{s,t,r}$ for the backup (57) with H_t replaced by $H_{t,r}$. The normalization by s_r preserves column mass one and therefore preserves the centered identity $F^{\mathrm{HK}}_{s,t,r}(0) = 0$. Using the unnormalized head would instead add the value offset $T_0 \log s_r$.

Thus r is a maximum number of action-graph steps, not another diffusion time. This conditioning creates literal locality. A walk of at most r steps cannot leave $B_G(j, r)$, so $H_{t,r}(:, j)$ has no mass outside that ball. On the seven-action chain above, $r = 2$ requires visiting only $\{2, 3, 4, 5, 6\}$ when $j = 4$, regardless of the total number of actions in a longer chain.

A numerical path example. Consider the unit-weight path $1 - 2 - \cdots - 9$, anchor $j = 5$, and choose $\nu_{\mathrm{u}} = 2$, $t = 1/2$, so $\theta = 1$. Conditional on at most two steps, the normalized Poisson weights for lengths $0, 1, 2$ are $2/5, 2/5, 1/5$. Combining the corresponding endpoint laws gives

$$H_{t,2}(:, 5) = 0.05 e_3 + 0.20 e_4 + 0.50 e_5 + 0.20 e_6 + 0.05 e_7.$$

The exact column $H_t(:, 5)$ is positive on all nine actions, but this local surrogate touches only the two-hop ball $\{3, 4, 5, 6, 7\}$. Its omitted walk-length mass is

$$\beta_2 = \Pr\{\mathrm{Pois}(1) \geq 3\} = 1 - \frac{5}{2e} \approx 0.0803.$$

If $T_0 = 1$ and the values on actions $3, \ldots, 7$ are $(0, 1, 2, 1, 0)$, the anchor-specific partition is

$$0.05 + 0.20e + 0.50e^2 + 0.20e + 0.05.$$

The local response assigns approximately 75.7% to action 5, 11.1% to each of actions 4 and 6, and 1.0% to each of actions 3 and 7. This illustrates the two forces in the backup: heat defines the geometric neighborhood, while exponentiated Q-values rank candidates within it.

At the end of iteration k, the variables have the exact meanings $v = P_G^k e_j, \omega = a_k = e^{-\theta} \frac{\theta^k}{k!}$, and $y = \sum_{\ell=0}^{k} a_\ell P_G^\ell e_j, Z_{\mathrm{P}} = \sum_{\ell=0}^{k} a_\ell$. Consequently, the returned vector is exactly $h^{(j,r)} =$

Algorithm 4 LOCALHEATCOLUMN: one column of $H_{t,r}$ from local graph access

Input: anchor j, radius r, diffusion time t, uniformization rate $\nu_{\mathrm{u}} > 0$, and local weighted-adjacency lists
Representation: sparse maps on discovered actions; an absent key has value zero
$\theta \leftarrow \nu_{\mathrm{u}} t$; $\omega \leftarrow e^{-\theta}$; $v \leftarrow e_j$; $y \leftarrow \omega v$; $Z_{\mathrm{P}} \leftarrow \omega$
for $k = 1, \dots, r$ **do**
  $\widetilde{v} \leftarrow 0$ as a sparse vector
  for each $i \in \mathrm{supp}(v)$ **do**
    Query D_{ii} and $\{(\ell, W_{\ell i}) : W_{\ell i} > 0\}$
    $\widetilde{v}_i \leftarrow \widetilde{v}_i + (1 - D_{ii}/\nu_{\mathrm{u}}) v_i$
    for each positive-weight neighbor ℓ of i **do**
      $\widetilde{v}_\ell \leftarrow \widetilde{v}_\ell + (W_{\ell i}/\nu_{\mathrm{u}}) v_i$
    end for
  end for
  $v \leftarrow \widetilde{v}$; $\omega \leftarrow \omega\theta/k$; $y \leftarrow y + \omega v$; $Z_{\mathrm{P}} \leftarrow Z_{\mathrm{P}} + \omega$
end for
$I_j^{(r)} \leftarrow \mathrm{supp}(y)$
$h_i^{(j,r)} \leftarrow y_i / Z_{\mathrm{P}}$ for every $i \in I_j^{(r)}$
return $(I_j^{(r)}, h^{(j,r)})$

$\frac{\sum_{k=0}^{r} a_k P_G^k e_j}{\sum_{k=0}^{r} a_k} = H_{t,r}(:, j)$, rather than a Monte Carlo estimate. For $t > 0$, $I_j^{(r)} = B_G(j, r)$. At the degenerate endpoint $t = 0$, we use $r = 0$, so $H_{0,0} = I$ and $I_j^{(0)} = \{j\}$.

The sparse multiplication $v \leftarrow P_G v$ expands the discovered set by at most one graph hop. Hence the algorithm never visits an action outside $B_G(j, r)$. For large θ, an implementation may omit the common factor $e^{-\theta}$, periodically rescale $(\omega, y, Z_{\mathrm{P}})$, or compute the normalized Poisson coefficients in log space.

The pseudocode is an exact-arithmetic presentation. For large θ, an implementation may drop the common factor $e^{-\theta}$, use the relative coefficients $\theta^k/k!$, and periodically rescale; equivalently, it may evaluate normalized Poisson weights in log space. Once $h^{(j,r)} = H_{t,r}(:, j)$ is available, one additional pass over $I = B_G(j, r)$ computes the stable local softmax

$$g_i = h_i^{(j,r)} e^{(Q_i - m_j)/T_0}, \quad m_j = \max_{i \in I} Q_i, \quad z_j = \sum_{i \in I} g_i, \quad p_i^{(j,r)} = g_i / z_j, \tag{62}$$

and the anchor-specific value is $m_j + T_0 \log z_j$. When some $h_i^{(j,r)}$ are extremely small, implementations should evaluate $\log h_i^{(j,r)} + (Q_i - m_j)/T_0$ by log-sum-exp rather than form g_i directly.

How the planner uses the local column. For one backup call at state s, the computation is:

1. sample an anchor $J \sim \mu_s$;
2. construct or retrieve the cached column $H_{t,r}(:, J)$ using Algorithm 4;
3. inspect only $i \in B_G(J, r)$ and form the local Gibbs probabilities in (62);
4. sample retained candidate actions from this local distribution and make the recursive generative-model calls required by the second-order estimator; and
5. apply the outer median-of-means wrapper to independent recursive estimates.

The graph walk determines which action coordinates are available to the backup; the MDP generative model is queried only afterward to simulate the consequences of the selected actions. These are two separate sources of computation.

The heat construction changes only the normalized local kernel supplied to the recursive cascade. It bypasses the top-weight pruning stage of LOCALLOTPLAN, because the complete finite-walk column is already supported on a locally computable graph ball; it does not change the Taylor recursion or its error schedule. One complete recursive output has controlled bias and moments but need not be sub-Gaussian because its final value is unclipped. The planner therefore averages outputs within independent blocks and takes a median only at the root. Averaging controls variance, while the outer median yields $O(\log(1/\delta))$ confidence amplification. Keeping the median outside

the recursion is important: a median is nonlinear and would disturb the expectation and independent cross-product identities used by the second-order correction. The full schedule and constants are defined in Appendix G.8.

Table 7 summarizes the roles of the main parameters.

Table 7: Parameters of local heat planning

Parameter	Meaning	Main tradeoff
t	diffusion scale of the exact target H_t	larger t shares mass farther across the action graph and usually requires a larger local radius;
r	maximum graph-walk length in $H_{t,r}$	larger r reduces heat approximation bias but visits a larger ball;
T_0	LOT temperature	larger T_0 softens the influence of Q and reduces amplification of omitted kernel mass;
ν_{u}	uniformization rate used to represent H_t	any $\nu_{\mathrm{u}} \geq d_{\max}$ is valid; it changes θ and the sufficient truncation radius, not the exact heat target.

The next theorem connects the construction to planning. Part (i) proves that the surrogate is exactly and locally computable. Part (ii) compares it with exact heat. Part (iii) transfers this kernel error through the soft backup and Bellman fixed point. Part (iv) composes the resulting local backup with the generic second-order root-value estimator.

Theorem 8 (Local heat construction and error propagation)**.** *Assume G is available through local weighted-adjacency lists and write $d_r = \max_j |B_G(j, r)|$. Then the following statements hold.*

(i) Exact construction of the finite-walk kernel. *P_G and $H_{t,r}$ are symmetric, nonnegative, and stochastic, and*

$$\mathrm{supp}(H_{t,r}(:, j)) = B_G(j, r), \qquad H_{t,r}(j, j) > 0. \tag{63}$$

Algorithm 4 never visits an action outside this ball and computes one column using

$$O((r+1)(\Delta_G+1)d_r) \text{ arithmetic}, \qquad O((\Delta_G+1)d_r) \text{ working memory}. \tag{64}$$

The support identity follows because every walk of length at most r stays inside the ball, while a shortest positive-weight path reaches every action in the ball. The $a_0 I$ term guarantees positive mass at the anchor. The cost is local when vectors and accumulators are stored as sparse maps over visited actions.

(ii) Approximation of exact heat. *Let $R_r = \sum_{k=r+1}^{\infty} a_k P_G^k$. Then*

$$H_t = e^{-tL} = s_r H_{t,r} + R_r, \tag{65}$$

every row and column sum of R_r is β_r, and, with $\|M\|_\infty = \max_i \sum_j |M_{ij}|$,

$$\begin{aligned} \|H_t - H_{t,r}\|_\infty &\leq 2\beta_r, \\ \max_j \sum_i |H_t(i,j) - H_{t,r}(i,j)| &\leq 2\beta_r, \\ \sum_{i \notin B_G(j,r)} H_t(i,j) &\leq \beta_r. \end{aligned} \tag{66}$$

Importantly, $H_{t,r}$ is not H_t cropped to the ball. Walks longer than r can end inside the ball, and their contribution belongs to R_r. The factor $2\beta_r$ has a simple origin: at most β_r mass is removed with the Poisson tail, and at most another β_r change is caused by renormalizing the retained head back to total mass one. Equivalently, the total-variation distance between an exact and truncated column is at most β_r, because total variation is one half of the ℓ_1 distance.

(iii) Backup and fixed-point errors. *Define*

$$F^{\mathrm{HK}}_{s,t,r}(Q) = T_0 \sum_j \mu_s(j) \log \sum_i H_{t,r}(i,j) e^{Q_i/T_0} \tag{67}$$

To see where the error bound comes from, suppose first that $\beta_r > 0$ and define the conditional long-walk kernel $\overline{R}_r = R_r/\beta_r$. Equation (65) *becomes*

$$H_t = (1-\beta_r)H_{t,r} + \beta_r \overline{R}_r,$$

where both $H_{t,r}$ and $\overline{R}_r$ are stochastic. For one anchor, set $u_i = e^{Q_i/T_0}$ and $\kappa = u_{\max}/u_{\min} = e^{\mathrm{osc}(Q)/T_0}$. The expectations of u under the short- and long-walk laws both lie in $[u_{\min}, u_{\max}]$. Consequently, mixing in β_r long-walk mass can change their partition-function ratio by at most

$$1 + \beta_r(\kappa - 1) \le 1 + \beta_r(e^{S/T_0} - 1) \qquad \textit{when } \mathrm{osc}(Q) \le S.$$

The corresponding lower ratio gives no larger absolute logarithmic deviation. When $\beta_r = 0$, the two kernels coincide. Taking T_0 times the logarithm motivates

$$\delta_r(S) = T_0 \log\left(1 + \beta_r(e^{S/T_0} - 1)\right). \tag{68}$$

Uniformly over $\mathrm{osc}(Q) \le S$,

$$|F^{\mathrm{HK}}_{s,t}(Q) - F^{\mathrm{HK}}_{s,t,r}(Q)| \le \delta_r(S). \tag{69}$$

Let $Q^\star_t$ and $Q^\star_{t,r}$ be the fixed points of the exact- and finite-walk Bellman operators, and define their state values by

$$V^\star_t(s) = F^{\mathrm{HK}}_{s,t}(Q^\star_t(s,\cdot)), \qquad V^\star_{t,r}(s) = F^{\mathrm{HK}}_{s,t,r}(Q^\star_{t,r}(s,\cdot)).$$

If $\sup_s \mathrm{osc}(Q^\star_t(s,\cdot)) \le S$, then

$$\|Q^\star_{t,r} - Q^\star_t\|_\infty \le \frac{\gamma\delta_r(S)}{1-\gamma}, \qquad \|V^\star_{t,r} - V^\star_t\|_\infty \le \frac{\delta_r(S)}{1-\gamma}. \tag{70}$$

Here $\delta_r(S)$ converts a kernel error into the error of one soft backup. Bellman recursion then amplifies persistent one-step value errors by the geometric factor $1/(1-\gamma)$; the Q-value bound has the additional factor γ because the aggregator appears after one discounted transition.

(iv) Composition with the localized planner. *The finite-walk backup uses $w^{(s,j)} = H_{t,r}(:,j)$ and $I_{s,j} = B_G(j,r)$, so its local action count is d_r. Each column already sums to one, hence no normalization offset is required. Under an independent generative model and $|R| \le R_{\max}$, put $\bar{B}_Q = (R_{\max} + \gamma T_0 \log 2)/(1-\gamma)$; this is the conservative action-value magnitude bound used by the generic recursion. There are $\varepsilon_0, C_{\mathrm{u}}, c_{\mathrm{u}} > 0$, depending only on $(R_{\max}, \gamma, T_0)$ and the recursion parameters fixed in Appendix G.8, such that, for every $0 < \varepsilon_{\mathrm{stat}} \le \varepsilon_0$ and $\delta \in (0,1)$, the recursive estimator in Algorithm 3, using the schedules in* (45)*, and the outer median-of-means step of Algorithm 2 estimate $V^\star_{t,r}(s_0)$ to accuracy $\varepsilon_{\mathrm{stat}}$ with probability at least $1-\delta$ using*

$$n_r(\varepsilon_{\mathrm{stat}}, \delta) \le \frac{C_{\mathrm{u}}}{\varepsilon^3_{\mathrm{stat}}} (C_{\mathrm{u}} d_r)^{c_{\mathrm{u}}[1 + \log\log(e + \bar{B}_Q/\varepsilon_{\mathrm{stat}})]} \log\frac{2}{\delta} \tag{71}$$

transition samples. After anchor J is sampled, no transition is queried and no value coordinate is inspected outside $B_G(J,r)$. Here $\varepsilon_{\mathrm{stat}}$ is the desired error relative to the finite-walk target, and n_r counts only MDP generative-model calls. The graph arithmetic in (64) *is separate from this transition count; a computed column may be cached and reused. If N_{anc} distinct anchors are encountered and none is cached initially, total column-construction work is at most $O(N_{\mathrm{anc}}(r+1)(\Delta_G+1)d_r)$, with $O(N_{\mathrm{anc}} d_r)$ storage for the cached sparse columns, in addition to the graph representation. The leading $\varepsilon^{-3}_{\mathrm{stat}}$ is the statistical dependence of the second-order recursive planner. The remaining dependence enters through d_r, the largest local action count, rather than directly through the global action count K. Algorithms 6 and 7 implement this heat specialization of the generic recursive estimator. The complete truncated column is retained, so no additional top-weight pruning error is introduced.*

The fixed-point factors in part (iii) follow from the two transparent inequalities

$$\|V^\star_{t,r} - V^\star_t\|_\infty \le \|Q^\star_{t,r} - Q^\star_t\|_\infty + \delta_r(S), \qquad \|Q^\star_{t,r} - Q^\star_t\|_\infty \le \gamma\|V^\star_{t,r} - V^\star_t\|_\infty.$$

Solving them gives the displayed V- and Q-error bounds. In particular, only the span of the exact fixed point is needed in the one-step comparison.

The error chain behind part (iii) is

$$\beta_r \text{ (Poisson tail)} \quad \Longrightarrow \quad \delta_r(S) \text{ (one backup)} \quad \Longrightarrow \quad \frac{\delta_r(S)}{1-\gamma} \text{ (root value).}$$

The exponential factor e^{S/T_0} appears because the log partition can amplify a small amount of omitted kernel mass when the omitted actions have values up to S larger than the retained actions. Only the span $S = \mathrm{osc}(Q)$ matters because adding the same constant to every action value shifts both backups equally. For a small Poisson tail,

$$\delta_r(S) = T_0\beta_r(e^{S/T_0} - 1) + O(\beta_r^2).$$

Thus a large value span or a low temperature demands a smaller tail and, consequently, a larger radius.

C.3 Choosing the radius for the exact heat target

The final estimator has two errors: it estimates the finite-walk value $V^\star_{t,r}(s_0)$ statistically, and $V^\star_{t,r}$ approximates the exact heat value $V^\star_t$. We allocate $\varepsilon/2$ to each. To match the generic localized-planning theorem, use the conservative envelope

$$\bar{B}_Q = \frac{R_{\max} + \gamma T_0 \log 2}{1-\gamma}, \qquad S_Q = 2\bar{B}_Q. \tag{72}$$

The term $T_0 \log 2$ is not intrinsic to graph heat. It is inherited from the generic planner, which requires every pruned kernel column to retain at least half its mass; removing at most half the mass can shift a log partition by at most $T_0 \log 2$. In the constructive heat route we use the entire normalized support of $H_{t,r}$, so $F^{\mathrm{HK}}_{s,t,r}(0) = 0$ and a tighter envelope is possible. Because the full exact and finite-walk heat columns have mass one and their centered backups vanish at $Q = 0$, the tighter span $2R_{\max}/(1-\gamma)$ is also valid for the deterministic heat-bias term; (72) is retained here for direct compatibility with the main theorem and the generic cascade constants.

To allocate at most $\varepsilon/2$ to deterministic heat approximation, we require

$$\frac{\delta_r(S_Q)}{1-\gamma} = \frac{T_0}{1-\gamma}\log\Big(1 + \beta_r(e^{S_Q/T_0} - 1)\Big) \le \frac{\varepsilon}{2}.$$

Solving this inequality for β_r gives the tail budget b_ε below. The only remaining task is therefore to choose r so that a Poisson random variable with mean $\theta = \nu_{\mathrm{u}} t$ exceeds r with probability at most b_ε.

Corollary 2 (Exact-heat root-value estimation)**.** *There are $\varepsilon_0, C, c > 0$, depending only on $(R_{\max}, \gamma, T_0)$ and the recursion parameters fixed in Appendix G.8, such that the following holds for $0 < \varepsilon \le \varepsilon_0$ and $\delta \in (0,1)$. With $\bar{B}_Q, S_Q$ from* (72)*, define the permitted Poisson-tail mass*

$$b_\varepsilon = \min\left\{\frac{1}{2}, \frac{\exp((1-\gamma)\varepsilon/(2T_0)) - 1}{\exp(S_Q/T_0) - 1}\right\}; \tag{73}$$

and the explicit radius

$$r = \left\lceil \max\left\{1, 2e\nu_{\mathrm{u}} t, \log_2 \frac{1}{b_\varepsilon}\right\}\right\rceil - 1 \tag{74}$$

ensures $\beta_r \le b_\varepsilon$ and therefore $\delta_r(S_Q)/(1-\gamma) \le \varepsilon/2$. Let $\widehat{V}_{\mathrm{HK}}(s_0)$ be the output of Algorithm 7. Then

$$\Pr\left\{\left|\widehat{V}_{\mathrm{HK}}(s_0) - V^\star_t(s_0)\right| \le \varepsilon\right\} \ge 1 - \delta, \tag{75}$$

using at most

$$n_{\mathrm{heat}}(\varepsilon, \delta) \le \frac{C}{\varepsilon^3}(Cd_r)^{c[1+\log\log(e+\bar{B}_Q/\varepsilon)]}\log\frac{2}{\delta} \tag{76}$$

MDP transition samples.

If, uniformly over the graph family, for some volume exponent $d_g \ge 1$,

$$|B_G(j,m)| \le C_{\mathrm{vol}}(m+1)^{d_g}, \tag{77}$$

then $d_r \le \min\{K, C_{\mathrm{vol}}(r+1)^{d_g}\}$. *For fixed structural parameters, bounded* $\nu_{\mathrm{u}}t$, *and nonsaturated balls,*

$$\begin{aligned} r &= O(\log(1/\varepsilon)), \qquad & d_r &= O((\log(1/\varepsilon))^{d_g}), \\ n_{\mathrm{heat}}(\varepsilon,\delta) &\le \varepsilon^{-3}\exp\left(O(d_g[\log\log(1/\varepsilon)]^2)\right)\log\frac{2}{\delta} \\ &= (1/\varepsilon)^{3+o(1)}\log\frac{2}{\delta}. \end{aligned} \tag{78}$$

The graph arithmetic for each uncached column is $O((\Delta_G+1)\log^{d_g+1}(1/\varepsilon))$.

Where the explicit radius comes from. For $r+1>\theta$, the Poisson Chernoff inequality gives

$$\beta_r \le e^{-\theta}\left(\frac{e\theta}{r+1}\right)^{r+1}.$$

The condition $r+1 \ge 2e\theta$ makes the ratio in parentheses at most $1/2$, after which $r+1 \ge \log_2(1/b_\varepsilon)$ makes the resulting geometric upper bound at most b_ε. The extra lower bound 1 handles the small-radius edge cases. This choice is deliberately simple; it is sufficient rather than minimal.

Practical radius choice. The exact smallest admissible radius is the Poisson quantile

$$r_\varepsilon^{\mathrm{q}} = \min\{r \in \mathbb{N}_0 : \Pr(\mathrm{Pois}(\nu_{\mathrm{u}}t) \ge r+1) \le b_\varepsilon\}.$$

It can replace (74) and is never larger than that sufficient Chernoff choice. The explicit formula is retained in the corollary because it is simple and makes the stated $O(\log(1/\varepsilon))$ bound immediate.

C.4 Complete end-to-end heat-kernel planning algorithm

We now assemble the preceding ingredients into a complete planner. There are three computational layers:

$$\boxed{\text{local graph diffusion}} \longrightarrow \boxed{\text{one recursive value sample}} \longrightarrow \boxed{\text{root confidence amplification}}.$$

The first layer constructs the finite-walk heat weights associated with a sampled anchor. The second layer uses those weights in the localized second-order recursion. The third layer averages independent recursive outputs and applies a median only at the root.

For a fixed radius r, define

$$I_j^{(r)} := \mathrm{supp}\big(H_{t,r}(:,j)\big), \qquad w_i^{\mathcal{I},(s,j)} := H_{t,r}(i,j), \quad i \in I_j^{(r)}. \tag{79}$$

For $t>0$, Theorem 8 gives

$$I_j^{(r)} = B_G(j,r).$$

The heat graph is state-independent in the present construction, so the same set and weights are used whenever anchor j is sampled at any MDP state. For a state-dependent graph family G_s, they would instead depend on (s,j).

Because $H_{t,r}(:,j)$ is already a probability distribution,

$$m_{s,j} := \sum_{i \in I_j^{(r)}} w_i^{\mathcal{I},(s,j)} = 1, \qquad b_{s,j} = 0, \qquad F_{s,t,r}^{\mathrm{HK}}(0) = 0. \tag{80}$$

Thus there is no retained-mass correction and no anchor-dependent value offset.

Lazy access to deterministic heat columns. The planner does not construct all K columns in advance. It constructs a column only when its anchor is first sampled and may cache the resulting sparse vector. Since the graph, t, r, and ν_{u} are fixed before MDP transition sampling, the cached objects contain no reward, successor state, or random value estimate.

The cache is indexed only by anchor j because the present appendix fixes one state-independent graph and one heat configuration. For state-dependent graphs, the cache must instead be indexed by at least (s,j). Caching is optional: recomputing and discarding a column changes graph arithmetic but does not change the statistical guarantee.

Algorithm 5 HEATCOLUMNACCESS: obtain one cached local heat column

Input: anchor j, radius r, diffusion time t, uniformization rate ν_{u}, local weighted-adjacency access, and cache $\mathcal{C}_r$
Output: local set $I_j^{(r)}$ and normalized weights $h^{(j,r)} = H_{t,r}(:, j)$
if $j \in \mathrm{dom}(\mathcal{C}_r)$ **then**
  return $\mathcal{C}_r[j]$
end if
$(I_j^{(r)}, h^{(j,r)}) \leftarrow \textsc{LocalHeatColumn}(j, r, t, \nu_{\mathrm{u}})$
$\mathcal{C}_r[j] \leftarrow (I_j^{(r)}, h^{(j,r)})$
return $(I_j^{(r)}, h^{(j,r)})$

Heat-specialized recursive sample. Since the centered normalized heat backup satisfies $F^{\mathrm{HK}}_{s,t,r}(0) = 0$, its fixed point obeys the tighter magnitude bound

$$B_Q^{\mathrm{HK}} := \frac{R_{\max}}{1-\gamma}. \tag{81}$$

This is the clipping radius used for the local coordinate baselines below. It should not be confused with

$$\bar{B}_Q = \frac{R_{\max} + \gamma T_0 \log 2}{1-\gamma},$$

which is the conservative generic envelope used to define the span S_Q and the common cascade constants.

Algorithm 6 specializes Algorithm 3 to the heat family (79). It makes explicit where graph computation ends and MDP transition sampling begins.

Here $\Pi_{[-B,B]}(x) = \min\{B, \max\{-B, x\}\}$ denotes scalar clipping. The quantity $r_\eta = \eta^{1/3}$ in (45) is an accuracy scale used by the recursion; it is unrelated to the graph radius r.

All calls denoted by different superscripts, all recursive subtrees, and the two quadratic batches are mutually independent. The deterministic heat cache may be shared across all these calls without violating independence. The planner must not cache Q-dependent Gibbs probabilities, sampled actions, rewards, successor states, or stochastic recursive values.

The coarse base routine $\mathsf{Coarse}^{\mathrm{HK},r}_{\eta_0}$ is Algorithm 9 specialized to

$$I_{s,j} = I_j^{(r)}, \qquad w^{\mathcal{I},(s,j)} = H_{t,r}(:, j), \qquad m_{s,j} = 1, \qquad b_{s,j} = 0.$$

Therefore the coarse routine, the fine recursion, and every recursive child use the same fixed finite-walk Bellman operator.

Meaning of the three correction terms. The output of Algorithm 6 contains:

1. the baseline heat backup $\widehat{f}_J$, obtained from estimated action values on the sampled anchor's local graph ball;
2. the linear correction G_{lin}, which corrects the first-order error in the baseline action values; and
3. the product $\overline{D}_1\overline{D}_2/(4T_0)$, which supplies the second-order curvature correction.

The two quadratic batches are independent so that their product estimates a product of conditional means without introducing the diagonal variance term that would appear if one squared a single noisy batch. The returned recursive sample is intentionally not clipped; the proof instead controls its bias and third moment.

Full exact-heat root-value estimator. Algorithm 7 first chooses a radius that makes the deterministic finite-walk approximation error at most $\varepsilon/2$. It then estimates the finite-walk value to statistical accuracy $\varepsilon/2$.

Algorithm 6 One heat-specialized recursive sample $Y_\eta^{\mathrm{HK},r}(s)$

Input: MDP state s, tolerance η, fixed radius r, cache $\mathcal{C}_r$, graph parameters (t, ν_{u}), independent MDP generative model, and fixed cascade constants

if $\eta \geq \eta_0$ **then**
  return $\mathsf{Coarse}_{\eta_0}^{\mathrm{HK},r}(s)$
end if

Set $(r_\eta, N_\eta, \eta_b, \eta_\ell, \eta_q)$ according to (45)

Draw an anchor $J \sim \mu_s$

$(I, h) \leftarrow \textsc{HeatColumnAccess}(J, r, t, \nu_{\mathrm{u}}, \mathcal{C}_r)$

For a fresh generative-model call at (s, i), define

$$X_\vartheta(s,i) := R + \gamma Y_\vartheta^{\mathrm{HK},r}(S'), \qquad (R, S') \sim \mathsf{Gen}(s,i)$$

for each $i \in I$ **do**
  Draw mutually independent copies $X_{\eta_b}^{(1)}(s,i), \ldots, X_{\eta_b}^{(N_\eta)}(s,i)$

$$\widehat{Q}_i \leftarrow \Pi_{[-B_Q^{\mathrm{HK}}, B_Q^{\mathrm{HK}}]} \left(\frac{1}{N_\eta} \sum_{n=1}^{N_\eta} X_{\eta_b}^{(n)}(s,i) \right)$$

end for

$\widehat{m}_J \leftarrow \max_{i \in I} \widehat{Q}_i$

for each $i \in I$ **do**

$$g_i \leftarrow h_i \exp\left(\frac{\widehat{Q}_i - \widehat{m}_J}{T_0} \right)$$

end for

$z_J \leftarrow \sum_{i \in I} g_i$

$\widehat{f}_J \leftarrow \widehat{m}_J + T_0 \log z_J$

$p_i \leftarrow g_i / z_J$ for every $i \in I$

Draw $A_0, A, A' \overset{\mathrm{iid}}{\sim} p$ conditionally on $(J, \widehat{Q})$

$G_{\mathrm{lin}} \leftarrow X_{\eta_\ell}(s, A_0) - \widehat{Q}_{A_0}$

for $q = 1, 2$ **do**

$$\overline{D}_q \leftarrow \frac{1}{m_2} \sum_{h=1}^{m_2} \left[X_{\eta_q}^{(qh)}(s, A) - \widehat{Q}_A - X_{\eta_q}^{\prime(qh)}(s, A') + \widehat{Q}_{A'} \right]$$

end for

return $\widehat{f}_J + G_{\mathrm{lin}} + \dfrac{\overline{D}_1 \overline{D}_2}{4T_0}$

Reading Algorithm 7. The procedure has the following interpretation.

1. It chooses r_ε before observing any MDP transition.
2. At a particular recursive invocation, it samples its own anchor $J \sim \mu_s$.
3. Only the heat column associated with that sampled anchor is constructed or retrieved.
4. Algorithm 4 uses only graph adjacency lists inside $B_G(J, r_\varepsilon)$.
5. The heat weights and the current baseline estimates determine the local Gibbs distribution.
6. Every baseline action, Gibbs-sampled action, and MDP query made at that invocation lies in $B_G(J, r_\varepsilon)$.
7. Recursive child calls sample their own anchors and use their own local balls.
8. Independent recursive outputs are averaged within blocks, while the median is applied only once at the root.

The last distinction is important. Locality is relative to the anchor drawn at each recursive call; it does not mean that the entire recursion remains inside the graph ball of the root call's anchor.

Why the final error is at most ε. The statistical target of the recursive estimator is the finite-walk fixed point. By the cascade bias and moment guarantees,

$$\Pr\left\{ \left| \widehat{V}_{\mathrm{HK}}(s_0) - V_{t,r_\varepsilon}^\star(s_0) \right| \leq \frac{\varepsilon}{2} \right\} \geq 1 - \delta.$$

Algorithm 7 HEATKERNELPLAN: certified local estimation of the exact centered heat value

Input: root state s_0, accuracy ε, confidence δ, action graph G, diffusion time $t \geq 0$, temperature $T_0 > 0$, discount $\gamma \in (0,1)$, reward bound $R_{\max}$, known $\nu_{\mathrm{u}} > 0$ satisfying $\nu_{\mathrm{u}} \geq d_{\max}$, local weighted-adjacency access, independent MDP generative model, samplers for $J \sim \mu_s$, and constants $(\eta_0, C_{\mathrm{b}}, C_3, c_b, c_q, c_N, a, m_2)$

$\bar{B}_Q \leftarrow \dfrac{R_{\max} + \gamma T_0 \log 2}{1-\gamma}$

$S_Q \leftarrow 2\bar{B}_Q$

if $t = 0$ **then**

  $r_\varepsilon \leftarrow 0$

else

  $b_\varepsilon \leftarrow \min\left\{\dfrac{1}{2}, \dfrac{\exp((1-\gamma)\varepsilon/(2T_0)) - 1}{\exp(S_Q/T_0) - 1}\right\}$

  $r_\varepsilon \leftarrow \lceil \max\{1,\ 2e\nu_{\mathrm{u}} t,\ \log_2(1/b_\varepsilon)\} \rceil - 1$

end if

Initialize an empty deterministic cache $\mathcal{C}_{r_\varepsilon}$

$u \leftarrow \varepsilon/2$

$\eta \leftarrow u/(2C_{\mathrm{b}})$; require $\eta \leq \eta_0$

$M \leftarrow \left\lceil \dfrac{16 C_3^{2/3}}{u^2} \right\rceil$

$B \leftarrow 2\left\lceil 4 \log \dfrac{2}{\delta} \right\rceil + 1$

for $b = 1, \ldots, B$ **do**

  for $m = 1, \ldots, M$ **do**

    $Z_{bm} \leftarrow Y_\eta^{\mathrm{HK}, r_\varepsilon}(s_0)$ from Algorithm 6, using a fresh recursion tree and the shared deterministic cache $\mathcal{C}_{r_\varepsilon}$

  end for

  $\overline{Z}_b \leftarrow \dfrac{1}{M} \sum_{m=1}^{M} Z_{bm}$

end for

return $\widehat{V}_{\mathrm{HK}}(s_0) \leftarrow \mathrm{median}\{\overline{Z}_1, \ldots, \overline{Z}_B\}$

The radius choice is deterministic and gives

$$\left|V_{t,r_\varepsilon}^\star(s_0) - V_t^\star(s_0)\right| \leq \frac{\delta_{r_\varepsilon}(S_Q)}{1-\gamma} \leq \frac{\varepsilon}{2}.$$

Consequently,

$$\begin{aligned}\left|\widehat{V}_{\mathrm{HK}}(s_0) - V_t^\star(s_0)\right| &\leq \underbrace{\left|\widehat{V}_{\mathrm{HK}}(s_0) - V_{t,r_\varepsilon}^\star(s_0)\right|}_{\text{statistical error}} \\ &\quad + \underbrace{\left|V_{t,r_\varepsilon}^\star(s_0) - V_t^\star(s_0)\right|}_{\text{heat-truncation bias}} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\end{aligned} \tag{82}$$

with probability at least $1 - \delta$. No split of the failure probability is needed because the second term is deterministic.

What is and is not counted. Algorithm 7 never forms the dense matrix H_t, the full matrix H_{t,r_ε}, or

$$d_\varepsilon = \max_j |B_G(j, r_\varepsilon)|.$$

The quantity d_ε is used only to state the complexity bound; it is not computed by the planner.

Algorithm 4 uses no MDP transition samples. Its graph cost per uncached column is

$$O\left((r_\varepsilon + 1)(\Delta_G + 1) d_\varepsilon\right).$$

If N_{anc} different anchors are cached, total column-construction work and storage are bounded respectively by

$$O\left(N_{\mathrm{anc}}(r_\varepsilon+1)(\Delta_G+1)d_\varepsilon\right), \qquad O(N_{\mathrm{anc}}d_\varepsilon).$$

In contrast, (76) counts only calls to the MDP generative model.

The algorithm passes $H_{t,r_\varepsilon}(:,j)$ directly as normalized weights. It does not construct $-\lambda\log H_{t,r_\varepsilon}(i,j)$, because the truncated kernel is zero outside its graph ball and the corresponding logarithmic entries would be $+\infty$.

Finally, the returned quantity estimates the centered heat value $V_t^\star(s_0)$. If the desired target is the ordinary uncentered uniform-prior LOT value from Theorem 7, use

$$V_{\mathrm{logH}}^\star(s_0)=V_t^\star(s_0)-\frac{T_0\log K}{1-\gamma}.$$

End-to-end interpretation. To estimate the exact heat-regularized root value to accuracy ε, choose r so that the finite-walk bias is at most $\varepsilon/2$, then run the localized planner on $H_{t,r}$ to statistical accuracy $\varepsilon/2$. If an r-hop ball contains only $O((r+1)^{d_g})$ actions, the required radius grows logarithmically with $1/\varepsilon$ and the local action count grows polylogarithmically. The planner can then avoid an explicit factor K until the ball saturates. The claim is not that heat always helps: it helps when the action graph is locally accessible, its balls grow slowly, and the chosen diffusion scale does not spread mass across the whole graph.

Scope. The construction computes $H_{t,r}$ exactly and approximates H_t; it does not recover exact entries of H_t inside a finite ball. In particular, $H_{t,r}$ is not cropped exact heat, not heat on the subgraph induced by $B_G(j,r)$, and not itself a semigroup in t; it is the endpoint law conditioned on a Poisson walk length of at most r. The transition bound targets the scalar centered heat-regularized root value, not the hard-max value, best-action identification, simple regret, or policy return. A K-free arithmetic interpretation additionally requires local adjacency access, efficient sampling from μ_s (for example an alias table), a supplied positive upper bound $\nu_{\mathrm{u}}\geq d_{\max}$, bounded $\nu_{\mathrm{u}}t$, polynomial ball growth, and balls that have not saturated at all K actions. Finding $d_{\max}$ or building an anchor sampler may itself require one-time global preprocessing; these costs are not included in the transition count. A loose ν_{u} increases $\theta=\nu_{\mathrm{u}}t$ and therefore the sufficient radius, so the tightest cheaply available bound is preferable. The direct-heat statements use only that μ_s is a probability distribution; they do not require a K-independent lower bound on every anchor weight.

D Additional Assumptions and Full Theorem Statements

The appendix follows the same RL questions as the main paper. It first gives the complete learning assumptions and finite-sample estimators (Section D.1), followed by the smoothness identities used by the root-value estimator (Section D.2) and the fixed action-pruning results (Section D.4). Sections C.1 and C.2 separate small exact-heat neighborhoods from the locally computable truncated-heat approximation. The remaining statements cover policy geometry, perturbations, trajectory sampling, spectral approximation, and continuous actions. Complete proofs then follow.

Proof coverage. Sections F–G contain one explicit proof route for each active theorem, lemma, proposition, and corollary. The proof map at the start of Section G lists every result and the thematic proof subsection in which it is established.

D.1 Additional tabular and finite-sample learning results

Sharpness and a single trajectory. The two parts of Theorem 4 answer different questions: part (i) gives a finite-sample error after estimating the geometry once and then freezing it, whereas part (ii) proves asymptotic convergence while the geometry and values are updated together. Neither part gives a joint finite-time rate for the complete online procedure. The $n^{-1/2}$ dependence is minimax-sharp on the explicit zero-reward, three-state, two-action family of Theorem 11, even when transition queries are chosen adaptively; this proves sharpness on that subclass, not that every dependence in (35) is globally optimal. A separate result treats data from one trajectory rather than independent resets: under

stationarity, mixing, and sufficient coverage, Theorem 19 gives a high-probability geometry-only fixed-point bound with leading rate $\widetilde{O}((n\underline{\nu})^{-1/2})$, where $\underline{\nu}$ lower-bounds the stationary probability of each state–action pair, so the least frequently visited pairs determine the rate. Corollary 6 supplies the predictable empirical-diffusion construction used in Part (ii).

Predictable online implementation. In this paragraph, the subscript t denotes the online iteration, not the diffusion time used in (34). Let $\widehat{P}_t$ contain only transitions observed before iteration t, let Geom be a bounded nonnegative map, define $\widehat{C}_t = (\widehat{C}_{s,t})_s := \mathsf{Geom}(\widehat{P}_t)$ and $C = (C_s)_s := \mathsf{Geom}(P)$, and assume (36). Here Geom converts empirical transition frequencies into pairwise action costs. Concretely, we use

$$(\widehat{C}_{s,t})_{ab} = \kappa_C \ \mathrm{TV}\Big(\widehat{P}_t(\cdot \mid s,a), \widehat{P}_t(\cdot \mid s,b)\Big)\,,$$

where $\widehat{P}_t$ contains only transitions observed before iteration t. Under infinite visitation, $\widehat{P}_t \to P$ and therefore $\widehat{C}_t \to C := \mathsf{Geom}(P)$ uniformly almost surely. The update order is essential: the transition observed at iteration t enters $\widehat{P}_{t+1}$ only after the temporal-difference target has been formed, hence $\widehat{C}_t$ is determined by the past and cannot depend on the random successor state used in that same target. This past-measurability is the meaning of *predictable*. The complete ONLINELOTQ loop is Algorithm 8 in Appendix D.1. At state s, the plug-in policy is the distribution returned by LOTBACKUP for $Q_t(s,\cdot)$ and $\widehat{C}_{s,t}$; it is an analytic readout of the backup already being computed, not a separately trained actor. In the proof, geometry mismatch contributes a predictable temporal-difference drift bounded by $\gamma\tau \sup_s \|\widehat{C}_{s,t} - C_s\|_{\max}$, which tends to zero almost surely by (36). The remaining sampling noise is a bounded martingale difference, and asynchronous stochastic approximation absorbs the vanishing drift. Exact constants, assumptions, and proofs are in Appendix D.1.

Algorithm 8 gives the online loop analyzed in Theorem 4(ii).

Algorithm 8 ONLINELOTQ: tabular Q-learning with same-stream geometry

Input: Geom, full-support $\{\rho_s, \mu_s\}_s$, τ, λ; set $T_0 \leftarrow \tau\lambda$; initialize Q_0, $N_0 \equiv 0$, $\widehat{P}_0$
for $t = 0, 1, 2, \ldots$ **do**
  $\widehat{C}_t \leftarrow \mathsf{Geom}(\widehat{P}_t)$ *(uses transitions from times $k < t$ only)*
  Choose a_t (behavior policy ensuring infinite visitation); observe r_t, s_{t+1}; $s' \leftarrow s_{t+1}$
  for each anchor j at s' **do**
    $\kappa_{s',j,t} \leftarrow \sum_i \rho_{s'}(i) e^{-(\widehat{C}_{s',t})_{ij}/\lambda}$; $w_{i,t}^{(s',j)} \leftarrow \rho_{s'}(i) e^{-(\widehat{C}_{s',t})_{ij}/\lambda} / \kappa_{s',j,t}$; $b_{s',j,t} \leftarrow T_0 \log \kappa_{s',j,t}$
  end for
  $(F_t, \pi_t(s')) \leftarrow \mathrm{LOTBACKUP}(Q_t(s',\cdot), \{(w_t^{(s',j)}, b_{s',j,t})\}_j, \mu_{s'}, T_0)$
  $Q_{t+1} \leftarrow Q_t$; $Q_{t+1}(s_t,a_t) \leftarrow (1-\alpha_t) Q_t(s_t,a_t) + \alpha_t[r_t + \gamma F_t]$, $\alpha_t = \alpha_{N_t(s_t,a_t)}(s_t,a_t)$
  Update N_{t+1} and $\widehat{P}_{t+1}$ using (s_t, a_t, s_{t+1})
end for

The results below complement Theorem 4: the standing sampling assumption, the same-stream convergence theorem in full, the explicit reward-free batch estimator with its constants, and the sharp explicit subclass.

Interpreting the reward-free batch construction. In (33), the first inequality holds because every transport plan has total mass one, so changing every cost entry by at most η changes its objective by at most $\tau\eta$; the Bellman continuation contributes the factor γ, and contraction accumulates the one-step error by $1/(1-\gamma)$. Theorem 18 also controls the induced policy error.

Using the same distribution ν for every action makes the comparison fair: differences in the samples can then be attributed to the actions rather than to different starting-state distributions, and the choice of ν also determines which states receive more importance.

For one transition $s \to s'$, the vector $x = e_{s'} - e_s$ removes one unit of state occupancy from s and adds it to s'; if P_a is the row-stochastic transition matrix satisfying $P_a(s,s') = P(s' \mid s,a)$ and ν

is viewed as a column vector, then $z_a = P_a^\top \nu - \nu$ is the ν-averaged occupancy change produced by action a. It summarizes the average effect of the action, not its complete state-conditional transition kernel. For a sampled transition $s \to s'$, the matrix $xx^\top/2$ is a normalized Laplacian contribution from the undirected state edge $\{s, s'\}$, the factor $1/2$ normalizing its operator norm as $\|xx^\top/2\|_{\mathrm{op}} \le 1$. Consequently, L_ν pools transition edges from all actions into one symmetric state-space diffusion geometry: the outer product $xx^\top$ deliberately removes transition direction, while the signed mean z_a retains the action-specific net occupancy change.

Equivalently, $\phi_a^t := \sqrt{\kappa_C}\, e^{-tL_\nu} z_a$ is the learned embedding of action a, and $C_{ab}^t = \|\phi_a^t - \phi_b^t\|_2^2$ is the squared Euclidean distance between the two action embeddings. At $t = 0$, the construction compares raw average occupancy changes; larger t smooths differences across states that are strongly connected by the dynamics, although an excessively large t can erase useful distinctions. A small value of C_{ab}^t therefore means that actions a and b have similar smoothed, ν-averaged effects; it does not imply $P(\cdot \mid s, a) = P(\cdot \mid s, b)$ for every state s. Because the construction uses no reward observations, the same geometry can be reused across reward functions sharing the same state–action system, dynamics, distribution ν, and geometry settings.

Assumption 3 (Asynchronous tabular sampling and predictable geometry)**.** For Theorem 9, let $\mathcal{S}, \mathcal{A}$ be finite, the action priors and reference distributions be fixed and have full support, and C be a bounded nonnegative target cost family. Before observing (r_t, s_{t+1}), $(Q_t, s_t, a_t, \widehat{C}_t)$ is $\mathcal{F}_t$-measurable, and

$$|r_t| \le R_{\max}, \qquad \mathbb{E}[r_t \mid \mathcal{F}_t, s_t, a_t] = r(s_t, a_t), \qquad \Pr(s_{t+1} = s' \mid \mathcal{F}_t, s_t, a_t) = P(s' \mid s_t, a_t).$$

For some deterministic $\bar{C} < \infty$,

$$0 \le (\widehat{C}_{s,t})_{ij} \le \bar{C}, \qquad \varepsilon_t := \sup_s \left\| \widehat{C}_{s,t} - C_s \right\|_\infty \longrightarrow 0 \quad \text{a.s.} \tag{83}$$

Every state–action pair is visited infinitely often. If $N_t(s, a)$ counts its previous updates, its stepsizes satisfy $\alpha_n(s, a) \in [0, 1]$, $\sum_n \alpha_n(s, a) = \infty$, and $\sum_n \alpha_n(s, a)^2 < \infty$.

Theorem 9 (Tabular LOT Q-learning with online geometry)**.** *Under Assumptions 1 and 3, consider*

$$\begin{aligned} Q_{t+1}(s, a) &= Q_t(s, a), && (s, a) \ne (s_t, a_t), \\ Q_{t+1}(s_t, a_t) &= (1 - \alpha_t) Q_t(s_t, a_t) \\ &\quad + \alpha_t \left[r_t + \gamma F_{s_{t+1}}^{\widehat{C}_{s_{t+1},t}}(Q_t(s_{t+1}, \cdot)) \right], \end{aligned} \tag{84}$$

where $\alpha_t = \alpha_{N_t(s_t,a_t)}(s_t, a_t)$. From any finite initialization,

$$Q_t \longrightarrow Q_C^\star \qquad \text{a.s.} \tag{85}$$

Moreover, the induced tabular policy satisfies, for every s,

$$\nabla_Q F_s^{\widehat{C}_{s,t}}(Q_t(s, \cdot)) \longrightarrow \nabla_Q F_s^{C_s}(Q_C^\star(s, \cdot)) \qquad \text{a.s.} \tag{86}$$

Fixed geometry is the case $\widehat{C}_t \equiv C$; neither $\sum_t \alpha_t \varepsilon_t < \infty$ nor separated learning timescales are required.

Indeed, the predictable backup and target errors are at most $\tau\varepsilon_t$ and $\gamma\tau\varepsilon_t \to 0$. The remaining noise is a bounded martingale difference and the limiting operator a γ-contraction, so asynchronous stochastic approximation absorbs the vanishing drift (Appendix D.1).

Theorem 10 (Reward-free transition samples to LOT targets)**.** *Let $D = |\mathcal{S}|$ and fix a task-independent distribution ν. For each action a, draw from a generative model n independent samples $S_{a\ell} \sim \nu$ and $S'_{a\ell} \sim P(\cdot \mid S_{a\ell}, a)$, independently over (a, ℓ). With $x_{a\ell} = e_{S'_{a\ell}} - e_{S_{a\ell}}$, define*

$$L_\nu = \frac{1}{K} \sum_a \mathbb{E} \frac{x_{a1} x_{a1}^\top}{2}, \quad z_a = \mathbb{E} x_{a1}, \quad \widehat{L}_\nu = \frac{1}{Kn} \sum_{a,\ell} \frac{x_{a\ell} x_{a\ell}^\top}{2}, \quad \widehat{z}_a = \frac{1}{n} \sum_\ell x_{a\ell}.$$

For deterministic $t \ge 0$ and $\kappa_C > 0$, let

$$C_{ab}^t = \kappa_C \left\| e^{-tL_\nu}(z_a - z_b) \right\|_2^2, \qquad \widehat{C}_{ab}^t = \kappa_C \left\| e^{-t\widehat{L}_\nu}(\widehat{z}_a - \widehat{z}_b) \right\|_2^2.$$

For $N = Kn$ and $\delta \in (0,1)$, set

$$\eta_L = \sqrt{\frac{\log(4D/\delta)}{2N}} + \frac{2\log(4D/\delta)}{3N}, \qquad \eta_z = \sqrt{\frac{2}{n}} + 2\sqrt{\frac{\log(2K/\delta)}{n}},$$

and $\varepsilon_{n,\delta} = \kappa_C(16t\eta_L + 8\sqrt{2}\,\eta_z)$. With probability at least $1-\delta$,

$$\begin{aligned} \left\|\widehat{C}^t - C^t\right\|_\infty &\le \varepsilon_{n,\delta}, \\ \sup_{s,Q} |F_s^{\widehat{C}}(Q) - F_s^C(Q)| &\le \tau\varepsilon_{n,\delta}, \\ \left\|Q^\star_{\widehat{C}} - Q^\star_C\right\|_\infty &\le \frac{\gamma\tau}{1-\gamma}\varepsilon_{n,\delta}. \end{aligned} \tag{87}$$

At the respective fixed points, both policy-KL directions are at most $2\varepsilon_{n,\delta}/[\lambda(1-\gamma)]$. On this reward-independent event, the same frozen cost and all displayed bounds hold simultaneously for every bounded-reward task sharing P.

Thus the fixed-point error is $\widetilde{O}(\kappa_C\gamma\tau(1-\gamma)^{-1}[t/\sqrt{Kn} + 1/\sqrt{n}])$, using Kn samples. The data estimate only C; both fixed points still use the true transition kernel. This is therefore a geometry-learning guarantee, not a model-learning bound or a rate for the simultaneous recursion in Theorem 9.

Theorem 11 (Sharp batch rate on an explicit diffusion family)**.** *For fixed $\gamma \in (0,1)$ and $\tau, \lambda, h > 0$, there is an explicit zero-reward, three-state, two-action transition-induced diffusion family whose minimax risk over estimators using at most n adaptive generative-model transition samples, $\mathfrak{R}_n := \inf_{\widehat{Q}} \sup_p \mathbb{E}_p \left\|\widehat{Q} - Q^\star_p\right\|_\infty$, satisfies*

$$\frac{a_{\gamma,\tau,\lambda,h}}{2}\left(1 - \frac{1}{\sqrt{15}}\right) n^{-1/2} \le \mathfrak{R}_n \le \frac{2\gamma\tau h}{1-\gamma^2}\, n^{-1/2}, \tag{88}$$

where

$$a_{\gamma,\tau,\lambda,h} = \frac{2\gamma\tau h e^{-2h}}{(1-\gamma^2)\{1 + \exp(2e^{-3h/2}/\lambda)\}} > 0.$$

This subclass result is not a matched lower bound for the estimator in Theorem 10.

Theorems 10, 11, and 19 cover an independent batch, an explicit subclass with sharp $n^{-1/2}$ rate, and one exploratory trajectory. None gives a finite-time rate for simultaneous geometry and Q-learning; the $n^{-1/2}$ geometry rate is also distinct from the $n_{\text{plan}}^{-1/3}$ root-planning accuracy scale.

Successor-feature costs and spectral weights. Successor features provide one dynamics-derived action embedding, but the spectral weight must match the chosen feature map. For a reversible policy chain, $L_{\text{rw}} = I - P^\pi$ is self-adjoint in the stationary weighted inner product (or after the usual symmetric similarity transform), and $M^\pi = (I - \gamma P^\pi)^{-1} = ((1-\gamma)I + \gamma L_{\text{rw}})^{-1}$ Dayan (1993); Stachenfeld et al. (2017). Squared Euclidean distances between rows of M^π therefore use the squared spectral weight $((1-\gamma) + \gamma\lambda)^{-2}$. If instead we want $g(\lambda) = ((1-\gamma) + \gamma\lambda)^{-1}$, we explicitly use the square-root successor feature $((1-\gamma)I + \gamma L_{\text{rw}})^{-1/2}$. For a nonreversible chain, P^π can be nonnormal, so either spectral construction requires an explicit reversible symmetrization.

D.2 Smoothness and second-order identities used by the planner

Proposition 3 (Bellman contraction)**.** *Under Assumption 1, for every bounded cost family $\{C_s\}_s$, F_s^{LOT} is monotone and 1-Lipschitz in sup norm, so $\mathcal{T}^{\text{LOT}}$ is a γ-contraction with unique fixed point $Q^\star_{\text{LOT}}$. Writing $\bar{W} = \sup_{s,Q} |F_s^{\text{LOT}}(Q) - \max_a Q_a|/\tau < \infty$,*

$$\|Q^\star_{\text{LOT}} - Q^\star\|_\infty \le \frac{\gamma\tau\bar{W}}{1-\gamma}. \tag{89}$$

Proposition 4 (Matched-MaxEnt limit and value ordering)**.** *Let $T_0 = \tau\lambda$ and $F_s^\rho(Q) = T_0 \log\sum_i \rho_s(i) e^{Q_i/T_0}$, and let $Q^\star_C$, $Q^\star_\rho$, $Q^\star$ be the Bellman fixed points of the LOT aggregator*

F_s^C, of F_s^ρ, and of the hard maximum, respectively. At $C_s = 0$, LOT equals F_s^ρ and its policy is the ρ_s-tilted softmax. For every nonnegative cost family with $\bar{C} = \sup_s \|C_s\|_\infty < \infty$,

$$F_s^C(Q) \le F_s^\rho(Q) \le \max_i Q_i, \tag{90}$$

$$0 \le F_s^\rho(Q) - F_s^C(Q) \le \tau \bar{C}, \tag{91}$$

$$Q_C^\star \le Q_\rho^\star \le Q^\star, \|Q_\rho^\star - Q_C^\star\|_\infty \le \frac{\gamma\tau\bar{C}}{1-\gamma}. \tag{92}$$

This is value ordering, not policy-return ordering; it applies to ordinary uncentered LOT, while centered direct heat must first remove its known $T_0 \log K$ offset. Geometry is thus not claimed to reduce regularization bias.

For completeness, this section records the LOT derivatives inherited from Dam (2026): smoothness of the soft policy, the graph-Laplacian form of its Hessian, a Lipschitz-Hessian Taylor bound, and the paired-sample estimator for the quadratic Taylor term. These are exactly the quantities used by the recursive root-value estimator.

Proposition 5 (Smooth soft policy)**.** *Under Assumption 1, F_s^{LOT} is convex and C^1 in Q, with $\nabla_Q F_s^{\mathrm{LOT}}(Q) = \pi^{\mathrm{LOT}}$, the unique optimizer of* (8)*; moreover, $Q \mapsto \pi^{\mathrm{LOT}}$ is smooth.*

Define the *transport co-occurrence graph*

$$A_{ik} = \sum_j \frac{\Gamma^\star_{ij}\Gamma^\star_{kj}}{\mu_s(j)}, \qquad L_A = \mathrm{diag}(A\mathbf{1}) - A, \tag{93}$$

where A_{ik} is large when actions i and k send substantial mass to the same anchors.

Theorem 12 (Curvature is a graph Laplacian)**.** *Under Assumption 1,*

$$\begin{aligned}\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) &= \frac{1}{\tau\lambda}\left(\mathrm{diag}(\pi^{\mathrm{LOT}}) - \Gamma^\star \,\mathrm{diag}(\mu_s)^{-1}\Gamma^{\star\top}\right) \\ &= \frac{1}{\tau\lambda} L_A.\end{aligned} \tag{94}$$

Hence $A\mathbf{1} = \pi^{\mathrm{LOT}}$, L_A is symmetric PSD, and $L_A\mathbf{1} = 0$. Thus F_s^{LOT} is convex, $F_s^{\mathrm{LOT}}(Q + c\mathbf{1}) = F_s^{\mathrm{LOT}}(Q) + c$, and

$$\left\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\right\|_{\mathrm{op}} = \frac{\lambda_{\max}(L_A)}{\tau\lambda}, \qquad \lambda_{\max}(L_A) \le \frac{1}{2}$$

by Theorem 1.

Corollary 3 (Generalizing the softmax Jacobian)**.** *If $C^L = 0$, then $\Gamma^\star = \pi^{\mathrm{LOT}}\mu_s^\top$, $A = \pi^{\mathrm{LOT}}\pi^{\mathrm{LOT}\top}$, and* (94) *reduces to*

$$\frac{1}{\tau\lambda}\left(\mathrm{diag}(\pi^{\mathrm{LOT}}) - \pi^{\mathrm{LOT}}\pi^{\mathrm{LOT}\top}\right),$$

the prior-weighted softmax Jacobian. LOT therefore replaces the multinomial covariance by a geometry-aware transport-graph Laplacian.

Equations (13) and (94) describe the same Hessian as a covariance mixture and a graph Laplacian. For a diffusion-derived cost C_t, t, τ, and λ determine which local action differences receive curvature.

D.3 Second-order smooth planning implications

The covariance form supplies the derivatives used by Dam (2026). Recursive estimation additionally requires an action-count-free remainder bound and an unclipped recursion with per-child accuracies.

Theorem 13 (Lipschitz Hessian, cubic remainder, and quadratic estimator)**.** *Under Assumption 1 and finite $\mathcal{A}$, for all $Q, \widehat{Q} \in \mathbb{R}^K$,*

$$\left\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) - \nabla_Q^2 F_s^{\mathrm{LOT}}(\widehat{Q})\right\|_{\mathrm{op}} \le \frac{3}{2\tau^2\lambda^2}\left\|Q - \widehat{Q}\right\|_2. \tag{95}$$

Consequently, for $\Delta = Q - \widehat{Q}$,

$$\left| F_s^{\mathrm{LOT}}(Q) - F_s^{\mathrm{LOT}}(\widehat{Q}) - \langle \nabla F_s^{\mathrm{LOT}}(\widehat{Q}),\, \Delta \rangle - \frac{1}{2}\Delta^\top \nabla^2 F_s^{\mathrm{LOT}}(\widehat{Q})\Delta \right| \le \frac{\|\Delta\|_2^3}{4\tau^2\lambda^2}. \tag{96}$$

If $J \sim \mu_s$ and, conditionally on J, $A, A' \overset{\text{i.i.d.}}{\sim} p^{(J)}(\widehat{Q})$, then

$$\begin{aligned} \Delta^\top \nabla^2 F_s^{\mathrm{LOT}}(\widehat{Q})\Delta &= \frac{1}{\tau\lambda}\mathbb{E}\Big[\mathrm{Var}_{p^{(J)}(\widehat{Q})}(\Delta_A)\Big] \\ &= \frac{1}{2\tau\lambda}\mathbb{E}\big[(\Delta_A - \Delta_{A'})^2\big]. \end{aligned} \tag{97}$$

If $\widetilde{\Delta}^{(1)}, \widetilde{\Delta}^{(2)}$ are independent conditional on (J, A, A') and each sampled coordinate is conditionally unbiased for the corresponding coordinate of Δ, then

$$\widehat{q}_2 = \frac{1}{4\tau\lambda}\big(\widetilde{\Delta}_A^{(1)} - \widetilde{\Delta}_{A'}^{(1)}\big)\big(\widetilde{\Delta}_A^{(2)} - \widetilde{\Delta}_{A'}^{(2)}\big) \tag{98}$$

is unbiased for $\frac{1}{2}\Delta^\top \nabla^2 F_s^{\mathrm{LOT}}(\widehat{Q})\Delta$.

D.4 Error-controlled action pruning and the recursive root-value estimator

Write $T_0 = \tau\lambda$. For each (s, j), let $w^{(s,j)} \in \Delta_{\mathcal{A}}$ and let $b_{s,j}$ depend only on the geometry. The planner uses the normalized-kernel form

$$\begin{aligned} f_{s,j}(Q) &= b_{s,j} + T_0 \log \sum_i w_i^{(s,j)} e^{Q_i/T_0}, \\ \mathcal{F}_s(Q) &= \sum_j \mu_s(j) f_{s,j}(Q), \\ p_i^{(s,j)}(Q) &= \frac{w_i^{(s,j)} e^{Q_i/T_0}}{\sum_k w_k^{(s,j)} e^{Q_k/T_0}}. \end{aligned} \tag{99}$$

For ordinary LOT,

$$\begin{aligned} G_{s,j} &= \sum_k \rho_s(k) e^{-C_{s,kj}/\lambda}, \\ w_i^{(s,j)} &= \rho_s(i) e^{-C_{s,ij}/\lambda} / G_{s,j}, \\ b_{s,j} &= T_0 \log G_{s,j}, \end{aligned} \tag{100}$$

so $\mathcal{F}_s = F_s^{\mathrm{LOT}}$; centered heat-kernel LOT instead uses $w^{(s,j)} = H_t(:, j)$ and $b_{s,j} = 0$. This normalization makes the planning constants depend on a centered value envelope rather than directly on $C_{\max}$ or the smallest kernel entry.

Lemma 3 (Gibbs-weighted cubic remainder)**.** *Let $I \subset \mathbb{R}$ have span S_I and $Q, \widehat{Q} \in I^K$. The second-order Taylor remainder of $f_{s,j}$ satisfies*

$$|\mathcal{R}_{s,j}(Q, \widehat{Q})| \le \frac{4e^{S_I/T_0}}{3T_0^2} \sum_i w_i^{(s,j)} |Q_i - \widehat{Q}_i|^3. \tag{101}$$

Thus coordinatewise $O(\eta)$ third-moment error gives $O(\eta)$ expected Taylor bias without an additional factor of K.

For fixed nonempty neighborhoods $\mathcal{I} = \{I_{s,j} \subseteq \mathcal{A}\}$, define

$$\begin{aligned} Z_{s,j}^{\mathcal{I}}(Q) &= \sum_{i \in I_{s,j}} w_i^{(s,j)} e^{Q_i/T_0}, \\ f_{s,j}^{\mathcal{I}}(Q) &= b_{s,j} + T_0 \log Z_{s,j}^{\mathcal{I}}(Q), \\ p_i^{\mathcal{I},(s,j)}(Q) &= \frac{\mathbf{1}\{i \in I_{s,j}\} w_i^{(s,j)} e^{Q_i/T_0}}{Z_{s,j}^{\mathcal{I}}(Q)}, \\ F_s^{\mathcal{I}}(Q) &= \sum_j \mu_s(j) f_{s,j}^{\mathcal{I}}(Q). \end{aligned} \tag{102}$$

These neighborhoods are fixed from geometry before transition sampling and never depend on sampled or estimated values.

Lemma 4 (Uniform error of fixed action pruning)**.** *Let* $\alpha_{s,j} = w^{(s,j)}(I^c_{s,j}) < 1$ *and* $\operatorname{osc}(Q) \le S$. *Then*

$$\zeta_{s,j}(S) = e^{S/T_0}\frac{\alpha_{s,j}}{1-\alpha_{s,j}}, \tag{103}$$

$$0 \le f_{s,j}(Q) - f^{\mathcal{I}}_{s,j}(Q) \le T_0 \log(1+\zeta_{s,j}(S)). \tag{104}$$

Consequently,

$$0 \le \mathcal{F}_s(Q) - F^{\mathcal{I}}_s(Q) \le \Delta_{\mathcal{I}}(S), \qquad \Delta_{\mathcal{I}}(S) := T_0 \sup_s \sum_j \mu_s(j)\log(1+\zeta_{s,j}(S)). \tag{105}$$

Appendix D.4 also gives policy-mass, total-variation, and ordinary-LOT cost-ball consequences.

Let $\mathcal{T}_{\mathcal{F}}$ and $\mathcal{T}^{\mathcal{I}}$ use $\mathcal{F}_s$ and $F^{\mathcal{I}}_s$ in the Bellman backup, with fixed points $Q^\star_{\mathcal{F}}, Q^\star_{\mathcal{I}}$ and state values

$$V^\star_{\mathcal{F}}(s) = \mathcal{F}_s(Q^\star_{\mathcal{F}}(s,\cdot)), \qquad V^\star_{\mathcal{I}}(s) = F^{\mathcal{I}}_s(Q^\star_{\mathcal{I}}(s,\cdot)).$$

Theorem 14 (Bellman error from action pruning)**.** *If* $\operatorname{osc}(Q^\star_{\mathcal{F}}(s,\cdot)) \le S_Q$ *for every* s, *then, pointwise,*

$$\begin{aligned} 0 \le Q^\star_{\mathcal{F}} - Q^\star_{\mathcal{I}}, \qquad \|Q^\star_{\mathcal{F}} - Q^\star_{\mathcal{I}}\|_\infty &\le \frac{\gamma\Delta_{\mathcal{I}}(S_Q)}{1-\gamma}, \\ 0 \le V^\star_{\mathcal{F}}(s) - V^\star_{\mathcal{I}}(s) &\le \frac{\Delta_{\mathcal{I}}(S_Q)}{1-\gamma}. \end{aligned} \tag{106}$$

Under the envelope below, $S_Q = 2\bar{B}_Q$ *is valid.*

Theorem 15 (Localized second-order recursive estimator)**.** *Assume* (99), $T_0 > 0$, *an independent generative model,* $|R| \le R_{\max}$, $\gamma < 1$, *and* $B_F := \sup_s |\mathcal{F}_s(0)| < \infty$. *Fix* $\mathcal{I}$ *with* $w^{(s,j)}(I_{s,j}) \ge 1/2$, *let* $d_{\mathcal{I}} = \sup_{s,j}|I_{s,j}|$, *and put*

$$\bar{B}_F = B_F + T_0\log 2, \qquad \bar{B}_Q = \frac{R_{\max} + \gamma\bar{B}_F}{1-\gamma}.$$

There are constants $\eta_0, C_{\mathrm{b}}, C_3, c_0, c_1 > 0$, *depending only on* $(R_{\max}, \gamma, T_0, B_F)$ *and the fixed cascade choices, such that the unclipped sample* $Y^{\mathcal{I}}_\eta(s)$ *in Algorithm 3 terminates and, for* $0 < \eta \le \eta_0$,

$$\left|\mathbb{E}Y^{\mathcal{I}}_\eta(s) - V^\star_{\mathcal{I}}(s)\right| \le C_{\mathrm{b}}\eta, \qquad \sup_{s,\eta}\mathbb{E}|Y^{\mathcal{I}}_\eta(s)|^3 \le C_3. \tag{107}$$

Its worst-case transition cost satisfies

$$C(\eta) \le \frac{c_0}{\eta}(c_0 d_{\mathcal{I}})^{c_1[1+\log\log(e+\bar{B}_Q/\eta)]}. \tag{108}$$

Apart from B_F, *the constants do not depend on* $K, d_{\mathcal{I}}, C_{\max}, \rho_{\min}, \mu_{\min}$, *or* $\mathcal{I}$. *For uncentered LOT,* B_F *itself may depend on the cost scale. Bias, third moment, termination, and cost are conclusions, not recursive assumptions. The fine sample is not clipped;* $I_{s,j} = \mathcal{A}$ *recovers the dense result with* $d_{\mathcal{I}} = K$.

D.5 Additional results for the local policy metric

Corollary 4 (Fisher–Rao is the zero-cost case)**.** *If* $C_s = 0$, *then* $\Omega_s(\pi) = T_0\mathrm{KL}(\pi\|\rho_s)$ *and* $L_Q = \operatorname{diag}(\pi_Q) - \pi_Q\pi_Q^\top$, *so for every* $v \in T$,

$$v^\top\nabla^2_T\Omega_s(\pi_Q)v = T_0\sum_i \frac{v_i^2}{\pi_{Q,i}}. \tag{109}$$

Thus ordinary entropy regularization gives Fisher–Rao geometry, while LOT deforms it into effective-resistance geometry determined by the optimal transport plan.

Theorem 5 identifies the statewise dual metric. The following pointwise conditioning result and behavioral locality result are useful consequences, but neither is a global policy-optimization rate.

Proposition 6 (Local spectral conditioning)**.** *At the matched policy π^{LOT}, the local regularizer Hessian on T has spectral moduli*

$$\sigma = \frac{\tau\lambda}{\lambda_{\max}(L_A)} \geq 2\tau\lambda, \qquad \beta = \frac{\tau\lambda}{\lambda_2(L_A)}, \tag{110}$$

where $\lambda_2(L_A)$ is the algebraic connectivity of the transport graph A and $\lambda_{\max}(L_A) \leq \frac{1}{2}$ (Thm. 1). Thus the local Hessian condition number is

$$\kappa_A = \frac{\beta}{\sigma} = \frac{\lambda_{\max}(L_A)}{\lambda_2(L_A)}. \tag{111}$$

This is a pointwise condition number because L_A changes with the matched policy. It does not by itself establish a global policy-mirror-descent rate; such a result additionally requires a specified iterative algorithm and uniform connectivity and curvature bounds along its trajectory.

Equation (41) identifies the local dual geometry as effective resistance on the transport graph A. The relevant pointwise conditioning is that of this *output* graph, not of the input graph L. Turning this local response geometry into a policy-optimization rate is left for an explicitly specified iterative algorithm.

Transport locality. The local-softmax form also gives a behavioral guarantee: probability does not move across a costly part of the graph unless the value advantage is large enough to pay the transport price.

Theorem 16 (Transport locality / barrier crossing)**.** *Fix s, Q. For any reference atom j and nonempty disjoint sets $B, R \subseteq \mathcal{A}$, let*

$$S_{ij}(Q) = Q_i + \tau\lambda \log \rho_s(i) - \tau C^L_{ij} \tag{112}$$

be the prior-adjusted transport score, and define

$$\Delta_j(B, R; Q) = \max_{k \in R} S_{kj}(Q) - \max_{i \in B} S_{ij}(Q), \tag{113}$$

and let $a_j(B, R; Q) = |B| \exp[-\Delta_j(B, R; Q)/(\tau\lambda)]$. Then the j-local policy in (10) *satisfies*

$$p^{(j)}(B) \leq \frac{a_j(B, R; Q)}{1 + a_j(B, R; Q)} \leq a_j(B, R; Q). \tag{114}$$

Consequently,

$$\pi^{\mathrm{LOT}}(B) \leq \sum_{\ell} \mu_s(\ell)\, \frac{a_\ell(B, R_\ell; Q)}{1 + a_\ell(B, R_\ell; Q)}. \tag{115}$$

for any choice of nonempty comparison sets with $R_\ell \cap B = \emptyset$. In particular, write $U_i(Q) = Q_i + \tau\lambda \log \rho_s(i)$ and $b_j(B, R) = \min_{i \in B} C^L_{ij} - \max_{k \in R} C^L_{kj}$. If $b_j(B, R) > 0$ and

$$\max_{i \in B} U_i(Q) - \max_{k \in R} U_k(Q) \leq \tau b_j(B, R) - \eta, \tag{116}$$

then $p^{(j)}(B) \leq |B| e^{-\eta/(\tau\lambda)}$.

This theorem is the precise locality claim suggested by the LOT mechanism. A far set B can receive substantial local mass only when its prior-adjusted value advantage offsets the transport barrier τb_j; otherwise the leakage is exponentially small in the margin η and in the inverse temperature $1/(\tau\lambda)$.

Diffusion-time convergence. For diffusion distances, the same knob also has a fixed-point interpretation: as the diffusion time grows, the spectral cost itself vanishes and LOT converges back to MaxEnt at an explicit spectral rate.

Theorem 17 (Diffusion-time convergence to MaxEnt)**.** *Let G be connected and write its Laplacian eigendecomposition as $L = \sum_{k=1}^{K} \lambda_k \phi_k \phi_k^\top$, with $0 = \lambda_1 < \lambda_2 \leq \cdots$. Define the diffusion-distance cost*

$$(C^L_t)_{ij} = \sum_{k=2}^{K} e^{-2t\lambda_k} \big(\phi_k(i) - \phi_k(j)\big)^2. \tag{117}$$

Let $F^{\mathrm{LOT}}_{s,t}$, $\mathcal{T}^{\mathrm{LOT}}_t$, and $Q^\star_t$ denote the LOT aggregator, Bellman operator, and fixed point built from C^L_t, a fixed full-support action prior ρ_s, and any full-support reference $\mu_{s,t}$. Let F^ρ_s and $Q^\star_\rho$ denote the geometry-free action-prior MaxEnt backup and fixed point from Prop. 4. Then, with $\delta_t = 2e^{-2t\lambda_2}$,

$$\left\|C^L_t\right\|_\infty \le \delta_t, \tag{118}$$

and, for every s, Q,

$$\left|F^{\mathrm{LOT}}_{s,t}(Q) - F^\rho_s(Q)\right| \le \tau\delta_t. \tag{119}$$

Consequently,

$$\left\|Q^\star_t - Q^\star_\rho\right\|_\infty \le \frac{\gamma\tau}{1-\gamma}\delta_t = \frac{2\gamma\tau}{1-\gamma}e^{-2t\lambda_2}. \tag{120}$$

Moreover, at any fixed Q, writing $\pi^{\mathrm{LOT}}_{s,t}(Q)$ for the LOT policy and $\pi^{\mathrm{soft}}_{\rho_s}(Q)$ for the action-prior softmax at temperature $\tau\lambda$,

$$D_{\mathrm{KL}}\left(\pi^{\mathrm{LOT}}_{s,t}(Q) \,\|\, \pi^{\mathrm{soft}}_{\rho_s}(Q)\right) \le \frac{\delta_t}{\lambda} = \frac{2}{\lambda}e^{-2t\lambda_2}, \tag{121}$$

and

$$\left\|\pi^{\mathrm{LOT}}_{s,t}(Q) - \pi^{\mathrm{soft}}_{\rho_s}(Q)\right\|_1 \le \exp\left(\frac{\delta_t}{\lambda}\right) - 1. \tag{122}$$

The theorem turns the empirical diffusion knob into an interpolation guarantee: small t keeps high-frequency geometry in the backup, while large t provably removes geometry and recovers the action-prior MaxEnt operator. The reference marginal may itself be heat-diffused with t; the zero-cost value and policy remain governed by the action prior ρ_s, so the bounds above compare $F^{\mathrm{LOT}}_{s,t}$ with $C = 0$ at the same reference $\mu_{s,t}$.

Learned-geometry perturbations. In the successor-geometry regime, the cost is estimated from a learned embedding, Laplacian, or successor representation. The next result makes this safe: uniform transport-cost error transfers linearly to the backup and to the Bellman fixed point, and it also gives a policy-stability bound.

Theorem 18 (Learned-geometry perturbation)**.** *Fix the same full-support ρ_s, μ_s and temperatures (τ, λ), and let $F^C_s, \mathcal{T}^C$ and $F^{\widehat{C}}_s, \mathcal{T}^{\widehat{C}}$ be LOT aggregators and Bellman operators built from bounded costs C_s and $\widehat{C}_s$. If*

$$\varepsilon_C = \sup_s \left\|\widehat{C}_s - C_s\right\|_\infty,$$

then for every s, Q,

$$\left|F^{\widehat{C}}_s(Q) - F^C_s(Q)\right| \le \tau\varepsilon_C. \tag{123}$$

Consequently, their fixed points satisfy

$$\left\|Q^\star_{\widehat{C}} - Q^\star_C\right\|_\infty \le \frac{\gamma\tau}{1-\gamma}\,\varepsilon_C. \tag{124}$$

Moreover, at any fixed s, Q, the corresponding LOT policies obey

$$\max\left\{D_{\mathrm{KL}}(\pi^C_s \| \pi^{\widehat{C}}_s), D_{\mathrm{KL}}(\pi^{\widehat{C}}_s \| \pi^C_s)\right\} \le \frac{2\varepsilon_C}{\lambda}, \tag{125}$$

and

$$\left\|\pi^{\widehat{C}}_s - \pi^C_s\right\|_1 \le \exp\left(\frac{2\varepsilon_C}{\lambda}\right) - 1. \tag{126}$$

Corollary 5 (Embedding and Laplacian error propagation)**.** *Suppose the learned cost is induced by action features, $C^\Phi_{ij} = \|\Phi_i - \Phi_j\|^2_M$ and $C^{\widehat{\Phi}}_{ij} = \left\|\widehat{\Phi}_i - \widehat{\Phi}_j\right\|^2_M$, where $\|x\|^2_M = x^\top Mx$ and $M \succeq 0$. If $\sup_i \left\|\widehat{\Phi}_i - \Phi_i\right\|_M \le \eta$ and $\sup_i \max\{\|\Phi_i\|_M, \left\|\widehat{\Phi}_i\right\|_M\} \le B$, then*

$$\left\|C^{\widehat{\Phi}} - C^\Phi\right\|_\infty \le 8B\eta + 4\eta^2,$$

and therefore

$$\left\|Q^\star_{\widehat{\Phi}} - Q^\star_{\Phi}\right\|_\infty \le \frac{\gamma\tau}{1-\gamma}(8B\eta + 4\eta^2).$$

Alternatively, if $C^L_{ij} = (e_i - e_j)^\top g(L)(e_i - e_j)$ *and* $C^{\widehat{L}}_{ij} = (e_i - e_j)^\top g(\widehat{L})(e_i - e_j)$ *with* $\left\|g(\widehat{L}) - g(L)\right\|_2 \le \varepsilon_g$*, then* $\left\|C^{\widehat{L}} - C^L\right\|_\infty \le 2\varepsilon_g$ *and* $\left\|Q^\star_{\widehat{L}} - Q^\star_L\right\|_\infty \le 2\gamma\tau\varepsilon_g/(1-\gamma)$.

Corollary 6 (Online diffusion geometry from the same stream)**.** *Let* $N_t(s,a) = \sum_{\ell<t} \mathbf{1}\{(s_\ell, a_\ell) = (s,a)\}$ *and, when* $N_t(s,a) > 0$*, define the empirical transition row*

$$\widehat{P}_t(\cdot \mid s,a) = \frac{1}{N_t(s,a)} \sum_{\ell<t} \mathbf{1}\{(s_\ell, a_\ell) = (s,a)\} e_{s_{\ell+1}}, \tag{127}$$

using any probability vector before the first visit. Let $P \mapsto W_s(P)$ *be a continuous map from the finite transition table to a symmetric, nonnegative action-similarity matrix with zero diagonal, and set* $L_s(P) = \mathrm{diag}(W_s(P)\mathbf{1}) - W_s(P)$*. For example,*

$$(W_s(P))_{ab} = \mathbf{1}\{a \ne b\}\, \kappa\left(\|P(\cdot \mid s,a) - P(\cdot \mid s,b)\|_2\right) \tag{128}$$

for any bounded, nonnegative, continuous κ*. Write* $\widehat{L}_{s,t} = L_s(\widehat{P}_t)$ *and* $L_s = L_s(P)$*. At fixed diffusion time* $h > 0$*, define*

$$\begin{aligned}(\widehat{C}^{\,h}_{s,t})_{ab} &= (e_a - e_b)^\top e^{-2h\widehat{L}_{s,t}}(e_a - e_b),\\ (C^h_s)_{ab} &= (e_a - e_b)^\top e^{-2hL_s}(e_a - e_b).\end{aligned} \tag{129}$$

If every state–action pair is visited infinitely often, then

$$\widehat{P}_t \to P, \qquad \sup_s \left\|\widehat{C}^{\,h}_{s,t} - C^h_s\right\|_\infty \to 0 \quad \text{a.s.}, \tag{130}$$

and, statewise,

$$\left\|\widehat{C}^{\,h}_{s,t} - C^h_s\right\|_\infty \le 4h \left\|\widehat{L}_{s,t} - L_s\right\|_2. \tag{131}$$

Consequently the predictable ordering "use $\widehat{C}^h_t$ *in the TD target, then incorporate* (s_t, a_t, s_{t+1}) *into* $\widehat{P}_{t+1}$*" satisfies Theorem 9; both critic and actor converge almost surely to those of the target diffusion geometry.*

Theorem 19 (Finite-sample geometry learning from one trajectory)**.** *Let* $S_0 = |\mathcal{S}|$*,* $K = |\mathcal{A}|$*, and* $M_0 = S_0 K$*. A fixed stationary behavior policy* β *induces the homogeneous state–action chain* $Z_t = (s_t, a_t)$ *with kernel*

$$K_\beta((s,a),(s',a')) = P(s' \mid s,a)\beta(a' \mid s').$$

Assume $Z_0 \sim \nu_\beta$*, the chain has pseudo-spectral gap* $\gamma_{\mathrm{ps}} > 0$*, and a known lower bound satisfies* $\min_{s,a} \nu_\beta(s,a) \ge \underline{\nu} > 0$*. Given* n *observed transitions, let*

$$N_n(x) = \sum_{t=0}^{n-1} \mathbf{1}\{Z_t = x\}, \qquad m = \left\lfloor \frac{n\underline{\nu}}{2} \right\rfloor,$$

and let $T_{x,k}$ *be the time of the* k*-th visit to* $x = (s,a)$*. If* $N_n(x) \ge m$*, form the equal-count row estimator*

$$\widetilde{P}_n(\cdot \mid x) = \frac{1}{m} \sum_{k=1}^{m} e_{s_{T_{x,k}+1}}, \tag{132}$$

using any probability vector for an unavailable row. For $\delta \in (0,1)$*, suppose*

$$n\underline{\nu} \ge 4, \qquad n\gamma_{\mathrm{ps}}\underline{\nu} \ge 104 \log \frac{4M_0}{\delta}. \tag{133}$$

Then, with probability at least $1 - \delta$*, every row is available and*

$$\left\|\widetilde{P}_n - P\right\|_{2,\infty} := \max_{s,a} \left\|\widetilde{P}_n(\cdot \mid s,a) - P(\cdot \mid s,a)\right\|_2 \le r_n(\delta), \tag{134}$$

$$\begin{aligned} r_n(\delta) &:= \sqrt{\frac{2\ell_\delta}{m}} + \frac{2\sqrt{2}\,\ell_\delta}{3m},\\ \ell_\delta &:= \log \frac{4(S_0+1)M_0}{\delta}. \end{aligned}$$

More generally, suppose the bounded nonnegative geometry map $P \mapsto \mathcal{C}(P) = \{C_s(P)\}_s$ obeys

$$\max_s \max_{a,b} |C_s(P')_{ab} - C_s(P)_{ab}| \le L_{\mathcal{C}} \|P' - P\|_{2,\infty}. \tag{135}$$

Under Assumption 1, write $Q^\star_{P,C}$ for the LOT fixed point using transition kernel P and cost family C. On the same event, the geometry-only plug-in error satisfies

$$\left\| Q^\star_{P,\mathcal{C}(\widetilde{P}_n)} - Q^\star_{P,\mathcal{C}(P)} \right\|_\infty \le \frac{\gamma\tau L_{\mathcal{C}}}{1-\gamma} r_n(\delta). \tag{136}$$

Corollary 7 (Single-trajectory diffusion geometry)**.** *Under Theorem 19, suppose*

$$\max_s \|L_s(P') - L_s(P)\|_2 \le L_{\mathcal{L}} \|P' - P\|_{2,\infty}. \tag{137}$$

For the squared diffusion-distance costs in (129)*, with probability at least $1-\delta$,*

$$\max_s \left\| C^h_s(\widetilde{P}_n) - C^h_s(P) \right\|_\infty \le 4hL_{\mathcal{L}} r_n(\delta), \tag{138}$$

$$\left\| Q^\star_{P,C^h(\widetilde{P}_n)} - Q^\star_{P,C^h(P)} \right\|_\infty \le \frac{4\gamma\tau h L_{\mathcal{L}}}{1-\gamma} r_n(\delta). \tag{139}$$

If the similarity construction in (128) *is restricted to a fixed symmetric edge mask of maximum degree D and κ is L_κ-Lipschitz, then $L_{\mathcal{L}} = 4DL_\kappa$ is valid. Consequently, the right sides of* (138) *and* (139) *become respectively*

$$16hDL_\kappa r_n(\delta) \quad \textit{and} \quad \frac{16\gamma\tau hDL_\kappa}{1-\gamma} r_n(\delta).$$

For the complete similarity graph, take $D = K - 1$.

Since $m \ge n\underline{\nu}/4$, the leading rate is $\widetilde{O}((n\underline{\nu})^{-1/2})$. Mixing pays for coverage, not for an additional leading $\sqrt{t_{\mathrm{mix}}}$ factor: for total-variation mixing time $t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$, the standard comparison $\gamma_{\mathrm{ps}} \ge 1/(2t_{\mathrm{mix}})$ (Paulin, 2015) makes

$$n \ge \frac{208 t_{\mathrm{mix}}}{\underline{\nu}} \log \frac{4S_0 K}{\delta}$$

sufficient for the second condition in (133). Direct Markov-dependent matrix moments would instead call for a Markov matrix Bernstein inequality (Neeman et al., 2024); they are not the row-normalized estimator analyzed here.

Remark 1 (Scope of online geometry). Corollary 6 concerns a bounded squared diffusion-distance ground cost, not the centered direct heat-kernel branch of Theorem 6. Heat filtering requires neither an eigengap nor connectedness. Effective-resistance costs using $L^\dagger$ require stable rank (typically a uniform spectral gap), while hard spectral truncation requires an eigengap. The limiting geometry here is a fixed environment quantity; an actor- or critic-dependent population target would require a genuinely coupled stochastic-approximation analysis. Theorem 19 is a batch post-processing guarantee for one stationary trajectory; a burn-in gives the usual extension to a nonstationary start. Both fixed points in (136) retain the true dynamics P. Using $\widetilde{P}_n$ as the planning model adds the standard transition-model error, and obtaining a finite-time rate for the simultaneous recursion in Theorem 9 remains separate.

Finally, the spectral cost may be truncated to the leading proto-value modes—the same low-dimensional representation used for value—at controlled cost.

Theorem 20 (Spectral truncation)**.** *Let $\widehat{C^L}$ use only the first d retained spectral modes in* (6)*, and set $\varepsilon_d = \sum_{k>d} g(\lambda_k) \max_{ij} (\phi_k(i) - \phi_k(j))^2$, so $\left\| \widehat{C^L} - C^L \right\|_\infty \le \varepsilon_d$. Then, with $\widehat{\mathcal{T}^{\mathrm{LOT}}}$ the operator built from $\widehat{C^L}$,*

$$\left| F^{\mathrm{LOT}}_s(Q) - \widehat{F}^{\mathrm{LOT}}_s(Q) \right| \le \tau\, \varepsilon_d$$
$$\Longrightarrow \left\| \widehat{Q}^\star_{\mathrm{LOT}} - Q^\star_{\mathrm{LOT}} \right\|_\infty \le \frac{\gamma\tau\varepsilon_d}{1-\gamma}.$$

This is the learned-geometry theorem specialized to deterministic spectral approximation; the same 1-Lipschitz-in-cost mechanism for entropic OT Eckstein & Nutz (2022) propagates through the contraction. For the effective-resistance filter, ε_d is the omitted cost tail. Its smallness is logically separate from approximation of the value function by the leading proto-value modes; Theorem 21 below keeps the two errors distinct.

Joint value–geometry spectral approximation. The previous theorem controls approximation of the *transport cost* alone. In a spectral implementation, however, the same leading modes are often used twice: to approximate the cost and to represent the value function. The next theorem separates these two errors.

Theorem 21 (Joint spectral approximation)**.** *Let $\mathcal{T}^C$ be the exact LOT Bellman operator built from a bounded target cost C, and let $\mathcal{T}^{C_d}$ be the operator built from a rank-d spectral approximation C_d. Assume*

$$\varepsilon_C(d) = \sup_s \|C_{d,s} - C_s\|_\infty < \infty. \tag{140}$$

Let $\mathcal{X}$ be the Banach space of bounded action-value functions under the sup norm, let $\mathcal{V}_d \subset \mathcal{X}$ be nonempty and closed, and let $\Pi_d : \mathcal{X} \to \mathcal{V}_d$ be a single-valued statewise projection using the same leading modes. Assume $\Pi_d Q = Q$ for $Q \in \mathcal{V}_d$ and that it is κ_d-Lipschitz in sup norm,

$$\|\Pi_d X - \Pi_d Y\|_\infty \le \kappa_d \|X - Y\|_\infty, \qquad \gamma\kappa_d < 1. \tag{141}$$

Define the projected approximate operator

$$\widehat{\mathcal{T}}_d Q = \Pi_d \mathcal{T}^{C_d} Q. \tag{142}$$

Then $\widehat{\mathcal{T}}_d$ is a $\gamma\kappa_d$-contraction and has a unique fixed point $\widehat{Q}_d^\star \in \mathcal{V}_d$. If $Q_C^\star$ is the fixed point of $\mathcal{T}^C$ and

$$\varepsilon_V^\star(d) = \left\|\Pi_d \mathcal{T}^C Q_C^\star - \mathcal{T}^C Q_C^\star\right\|_\infty, \tag{143}$$

then

$$\left\|\widehat{Q}_d^\star - Q_C^\star\right\|_\infty \le \frac{\varepsilon_V^\star(d) + \kappa_d\gamma\tau\varepsilon_C(d)}{1 - \gamma\kappa_d}. \tag{144}$$

In particular, if a uniform projection residual bound holds on a bounded invariant set $\mathcal{B}$ containing $Q_C^\star$,

$$\varepsilon_V(d) = \sup_{Q\in\mathcal{B}} \left\|\Pi_d \mathcal{T}^C Q - \mathcal{T}^C Q\right\|_\infty, \tag{145}$$

then the same conclusion holds with $\varepsilon_V(d)$ in place of $\varepsilon_V^\star(d)$. For the spectral cost in (6)*, one may take $\varepsilon_C(d) = \varepsilon_d$ from Thm. 20; if Π_d is nonexpansive ($\kappa_d = 1$), the bound becomes*

$$\left\|\widehat{Q}_d^\star - Q_C^\star\right\|_\infty \le \frac{\varepsilon_V(d) + \gamma\tau\varepsilon_d}{1 - \gamma}. \tag{146}$$

The specialization $\kappa_d = 1$ is conditional. An orthogonal spectral projection in L^2 is not generally nonexpansive in the sup norm.

Thus the approximation error has two interpretable pieces: ε_V is the price of representing the Bellman image in the first d modes, while $\gamma\tau\varepsilon_C$ is the price of using those same modes to approximate the transport geometry. This is the formal version of the representation-as-geometry claim: when the leading spectral basis is good for value and its cost tail is small, the full LOT fixed point is recovered up to the sum of these errors.

Moving-reference stability. The heat-diffused reference $\mu_s = (1-\xi)H_{t_\mu}\pi_s^{\mathrm{prior}} + \xi\mathrm{Unif}$ may be refreshed during learning, for example from a previous actor. Since the action prior ρ_s is separated from the transport reference, we can quantify the effect of changing μ_s while holding C_s, ρ_s fixed.

Theorem 22 (Moving-reference stability and tracking)**.** *Fix bounded costs C_s, full-support action priors ρ_s, and temperatures (τ, λ). Define the reference-cost oscillation*

$$\Delta_C = \sup_s \max_{j,k} \max_i |(C_s)_{ij} - (C_s)_{ik}|. \tag{147}$$

Let μ and ν be two full-support reference families, and write $F_s^\mu, \mathcal{T}^\mu, Q_\mu^\star$ and $F_s^\nu, \mathcal{T}^\nu, Q_\nu^\star$ for the corresponding aggregators, Bellman operators, and fixed points. With

$$\delta_\mu = \sup_s \|\mu_s - \nu_s\|_1, \tag{148}$$

we have, for every s, Q,

$$|F_s^\mu(Q) - F_s^\nu(Q)| \le \frac{\tau\Delta_C}{2} \|\mu_s - \nu_s\|_1, \tag{149}$$

and therefore

$$\left\|Q_\mu^\star - Q_\nu^\star\right\|_\infty \le \frac{\gamma\tau\Delta_C}{2(1-\gamma)}\,\delta_\mu. \tag{150}$$

At any fixed Q, the induced policies satisfy

$$\left\|\pi_s^\mu(Q) - \pi_s^\nu(Q)\right\|_1 \le \left\|\mu_s - \nu_s\right\|_1. \tag{151}$$

Finally, consider exact value iteration with a changing reference sequence: $Q_{k+1} = \mathcal{T}^{\mu^k} Q_k$, and let $Q_k^\star$ be the fixed point of $\mathcal{T}^{\mu^k}$. If

$$\delta_k = \sup_s \left\|\mu_s^{k+1} - \mu_s^k\right\|_1, \qquad B_\mu = \frac{\gamma\tau\Delta_C}{2(1-\gamma)}, \tag{152}$$

then the tracking error $e_k = \|Q_k - Q_k^\star\|_\infty$ obeys

$$e_{k+1} \le \gamma e_k + B_\mu \delta_k, \tag{153}$$

and hence

$$e_n \le \gamma^n e_0 + B_\mu \sum_{k=0}^{n-1} \gamma^{n-1-k}\delta_k. \tag{154}$$

Consequently, if $\delta_k \to 0$, exact value iteration tracks the moving LOT fixed point ($e_k \to 0$); if $\delta_k \le \bar{\delta}$ uniformly, then

$$\limsup_{k\to\infty} e_k \le \frac{B_\mu \bar{\delta}}{1-\gamma}. \tag{155}$$

For heat-diffused references $\mu_s^k = (1-\xi)H_{t_\mu}\pi_s^k + \xi\,\mathrm{Unif}$ with a Markov heat kernel H_{t_μ},

$$\delta_k \le (1-\xi)\sup_s \left\|\pi_s^{k+1} - \pi_s^k\right\|_1. \tag{156}$$

Corollary 8 (Joint predictable cost and reference learning)**.** *The conclusion of Theorem 9 remains valid when its fixed reference is replaced in the sampled target by an $\mathcal{F}_t$-measurable $\widehat{\mu}_t$ with uniform full support and $\varepsilon_{\mu,t} := \sup_s \|\widehat{\mu}_{s,t} - \mu_s\|_1 \to 0$ almost surely. Indeed,*

$$\sup_Q \left\|\mathcal{T}^{\widehat{C}_t,\widehat{\mu}_t} Q - \mathcal{T}^{C,\mu} Q\right\|_\infty \le \gamma\tau\left(\varepsilon_t + \frac{\Delta_C}{2}\varepsilon_{\mu,t}\right) \to 0. \tag{157}$$

The corresponding plug-in actor converges to the actor induced by (C,μ).

The theorem justifies a practical critic/actor schedule: freeze μ_s for several critic backups, or update it slowly from the previous actor. The Bellman fixed point then drifts only linearly in the L^1 reference change, and the critic tracks this drift when the update size tends to zero.

Continuous-action LOT. The finite-action derivation is not tied to a simplex. The only structural requirement is that the *reference marginal* is fixed while the source/action marginal is optimized. On a compact action space this yields the same one-shot object, now as a mixture of cost-shifted Boltzmann measures.

Theorem 23 (Continuous-action LOT backup)**.** *Let $(\mathcal{A}, d)$ be a compact metric action space. For each state s, let ρ_s and μ_s be probability measures on $\mathcal{A}$, let $c_s : \mathcal{A}\times\mathcal{A} \to \mathbb{R}_{\ge 0}$ be bounded and continuous, and let $Q \in C(\mathcal{A})$. Define the semi-relaxed continuous LOT aggregator*

$$\begin{aligned}\mathcal{F}_s^c(Q) = \sup_{\Gamma:\Gamma_2=\mu_s} \Big\{ &\int [Q(a) - \tau c_s(a,b)]\, d\Gamma(a,b) \\ &- \tau\lambda\,\mathrm{KL}(\Gamma \,\|\, \rho_s\otimes\mu_s)\Big\}.\end{aligned} \tag{158}$$

where the supremum is over probability couplings on $\mathcal{A}\times\mathcal{A}$ with second marginal μ_s. For $b \in \mathcal{A}$, set

$$Z_{s,b}(Q) = \int_{\mathcal{A}} \exp\left(\frac{Q(a) - \tau c_s(a,b)}{\tau\lambda}\right) d\rho_s(a), \tag{159}$$

and define the local Boltzmann measure

$$p^Q_{s,b}(da) = \frac{\exp((Q(a) - \tau c_s(a,b))/(\tau\lambda))}{Z_{s,b}(Q)}\, d\rho_s(a). \tag{160}$$

Then the unique maximizing plan among plans absolutely continuous with respect to $\rho_s \otimes \mu_s$ is

$$\Gamma^{Q,\star}_s(da, db) = p^Q_{s,b}(da)\, d\mu_s(b), \tag{161}$$

the induced LOT policy is the mixture

$$\pi^{Q,c}_s(da) = \int_{\mathcal{A}} p^Q_{s,b}(da)\, d\mu_s(b), \tag{162}$$

and the value has the closed form

$$\mathcal{F}^c_s(Q) = \tau\lambda \int_{\mathcal{A}} \log Z_{s,b}(Q)\, d\mu_s(b). \tag{163}$$

Moreover, $\mathcal{F}^c_s$ is convex, monotone, translation equivariant, and 1-Lipschitz on $C(\mathcal{A})$ under $\|\cdot\|_\infty$. Its Frechet derivative is the policy measure:

$$D\mathcal{F}^c_s(Q)[h] = \int_{\mathcal{A}} h(a)\, d\pi^{Q,c}_s(a), \qquad h \in C(\mathcal{A}), \tag{164}$$

and its second derivative is the covariance mixture

$$D^2\mathcal{F}^c_s(Q)[h,g] = \frac{1}{\tau\lambda} \int_{\mathcal{A}} \mathrm{Cov}_{p^Q_{s,b}}(h,g)\, d\mu_s(b). \tag{165}$$

Consequently $D^2\mathcal{F}^c_s(Q)[h,h] \geq 0$ and $D^2\mathcal{F}^c_s(Q)[h,h] \leq \mathrm{osc}(h)^2/(4\tau\lambda)$. If $c_s \equiv 0$, then $p^Q_{s,b}$ is independent of b and

$$\pi^{Q,c=0}_s(da) = \frac{e^{Q(a)/(\tau\lambda)}}{\int e^{Q(a')/(\tau\lambda)} d\rho_s(a')}\, d\rho_s(a), \tag{166}$$

the action-prior MaxEnt policy. If ρ_s and μ_s are finite atomic measures, (160)–(163) *reduce exactly to the finite LOT formulas in Lemma 1 and Theorem 1. The value formula* (163) *also extends to bounded measurable Q. Finally, consider a discounted MDP with a general measurable state space, compact continuous action space $\mathcal{A}$, bounded measurable rewards, and a stochastic kernel $P(ds' \mid s, a)$. Assume that $s \mapsto \rho_s$ and $s \mapsto \mu_s$ are probability kernels, that $(s,a,b) \mapsto c_s(a,b)$ is bounded and jointly measurable, and that $c_s(\cdot,\cdot)$ is continuous for each s. Using the bounded-measurable extension of* (163)*, these hypotheses make $s \mapsto \mathcal{F}^c_s(Q(s,\cdot))$ measurable for every bounded jointly measurable Q. On the Banach space of bounded measurable action-value functions, the operator*

$$(\mathcal{T}^c Q)(s,a) = r(s,a) + \gamma \int \mathcal{F}^c_{s'}(Q(s',\cdot))\, P(ds' \mid s, a) \tag{167}$$

is a γ-contraction in the sup norm and therefore has a unique fixed point.

The continuous law has the same barrier-crossing behavior as the finite locality theorem. For measurable sets $B, R \subset \mathcal{A}$ with $\rho_s(R) > 0$ and any reference point b,

$$\begin{aligned} p^Q_{s,b}(B) &\leq \frac{\rho_s(B)}{\rho_s(R)} \exp\left(-\frac{\Delta_b(B,R;Q)}{\tau\lambda}\right), \\ \Delta_b(B,R;Q) &= \inf_{a\in R}[Q(a) - \tau c_s(a,b)] \\ &\quad - \sup_{a\in B}[Q(a) - \tau c_s(a,b)]. \end{aligned} \tag{168}$$

Thus, without any grid, probability assigned to a far measurable region is exponentially small unless its value advantage compensates for the transport price.

Corollary 9 (Finite-grid consistency)**.** *Fix a state and the compact-action assumptions of Theorem 23. Let $r_m : \mathcal{A} \to \mathcal{A}_m$ be a measurable quantizer onto a finite grid with mesh η_m, meaning $\sup_{a\in\mathcal{A}} d(a, r_m(a)) \leq \eta_m$. Let $\rho_{s,m} = (r_m)_\# \rho_s$, $\mu_{s,m} = (r_m)_\# \mu_s$, $Q_m(x) = Q(x)$,*

and $C_m(x,y) = c_s(x,y)$ for $x, y \in \mathcal{A}_m$. If $\mathcal{F}^{\mathrm{LOT}}_{s,m}$ is the finite LOT aggregator built from $(\rho_{s,m}, \mu_{s,m}, Q_m, C_m)$, and

$$\omega_Q(\eta) = \sup_{d(a,a')\le\eta} |Q(a) - Q(a')|,$$
$$\omega_c(\eta) = \sup_{\substack{d(a,a')\le\eta \\ d(b,b')\le\eta}} |c_s(a,b) - c_s(a',b')|,$$

where the finite log-partition definition is used directly; it remains well-defined when either pushforward has zero-mass atoms. Then

$$\left|\mathcal{F}^{\mathrm{LOT}}_{s,m}(Q_m) - \mathcal{F}^{c}_{s}(Q)\right| \le \omega_Q(\eta_m) + \tau\omega_c(\eta_m). \tag{169}$$

Consequently $\mathcal{F}^{\mathrm{LOT}}_{s,m}(Q_m) \to \mathcal{F}^{c}_{s}(Q)$ as the mesh goes to zero, and the finite-grid LOT policies, viewed as measures on $\mathcal{A}$, converge weakly to $\pi^{Q,c}_s$.

Proposition 7 (Quadratic-Gaussian local policies: noncompact extension)**.** *Separately from the compact-space theorem, let $\mathcal{A} = \mathbb{R}^m$, suppose $\rho_s = \mathcal{N}(m_0, \Sigma_0)$ with $\Sigma_0 \succ 0$, let μ_s have a finite second moment, and suppose $c_s(a,b) = \frac{1}{2}(a-b)^\top M(a-b)$ with $M \succeq 0$, and take the* exact *quadratic $Q(a) = q_0 + q^\top a - \frac{1}{2}a^\top Ha$ with $H \succeq 0$. If $\Lambda = \Sigma_0^{-1} + H/(\tau\lambda) + M/\lambda \succ 0$, then each local continuous-LOT component is Gaussian,*

$$\begin{aligned} p^{Q}_{s,b} &= \mathcal{N}(\Lambda^{-1}\eta_b, \Lambda^{-1}), \\ \eta_b &= \Sigma_0^{-1}m_0 + q/(\tau\lambda) + Mb/\lambda. \end{aligned} \tag{170}$$

Thus continuous LOT becomes a mixture of Gaussian components indexed by $b \sim \mu_s$. If a non-quadratic Q is only approximated locally by the displayed quadratic, the same formula is a local Laplace approximation, not an exact LOT policy. This provides a possible actor-critic implementation path whose components are shifted by the action cost.

E Extended Discussion and Limitations

The regularizer is not a graph-smoothness penalty. LOT uses relationships between actions inside a soft Bellman backup, but none of our results bounds $\pi^\top L\pi$ or proves that the resulting policy is smooth on the input action graph. The proved curvature statement is different: the backup Hessian is the Laplacian of a *value-dependent transport co-occurrence graph* (Theorem 12), and inverse policy curvature is effective resistance on that induced graph (Theorem 5). The behavioral result in Theorem 16 is also more specific: it bounds the probability placed across a transport-cost barrier until the prior-adjusted value advantage compensates for that cost.

The input graph and the transport co-occurrence graph play different roles. In particular, the spectral gap $\lambda_2(L)$ of the input graph controls the separate large-diffusion-time limit of the diffusion-distance cost; it does not control conditioning of the value-dependent curvature matrix L_Q.

LOT also does not have smaller hard-max regularization bias than a matched MaxEnt backup. For ordinary uncentered LOT with nonnegative cost, Proposition 4 gives $Q^\star_C \le Q^\star_\rho \le Q^\star$. This is an ordering of regularized fixed-point values, not an ordering of policy performance. It also does not apply directly to the additive centering used for the direct-heat planning target.

What the planning theorem guarantees. The operator results cover finite action sets and compact continuous action spaces with bounded costs. The unconditional finite-action planner assumes a fixed LOT geometry and independent access to a generative model. Its output guarantee is for an ε-accurate *regularized root value*. It does not provide a theorem for selecting the root action, simple regret, whole-policy quality, or online interaction regret.

The local-action complexity depends on deterministic, Q-independent sets that retain a prescribed amount of kernel mass. In the geometric result these sets come from the *centered direct-heat* construction, not the bounded squared diffusion-distance Gibbs cost. The Gaussian heat-tail bound is assumed only over a stated radius range, and polynomial ball growth is an assumption on the graph family. Bounded degree alone gives neither property, and every local neighborhood eventually saturates at K on a finite graph.

The transition-sample bound does not itself require an algorithm for locally accessing exact heat weights; the corresponding sub-K arithmetic claim does. The MaxEnt comparison applies only to deterministic, value-independent kernel-mass pruning with a uniform prior. It is not a lower bound for every MaxEnt planning or randomized-integration method. Even when $d_{\mathrm{ker}}(\varepsilon)$ is polylogarithmic, the present cascade has complexity $\varepsilon^{-3}\exp(O((\log\log(1/\varepsilon))^2))$, not literally $\widetilde{O}(\varepsilon^{-3})$.

What the learning theorems guarantee. The same-stream result allows a bounded, uniformly consistent cost estimate inside tabular Q-learning, provided the estimate used in the current TD target depends only on earlier data. It proves almost-sure convergence, not a finite-time rate for the simultaneous cost-and-Q recursion. The trajectory result instead bounds geometry-induced fixed-point error under one fixed stationary behavior chain. Mixing supplies coverage, and a burn-in handles a nonstationary initialization.

The independent-batch theorem constructs a global action cost from transition samples, but its target comparison still uses exact downstream dynamics. The separate matching $n^{-1/2}$ upper and lower bounds concern target estimation in an explicit one-parameter family. They are neither lower bounds for the global cost estimator nor general-MDP minimax results.

The heat-kernel and inverse-curvature assumptions do not all hold at once. No single construction in the paper simultaneously supplies learned geometry, exact finite support, and the full interior inverse-curvature formula. Learned diffusion-distance costs are generally dense. Exact direct heat may have small mass-retaining neighborhoods, but its columns are dense. The normalized truncated kernel $H_{t,r}$ is locally computable and exactly zero beyond r hops, but it approximates exact heat and incurs the Poisson-tail error. Moreover, those zeros can add null directions, whereas the full tangent-space effective-resistance identity assumes positive kernels induced by finite costs. Results can be combined through the normalized-column Bellman interface only after checking the assumptions of each result.

The $H_{t,r}$ construction computes normalized mixtures of random walks of length at most r; it does not compute the entries of exact H_t retained on the same ball, and it does not apply to diffusion-distance Gibbs kernels.

Open directions. Important next steps are endogenous representation targets, function approximation, coupled geometry and value learning, finite-time guarantees for the same-stream recursion, and an iterative policy-optimization method that uses the effective-resistance metric. None of these extensions is claimed here.

F Proof of the Closed-Form LOT Backup

Because (8) also optimizes the row marginal $\pi = \Gamma\mathbf{1}$, only $\Gamma^\top\mathbf{1} = \mu_s$ remains fixed, and with $\Gamma_{ij} = \mu_s(j)p_i^{(j)}$ the problem separates:

$$F_s^{\mathrm{LOT}}(Q) = \sum_j \mu_s(j) \max_{p^{(j)} \in \Delta_{\mathcal{A}}} \Big\{ \langle p^{(j)},\, Q - \tau C_{:j} \rangle - T_0 \mathrm{KL}(p^{(j)} \| \rho_s) \Big\}. \tag{171}$$

Result proved in this section. Lemma 1 states that the transport optimization inside a LOT Bellman backup separates over anchor actions. We prove the optimizer first and then substitute it back into the objective to obtain the value formula.

F.1 Proof of Lemma 1

Fix a state s and suppress its index: write $\rho = \rho_s$, $\mu = \mu_s$, $C = C_s$, and $T_0 = \tau\lambda$. By the assumptions of Lemma 1, $T_0 > 0$, the vectors $\rho, \mu \in \Delta_{\mathcal{A}}$ have full support, $Q \in \mathbb{R}^K$, and every entry of C is finite, so all exponential weights below are finite and strictly positive. We use the convention $0 \log 0 = 0$.

We first eliminate the optimized row marginal exactly.

Recall from (7) that

$$W_{\lambda,s}^{\rho}(\pi, \mu) = \min_{\Gamma \in \Pi(\pi,\mu)} \Big\{ \langle \Gamma,\, C \rangle + \lambda \sum_{i,j} \Gamma_{ij} \log \tfrac{\Gamma_{ij}}{\rho(i)\mu(j)} \Big\},$$
$$\Pi(\pi, \mu) = \big\{ \Gamma \geq 0 : \Gamma\mathbf{1} = \pi,\ \Gamma^\top\mathbf{1} = \mu \big\}.$$

For every $\pi \in \Delta_{\mathcal{A}}$, the set $\Pi(\pi, \mu)$ is nonempty (it contains $\pi\mu^\top$), closed, and bounded, and the transport objective is continuous on it under $0 \log 0 = 0$, so the minimum is attained. For each fixed π, multiplication by $-\tau < 0$ converts the inner minimum into a maximum over the same feasible set, $-\tau \min_\Gamma G(\Gamma) = \max_\Gamma [-\tau G(\Gamma)]$; this is not a minimax interchange. Substituting the definition of $W_{\lambda,s}^{\rho}$ into (8) gives

$$F_s^{\mathrm{LOT}}(Q) = \max_{\substack{\pi \in \Delta_{\mathcal{A}} \\ \Gamma \in \Pi(\pi,\mu)}} \Big\{ \langle \pi,\, Q \rangle - \tau \langle \Gamma,\, C \rangle - T_0 \sum_{i,j} \Gamma_{ij} \log \tfrac{\Gamma_{ij}}{\rho(i)\mu(j)} \Big\}. \tag{172}$$

For every feasible pair, the row-marginal constraint gives $\pi = \Gamma\mathbf{1}$, hence $\langle \pi,\, Q \rangle = \sum_{i,j} \Gamma_{ij} Q_i = \langle \Gamma,\, Q\mathbf{1}^\top \rangle$, where $(Q\mathbf{1}^\top)_{ij} = Q_i$ and the last inner product is the Frobenius one. Conversely, if $\Gamma \geq 0$ and $\Gamma^\top\mathbf{1} = \mu$, then $\pi := \Gamma\mathbf{1}$ satisfies $\pi \geq 0$ and $\mathbf{1}^\top\pi = (\Gamma^\top\mathbf{1})^\top\mathbf{1} = \mu^\top\mathbf{1} = 1$, so $\pi \in \Delta_{\mathcal{A}}$ and $\Gamma \in \Pi(\pi, \mu)$. This proves the exact identity $\bigcup_{\pi \in \Delta_{\mathcal{A}}} \Pi(\pi, \mu) = \{\Gamma \geq 0 : \Gamma^\top\mathbf{1} = \mu\}$. Eliminating the redundant variable π from (172) yields the semi-relaxed problem

$$F_s^{\mathrm{LOT}}(Q) = \max_{\Gamma \in \mathcal{D}_\mu} \mathcal{J}_Q(\Gamma), \qquad \mathcal{D}_\mu := \big\{ \Gamma \geq 0 : \Gamma^\top\mathbf{1} = \mu \big\}, \tag{173}$$

where

$$\mathcal{J}_Q(\Gamma) := \sum_{i,j} \Gamma_{ij}(Q_i - \tau C_{ij}) - T_0 \sum_{i,j} \Gamma_{ij} \log \tfrac{\Gamma_{ij}}{\rho(i)\mu(j)}. \tag{174}$$

Only the column marginal μ is prescribed; the row marginal is the policy produced by the optimizer.

Next, we establish existence, uniqueness, and strict positivity.

The set $\mathcal{D}_\mu$ is nonempty ($\rho\mu^\top \in \mathcal{D}_\mu$), convex, and compact, since each column lies in the scaled simplex $\{x \in \mathbb{R}_+^K : \mathbf{1}^\top x = \mu(j)\}$; in particular $0 \leq \Gamma_{ij} \leq \mu(j)$. The map $x \mapsto x \log(x/b)$, extended continuously at zero by $0 \log 0 = 0$, is continuous for every $b > 0$, so $\mathcal{J}_Q$ is continuous on the compact set $\mathcal{D}_\mu$ and a maximizer exists. Moreover, $x \mapsto x \log(x/b)$ is strictly convex on $[0, \infty)$; since $\rho(i)\mu(j) > 0$ and $T_0 > 0$, the negative-entropy term in (174) is strictly concave in Γ, and adding the linear score and cost terms preserves strict concavity. Thus the maximizer $\Gamma^\star$ is unique.

To see that $\Gamma^\star$ is entrywise positive, suppose $\Gamma^\star_{ij} = 0$ for some (i, j). Since $\sum_\ell \Gamma^\star_{\ell j} = \mu(j) > 0$, some k satisfies $\Gamma^\star_{kj} > 0$. For $0 < \varepsilon < \Gamma^\star_{kj}$, the within-column perturbation $\Gamma^\varepsilon = \Gamma^\star + \varepsilon e_i e_j^\top - \varepsilon e_k e_j^\top$ remains in $\mathcal{D}_\mu$, and with $A_{\ell j} = Q_\ell - \tau C_{\ell j}$ and $b_{\ell j} = \rho(\ell)\mu(j) > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{J}_Q(\Gamma^\varepsilon) = A_{ij} - A_{kj} - T_0\Big[\log\tfrac{\varepsilon}{b_{ij}} - \log\tfrac{\Gamma^\star_{kj}-\varepsilon}{b_{kj}}\Big] \longrightarrow +\infty \qquad (\varepsilon \downarrow 0),$$

because the first logarithm tends to $-\infty$ while all other terms stay finite. Hence a small perturbation strictly increases the objective, contradicting optimality, and

$$\Gamma^\star_{ij} > 0 \qquad \text{for all } i, j. \tag{175}$$

The nonnegativity constraints are therefore inactive at the optimizer, and the equality-constrained stationarity conditions apply.

We then solve the equality-constrained KKT system.

Associate a multiplier $g_j \in \mathbb{R}$ with the column constraint $\sum_i \Gamma_{ij} = \mu(j)$, with the sign convention

$$\mathcal{L}(\Gamma, g) = \mathcal{J}_Q(\Gamma) + \sum_j g_j\Big(\mu(j) - \sum_i \Gamma_{ij}\Big). \tag{176}$$

Since $\frac{\mathrm{d}}{\mathrm{d}x}[x\log(x/b)] = \log(x/b) + 1$ for $x > 0$, stationarity of (176) at the positive optimizer gives

$$Q_i - \tau C_{ij} - T_0\Big(\log\tfrac{\Gamma^\star_{ij}}{\rho(i)\mu(j)} + 1\Big) - g_j = 0, \tag{177}$$

and solving for $\Gamma^\star_{ij}$, using $\tau C_{ij}/T_0 = \tau C_{ij}/(\tau\lambda) = C_{ij}/\lambda$ for the cost factor, yields

$$\Gamma^\star_{ij} = \rho(i)\mu(j)\exp\Big(\tfrac{Q_i - \tau C_{ij} - g_j}{T_0} - 1\Big) = \mu(j)\,\rho(i)e^{Q_i/T_0}\,e^{-C_{ij}/\lambda}\,e^{-g_j/T_0 - 1}. \tag{178}$$

Define $u_i := \rho(i)e^{Q_i/T_0}$, $\mathsf{K}_{C,ij} := e^{-C_{ij}/\lambda}$, and

$$Z_j(Q) := \sum_i u_i \mathsf{K}_{C,ij} = (\mathsf{K}_C^\top u)_j = \sum_i \rho(i)\exp\Big(\tfrac{Q_i - \tau C_{ij}}{T_0}\Big), \tag{179}$$

which is finite and strictly positive for every j. Summing (178) over i and imposing the column constraint gives $\mu(j) = \mu(j)e^{-g_j/T_0 - 1}Z_j(Q)$; since $\mu(j) > 0$, this forces $e^{-g_j/T_0-1} = 1/Z_j(Q)$, i.e. $g_j = T_0(\log Z_j(Q) - 1)$. Substituting this normalizer back into (178) gives

$$\Gamma^\star_{ij} = \mu(j)\,\frac{\rho(i)\exp\big((Q_i - \tau C_{ij})/T_0\big)}{Z_j(Q)}. \tag{180}$$

Finally, with $v_j := \mu(j)/Z_j(Q) = \mu(j)/(\mathsf{K}_C^\top u)_j$, (180) becomes

$$\Gamma^\star_{ij} = u_i \mathsf{K}_{C,ij} v_j, \tag{181}$$

the scaling formula in Lemma 1. Because the objective is concave and the constraints affine, the KKT conditions are sufficient for global optimality; the strict concavity established above gives uniqueness.

We now recover the anchor policies and the LOT policy.

Since $\mu(j) > 0$, define the conditional action distribution in column j by $p_i^{(j)}(Q) := \Gamma^\star_{ij}/\mu(j)$. Equation (180) gives

$$p_i^{(j)}(Q) = \frac{\rho(i)\exp\big((Q_i - \tau C_{ij})/T_0\big)}{Z_j(Q)}, \tag{182}$$

with $p_i^{(j)}(Q) > 0$ and $\sum_i p_i^{(j)}(Q) = 1$ by the definition of $Z_j(Q)$, so $\Gamma^\star_{ij} = \mu(j)p_i^{(j)}(Q)$. The policy chosen in the original LOT optimization is the free row marginal of the optimal plan:

$$\pi_i^{\mathrm{LOT}}(Q) = (\Gamma^\star \mathbf{1})_i = \sum_j \mu(j)p_i^{(j)}(Q) = \sum_j \mu(j)\,\frac{\rho(i)\exp\big((Q_i - \tau C_{ij})/T_0\big)}{Z_j(Q)}, \tag{183}$$

or in matrix notation $\Gamma^\star = \mathrm{diag}(u)\mathsf{K}_C\,\mathrm{diag}(v)$ and $\pi^{\mathrm{LOT}} = \Gamma^\star\mathbf{1} = u \odot (\mathsf{K}_C v)$. Feasibility follows from $\sum_i \Gamma^\star_{ij} = v_j(\mathsf{K}_C^\top u)_j = \mu(j)$, and $\pi^{\mathrm{LOT}} \in \Delta_{\mathcal{A}}$ because $\sum_i \pi_i^{\mathrm{LOT}} = \sum_{i,j}\Gamma^\star_{ij} = \sum_j \mu(j) = 1$, with all entries strictly positive.

Uniqueness of $\Gamma^\star$ implies uniqueness of π^{LOT}: any maximizer of the original policy problem, together with one of its optimal transport plans, would be a maximizer of (172), whose plan is unique. Moreover, $\Gamma^\star$ is the minimizing plan in $W^\rho_{\lambda,s}(\pi^{\mathrm{LOT}}, \mu)$; otherwise, replacing it by a lower-cost plan with the same two marginals would strictly increase the joint objective in (172), contradicting optimality.

We next evaluate the optimal value by an exact KL identity.

Every feasible column has total mass $\mu(j) > 0$, so any $\Gamma \in \mathcal{D}_\mu$ has the unique representation $\Gamma_{ij} = \mu(j)p_i^{(j)}$ with $p^{(j)} \in \Delta_{\mathcal{A}}$, and conversely any collection $\{p^{(j)}\}_j \subset \Delta_{\mathcal{A}}$ defines a feasible plan. Substituting this into (174), the factor $\mu(j)$ cancels inside the logarithm—no additional $\mu(j)\log\mu(j)$ term appears—and the problem separates exactly:

$$F_s^{\mathrm{LOT}}(Q) = \sum_j \mu(j) \max_{p^{(j)} \in \Delta_{\mathcal{A}}} \Big\{ \sum_i p_i^{(j)}(Q_i - \tau C_{ij}) - T_0 \sum_i p_i^{(j)} \log \tfrac{p_i^{(j)}}{\rho(i)} \Big\}. \tag{184}$$

For a fixed anchor j, let $a_i^{(j)} := Q_i - \tau C_{ij}$ and $p_i^{\star(j)} := \rho(i)e^{a_i^{(j)}/T_0}/Z_j(Q)$. For any $p \in \Delta_{\mathcal{A}}$, expanding $\mathrm{KL}(p\|p^{\star(j)}) = \sum_i p_i \log \frac{p_i}{\rho(i)} - \frac{1}{T_0}\sum_i p_i a_i^{(j)} + \log Z_j(Q)$ and rearranging yields the exact Gibbs variational identity

$$\sum_i p_i a_i^{(j)} - T_0 \sum_i p_i \log \tfrac{p_i}{\rho(i)} = T_0 \log Z_j(Q) - T_0\,\mathrm{KL}(p\|p^{\star(j)}). \tag{185}$$

Since KL divergence is nonnegative and vanishes only at $p = p^{\star(j)}$, the unique optimizer of column j is precisely $p^{\star(j)} = p^{(j)}(Q)$, with optimal value $T_0 \log Z_j(Q)$. More generally, combining (184) and (185) gives the global gap identity

$$\mathcal{J}_Q(\Gamma) = T_0 \sum_j \mu(j) \log Z_j(Q) - T_0 \sum_j \mu(j)\,\mathrm{KL}\big(p^{(j)}\|p^{\star(j)}\big), \tag{186}$$

so the objective attains its maximum if and only if every column equals its Gibbs optimizer. Substituting the optimal column values into (184) gives

$$F_s^{\mathrm{LOT}}(Q) = T_0 \sum_j \mu(j)\log Z_j(Q) = \tau\lambda \sum_j \mu_s(j)\log Z_j(Q),$$
$$Z_j(Q) = \sum_i \rho_s(i) \exp\Big(\tfrac{Q_i - \tau C_{ij}}{\tau\lambda}\Big).$$

which is exactly (9). The same value follows by direct substitution: (180) gives $\log \frac{\Gamma^\star_{ij}}{\rho(i)\mu(j)} = \frac{Q_i - \tau C_{ij}}{T_0} - \log Z_j(Q)$, so in (174) the score-and-cost term cancels entry by entry and

$$\mathcal{J}_Q(\Gamma^\star) = T_0 \sum_j \Big(\sum_i \Gamma^\star_{ij}\Big) \log Z_j(Q) = T_0 \sum_j \mu(j)\log Z_j(Q),$$

in agreement with the KL argument.

Finally, we verify consistency and explain why Sinkhorn is unnecessary.

Differentiating the closed-form value checks that the value and policy formulas agree:

$$\frac{\partial F_s^{\mathrm{LOT}}(Q)}{\partial Q_i} = T_0 \sum_j \mu(j) \frac{1}{Z_j(Q)} \frac{\rho(i)}{T_0} \exp\Big(\tfrac{Q_i - \tau C_{ij}}{T_0}\Big) = \sum_j \mu(j) p_i^{(j)}(Q) = \pi_i^{\mathrm{LOT}}(Q).$$

In balanced entropic OT, both marginals are prescribed, so the two unknown scaling vectors satisfy coupled row- and column-marginal equations, typically solved by alternating Sinkhorn updates. Here the row marginal is optimized rather than prescribed: the vector $u = \rho \odot \exp(Q/T_0)$ is fixed directly by the prior and the input action values, not an unknown balancing vector. One computes $Z = \mathsf{K}_C^\top u$ and $v = \mu \oslash Z$ once, after which $\Gamma^\star = \mathrm{diag}(u)\mathsf{K}_C\,\mathrm{diag}(v)$ and $\pi^{\mathrm{LOT}} = u \odot (\mathsf{K}_C v)$ follow by direct matrix–vector operations. There is no second marginal equation and hence no alternating row–column balancing. This proves every claim of Lemma 1. □

G Proofs of the Remaining Results

G.1 How to navigate the proofs

Every active theorem, lemma, proposition, and corollary is accounted for in Sections F and G. A result with several claims may use short transition paragraphs, but the argument remains in one proof block. When a main-text theorem is only a readable summary of more detailed results, we give a separate short consequence proof and identify exactly which detailed statements imply each part. Likewise, expanded appendix statements are explicitly named in the same proof block as their main-text versions rather than being left implicit.

The proof routes are largely independent after the closed form in Lemma 1. Table 8 lists every active formal result.

Table 8: Proof routes for active formal results

Proof route	**Results proved**
Closed form and Bellman operator	Lemma 1; Theorem 1; Proposition 3.
Online and batch learning	Theorems 9, 10, 11, and 4.
Smoothness and second order	Proposition 5; Theorems 12 and 13; Lemma 3.
Fixed action pruning	Lemmas 4 and 2; Theorem 14.
Recursive root-value estimation	Theorems 15 and 2.
Exact and truncated heat	Theorems 6, 7, 3, and 8; Propositions 1 and 2; Corollaries 2 and 1.
Local policy geometry	Theorems 5, 16, and 17; Propositions 4 and 6; Corollaries 3 and 4.
Learned geometry	Theorems 18 and 19; Corollaries 5, 6, and 7.
Spectral and moving-reference results	Theorems 20, 21, and 22; Corollary 8.
Continuous actions	Theorem 23; Corollary 9; Proposition 7.

G.2 Bellman backup and contraction

G.2.1 Proof of Theorem 1

Fix the state s and suppress its index when no confusion can arise. Write

$$\rho_i := \rho_s(i), \qquad \mu_j := \mu_s(j), \qquad T_0 := \tau\lambda > 0.$$

Under the assumptions of Lemma 1, we have $\rho_i > 0$, $\mu_j > 0$, and $C_{ij} < \infty$ for every i, j. Consequently, all exponential weights below are finite and strictly positive. In particular, every partition function $Z_j(Q)$ is positive, and the functions differentiated below are infinitely differentiable on $\mathbb{R}^K$.

We first rewrite the optimizer from Lemma 1.

Lemma 1 gives

$$\Gamma^\star_{ij} = u_i \mathsf{K}_{C,ij} v_j, \qquad u_i = \rho_i e^{Q_i/T_0}, \qquad \mathsf{K}_{C,ij} = e^{-C_{ij}/\lambda}, \qquad v_j = \frac{\mu_j}{(\mathsf{K}_C^\top u)_j}. \tag{187}$$

The denominator in the column scaling is

$$(\mathsf{K}_C^\top u)_j = \sum_k \mathsf{K}_{C,kj} u_k = \sum_k e^{-C_{kj}/\lambda} \rho_k e^{Q_k/T_0} = \sum_k \rho_k \exp\left(\frac{Q_k - \tau C_{kj}}{T_0}\right) = Z_j(Q), \tag{188}$$

where we used

$$-\frac{C_{kj}}{\lambda} = -\frac{\tau C_{kj}}{\tau\lambda} = -\frac{\tau C_{kj}}{T_0}.$$

Substituting (188) into (187) yields

$$\Gamma^\star_{ij} = \rho_i e^{Q_i/T_0} e^{-C_{ij}/\lambda} \frac{\mu_j}{Z_j(Q)} = \mu_j \frac{\rho_i \exp\left((Q_i - \tau C_{ij})/T_0\right)}{Z_j(Q)}. \tag{189}$$

Next, we identify each normalized column as a local softmax.

Because $\mu_j > 0$, the conditional distribution of the candidate action given anchor j is well defined by

$$p_i^{(j)}(Q) := \frac{\Gamma^\star_{ij}}{\mu_j}.$$

Equation (189) gives

$$p_i^{(j)}(Q) = \frac{\rho_i \exp\left((Q_i - \tau C_{ij})/T_0\right)}{\sum_k \rho_k \exp\left((Q_k - \tau C_{kj})/T_0\right)}, \tag{190}$$

which is precisely (10). Moreover,

$$p_i^{(j)}(Q) > 0$$

and

$$\sum_i p_i^{(j)}(Q) = \frac{\sum_i \rho_i \exp\left((Q_i - \tau C_{ij})/T_0\right)}{Z_j(Q)} = 1.$$

Thus $p^{(j)}(Q) \in \Delta_{\mathcal{A}}$.

The distribution $p^{(j)}(Q)$ is an ordinary prior-weighted softmax with prior ρ and temperature $T_0 = \tau\lambda$, applied to the anchor-dependent scores

$$Q_i - \tau C_{ij}.$$

Equivalently, its unnormalized weight is

$$\rho_i e^{Q_i/T_0} e^{-C_{ij}/\lambda}.$$

Thus the action cost appears in value units as τC_{ij} and in the exponential kernel as C_{ij}/λ.

For completeness, this local-softmax characterization also follows directly from the associated variational problem. For any $q \in \Delta_{\mathcal{A}}$, expanding the KL divergence from q to $p^{(j)}(Q)$ gives

$$\mathrm{KL}\left(q \middle\| p^{(j)}(Q)\right) = \sum_i q_i \left[\log \frac{q_i}{\rho_i} - \frac{Q_i - \tau C_{ij}}{T_0} + \log Z_j(Q)\right] \tag{191}$$

$$= \sum_i q_i \log \frac{q_i}{\rho_i} - \frac{1}{T_0} \sum_i q_i (Q_i - \tau C_{ij}) + \log Z_j(Q), \tag{192}$$

where we used $\sum_i q_i = 1$. Rearranging,

$$\sum_i q_i (Q_i - \tau C_{ij}) - T_0 \mathrm{KL}(q\|\rho) = T_0 \log Z_j(Q) - T_0 \mathrm{KL}\left(q \middle\| p^{(j)}(Q)\right). \tag{193}$$

Since KL divergence is nonnegative,

$$\sum_i q_i (Q_i - \tau C_{ij}) - T_0 \mathrm{KL}(q\|\rho) \le T_0 \log Z_j(Q).$$

Equality holds if and only if

$$q = p^{(j)}(Q).$$

Therefore $p^{(j)}(Q)$ is the unique maximizer of the local prior-regularized problem, and the corresponding optimal value is $T_0 \log Z_j(Q)$.

We continue to recover the policy mixture.

The LOT policy is the row marginal of the optimal coupling. Hence, for every candidate action i,

$$\pi_i^{\mathrm{LOT}}(Q) = (\Gamma^\star \mathbf{1})_i = \sum_j \Gamma_{ij}^\star = \sum_j \mu_j p_i^{(j)}(Q). \tag{194}$$

Stacking these componentwise identities gives

$$\Gamma_{ij}^\star = \mu_s(j) p_i^{(j)}(Q), \qquad \pi^{\mathrm{LOT}}(Q) = \sum_j \mu_s(j) p^{(j)}(Q),$$

which proves (11).

This is a convex mixture of probability distributions. Indeed, $\pi_i^{\mathrm{LOT}}(Q) \geq 0$, and

$$\sum_i \pi_i^{\mathrm{LOT}}(Q) = \sum_i \sum_j \mu_j p_i^{(j)}(Q) = \sum_j \mu_j \sum_i p_i^{(j)}(Q) = \sum_j \mu_j = 1.$$

Therefore $\pi^{\mathrm{LOT}}(Q) \in \Delta_{\mathcal{A}}$.

We then recover the value mixture.

The value formula in Lemma 1 is

$$F_s^{\mathrm{LOT}}(Q) = T_0 \sum_j \mu_j \log Z_j(Q).$$

Define the anchor-specific value

$$f_j(Q) := T_0 \log Z_j(Q) = \tau\lambda \log Z_j(Q).$$

Then

$$F_s^{\mathrm{LOT}}(Q) = \sum_j \mu_s(j) f_j(Q), \tag{195}$$

which proves (12).

Identity (193) shows more explicitly that

$$f_j(Q) = \max_{q \in \Delta_{\mathcal{A}}} \left\{ \sum_i q_i (Q_i - \tau C_{ij}) - T_0 \mathrm{KL}(q \| \rho) \right\}.$$

Thus the LOT value is an average of ordinary prior-weighted softmax values, one for each anchor.

We now differentiate each local log partition.

Define the unnormalized local weight

$$n_i^{(j)}(Q) := \rho_i \exp\left(\frac{Q_i - \tau C_{ij}}{T_0} \right).$$

Then

$$Z_j(Q) = \sum_i n_i^{(j)}(Q), \qquad p_i^{(j)}(Q) = \frac{n_i^{(j)}(Q)}{Z_j(Q)}.$$

For an arbitrary coordinate ℓ,

$$\frac{\partial n_i^{(j)}(Q)}{\partial Q_\ell} = \frac{1}{T_0} \mathbf{1}\{i = \ell\} n_i^{(j)}(Q), \tag{196}$$

$$\frac{\partial Z_j(Q)}{\partial Q_\ell} = \sum_i \frac{\partial n_i^{(j)}(Q)}{\partial Q_\ell} = \frac{1}{T_0} n_\ell^{(j)}(Q). \tag{197}$$

It follows that

$$\frac{\partial f_j(Q)}{\partial Q_\ell} = T_0 \frac{1}{Z_j(Q)} \frac{\partial Z_j(Q)}{\partial Q_\ell} = T_0 \frac{1}{Z_j(Q)} \frac{n_\ell^{(j)}(Q)}{T_0} = \frac{n_\ell^{(j)}(Q)}{Z_j(Q)} = p_\ell^{(j)}(Q). \tag{198}$$

Since this holds for every coordinate ℓ,

$$\nabla_Q f_j(Q) = p^{(j)}(Q). \tag{199}$$

Because there are finitely many anchors, we may differentiate (195) term by term. Thus,

$$\nabla_Q F_s^{\mathrm{LOT}}(Q) = \sum_j \mu_j \nabla_Q f_j(Q) = \sum_j \mu_j p^{(j)}(Q) = \pi^{\mathrm{LOT}}(Q), \tag{200}$$

where the last equality follows from (194). This proves the gradient identity directly from the closed-form value, without requiring an additional envelope theorem.

We next differentiate the local softmax.

Applying the quotient rule, together with (196) and (197), gives

$$\frac{\partial p_i^{(j)}(Q)}{\partial Q_\ell} = \frac{(\partial_\ell n_i^{(j)}) Z_j - n_i^{(j)} (\partial_\ell Z_j)}{Z_j^2} \tag{201}$$

$$= \frac{1}{T_0}\left[\mathbf{1}\{i=\ell\}\frac{n_i^{(j)}}{Z_j} - \frac{n_i^{(j)}}{Z_j}\frac{n_\ell^{(j)}}{Z_j}\right] \tag{202}$$

$$= \frac{1}{T_0} p_i^{(j)}(Q)\left(\mathbf{1}\{i=\ell\} - p_\ell^{(j)}(Q)\right). \tag{203}$$

In matrix form,

$$\nabla_Q p^{(j)}(Q) = \frac{1}{T_0}\left[\mathrm{diag}\big(p^{(j)}(Q)\big) - p^{(j)}(Q) p^{(j)}(Q)^\top\right]. \tag{204}$$

Since $\nabla_Q f_j(Q) = p^{(j)}(Q)$, we obtain

$$\nabla_Q^2 f_j(Q) = \frac{1}{T_0}\left[\mathrm{diag}\big(p^{(j)}(Q)\big) - p^{(j)}(Q) p^{(j)}(Q)^\top\right]. \tag{205}$$

Differentiating the finite mixture once more and using $T_0 = \tau\lambda$ gives

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \sum_j \mu_j \nabla_Q^2 f_j(Q) = \frac{1}{\tau\lambda}\sum_j \mu_s(j)\left[\mathrm{diag}\big(p^{(j)}(Q)\big) - p^{(j)}(Q) p^{(j)}(Q)^\top\right], \tag{206}$$

which is exactly (13).

We continue to verify the covariance structure.

For any probability vector $p \in \Delta_{\mathcal{A}}$, define

$$M(p) := \mathrm{diag}(p) - pp^\top.$$

Let I be a categorical random variable satisfying

$$\mathbb{P}(I = i) = p_i.$$

For any $x \in \mathbb{R}^K$,

$$x^\top M(p) x = \sum_i p_i x_i^2 - \left(\sum_i p_i x_i\right)^2 \tag{207}$$

$$= \sum_i p_i (x_i - \bar{x}_p)^2 \tag{208}$$

$$= \frac{1}{2}\sum_{i,k} p_i p_k (x_i - x_k)^2 \geq 0, \tag{209}$$

where

$$\bar{x}_p := \sum_i p_i x_i.$$

The first expression is exactly

$$\mathrm{Var}(x_I).$$

Equivalently, $M(p)$ is the covariance matrix of the one-hot random vector e_I. In particular, $M(p)$ is symmetric and positive semidefinite.

Furthermore,

$$M(p)\mathbf{1} = \mathrm{diag}(p)\mathbf{1} - p(p^\top \mathbf{1}) = p - p = 0.$$

Thus the all-ones vector is always a zero-curvature direction. When p has full support, equality in (209) holds if and only if

$$x_i = x_k \qquad \text{for every } i, k.$$

Therefore, for $K \geq 2$,

$$\ker M(p) = \mathrm{span}\{\mathbf{1}\}.$$

Every $p^{(j)}(Q)$ in the present setting has full support. Hence every local Hessian is positive semidefinite and has only the additive-constant direction in its nullspace. The same is true of their positive mixture:

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) \succeq 0, \qquad \nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\mathbf{1} = 0.$$

This is consistent with the translation identities

$$p^{(j)}(Q + c\mathbf{1}) = p^{(j)}(Q),$$

$$f_j(Q + c\mathbf{1}) = f_j(Q) + c,$$

and

$$F_s^{\mathrm{LOT}}(Q + c\mathbf{1}) = F_s^{\mathrm{LOT}}(Q) + c.$$

Thus adding the same constant to all action values changes the backed-up value by that constant but leaves the policy unchanged.

We then prove the sharp covariance bound.

We now prove

$$\left\|\mathrm{diag}(p) - pp^\top\right\|_2 \leq \frac{1}{2} \qquad \text{for every } p \in \Delta_{\mathcal{A}}. \tag{210}$$

Because $M(p)$ is symmetric and positive semidefinite,

$$\|M(p)\|_2 = \lambda_{\max}(M(p)) = \max_{\|x\|_2=1} x^\top M(p) x. \tag{211}$$

Fix a vector x with $\|x\|_2 = 1$, and define

$$a := \min_i x_i, \qquad b := \max_i x_i, \qquad m := \sum_i p_i x_i.$$

Since $x_i \in [a, b]$, we have

$$(x_i - a)(b - x_i) \geq 0$$

for every i. Expanding and averaging with respect to p gives

$$0 \leq \sum_i p_i (x_i - a)(b - x_i) = \sum_i p_i \left[(a+b)x_i - x_i^2 - ab\right] = (a+b)m - \sum_i p_i x_i^2 - ab.$$

Therefore,

$$\sum_i p_i x_i^2 \leq (a+b)m - ab.$$

Substituting this bound into the variance representation yields

$$x^\top M(p) x = \sum_i p_i x_i^2 - m^2 \leq (a+b)m - ab - m^2 = \frac{(b-a)^2}{4} - \left(m - \frac{a+b}{2}\right)^2 \leq \frac{(b-a)^2}{4}. \tag{212}$$

This is the usual range bound on variance, derived here directly.

Let r and t be indices attaining the maximum and minimum of x, respectively. If $r = t$, then $b - a = 0$. Otherwise,

$$b - a = x_r - x_t = \langle x,\, e_r - e_t \rangle \le \|x\|_2 \|e_r - e_t\|_2 = \sqrt{2}.$$

Thus

$$(b - a)^2 \le 2.$$

Combining this with (212) gives

$$x^\top M(p) x \le \frac{2}{4} = \frac{1}{2}.$$

Taking the maximum over unit vectors in (211) proves (210).

The constant $1/2$ is sharp. For example, take

$$K = 2, \qquad p = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \qquad x = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Then

$$M(p) = \begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{pmatrix}$$

and

$$M(p) x = \frac{1}{2} x.$$

Hence $\|M(p)\|_2 = 1/2$ in this example.

As an independent verification, the absolute row sum of row i is

$$\sum_k |M(p)_{ik}| = p_i(1 - p_i) + \sum_{k \ne i} p_i p_k = 2p_i(1 - p_i) \le \frac{1}{2}.$$

Since $M(p)$ is symmetric,

$$\|M(p)\|_1 = \|M(p)\|_\infty \le \frac{1}{2},$$

and therefore

$$\|M(p)\|_2 \le \sqrt{\|M(p)\|_1 \|M(p)\|_\infty} \le \frac{1}{2}.$$

We now pass the covariance bound through the mixture.

For every Q, the triangle inequality for the Euclidean operator norm, together with $\mu_j \ge 0$, $\sum_j \mu_j = 1$, and (210), gives

$$\left\| \nabla_Q^2 F_s^{\mathrm{LOT}}(Q) \right\|_2 \le \frac{1}{T_0} \sum_j \mu_j \left\| M\left(p^{(j)}(Q) \right) \right\|_2 \le \frac{1}{T_0} \sum_j \mu_j \frac{1}{2} = \frac{1}{2T_0} = \frac{1}{2\tau\lambda}. \tag{213}$$

Because all matrices in the mixture are positive semidefinite, the same conclusion can equivalently be obtained directly from their Rayleigh quotients.

Finally, we deduce global Euclidean smoothness.

Let $Q, Q' \in \mathbb{R}^K$ and set

$$d := Q' - Q.$$

The fundamental theorem of calculus applied to the continuously differentiable gradient gives

$$\nabla_Q F_s^{\mathrm{LOT}}(Q') - \nabla_Q F_s^{\mathrm{LOT}}(Q) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \nabla_Q F_s^{\mathrm{LOT}}(Q + td)\, \mathrm{d}t = \int_0^1 \nabla_Q^2 F_s^{\mathrm{LOT}}(Q + td) d\, \mathrm{d}t. \tag{214}$$

Using (213),

$$\begin{aligned}\left\|\nabla_Q F_s^{\mathrm{LOT}}(Q') - \nabla_Q F_s^{\mathrm{LOT}}(Q)\right\|_2 &\le \int_0^1 \left\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q+td)\right\|_2 \|d\|_2 \,\mathrm{d}t \\ &\le \int_0^1 \frac{1}{2\tau\lambda}\|d\|_2 \,\mathrm{d}t \\ &= \frac{1}{2\tau\lambda}\|Q'-Q\|_2.\end{aligned}$$

Therefore $\nabla_Q F_s^{\mathrm{LOT}}$ is globally $1/(2\tau\lambda)$-Lipschitz in Euclidean norm. This is precisely the claimed Euclidean smoothness.

Equivalently, the standard smoothness inequality gives

$$F_s^{\mathrm{LOT}}(Q') \le F_s^{\mathrm{LOT}}(Q) + \langle \nabla_Q F_s^{\mathrm{LOT}}(Q),\, Q'-Q\rangle + \frac{1}{4\tau\lambda}\|Q'-Q\|_2^2.$$

This proves every claim of Theorem 1. □

G.2.2 Proof of Proposition 3

Let $\mathbb{B} := \ell_\infty(\mathcal{S}\times\mathcal{A}) = \{Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R} : \|Q\|_\infty < \infty\}$ with $\|Q\|_\infty := \sup_{s,a}|Q(s,a)|$; the space $(\mathbb{B}, \|\cdot\|_\infty)$ is complete. Define $R_{\max} := \sup_{s,a}|r(s,a)| < \infty$ and $C_{\max} := \sup_s \|C_s\|_\infty < \infty$, and write $\mathbb{E}_{s'}$ for $\mathbb{E}_{s'\sim P(\cdot|s,a)}$. For a fixed state s, abbreviate $W_s(\pi) := W_{\lambda,s}^{\rho}(\pi,\mu_s)$ and $G_{s,\pi}(q) := \langle \pi,\, q\rangle - \tau W_s(\pi)$ for $q \in \mathbb{R}^{\mathcal{A}}$, so that the LOT aggregator is $F_s^{\mathrm{LOT}}(q) = \max_{\pi\in\Delta_{\mathcal{A}}} G_{s,\pi}(q)$. Throughout this state-local maximization, C_s, ρ_s, and μ_s are held fixed.

We first verify that the state-local optimization is finite.

For any $\pi \in \Delta_{\mathcal{A}}$, the independent coupling $\Gamma_{ij}^{\pi} := \pi_i\mu_s(j)$ belongs to $\Pi(\pi,\mu_s)$, and its relative-entropy term is

$$\sum_{i,j}\pi_i\mu_s(j)\log\tfrac{\pi_i\mu_s(j)}{\rho_s(i)\mu_s(j)} = \sum_i \pi_i \log\tfrac{\pi_i}{\rho_s(i)} = \mathrm{KL}(\pi\|\rho_s),$$

where terms with $\pi_i = 0$ are interpreted using $0\log 0 = 0$. Assumption 1 gives $\rho_s(i) > 0$, so this quantity is finite, and since C_s is finite the transport cost of this coupling is also finite. Hence $W_s(\pi) < \infty$ for every $\pi \in \Delta_{\mathcal{A}}$. Moreover, the action cost and the KL divergence are both nonnegative, so

$$W_s(\pi) \ge 0 \qquad \text{for every } \pi \in \Delta_{\mathcal{A}}. \tag{215}$$

Consequently, $G_{s,\pi}(q)$ and $F_s^{\mathrm{LOT}}(q)$ are finite for every finite q, and the closed form in Lemma 1 also shows directly that the maximum is attained.

Next, we prove monotonicity of the state aggregator.

Let $q \le q'$ componentwise. For every $\pi \in \Delta_{\mathcal{A}}$, nonnegativity of the policy coordinates gives $G_{s,\pi}(q') - G_{s,\pi}(q) = \langle \pi,\, q'-q\rangle = \sum_i \pi_i(q_i'-q_i) \ge 0$; the regularizer cancels because it does not depend on q. Hence $G_{s,\pi}(q) \le G_{s,\pi}(q')$ for every π, and taking the maximum over the same feasible simplex on both sides proves

$$q \le q' \quad\Longrightarrow\quad F_s^{\mathrm{LOT}}(q) \le F_s^{\mathrm{LOT}}(q'), \tag{216}$$

i.e. F_s^{LOT} is monotone in the componentwise order.

We next prove sup-norm nonexpansiveness.

For any fixed $\pi \in \Delta_{\mathcal{A}}$,

$$\left|G_{s,\pi}(q) - G_{s,\pi}(q')\right| = \left|\langle \pi,\, q-q'\rangle\right| \le \sum_i \pi_i|q_i-q_i'| \le \|q-q'\|_\infty \sum_i \pi_i = \|q-q'\|_\infty . \tag{217}$$

Passing this bound through the maximization via $\sup_x f(x) - \sup_x g(x) \le \sup_x(f(x)-g(x))$ gives $F_s^{\mathrm{LOT}}(q) - F_s^{\mathrm{LOT}}(q') \le \sup_{\pi\in\Delta_{\mathcal{A}}}[G_{s,\pi}(q) - G_{s,\pi}(q')] \le \|q-q'\|_\infty$, and interchanging q and q' yields

$$\left|F_s^{\mathrm{LOT}}(q) - F_s^{\mathrm{LOT}}(q')\right| \le \|q-q'\|_\infty : \tag{218}$$

the LOT aggregator is 1-Lipschitz under the sup norm. The same structure gives translation equivariance: for every $c \in \mathbb{R}$, using $\langle \pi,\, \mathbf{1} \rangle = 1$,

$$F_s^{\mathrm{LOT}}(q + c\mathbf{1}) = \max_{\pi \in \Delta_{\mathcal{A}}} \big\{ \langle \pi,\, q \rangle + c - \tau W_s(\pi) \big\} = F_s^{\mathrm{LOT}}(q) + c, \tag{219}$$

so adding a common constant to all action values changes the backed-up value by the same constant.

We continue to bound the difference from the hard maximum.

Set $M(q) := \max_i q_i$. By (215), $G_{s,\pi}(q) = \langle \pi,\, q \rangle - \tau W_s(\pi) \le \langle \pi,\, q \rangle \le M(q)$ for every $\pi \in \Delta_{\mathcal{A}}$, so taking the maximum over π gives

$$F_s^{\mathrm{LOT}}(q) \le M(q), \tag{220}$$

and the gap from the hard maximum is one-sided. For a uniform lower bound on $F_s^{\mathrm{LOT}}(q)$, let $e_a \in \Delta_{\mathcal{A}}$ denote the point mass at action a. The only coupling with row marginal e_a and column marginal μ_s is $\Gamma^{(a)}_{ij} := \mathbf{1}\{i = a\}\mu_s(j)$, whose transport objective equals

$$W_s(e_a) = \sum_j \mu_s(j) C_{s,aj} + \lambda \sum_j \mu_s(j) \log \tfrac{\mu_s(j)}{\rho_s(a)\mu_s(j)} = \sum_j \mu_s(j) C_{s,aj} + \lambda \log \tfrac{1}{\rho_s(a)} =: B_s(a). \tag{221}$$

Evaluating the policy optimization at $\pi = e_{a_q}$ for $a_q \in \arg\max_i q_i$ gives $F_s^{\mathrm{LOT}}(q) \ge G_{s,e_{a_q}}(q) = q_{a_q} - \tau W_s(e_{a_q}) = M(q) - \tau B_s(a_q)$, which combined with (220) yields

$$0 \le M(q) - F_s^{\mathrm{LOT}}(q) \le \tau B_s(a_q) \le \tau \max_a B_s(a). \tag{222}$$

We then show that the statewise gap bound is exact.

The preceding bound is exact after taking the supremum over q. Fix an action a and, for $L > 0$, set $q_a^{(L)} := 0$ and $q_i^{(L)} := -L$ for $i \ne a$, so that $M(q^{(L)}) = 0$. The closed value formula from Lemma 1, $F_s^{\mathrm{LOT}}(q^{(L)}) = \tau\lambda \sum_j \mu_s(j) \log[\sum_i \rho_s(i) \exp((q_i^{(L)} - \tau C_{s,ij})/(\tau\lambda))]$, here reads

$$F_s^{\mathrm{LOT}}(q^{(L)}) = \tau\lambda \sum_j \mu_s(j) \log\Big[\rho_s(a) e^{-C_{s,aj}/\lambda} + e^{-L/(\tau\lambda)} \sum_{i \ne a} \rho_s(i) e^{-C_{s,ij}/\lambda} \Big],$$

and letting $L \to \infty$,

$$\begin{aligned} \lim_{L\to\infty} F_s^{\mathrm{LOT}}(q^{(L)}) &= \tau\lambda \sum_j \mu_s(j) \log\big(\rho_s(a) e^{-C_{s,aj}/\lambda}\big) \\ &= \tau\lambda \log \rho_s(a) - \tau \sum_j \mu_s(j) C_{s,aj} = -\tau B_s(a). \end{aligned} \tag{223}$$

Since $M(q^{(L)}) = 0$, this proves $\sup_q [M(q) - F_s^{\mathrm{LOT}}(q)]/\tau \ge B_s(a)$ for every action a. Combining this lower bound with (222) gives

$$\sup_{q \in \mathbb{R}^{\mathcal{A}}} \frac{\big| F_s^{\mathrm{LOT}}(q) - \max_i q_i \big|}{\tau} = \max_a \Big\{ \sum_j \mu_s(j) C_{s,aj} + \lambda \log \tfrac{1}{\rho_s(a)} \Big\}, \tag{224}$$

the absolute value being harmless because $F_s^{\mathrm{LOT}}(q) \le \max_i q_i$. Consequently, using $\sum_j \mu_s(j) = 1$, $C_{s,aj} \le C_{\max}$, and $\rho_s(a) \ge \rho_{\min} > 0$,

$$\bar{W} = \sup_s \max_a \Big\{ \sum_j \mu_s(j) C_{s,aj} + \lambda \log \tfrac{1}{\rho_s(a)} \Big\} \le C_{\max} + \lambda \log \tfrac{1}{\rho_{\min}} < \infty. \tag{225}$$

Thus the uniform finiteness of $\bar{W}$ follows from the uniform cost bound and the uniform lower bound on the action prior; pointwise full support alone would not be sufficient if the state space were infinite. Equation (225) also implies

$$|F_s^{\mathrm{LOT}}(q)| \le |M(q)| + |F_s^{\mathrm{LOT}}(q) - M(q)| \le \|q\|_\infty + \tau\bar{W}. \tag{226}$$

We now lift monotonicity to the Bellman operator.

Let $Q, Q' \in \mathbb{B}$ with $Q(s,a) \le Q'(s,a)$ for every (s,a). Then $Q(s',\cdot) \le Q'(s',\cdot)$ for every next state s', so (216) gives $F^{\mathrm{LOT}}_{s'}(Q(s',\cdot)) \le F^{\mathrm{LOT}}_{s'}(Q'(s',\cdot))$. Taking the conditional expectation, multiplying by $\gamma > 0$, and adding the same immediate reward gives $(\mathcal{T}^{\mathrm{LOT}}Q)(s,a) \le (\mathcal{T}^{\mathrm{LOT}}Q')(s,a)$. Thus $\mathcal{T}^{\mathrm{LOT}}$ is monotone.

We next show that the Bellman operator is a self-map.

For every $Q \in \mathbb{B}$, (226) gives $|(\mathcal{T}^{\mathrm{LOT}}Q)(s,a)| \le |r(s,a)| + \gamma\,\mathbb{E}_{s'}\left[|F^{\mathrm{LOT}}_{s'}(Q(s',\cdot))|\right] \le R_{\max} + \gamma(\|Q\|_\infty + \tau\bar{W})$, and taking the supremum over (s,a),

$$\left\|\mathcal{T}^{\mathrm{LOT}}Q\right\|_\infty \le R_{\max} + \gamma\left(\|Q\|_\infty + \tau\bar{W}\right) < \infty. \tag{227}$$

Therefore $\mathcal{T}^{\mathrm{LOT}} : \mathbb{B} \to \mathbb{B}$.

We continue to prove Bellman contraction.

For any $Q, Q' \in \mathbb{B}$ and every (s,a), moving the absolute value inside the expectation and applying (218) pointwise in s' gives

$$\begin{aligned}\left|(\mathcal{T}^{\mathrm{LOT}}Q)(s,a) - (\mathcal{T}^{\mathrm{LOT}}Q')(s,a)\right| &= \gamma\left|\mathbb{E}_{s'}\left[F^{\mathrm{LOT}}_{s'}(Q(s',\cdot)) - F^{\mathrm{LOT}}_{s'}(Q'(s',\cdot))\right]\right| \\ &\le \gamma\mathbb{E}_{s'}\left[\|Q(s',\cdot) - Q'(s',\cdot)\|_\infty\right] \\ &\le \gamma\,\|Q - Q'\|_\infty .\end{aligned}$$

Taking the supremum over all state–action pairs gives

$$\left\|\mathcal{T}^{\mathrm{LOT}}Q - \mathcal{T}^{\mathrm{LOT}}Q'\right\|_\infty \le \gamma\,\|Q - Q'\|_\infty . \tag{228}$$

Since $\gamma \in (0,1)$, $\mathcal{T}^{\mathrm{LOT}}$ is a contraction on the complete metric space $\mathbb{B}$.

We then establish existence and uniqueness of the fixed point.

Banach's fixed-point theorem applied to (228) gives a unique bounded fixed point $Q^\star_{\mathrm{LOT}} \in \mathbb{B}$ satisfying $Q^\star_{\mathrm{LOT}} = \mathcal{T}^{\mathrm{LOT}}Q^\star_{\mathrm{LOT}}$. Moreover, for every $Q_0 \in \mathbb{B}$, the value-iteration sequence $Q_{n+1} = \mathcal{T}^{\mathrm{LOT}}Q_n$ converges geometrically:

$$\|Q_n - Q^\star_{\mathrm{LOT}}\|_\infty \le \gamma^n\,\|Q_0 - Q^\star_{\mathrm{LOT}}\|_\infty .$$

We now define the hard-max Bellman operator.

Define the ordinary Bellman optimality operator $(\mathcal{T}Q)(s,a) := r(s,a) + \gamma\,\mathbb{E}_{s'}[\max_b Q(s',b)]$. The hard maximum is monotone and 1-Lipschitz, $|\max_b q_b - \max_b q'_b| \le \|q - q'\|_\infty$, so $\mathcal{T}$ is monotone and $\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \le \gamma\,\|Q - Q'\|_\infty$. Furthermore, $\|\mathcal{T}Q\|_\infty \le R_{\max} + \gamma\,\|Q\|_\infty$, so $\mathcal{T}$ maps $\mathbb{B}$ into itself. Banach's theorem therefore gives its unique bounded fixed point $Q^\star = \mathcal{T}Q^\star$.

We next compare the two Bellman operators.

For any $Q \in \mathbb{B}$, the one-sided statewise gap gives

$$(\mathcal{T}Q)(s,a) - (\mathcal{T}^{\mathrm{LOT}}Q)(s,a) = \gamma\,\mathbb{E}_{s'}\left[\max_b Q(s',b) - F^{\mathrm{LOT}}_{s'}(Q(s',\cdot))\right] \in [0, \gamma\tau\bar{W}], \tag{229}$$

and consequently

$$\mathcal{T}^{\mathrm{LOT}}Q \le \mathcal{T}Q, \qquad \left\|\mathcal{T}^{\mathrm{LOT}}Q - \mathcal{T}Q\right\|_\infty \le \gamma\tau\bar{W}. \tag{230}$$

This comparison also implies the stronger componentwise fixed-point ordering $Q^\star_{\mathrm{LOT}} \le Q^\star$. Start both value iterations from the same bounded Q_0, i.e. $Q^{\mathrm{LOT}}_{n+1} = \mathcal{T}^{\mathrm{LOT}}Q^{\mathrm{LOT}}_n$ and $Q^{\max}_{n+1} = \mathcal{T}Q^{\max}_n$ with $Q^{\mathrm{LOT}}_0 = Q^{\max}_0$, and suppose inductively that $Q^{\mathrm{LOT}}_n \le Q^{\max}_n$. Then

$$Q^{\mathrm{LOT}}_{n+1} = \mathcal{T}^{\mathrm{LOT}}Q^{\mathrm{LOT}}_n \le \mathcal{T}Q^{\mathrm{LOT}}_n \le \mathcal{T}Q^{\max}_n = Q^{\max}_{n+1},$$

where the first inequality follows from (230) and the second from monotonicity of $\mathcal{T}$. Hence the ordering holds for every n, and taking the sup-norm limits gives $Q^\star_{\mathrm{LOT}} \le Q^\star$.

Finally, we propagate the one-step gap to the fixed points.

Using the fixed-point identities $Q^\star_{\mathrm{LOT}} = \mathcal{T}^{\mathrm{LOT}} Q^\star_{\mathrm{LOT}}$ and $Q^\star = \mathcal{T} Q^\star$, the triangle inequality (inserting $\mathcal{T}^{\mathrm{LOT}} Q^\star$) together with (228) and (230) gives

$$\begin{aligned}\|Q^\star_{\mathrm{LOT}} - Q^\star\|_\infty &\le \left\|\mathcal{T}^{\mathrm{LOT}} Q^\star_{\mathrm{LOT}} - \mathcal{T}^{\mathrm{LOT}} Q^\star\right\|_\infty + \left\|\mathcal{T}^{\mathrm{LOT}} Q^\star - \mathcal{T} Q^\star\right\|_\infty \\ &\le \gamma \|Q^\star_{\mathrm{LOT}} - Q^\star\|_\infty + \gamma\tau\bar{W}.\end{aligned}$$

Rearranging yields $(1-\gamma)\|Q^\star_{\mathrm{LOT}} - Q^\star\|_\infty \le \gamma\tau\bar{W}$, and since $1-\gamma > 0$,

$$\|Q^\star_{\mathrm{LOT}} - Q^\star\|_\infty \le \frac{\gamma\tau\bar{W}}{1-\gamma}, \tag{231}$$

which is exactly (89). This completes the proof. □

G.3 Online and batch learning of action geometry

G.3.1 Proof of Theorem 9

Let $\mathcal{X} := \mathcal{S} \times \mathcal{A}$ be the finite set of state–action pairs. For a bounded cost family $D = \{D_s\}_s$, define

$$(\mathcal{T}^D Q)(s,a) := r(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, F^{D_{s'}}_{s'}(Q(s',\cdot)). \tag{232}$$

By Proposition 3, $\mathcal{T}^D$ is a γ-contraction in $\|\cdot\|_\infty$; in particular, $\mathcal{T}^C$ has the unique fixed point $Q^\star_C$. Write $\widehat{C}_t := \{\widehat{C}_{s,t}\}_{s\in\mathcal{S}}$, $x_t := (s_t, a_t)$, and $\alpha_t := \alpha_{N_t(x_t)}(x_t)$. The entire cost family $\widehat{C}_t$, as well as Q_t and x_t, is $\mathcal{F}_t$-measurable.

We first identify the conditional TD target.

Define the sampled TD target

$$Y_{t+1} := r_t + \gamma F^{\widehat{C}_{s_{t+1},t}}_{s_{t+1}}\big(Q_t(s_{t+1},\cdot)\big). \tag{233}$$

Conditionally on $\mathcal{F}_t$, the table Q_t, the selected pair x_t, and the cost family $\widehat{C}_t$ are fixed, so the conditional reward mean and transition law in Assumption 3 give

$$\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] = r(s_t,a_t) + \gamma \sum_{s'} P(s' \mid s_t,a_t)\, F^{\widehat{C}_{s',t}}_{s'}\big(Q_t(s',\cdot)\big) = (\mathcal{T}^{\widehat{C}_t} Q_t)(x_t); \tag{234}$$

conditioning additionally on x_t is redundant because x_t is already $\mathcal{F}_t$-measurable. This calculation is the precise point at which predictability of $\widehat{C}_t$ is required: if the current transition were used to construct $\widehat{C}_t$ before forming the same TD target, (234) would not follow from the stated assumptions.

Next, we separate predictable geometry bias and sampling noise.

Define the predictable operator error

$$b_t := \mathcal{T}^{\widehat{C}_t} Q_t - \mathcal{T}^C Q_t. \tag{235}$$

For every next state s' and local value vector q, the learned-cost bound (123) gives $|F^{\widehat{C}_{s',t}}_{s'}(q) - F^{C_{s'}}_{s'}(q)| \le \tau \left\|\widehat{C}_{s',t} - C_{s'}\right\|_\infty \le \tau\varepsilon_t$, and applying this estimate inside the transition expectation in (232) yields $|b_t(s,a)| \le \gamma \sum_{s'} P(s' \mid s,a)\, \tau\varepsilon_t = \gamma\tau\varepsilon_t$. Consequently,

$$\|b_t\|_\infty \le \gamma\tau\varepsilon_t \longrightarrow 0 \qquad \text{almost surely}, \tag{236}$$

and b_t is $\mathcal{F}_t$-measurable because Q_t and $\widehat{C}_t$ are. Define the sampling innovation

$$M_{t+1} := Y_{t+1} - (\mathcal{T}^{\widehat{C}_t} Q_t)(x_t), \tag{237}$$

so that (234) implies

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = 0. \tag{238}$$

Thus b_t is a predictable vanishing drift, whereas M_{t+1} is martingale-difference reward and transition noise. Using $\mathcal{T}^{\widehat{C}_t}Q_t = \mathcal{T}^C Q_t + b_t$, the update at x_t can be written exactly as

$$Q_{t+1}(x_t) = Q_t(x_t) + \alpha_t\big[(\mathcal{T}^C Q_t)(x_t) - Q_t(x_t) + b_t(x_t) + M_{t+1}\big], \tag{239}$$

while $Q_{t+1}(x) = Q_t(x)$ for $x \neq x_t$.

We continue by uniformly bounding every learned-cost aggregator.

Assumption 3 gives $0 \le (\widehat{C}_{s,t})_{ij} \le \bar{C}$. At the zero vector, the closed-form LOT value is $F_s^{\widehat{C}_{s,t}}(0) = \tau\lambda \sum_j \mu_s(j) \log[\sum_i \rho_s(i) e^{-(\widehat{C}_{s,t})_{ij}/\lambda}]$, and since ρ_s is a probability distribution, every anchor j satisfies $e^{-\bar{C}/\lambda} \le \sum_i \rho_s(i) e^{-(\widehat{C}_{s,t})_{ij}/\lambda} \le 1$. Taking logarithms, multiplying by $\tau\lambda$, and averaging over j gives

$$-\tau\bar{C} \le F_s^{\widehat{C}_{s,t}}(0) \le 0. \tag{240}$$

By Proposition 3, the aggregator is 1-Lipschitz, so for every $q \in \mathbb{R}^{\mathcal{A}}$,

$$\big|F_s^{\widehat{C}_{s,t}}(q)\big| \le \big|F_s^{\widehat{C}_{s,t}}(q) - F_s^{\widehat{C}_{s,t}}(0)\big| + \big|F_s^{\widehat{C}_{s,t}}(0)\big| \le \|q\|_\infty + \tau\bar{C}, \tag{241}$$

uniformly over s and t.

We then prove pathwise boundedness of the iterates.

Set

$$B_Q := \max\Big\{\|Q_0\|_\infty,\ \frac{R_{\max} + \gamma\tau\bar{C}}{1-\gamma}\Big\}. \tag{242}$$

We prove by induction that

$$\|Q_t\|_\infty \le B_Q \qquad \text{for every } t. \tag{243}$$

The claim holds at $t = 0$. If it holds at time t, then (241) and $|r_t| \le R_{\max}$ imply $|Y_{t+1}| \le R_{\max} + \gamma(\|Q_t\|_\infty + \tau\bar{C}) \le R_{\max} + \gamma(B_Q + \tau\bar{C})$, and the definition of B_Q gives $R_{\max} + \gamma\tau\bar{C} \le (1-\gamma)B_Q$, hence

$$|Y_{t+1}| \le B_Q. \tag{244}$$

Since $\alpha_t \in [0,1]$, the updated coordinate is a convex combination, $|Q_{t+1}(x_t)| \le (1-\alpha_t)|Q_t(x_t)| + \alpha_t|Y_{t+1}| \le (1-\alpha_t)B_Q + \alpha_t B_Q = B_Q$, and every other coordinate is unchanged. This completes the induction and proves (243).

We now verify bounded martingale noise.

By (244), $|Y_{t+1}| \le B_Q$; conditional expectation preserves this bound, so $|(\mathcal{T}^{\widehat{C}_t}Q_t)(x_t)| = |\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t]| \le B_Q$. Consequently,

$$|M_{t+1}| \le 2B_Q, \qquad \mathbb{E}[M_{t+1}^2 \mid \mathcal{F}_t] \le 4B_Q^2, \tag{245}$$

so the martingale noise has uniformly bounded conditional second moment. For each $x \in \mathcal{X}$, define the effective asynchronous stepsize

$$a_t(x) := \alpha_t \mathbf{1}\{x_t = x\}. \tag{246}$$

If $t_0(x) < t_1(x) < \cdots$ are the successive update times of x, then $a_{t_n(x)}(x) = \alpha_n(x)$, so the infinite-visitation and stepsize assumptions give, almost surely,

$$\sum_{t=0}^{\infty} a_t(x) = \infty, \qquad \sum_{t=0}^{\infty} a_t(x)^2 < \infty \qquad \text{for every } x \in \mathcal{X}. \tag{247}$$

We next write the error recursion around the target fixed point.

Let $E_t := Q_t - Q_C^\star$. Since $Q_C^\star = \mathcal{T}^C Q_C^\star$, subtracting the fixed-point identity from (239) gives

$$E_{t+1}(x_t) = (1-\alpha_t)E_t(x_t) + \alpha_t\big[(\mathcal{T}^C Q_t)(x_t) - (\mathcal{T}^C Q_C^\star)(x_t) + b_t(x_t) + M_{t+1}\big], \tag{248}$$

while $E_{t+1}(x) = E_t(x)$ for $x \neq x_t$. The contraction property gives

$$\big\|\mathcal{T}^C Q_t - \mathcal{T}^C Q_C^\star\big\|_\infty \le \gamma\,\|E_t\|_\infty. \tag{249}$$

We continue to isolate and remove the martingale component.

Define an auxiliary noise-filter process $W_t \in \mathbb{R}^{\mathcal{X}}$ by $W_0 = 0$ and

$$W_{t+1}(x) = (1 - a_t(x))W_t(x) + a_t(x)M_{t+1} \qquad \text{for every } x \in \mathcal{X}. \tag{250}$$

When $x \neq x_t$, $a_t(x) = 0$ and that coordinate remains unchanged; when $x = x_t$, the filter uses exactly the same stepsize and martingale innovation as the Q-learning update. Fix a coordinate x. Because $a_t(x)$ and $W_t(x)$ are $\mathcal{F}_t$-measurable, (238) gives $\mathbb{E}[W_{t+1}(x)^2 \mid \mathcal{F}_t] = (1-a_t(x))^2 W_t(x)^2 + a_t(x)^2 \mathbb{E}[M_{t+1}^2 \mid \mathcal{F}_t]$, and using $(1-a)^2 \leq 1 - a$ for $a \in [0,1]$ together with (245),

$$\mathbb{E}[W_{t+1}(x)^2 \mid \mathcal{F}_t] \leq W_t(x)^2 - a_t(x)W_t(x)^2 + 4B_Q^2 a_t(x)^2. \tag{251}$$

The nonnegative almost-supermartingale lemma applied to (251), together with (247), implies that $W_t(x)^2$ converges almost surely and

$$\sum_{t=0}^{\infty} a_t(x)W_t(x)^2 < \infty \qquad \text{almost surely.} \tag{252}$$

The limit must be zero: if $W_t(x)^2 \to L_x > 0$, then eventually $W_t(x)^2 \geq L_x/2$, whence $\sum_t a_t(x)W_t(x)^2 \geq (L_x/2)\sum_t a_t(x) = \infty$, contradicting (252). Therefore $W_t(x) \to 0$ almost surely, and since $\mathcal{X}$ is finite this convergence holds simultaneously over all coordinates:

$$\|W_t\|_\infty \longrightarrow 0 \qquad \text{almost surely.} \tag{253}$$

We then reduce to a deterministic asynchronous contraction.

Define the residual process $Z_t := E_t - W_t$. Subtracting (250) from (248) cancels the martingale innovation exactly: at the updated coordinate,

$$Z_{t+1}(x_t) = (1 - \alpha_t)Z_t(x_t) + \alpha_t\big[(\mathcal{T}^C Q_t)(x_t) - (\mathcal{T}^C Q_C^\star)(x_t) + b_t(x_t)\big], \tag{254}$$

and every other coordinate remains unchanged. Because $E_t = Z_t + W_t$, the target in brackets satisfies, by (249),

$$\big|(\mathcal{T}^C Q_t)(x_t) - (\mathcal{T}^C Q_C^\star)(x_t) + b_t(x_t)\big| \leq \gamma \|E_t\|_\infty + \|b_t\|_\infty \leq \gamma \|Z_t\|_\infty + \gamma \|W_t\|_\infty + \|b_t\|_\infty .$$

Define

$$c_t := \gamma \|W_t\|_\infty + \|b_t\|_\infty , \tag{255}$$

so that (236) and (253) imply

$$c_t \longrightarrow 0 \qquad \text{almost surely.} \tag{256}$$

The process Z_t is bounded because Q_t is bounded, $Q_C^\star$ is finite, and $W_t \to 0$. On a probability-one sample path on which all preceding conclusions hold, define $L := \limsup_{t\to\infty} \|Z_t\|_\infty < \infty$ and fix any $\delta > 0$. By the definitions of L and c_t, there exists a finite time T_δ such that $\|Z_t\|_\infty \leq L + \delta$ and $c_t \leq \delta$ for every $t \geq T_\delta$. Therefore every target in the residual recursion after T_δ has absolute value at most

$$B_\delta := \gamma(L + \delta) + \delta. \tag{257}$$

Fix a coordinate x and consider its updates after T_δ. At every such update, (254) gives $|Z_{t+1}(x)| \leq (1 - a_t(x))|Z_t(x)| + a_t(x)B_\delta$, and consequently

$$\big(|Z_{t+1}(x)| - B_\delta\big)_+ \leq (1 - a_t(x))\big(|Z_t(x)| - B_\delta\big)_+. \tag{258}$$

Iterating this inequality yields $(|Z_t(x)| - B_\delta)_+ \leq (|Z_{T_\delta}(x)| - B_\delta)_+ \prod_{T_\delta \leq k < t}(1 - a_k(x))$, and since $a_k(x) \in [0,1]$ with $\sum_k a_k(x) = \infty$, the product tends to zero. Therefore $\limsup_{t\to\infty} |Z_t(x)| \leq B_\delta$, and because $\mathcal{X}$ is finite, $L = \limsup_{t\to\infty} \|Z_t\|_\infty \leq B_\delta = \gamma(L + \delta) + \delta$. Letting $\delta \downarrow 0$ gives $L \leq \gamma L$; since $L \geq 0$ and $\gamma < 1$, necessarily $L = 0$. Thus

$$\|Z_t\|_\infty \longrightarrow 0 \qquad \text{almost surely.} \tag{259}$$

Finally, $E_t = Z_t + W_t$, so $\|Q_t - Q_C^\star\|_\infty = \|E_t\|_\infty \leq \|Z_t\|_\infty + \|W_t\|_\infty \to 0$ almost surely. Hence

$$Q_t \longrightarrow Q_C^\star \qquad \text{almost surely,} \tag{260}$$

which proves (85).

We now explain why no summability or timescale separation is needed.

The proof uses $\|b_t\|_\infty \to 0$ only through the fact that c_t in (255) tends to zero; at no point do we require $\sum_t \alpha_t \|b_t\|_\infty < \infty$ or $\sum_t \alpha_t \varepsilon_t < \infty$. Thus $\widehat{C}_t$ may converge arbitrarily slowly, provided that it is predictable, uniformly bounded, and uniformly consistent, and the geometry estimate and value estimate need not evolve on separate timescales. If $\widehat{C}_t \equiv C$, then $\varepsilon_t = 0$ and $b_t \equiv 0$, so the same proof specializes immediately to ordinary tabular LOT Q-learning with the fixed contraction $\mathcal{T}^C$.

Finally, we prove convergence of the induced policies.

Fix a state s and define $\pi_{s,t} := \nabla_Q F_s^{\widehat{C}_{s,t}}(Q_t(s,\cdot))$ and $\pi_s^\star := \nabla_Q F_s^{C_s}(Q_C^\star(s,\cdot))$. Inserting the target-cost policy evaluated at the current value vector, $\tilde{\pi}_{s,t} := \nabla_Q F_s^{C_s}(Q_t(s,\cdot))$, the triangle inequality gives

$$\|\pi_{s,t} - \pi_s^\star\|_1 \le \|\pi_{s,t} - \tilde{\pi}_{s,t}\|_1 + \|\tilde{\pi}_{s,t} - \pi_s^\star\|_1 . \tag{261}$$

For the first term, the learned-cost policy bound (126), applied at the common input $Q_t(s,\cdot)$, gives

$$\|\pi_{s,t} - \tilde{\pi}_{s,t}\|_1 \le e^{2\varepsilon_t/\lambda} - 1. \tag{262}$$

For the second term, Theorem 1 gives the Euclidean smoothness estimate

$$\left\|\nabla_Q F_s^{C_s}(q) - \nabla_Q F_s^{C_s}(q')\right\|_2 \le \frac{1}{2\tau\lambda} \|q - q'\|_2 , \tag{263}$$

and applying the norm conversions $\|v\|_1 \le \sqrt{K}\,\|v\|_2$ and $\|v\|_2 \le \sqrt{K}\,\|v\|_\infty$ ($v \in \mathbb{R}^K$) around (263) gives

$$\|\tilde{\pi}_{s,t} - \pi_s^\star\|_1 \le \sqrt{K}\,\frac{1}{2\tau\lambda} \|Q_t(s,\cdot) - Q_C^\star(s,\cdot)\|_2 \le \frac{K}{2\tau\lambda} \|Q_t(s,\cdot) - Q_C^\star(s,\cdot)\|_\infty . \tag{264}$$

Combining (261)–(264),

$$\|\pi_{s,t} - \pi_s^\star\|_1 \le e^{2\varepsilon_t/\lambda} - 1 + \frac{K}{2\tau\lambda} \|Q_t(s,\cdot) - Q_C^\star(s,\cdot)\|_\infty . \tag{265}$$

Almost surely, the first term tends to zero because $\varepsilon_t \to 0$, and the second because $\|Q_t(s,\cdot) - Q_C^\star(s,\cdot)\|_\infty \le \|Q_t - Q_C^\star\|_\infty \to 0$. Therefore, for every state s, $\pi_{s,t} \to \pi_s^\star$ almost surely, which proves (86). Since the state set is finite and (265) is uniform in s, the proof actually gives uniform policy convergence over states. This completes the proof. □

G.3.2 Proof of Theorem 10

Recall that $D := |\mathcal{S}|$, $N := Kn$, $x_{a\ell} := e_{S'_{a\ell}} - e_{S_{a\ell}}$, and $B_{a\ell} := \frac{1}{2} x_{a\ell} x_{a\ell}^\top$. The samples are independent over (a,ℓ), although their distributions may depend on the action a; we construct one event that simultaneously controls the shared transition Laplacian and every action-specific transition mean. Throughout this proof, the infinity norm of a cost matrix denotes the entrywise maximum $\|C\|_{\max} := \max_{a,b} |C_{ab}|$, which is the cost norm denoted by $\|C\|_\infty$ in the theorem, and we use the standing full-support condition in Assumption 1 when comparing policy likelihood ratios.

We first establish the structure of each sampled Laplacian.

If $S'_{a\ell} = S_{a\ell}$, then $x_{a\ell} = 0$ and $B_{a\ell} = 0$. Otherwise $x_{a\ell} = e_u - e_v$ for two distinct states u, v, so $\|x_{a\ell}\|_2^2 = 2$; in both cases $\|x_{a\ell}\|_2 \le \sqrt{2}$. In the nonzero case, $v_{a\ell} := x_{a\ell}/\sqrt{2}$ is a unit vector and $B_{a\ell} = v_{a\ell} v_{a\ell}^\top$ is an orthogonal rank-one projection. In both the zero and nonzero cases,

$$B_{a\ell} \succeq 0, \qquad \|B_{a\ell}\|_2 \le 1, \qquad B_{a\ell}^2 = B_{a\ell}. \tag{266}$$

More explicitly, when $x_{a\ell} = e_u - e_v$, the only nonzero entries are $(B_{a\ell})_{uu} = (B_{a\ell})_{vv} = \frac{1}{2}$ and $(B_{a\ell})_{uv} = (B_{a\ell})_{vu} = -\frac{1}{2}$, so $B_{a\ell}$ is the combinatorial Laplacian of one undirected edge of weight $1/2$; in particular, $B_{a\ell} = B_{a\ell}^\top$ and $B_{a\ell}\mathbf{1} = 0$. Averages and expectations preserve symmetry, positive semidefiniteness, zero row sums, and nonpositive off-diagonal entries, so

$$L_\nu = \frac{1}{K} \sum_a \mathbb{E}[B_{a1}], \qquad \widehat{L}_\nu = \frac{1}{Kn} \sum_{a,\ell} B_{a\ell}$$

are symmetric positive-semidefinite weighted graph Laplacians. This construction does not require the transition kernel P to be reversible: each directed transition sample is converted into the symmetric edge Laplacian $\frac{1}{2}(e_{S'_{a\ell}} - e_{S_{a\ell}})(e_{S'_{a\ell}} - e_{S_{a\ell}})^\top$.

Next, we compute the matrix-Bernstein parameters.

Enumerate the $N = Kn$ independent matrices as $B_1, \dots, B_N$ and define $M_i := \mathbb{E}[B_i]$ and $Y_i := B_i - M_i$, so $\mathbb{E}[Y_i] = 0$. Since $0 \preceq B_i \preceq I$, taking expectations gives $0 \preceq M_i \preceq I$, hence $-I \preceq -M_i \preceq Y_i = B_i - M_i \preceq I - M_i \preceq I$ and

$$\|Y_i\|_2 \le 1. \tag{267}$$

Although the matrices need not be identically distributed across actions, they are centered around the correct population Laplacian: $\mathbb{E}[\widehat{L}_\nu] = \frac{1}{Kn}\sum_{a=1}^{K}\sum_{\ell=1}^{n}\mathbb{E}[B_{a\ell}] = \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}[B_{a1}] = L_\nu$. Expanding $(B_i - M_i)^2$, the cross terms combine with M_i^2 since $\mathbb{E}[B_i] = M_i$, and $B_i^2 = B_i$ from (266) gives

$$\mathbb{E}[Y_i^2] = \mathbb{E}[B_i^2] - M_i^2 = M_i - M_i^2. \tag{268}$$

Because the eigenvalues of M_i lie in $[0,1]$, those of $M_i - M_i^2$ have the form $\theta(1-\theta) \le \frac{1}{4}$, so

$$0 \preceq \mathbb{E}[Y_i^2] \preceq \tfrac{1}{4}I, \qquad \Big\|\sum_{i=1}^{N}\mathbb{E}[Y_i^2]\Big\|_2 \le \frac{N}{4}. \tag{269}$$

Finally, $\widehat{L}_\nu - L_\nu = \frac{1}{N}\sum_{i=1}^{N} Y_i$, so $\left\|\widehat{L}_\nu - L_\nu\right\|_2 \ge u$ exactly when $\|\sum_i Y_i\|_2 \ge Nu$. Self-adjoint matrix Bernstein (Tropp, 2012), with increment bound 1 and variance proxy $N/4$, bounds this probability by $2D\exp[-N^2u^2/(2(N/4) + (2/3)Nu)]$; simplifying, for every $u \ge 0$,

$$\Pr\Big\{\left\|\widehat{L}_\nu - L_\nu\right\|_2 \ge u\Big\} \le 2D\exp\Big[-\frac{Nu^2}{1/2 + (2/3)u}\Big]. \tag{270}$$

We next invert the matrix-Bernstein bound.

Set $\ell_L := \log\frac{4D}{\delta}$ and $\eta_L := \sqrt{\ell_L/(2N)} + 2\ell_L/(3N)$. The standard Bernstein inversion states that $\Pr\{\|\sum_i Y_i\|_2 \ge \sqrt{2\sigma^2\ell} + \frac{2\ell}{3}\} \le 2De^{-\ell}$, where here $\sigma^2 = \|\sum_i \mathbb{E}[Y_i^2]\|_2 \le N/4$, and dividing the threshold by N gives $\frac{1}{N}(\sqrt{2(N/4)\ell_L} + \frac{2\ell_L}{3}) = \sqrt{\ell_L/(2N)} + 2\ell_L/(3N) = \eta_L$. Therefore $\Pr\{\left\|\widehat{L}_\nu - L_\nu\right\|_2 > \eta_L\} \le 2De^{-\ell_L} = \delta/2$; equivalently, the event

$$\mathcal{E}_L := \Big\{\left\|\widehat{L}_\nu - L_\nu\right\|_2 \le \eta_L\Big\} \tag{271}$$

has probability at least $1 - \delta/2$.

We continue to bound the expected transition-mean error.

Fix an action a and abbreviate $x_\ell := x_{a\ell}$, $z := z_a = \mathbb{E}[x_\ell]$, and $\widehat{z} := \widehat{z}_a = \frac{1}{n}\sum_{\ell=1}^{n} x_\ell$. By Jensen's inequality, $\mathbb{E}\,\|\widehat{z} - z\|_2 \le \sqrt{\mathbb{E}\,\|\widehat{z} - z\|_2^2}$, and independence and centering eliminate all cross terms in $\mathbb{E}\|\frac{1}{n}\sum_\ell(x_\ell - z)\|_2^2$:

$$\mathbb{E}\,\|\widehat{z} - z\|_2^2 = \frac{1}{n^2}\sum_{\ell=1}^{n}\mathbb{E}\,\|x_\ell - z\|_2^2 = \frac{1}{n}\big(\mathbb{E}\,\|x_1\|_2^2 - \|z\|_2^2\big) \le \frac{2}{n}.$$

Hence

$$\mathbb{E}\,\|\widehat{z}_a - z_a\|_2 \le \sqrt{\frac{2}{n}}. \tag{272}$$

We then obtain simultaneous concentration of all action means.

Consider $f(x_1, \dots, x_n) := \|\frac{1}{n}\sum_{\ell=1}^{n} x_\ell - z\|_2$. If one observation x_ℓ is replaced by another admissible observation x'_ℓ, the reverse triangle inequality and $\|x_\ell\|_2, \|x'_\ell\|_2 \le \sqrt{2}$ give

$|f(\dots,x_\ell,\dots)-f(\dots,x'_\ell,\dots)| \le \frac{1}{n}\|x_\ell - x'_\ell\|_2 \le \frac{2\sqrt{2}}{n}$, so the sum of squared bounded-difference constants is $n(2\sqrt{2}/n)^2 = 8/n$. McDiarmid's inequality, together with (272), yields

$$\Pr\Big\{\|\widehat{z}_a - z_a\|_2 \ge \sqrt{\tfrac{2}{n}} + u\Big\} \le \exp\Big(-\frac{nu^2}{4}\Big). \tag{273}$$

Set $u_z := 2\sqrt{\log(2K/\delta)/n}$ and $\eta_z := \sqrt{2/n} + u_z$. For each fixed action, the failure probability in (273) at $u = u_z$ is $\exp(-nu_z^2/4) = \delta/(2K)$, so a union bound over the K actions shows that

$$\mathcal{E}_z := \Big\{\max_a \|\widehat{z}_a - z_a\|_2 \le \eta_z\Big\} \tag{274}$$

has probability at least $1-\delta/2$. Combining (271) and (274), the event

$$\mathcal{E} := \mathcal{E}_L \cap \mathcal{E}_z \tag{275}$$

has probability at least $1-\delta$. The events $\mathcal{E}_L$ and $\mathcal{E}_z$ are constructed from the same transition samples and need not be independent; the union bound does not require independence.

We now prove the heat-semigroup perturbation bound.

Let $A, B \succeq 0$ be symmetric and let $r \ge 0$. Duhamel's formula gives

$$e^{-rA} - e^{-rB} = -r\int_0^1 e^{-r(1-u)A}(A-B)e^{-ruB}\,du. \tag{276}$$

Since $A, B \succeq 0$, $\|e^{-vA}\|_2 \le 1$ and $\|e^{-vB}\|_2 \le 1$ for every $v \ge 0$, so taking operator norms in (276) gives

$$\|e^{-rA} - e^{-rB}\|_2 \le r\int_0^1 \|e^{-r(1-u)A}\|_2 \,\|A-B\|_2\, \|e^{-ruB}\|_2\,du \le r\,\|A-B\|_2. \tag{277}$$

Define $H := e^{-2tL_\nu}$ and $\widehat{H} := e^{-2t\widehat{L}_\nu}$. On $\mathcal{E}_L$, applying (277) with $r = 2t$ gives

$$\|\widehat{H} - H\|_2 \le 2t\,\|\widehat{L}_\nu - L_\nu\|_2 \le 2t\eta_L, \tag{278}$$

and both H and $\widehat{H}$ are symmetric positive-semidefinite contractions, $0 \preceq H \preceq I$ and $0 \preceq \widehat{H} \preceq I$.

We next rewrite the diffusion costs as quadratic forms.

For a pair of actions (a,b), write $w := z_a - z_b$ and $\widehat{w} := \widehat{z}_a - \widehat{z}_b$. Since $\|x_{a\ell}\|_2 \le \sqrt{2}$, both $\|z_a\|_2 = \|\mathbb{E}[x_{a1}]\|_2 \le \mathbb{E}\,\|x_{a1}\|_2 \le \sqrt{2}$ and $\|\widehat{z}_a\|_2 \le \frac{1}{n}\sum_\ell \|x_{a\ell}\|_2 \le \sqrt{2}$, so

$$\|w\|_2 \le 2\sqrt{2}, \qquad \|\widehat{w}\|_2 \le 2\sqrt{2}, \tag{279}$$

while on $\mathcal{E}_z$,

$$\|\widehat{w} - w\|_2 \le \|\widehat{z}_a - z_a\|_2 + \|\widehat{z}_b - z_b\|_2 \le 2\eta_z. \tag{280}$$

Because L_ν and $\widehat{L}_\nu$ are symmetric,

$$C_{ab}^t = \kappa_C\,\|e^{-tL_\nu}w\|_2^2 = \kappa_C w^\top e^{-2tL_\nu}w = \kappa_C w^\top H w, \tag{281}$$

$$\widehat{C}_{ab}^t = \kappa_C\,\|e^{-t\widehat{L}_\nu}\widehat{w}\|_2^2 = \kappa_C \widehat{w}^\top e^{-2t\widehat{L}_\nu}\widehat{w} = \kappa_C \widehat{w}^\top \widehat{H}\widehat{w}. \tag{282}$$

We continue to propagate the Laplacian and mean errors to the cost.

Adding and subtracting $\widehat{w}^\top H \widehat{w}$,

$$|\widehat{w}^\top\widehat{H}\widehat{w} - w^\top H w| \le |\widehat{w}^\top(\widehat{H}-H)\widehat{w}| + |\widehat{w}^\top H\widehat{w} - w^\top H w|.$$

The first term is at most $\|\widehat{w}\|_2^2\,\|\widehat{H}-H\|_2$. For the second, $\widehat{w}^\top H\widehat{w} - w^\top H w = (\widehat{w}-w)^\top H\widehat{w} + w^\top H(\widehat{w}-w)$ and $\|H\|_2 \le 1$ give $|\widehat{w}^\top H\widehat{w} - w^\top H w| \le \|\widehat{w}-w\|_2\,(\|\widehat{w}\|_2 + \|w\|_2)$. Combining these

bounds with (278)–(280), i.e. plugging in $\|\widehat{w}\|_2^2 \le (2\sqrt{2})^2$, $\left\|\widehat{H}-H\right\|_2 \le 2t\eta_L$, $\|\widehat{w}-w\|_2 \le 2\eta_z$, and $\|\widehat{w}\|_2 + \|w\|_2 \le 4\sqrt{2}$, on $\mathcal{E}$ we obtain simultaneously for every action pair (a,b),

$$\left|\widehat{C}^t_{ab} - C^t_{ab}\right| \le \kappa_C\big(16t\eta_L + 8\sqrt{2}\,\eta_z\big) = \varepsilon_{n,\delta}. \tag{283}$$

Therefore

$$\left\|\widehat{C}^t - C^t\right\|_\infty \le \varepsilon_{n,\delta} \tag{284}$$

on an event of probability at least $1-\delta$. This proves the first line of (87).

We then propagate the cost error to backups and fixed points.

On the same event, Theorem 18, applied with $\varepsilon_C = \varepsilon_{n,\delta}$, gives, uniformly over all states and local value vectors q,

$$\left|F_s^{\widehat{C}}(q) - F_s^C(q)\right| \le \tau\varepsilon_{n,\delta}. \tag{285}$$

Consequently, $\left\|\mathcal{T}^{\widehat{C}}Q - \mathcal{T}^C Q\right\|_\infty \le \gamma\tau\varepsilon_{n,\delta}$ for every Q, and since both Bellman operators are γ-contractions, their fixed points satisfy

$$\left\|Q^\star_{\widehat{C}} - Q^\star_C\right\|_\infty \le \frac{\gamma\tau}{1-\gamma}\varepsilon_{n,\delta}. \tag{286}$$

Equations (285) and (286) prove the remaining value statements in (87).

We now compare policies at their respective fixed points.

Fix a state s, an anchor j, and a candidate action i, and define the population and empirical dimensionless local logits

$$\theta_{s,ij} := \frac{Q^\star_C(s,i) - \tau C^t_{ij}}{\tau\lambda}, \qquad \widehat{\theta}_{s,ij} := \frac{Q^\star_{\widehat{C}}(s,i) - \tau\widehat{C}^t_{ij}}{\tau\lambda}.$$

By (284) and (286), using $\frac{\gamma}{1-\gamma}+1 = \frac{1}{1-\gamma}$, their difference satisfies

$$\left|\widehat{\theta}_{s,ij} - \theta_{s,ij}\right| \le \frac{\left\|Q^\star_{\widehat{C}} - Q^\star_C\right\|_\infty}{\tau\lambda} + \frac{\left\|\widehat{C}^t - C^t\right\|_\infty}{\lambda} \le \frac{\varepsilon_{n,\delta}}{\lambda(1-\gamma)} =: \beta_{n,\delta}. \tag{287}$$

For anchor j, define the unnormalized local weights $g_{s,ij} := \rho_s(i)e^{\theta_{s,ij}}$ and $\widehat{g}_{s,ij} := \rho_s(i)e^{\widehat{\theta}_{s,ij}}$. Equation (287) gives

$$e^{-\beta_{n,\delta}} \le \frac{\widehat{g}_{s,ij}}{g_{s,ij}} \le e^{\beta_{n,\delta}}. \tag{288}$$

Let $Z_{s,j} := \sum_i g_{s,ij}$ and $\widehat{Z}_{s,j} := \sum_i \widehat{g}_{s,ij}$. Summing the componentwise inequalities in (288) gives $e^{-\beta_{n,\delta}}Z_{s,j} \le \widehat{Z}_{s,j} \le e^{\beta_{n,\delta}}Z_{s,j}$, and hence

$$e^{-\beta_{n,\delta}} \le \frac{Z_{s,j}}{\widehat{Z}_{s,j}} \le e^{\beta_{n,\delta}}. \tag{289}$$

The local policies at their respective fixed points are $p^{C,(j)}_{s,i} := g_{s,ij}/Z_{s,j}$ and $p^{\widehat{C},(j)}_{s,i} := \widehat{g}_{s,ij}/\widehat{Z}_{s,j}$, so multiplying (288) and (289) yields

$$e^{-2\beta_{n,\delta}} \le \frac{p^{\widehat{C},(j)}_{s,i}}{p^{C,(j)}_{s,i}} \le e^{2\beta_{n,\delta}} \qquad \text{for every } i,j,s. \tag{290}$$

The mixed LOT policies use the same reference distribution μ_s, namely $\pi^C_s(i) = \sum_j \mu_s(j)p^{C,(j)}_{s,i}$ and $\pi^{\widehat{C}}_s(i) = \sum_j \mu_s(j)p^{\widehat{C},(j)}_{s,i}$. For example, the upper inequality in (290) implies $\pi^{\widehat{C}}_s(i) = \sum_j \mu_s(j)p^{\widehat{C},(j)}_{s,i} \le e^{2\beta_{n,\delta}}\sum_j \mu_s(j)p^{C,(j)}_{s,i} = e^{2\beta_{n,\delta}}\pi^C_s(i)$, and the lower inequality follows similarly. Therefore mixing preserves the likelihood-ratio interval:

$$e^{-2\beta_{n,\delta}} \le \frac{\pi^{\widehat{C}}_s(i)}{\pi^C_s(i)} \le e^{2\beta_{n,\delta}} \qquad \text{for every } i,s. \tag{291}$$

Full support of ρ_s and finiteness of the costs ensure that all denominators are positive. Taking logarithms, $|\log(\pi_s^{\widehat{C}}(i)/\pi_s^C(i))| \le 2\beta_{n,\delta}$, so both $D_{\mathrm{KL}}(\pi_s^{\widehat{C}}\|\pi_s^C) = \sum_i \pi_s^{\widehat{C}}(i)\log\frac{\pi_s^{\widehat{C}}(i)}{\pi_s^C(i)} \le 2\beta_{n,\delta}$ and, symmetrically, $D_{\mathrm{KL}}(\pi_s^C\|\pi_s^{\widehat{C}}) \le 2\beta_{n,\delta}$. Substituting the definition of $\beta_{n,\delta}$ proves, for every state s,

$$\max\Big\{D_{\mathrm{KL}}\big(\pi_s^{\widehat{C}}\|\pi_s^C\big),\, D_{\mathrm{KL}}\big(\pi_s^C\|\pi_s^{\widehat{C}}\big)\Big\} \le \frac{2\varepsilon_{n,\delta}}{\lambda(1-\gamma)}. \tag{292}$$

Finally, we justify simultaneous transfer across reward tasks.

The event $\mathcal{E}$ in (275) is defined entirely through the reward-free transition samples $\{(S_{a\ell}, S'_{a\ell})\}_{a,\ell}$: it depends on ν, P, t, κ_C, and the prescribed confidence level, but contains no reward observation and no reward function. On this single event, $\left\|\widehat{C}^t - C^t\right\|_\infty \le \varepsilon_{n,\delta}$ is a deterministic statement. For any bounded reward task sharing the same transition kernel P and the same fixed parameters $(\gamma, \tau, \lambda, \rho_s, \mu_s)$, Theorem 18 deterministically converts the same cost event into the backup, fixed-point, and policy bounds proved above. This implication holds for every such reward function on the same event, so no union bound over reward tasks is required, and all conclusions hold simultaneously for every bounded reward task sharing the dynamics. This completes the proof. □

G.3.3 Proof of Theorem 11

Fix $\gamma \in (0,1)$ and $\tau, \lambda, h > 0$ as in the theorem, and fix an integer $n \ge 1$. We construct the announced one-parameter family, compute its LOT fixed point exactly, and then prove the upper and lower bounds. The generative-model oracle returns only a sampled next state and its reward, which is identically zero here; it does not reveal the transition-induced graph or cost matrix. The family and its construction are known to the estimator, and only the parameter p is unknown.

We first define the transition family.

Let $\mathcal{S} = \{s, z_0, z_1\}$ and $\mathcal{A} = \{1, 2\}$, set $r(x, a) = 0$ for every (x, a), and let the parameter space be $\mathcal{P} = [\frac{3}{8}, \frac{5}{8}]$. For $p \in \mathcal{P}$, the only unknown transition row is

$$P_p(z_1 \mid s, 1) = p, \qquad P_p(z_0 \mid s, 1) = 1 - p, \tag{293}$$

while the remaining rows are known: $P_p(z_0 \mid s, 2) = 1$ and $P_p(s \mid z_b, a) = 1$ for $b \in \{0, 1\}$, $a \in \{1, 2\}$. Thus a query to $(s, 1)$ returns z_1 with probability p and z_0 with probability $1 - p$, whereas every other state–action query has the same deterministic response for all p. At every state x, take the action prior and reference-action distribution to be uniform, $\rho_x = \mu_x = (\frac{1}{2}, \frac{1}{2})$, and construct the two-node action-similarity graph by

$$(W_x(P_p))_{12} = (W_x(P_p))_{21} = \kappa\big(\|P_p(\cdot \mid x, 1) - P_p(\cdot \mid x, 2)\|_2\big), \qquad \kappa(d) = \Big(1 - \frac{d}{\sqrt{2}}\Big)_+,$$

with $(W_x)_{11} = (W_x)_{22} = 0$. Let $L_x(P_p) = \mathrm{diag}(W_x(P_p)\mathbf{1}) - W_x(P_p)$ be the corresponding combinatorial Laplacian, and define the squared diffusion-distance cost at time h by $(C_x^h(P_p))_{ab} = (e_a - e_b)^\top e^{-2hL_x(P_p)}(e_a - e_b)$. This is the construction in (128)–(129), specialized to the explicit Lipschitz similarity map κ.

Next, we compute the transition-induced costs.

In the ordered state basis (s, z_0, z_1), the two transition rows at s are $P_p(\cdot \mid s, 1) = (0, 1-p, p)$ and $P_p(\cdot \mid s, 2) = (0, 1, 0)$; their difference is $(0, -p, p)$, of norm $\sqrt{p^2 + p^2} = \sqrt{2}\,p$. Because $p \in [3/8, 5/8] \subset [0, 1]$, the positive-part operation in κ is inactive, so the off-diagonal graph weight is $w(p) = \kappa(\sqrt{2}\,p) = 1 - p$ and $L_s(p) = w(p)\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$. Let $v = e_1 - e_2$; since $\|v\|_2^2 = 2$ and $L_s(p)v = 2w(p)v$, spectral functional calculus gives $e^{-2hL_s(p)}v = e^{-4hw(p)}v$. The diagonal costs vanish and the two off-diagonal entries coincide, so

$$c(p) := (C_s^h(P_p))_{12} = v^\top e^{-2hL_s(p)}v = 2e^{-4h(1-p)}, \qquad c'(p) = 8he^{-4h(1-p)}; \tag{294}$$

in particular, c is strictly increasing in p. At either z_0 or z_1, the two action-conditioned transition rows are both the point mass at s, so their distance is zero and their graph weight is $\kappa(0) = 1$; the common, known off-diagonal cost at both states is thus $\bar{c} := 2e^{-4h}$.

We next evaluate a symmetric two-action LOT backup.

For $c \geq 0$, write $C^c = \left(\begin{smallmatrix} 0 & c \\ c & 0 \end{smallmatrix}\right)$, let F^c denote the LOT aggregator with this cost and with $\rho = \mu = (1/2, 1/2)$, and define $b(c) := \tau\lambda \log \frac{1+e^{-c/\lambda}}{2}$. For a constant action-value vector $q\mathbf{1}$, the closed form (9) gives, for either anchor j, $Z_j(q\mathbf{1}) = \frac{1}{2}e^{q/(\tau\lambda)} + \frac{1}{2}e^{(q-\tau c)/(\tau\lambda)} = e^{q/(\tau\lambda)}\frac{1+e^{-c/\lambda}}{2}$; both anchors have the same partition function and their weights sum to one, so

$$F^c(q\mathbf{1}) = \tau\lambda \sum_{j=1}^{2} \mu(j) \log Z_j(q\mathbf{1}) = q + \tau\lambda \log \frac{1+e^{-c/\lambda}}{2} = q + b(c).$$

Differentiating, $b'(c) = \tau\lambda \frac{-(1/\lambda)e^{-c/\lambda}}{1+e^{-c/\lambda}} = -\frac{\tau}{1+e^{c/\lambda}}$, so b is decreasing, and since $c \geq 0$, $|b'(c)| = \frac{\tau}{1+e^{c/\lambda}} \leq \frac{\tau}{2}$.

We continue to solve the regularized Bellman fixed point.

Let $Q_p^\star$ be the unique fixed point of the LOT Bellman operator for model P_p and the induced costs above. We first derive the symmetry of the fixed point directly from its Bellman equations. Since all four transition rows from (z_b, a) are the point mass at s and all rewards are zero, $Q_p^\star(z_b, a) = \gamma V_p^\star(s)$ for every $b \in \{0, 1\}$ and $a \in \{1, 2\}$; these four action values are equal, and we denote their common value by $y(p)$. The two z-states have the same cost $\bar{c}$, so $V_p^\star(z_0) = V_p^\star(z_1) = F^{\bar{c}}(y(p)\mathbf{1}) = y(p) + b(\bar{c})$. It follows that

$$Q_p^\star(s, 1) = \gamma\big[(1-p)V_p^\star(z_0) + pV_p^\star(z_1)\big] = \gamma\{y(p) + b(\bar{c})\},$$

and likewise $Q_p^\star(s, 2) = \gamma V_p^\star(z_0) = \gamma\{y(p) + b(\bar{c})\}$, so the two action values at s are also equal; denote their common value by $x(p)$. The regularized value at s is $V_p^\star(s) = F^{c(p)}(x(p)\mathbf{1}) = x(p) + b(c(p))$, and the fixed-point equations reduce to $x(p) = \gamma\{y(p)+b(\bar{c})\}$ and $y(p) = \gamma\{x(p)+b(c(p))\}$. Substituting the second equation into the first gives $x(p) = \gamma^2 x(p) + \gamma^2 b(c(p)) + \gamma b(\bar{c})$, and solving (similarly for y),

$$x(p) = \frac{\gamma^2 b(c(p)) + \gamma b(\bar{c})}{1-\gamma^2}, \qquad y(p) = \frac{\gamma b(c(p)) + \gamma^2 b(\bar{c})}{1-\gamma^2}. \tag{295}$$

Conversely, assigning these values to the six state–action pairs satisfies every Bellman equation, so uniqueness of the contraction fixed point shows that this is exactly $Q_p^\star$. This calculation also isolates the statistical difficulty: although the unknown transition row in (293) mixes z_0 and z_1, these states have identical continuation values, so p has no direct value effect through the transition expectation; its entire effect on $Q_p^\star$ is through the learned action cost $c(p)$.

We then establish a global Lipschitz bound.

Differentiating (295) gives $x'(p) = \frac{\gamma^2}{1-\gamma^2} b'(c(p))c'(p)$ and $y'(p) = \frac{\gamma}{1-\gamma^2} b'(c(p))c'(p)$; since $0 < \gamma < 1$, $|x'(p)| = \gamma|y'(p)|$, so the y coordinate has the larger derivative. For every $p \in [0, 1]$, $|b'(c(p))| \leq \tau/2$ and $c'(p) = 8he^{-4h(1-p)} \leq 8h$, hence

$$\max\{|x'(p)|, |y'(p)|\} = |y'(p)| \leq \frac{\gamma}{1-\gamma^2}\,\frac{\tau}{2}\,(8h) = \frac{4\gamma\tau h}{1-\gamma^2} =: L_{\gamma,\tau,h}.$$

The mean-value theorem therefore implies that, for any $p, q \in [0, 1]$,

$$\big\|Q_q^\star - Q_p^\star\big\|_\infty = \max\{|x(q) - x(p)|, |y(q) - y(p)|\} \leq L_{\gamma,\tau,h}|q - p|.$$

We now prove the plug-in upper bound.

Query $(s, 1)$ exactly n times and let X_ℓ indicate that the ℓth returned state is z_1, so that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ and $\widehat{p} = \frac{1}{n}\sum_{\ell=1}^n X_\ell$. The empirical frequency $\widehat{p}$ need not lie in the parameter interval $[3/8, 5/8]$; to ensure that the returned fixed point belongs to the stated family, define the projected estimate $\widetilde{p} = \Pi_{[3/8,5/8]}(\widehat{p})$ and return $Q_{\widetilde{p}}^\star$. Because the true p lies in the projection interval, $|\widetilde{p} - p| \leq |\widehat{p} - p|$. Using the Lipschitz bound, $\mathbb{E}_p|\widetilde{p} - p| \leq \mathbb{E}_p|\widehat{p} - p|$, Cauchy–Schwarz, and $\mathbb{E}_p(\widehat{p} - p)^2 = \text{Var}_p(\widehat{p}) = p(1-p)/n \leq 1/(4n)$,

$$\mathbb{E}_p\big\|Q_{\widetilde{p}}^\star - Q_p^\star\big\|_\infty \leq L_{\gamma,\tau,h}\sqrt{\mathbb{E}_p(\widehat{p} - p)^2} \leq \frac{4\gamma\tau h}{1-\gamma^2}\,\frac{1}{2\sqrt{n}} = \frac{2\gamma\tau h}{(1-\gamma^2)\sqrt{n}}.$$

This is a valid nonadaptive estimator using exactly n transition samples, so $\mathfrak{R}_n \le 2\gamma\tau h/((1-\gamma^2)\sqrt{n})$, which proves the upper side of (88). The same estimator satisfies a high-probability guarantee: Hoeffding's inequality gives $\Pr_p\{|\widehat{p}-p| > \sqrt{\log(2/\delta)/(2n)}\} \le \delta$, and projection cannot increase the error, so, with probability at least $1-\delta$,

$$\left\|Q^\star_{\widetilde{p}} - Q^\star_p\right\|_\infty \le \frac{4\gamma\tau h}{1-\gamma^2}\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{296}$$

We next choose two nearby models and separate their costs.

For the lower bound, define $d_n = 1/(8\sqrt{n})$ and $p_\pm = \frac{1}{2} \pm d_n$. Since $n \ge 1$, $d_n \le 1/8$, and therefore $p_-, p_+ \in [3/8, 5/8]$. Using (294) and $\sinh u \ge u$ for $u \ge 0$,

$$c(p_+) - c(p_-) = 2e^{-2h}\left(e^{4hd_n} - e^{-4hd_n}\right) = 4e^{-2h}\sinh(4hd_n) \ge 16he^{-2h}d_n = \frac{2he^{-2h}}{\sqrt{n}}.$$

Moreover, $p_+ \le 5/8$ gives $1-p_+ \ge 3/8$, and consequently $c(p_+) = 2e^{-4h(1-p_+)} \le 2e^{-3h/2}$.

We continue to convert cost separation into fixed-point separation.

Because c is increasing and b is decreasing, $|b(c(p_+)) - b(c(p_-))| = \int_{c(p_-)}^{c(p_+)} |b'(u)|\,du = \int_{c(p_-)}^{c(p_+)} \frac{\tau}{1+e^{u/\lambda}}\,du$, and since the integrand is decreasing in u,

$$|b(c(p_+)) - b(c(p_-))| \ge \frac{\tau\{c(p_+) - c(p_-)\}}{1 + e^{c(p_+)/\lambda}} \ge \frac{2\tau he^{-2h}}{1+\exp(2e^{-3h/2}/\lambda)}\,\frac{1}{\sqrt{n}}.$$

From (295), $|y(p_+) - y(p_-)| = \frac{\gamma}{1-\gamma^2}|b(c(p_+)) - b(c(p_-))|$ and $|x(p_+) - x(p_-)| = \frac{\gamma^2}{1-\gamma^2}|b(c(p_+)) - b(c(p_-))| = \gamma|y(p_+) - y(p_-)|$. Since $\gamma < 1$, the y coordinates determine the sup-norm separation, so $\Delta_n = |y(p_+) - y(p_-)|$ and

$$\Delta_n := \left\|Q^\star_{p_+} - Q^\star_{p_-}\right\|_\infty \ge \frac{2\gamma\tau he^{-2h}}{(1-\gamma^2)\{1+\exp(2e^{-3h/2}/\lambda)\}}\,\frac{1}{\sqrt{n}} = \frac{a_{\gamma,\tau,\lambda,h}}{\sqrt{n}}. \tag{297}$$

Thus the separation constant is exactly $a_{\gamma,\tau,\lambda,h} = \frac{2\gamma\tau he^{-2h}}{(1-\gamma^2)\{1+\exp(2e^{-3h/2}/\lambda)\}} > 0$, as stated in the theorem.

We then bound the information in an adaptive transcript.

Consider an arbitrary randomized algorithm using at most n adaptive generative-model queries; if it stops before n queries, pad the transcript with model-independent dummy observations, so it suffices to analyze exactly n rounds. Let H_{t-1} denote the complete history before round t, let I_t be the queried state–action pair, and let O_t be the returned next state, and include the algorithm's internal random seed in the transcript. Conditional on any common realized history, the algorithm uses the same query-selection kernel under p_- and p_+, so the query-selection decision itself contributes zero conditional KL divergence. Let $\mathbb{P}^{(n)}_-$ and $\mathbb{P}^{(n)}_+$ be the laws of the complete transcript under p_- and p_+. For every query other than $(s,1)$, the response distribution is identical under the two models; for a query to $(s,1)$, the response is Bernoulli with parameter p_- or p_+. Writing $N_{s,1} = \sum_{t=1}^n \mathbf{1}\{I_t = (s,1)\} \le n$ for the number of informative queries, the KL chain rule gives

$$D_{\mathrm{KL}}\left(\mathbb{P}^{(n)}_- \| \mathbb{P}^{(n)}_+\right) = \mathbb{E}_- \sum_{t=1}^n \mathbf{1}\{I_t = (s,1)\}\,\mathrm{kl}(p_-, p_+) = \mathbb{E}_-[N_{s,1}]\,\mathrm{kl}(p_-, p_+) \le n\,\mathrm{kl}(p_-, p_+) :$$

adaptivity cannot produce more information than querying $(s,1)$ on all n rounds. For the symmetric Bernoulli parameters $p_\pm = \frac{1}{2} \pm d_n$,

$$\mathrm{kl}(p_-, p_+) = \left(\tfrac{1}{2} - d_n\right)\log\frac{1/2 - d_n}{1/2 + d_n} + \left(\tfrac{1}{2} + d_n\right)\log\frac{1/2 + d_n}{1/2 - d_n} = 2d_n \log\frac{1+2d_n}{1-2d_n}.$$

For $u \in [0,1)$, $\log\frac{1+u}{1-u} = 2\int_0^u \frac{dv}{1-v^2} \le 2\int_0^u \frac{dv}{1-u^2} = \frac{2u}{1-u^2}$; applying this inequality with $u = 2d_n$,

$$n\,\mathrm{kl}(p_-, p_+) \le n(2d_n)\frac{4d_n}{1-4d_n^2} = \frac{8nd_n^2}{1-4d_n^2} = \frac{1/8}{1-1/(16n)} \le \frac{2}{15},$$

where the last inequality uses $n \geq 1$. Hence $D_{\mathrm{KL}}(\mathbb{P}_-^{(n)} \| \mathbb{P}_+^{(n)}) \leq 2/15$, and Pinsker's inequality gives

$$\mathrm{TV}\big(\mathbb{P}_-^{(n)}, \mathbb{P}_+^{(n)}\big) \leq \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}\big(\mathbb{P}_-^{(n)} \| \mathbb{P}_+^{(n)}\big)} \leq \frac{1}{\sqrt{15}}.$$

We now derive the expected-risk lower bound.

For completeness, we prove the metric two-point inequality used here. Fix an arbitrary estimator $\widehat{Q}$, include any internal randomization in its transcript, and define $d_\pm := \left\|\widehat{Q} - Q^\star_{p_\pm}\right\|_\infty$. By the triangle inequality, pointwise in the transcript, $d_- + d_+ \geq \left\|Q^\star_{p_+} - Q^\star_{p_-}\right\|_\infty = \Delta_n$. Let ν dominate both transcript laws and write $f_\pm = d\mathbb{P}_\pm^{(n)}/d\nu$. Then

$$\begin{aligned}
\mathbb{E}_{p_-} d_- + \mathbb{E}_{p_+} d_+ &= \int \big(d_- f_- + d_+ f_+\big) d\nu \\
&\geq \int \big(d_- + d_+\big) \min(f_-, f_+)\, d\nu \\
&\geq \Delta_n \int \min(f_-, f_+)\, d\nu.
\end{aligned}$$

and $\int \min(f_-, f_+)\, d\nu = 1 - \mathrm{TV}(\mathbb{P}_-^{(n)}, \mathbb{P}_+^{(n)})$, so $\mathbb{E}_{p_-} d_- + \mathbb{E}_{p_+} d_+ \geq \Delta_n\{1 - \mathrm{TV}(\mathbb{P}_-^{(n)}, \mathbb{P}_+^{(n)})\}$. Therefore

$$\max_{\sigma \in \{-,+\}} \mathbb{E}_{p_\sigma} \left\|\widehat{Q} - Q^\star_{p_\sigma}\right\|_\infty \geq \frac{\Delta_n}{2}\big\{1 - \mathrm{TV}\big(\mathbb{P}_-^{(n)}, \mathbb{P}_+^{(n)}\big)\big\} \geq \frac{a_{\gamma,\tau,\lambda,h}}{2\sqrt{n}}\Big(1 - \frac{1}{\sqrt{15}}\Big).$$

This is the metric two-point inequality (Polyanskiy & Wu, 2021), written explicitly for the sup norm. Since p_- and p_+ belong to the full parameter interval,

$$\mathfrak{R}_n = \inf_{\widehat{Q}} \sup_{p \in [3/8, 5/8]} \mathbb{E}_p \left\|\widehat{Q} - Q^\star_p\right\|_\infty \geq \frac{a_{\gamma,\tau,\lambda,h}}{2\sqrt{n}}\Big(1 - \frac{1}{\sqrt{15}}\Big).$$

Together with the plug-in upper bound, this proves (88).

We next derive the corresponding testing lower bound.

The same two models yield a probability lower bound. By the triangle inequality, every possible estimate is at distance at least $\Delta_n/2$ from at least one of $Q^\star_{p_-}$ and $Q^\star_{p_+}$. Define the test that selects the closer of the two fixed points: whenever this test makes an error, the estimation loss under the true model is at least $\Delta_n/2$. The minimum sum of the two testing errors is $1 - \mathrm{TV}(\mathbb{P}_-^{(n)}, \mathbb{P}_+^{(n)})$, so the sum of the two estimation-error probabilities at radius $\Delta_n/2$ is at least this quantity. Since $\Delta_n/2 \geq a_{\gamma,\tau,\lambda,h}/(2\sqrt{n})$, replacing $\Delta_n/2$ by the possibly smaller threshold on the right preserves the lower bound. Thus

$$\inf_{\widehat{Q}} \sup_{p \in [3/8, 5/8]} \Pr_p\Big\{\left\|\widehat{Q} - Q^\star_p\right\|_\infty \geq \frac{a_{\gamma,\tau,\lambda,h}}{2\sqrt{n}}\Big\} \geq \frac{1}{2}\big\{1 - \mathrm{TV}\big(\mathbb{P}_-^{(n)}, \mathbb{P}_+^{(n)}\big)\big\} \geq \frac{1}{2}\Big(1 - \frac{1}{\sqrt{15}}\Big).$$

Finally, we derive the mean-squared rate.

Although the theorem is stated for expected sup-norm loss, the same family also has a sharp mean-squared sup-norm rate. For the projected plug-in estimator, using $\mathbb{E}_p(\widetilde{p} - p)^2 \leq \mathbb{E}_p(\widehat{p} - p)^2$ again,

$$\sup_{p \in [3/8, 5/8]} \mathbb{E}_p \left\|Q^\star_{\widetilde{p}} - Q^\star_p\right\|_\infty^2 \leq L^2_{\gamma,\tau,h} \sup_p \mathbb{E}_p(\widehat{p} - p)^2 \leq \frac{L^2_{\gamma,\tau,h}}{4n} = \frac{4\gamma^2\tau^2h^2}{(1-\gamma^2)^2}\,\frac{1}{n}.$$

Conversely, since $\mathbb{E}X^2 \geq t^2 \Pr\{X \geq t\}$ with $t = a_{\gamma,\tau,\lambda,h}/(2\sqrt{n})$, the testing lower bound above gives, for every estimator,

$$\sup_{p \in [3/8, 5/8]} \mathbb{E}_p \left\|\widehat{Q} - Q^\star_p\right\|_\infty^2 \geq \frac{a^2_{\gamma,\tau,\lambda,h}}{8n}\Big(1 - \frac{1}{\sqrt{15}}\Big).$$

Taking the infimum over estimators gives

$$\frac{a^2_{\gamma,\tau,\lambda,h}}{8n}\Big(1-\frac{1}{\sqrt{15}}\Big) \le \inf_{\widehat{Q}} \sup_{p\in[3/8,5/8]} \mathbb{E}_p \left\|\widehat{Q}-Q^\star_p\right\|^2_\infty \le \frac{4\gamma^2\tau^2h^2}{(1-\gamma^2)^2}\,\frac{1}{n},$$

so the mean-squared sup-norm risk is $\Theta(n^{-1})$ on this family. The result is a sharpness statement for this explicit transition-induced subclass; it is not a matched lower bound for the estimator in Theorem 10. □

G.3.4 Proof of Theorem 4

Part (i), including the reward-free event, the rate $\widetilde{O}(\kappa_C[t/\sqrt{Kn}+1/\sqrt{n}])$, and the fixed-point perturbation bound, is Theorem 10. Part (ii), including predictable same-stream geometry, almost-sure Q convergence, and policy-readout convergence without a summable geometry bias, is Theorem 9. Combining the two independent statements proves the theorem. □

G.4 Smoothness, curvature, and second-order identities

G.4.1 Proof of Proposition 5

Fix a state s and suppress its index where no confusion can arise, writing $\rho=\rho_s$, $\mu=\mu_s$, $C=C_s$. Under the standing assumptions, $\mathcal{A}$ is finite, $\tau,\lambda>0$, the entries of C are finite, and ρ,μ have full support. We prove successively that the transport penalty is convex, that the LOT maximizer is unique, and that the resulting value and policy maps are smooth.

We first establish well-posedness of the transport problem.

Define the transport objective

$$\mathcal{R}(\Gamma) := \langle \Gamma,\, C\rangle + \lambda \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho(i)\mu(j)},$$

with the convention $0\log 0=0$. Since $\rho(i)\mu(j)>0$ for every i,j, the scalar function $\phi_{ij}(x) := x\log\frac{x}{\rho(i)\mu(j)} = x\log x - x\log(\rho(i)\mu(j))$, $x\ge 0$, is finite and continuous on $[0,\infty)$ and strictly convex there, being the strictly convex function $x\log x$ plus a linear function. It follows that $\mathcal{R}$ is continuous and strictly convex in Γ. For every $\pi\in\Delta_{\mathcal{A}}$, the coupling set $\Pi(\pi,\mu)=\{\Gamma\ge 0 : \Gamma\mathbf{1}=\pi,\ \Gamma^\top\mathbf{1}=\mu\}$ is nonempty, because the independent coupling $\pi\mu^\top$ belongs to it, and is convex and compact: it is a closed subset of the nonnegative orthant, and every feasible coupling has total mass $\sum_{i,j}\Gamma_{ij}=\sum_j\mu(j)=1$. Consequently, the minimum defining $W^\rho_{\lambda,s}(\pi,\mu)$ is attained, and since $\mathcal{R}$ is strictly convex, the minimizing coupling is unique for every fixed π. The regularizer is also finite on the whole simplex: evaluating the transport objective at the independent coupling, the $\mu(j)$ factor cancels inside the logarithm and

$$\mathcal{R}(\pi\mu^\top) = \sum_{i,j}\pi_i\mu(j)C_{ij} + \lambda\sum_i \pi_i\log\frac{\pi_i}{\rho(i)} < \infty.$$

Next, we prove convexity of the transport penalty.

Let $\pi^0,\pi^1\in\Delta_{\mathcal{A}}$, let Γ^0,Γ^1 be their respective optimal couplings, fix $\theta\in[0,1]$, and set $\pi^\theta := \theta\pi^0+(1-\theta)\pi^1$ and $\Gamma^\theta := \theta\Gamma^0+(1-\theta)\Gamma^1$. The marginal constraints are affine, so $\Gamma^\theta\mathbf{1}=\pi^\theta$ and $(\Gamma^\theta)^\top\mathbf{1}=\mu$, i.e. $\Gamma^\theta\in\Pi(\pi^\theta,\mu)$. Using feasibility of Γ^θ and convexity of $\mathcal{R}$,

$$W^\rho_{\lambda,s}(\pi^\theta,\mu) \le \mathcal{R}(\Gamma^\theta) \le \theta\mathcal{R}(\Gamma^0)+(1-\theta)\mathcal{R}(\Gamma^1) = \theta W^\rho_{\lambda,s}(\pi^0,\mu)+(1-\theta)W^\rho_{\lambda,s}(\pi^1,\mu),$$

so $\pi\mapsto W^\rho_{\lambda,s}(\pi,\mu)$ is convex on $\Delta_{\mathcal{A}}$. In fact, the same argument shows strict convexity: if $\pi^0\ne\pi^1$, their optimal couplings must satisfy $\Gamma^0\ne\Gamma^1$, since their row marginals differ, and for $\theta\in(0,1)$ strict convexity of $\mathcal{R}$ gives $W^\rho_{\lambda,s}(\pi^\theta,\mu) < \theta W^\rho_{\lambda,s}(\pi^0,\mu)+(1-\theta)W^\rho_{\lambda,s}(\pi^1,\mu)$. Only convexity is needed below, but this strictness is consistent with the uniqueness of the regularized-greedy policy.

We continue to rewrite the backup over a fixed feasible set.

Starting from (8), $F_s^{\mathrm{LOT}}(Q) = \max_{\pi\in\Delta_{\mathcal{A}}}\{\langle \pi,\, Q\rangle - \tau \min_{\Gamma\in\Pi(\pi,\mu)} \mathcal{R}(\Gamma)\}$, and since $\tau > 0$, for every fixed π the inner term equals $\max_{\Gamma\in\Pi(\pi,\mu)}\{-\tau\mathcal{R}(\Gamma)\}$. Moreover, if $\Gamma \geq 0$ satisfies $\Gamma^\top \mathbf{1} = \mu$, then $\pi := \Gamma\mathbf{1} \geq 0$ and $\mathbf{1}^\top\pi = \mathbf{1}^\top\mu = 1$, so $\pi \in \Delta_{\mathcal{A}}$; conversely, every coupling appearing in the original optimization satisfies this column constraint. We can therefore eliminate the optimized row marginal exactly:

$$F_s^{\mathrm{LOT}}(Q) = \max_{\Gamma\in\mathcal{G}_\mu} \mathcal{J}_Q(\Gamma), \qquad \mathcal{G}_\mu := \big\{\Gamma \geq 0 : \Gamma^\top\mathbf{1} = \mu\big\},$$

where

$$\mathcal{J}_Q(\Gamma) := \langle \Gamma,\, Q\mathbf{1}^\top - \tau C\rangle - \tau\lambda \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho(i)\mu(j)}.$$

This is an exact reparameterization, not a minimax interchange. The set $\mathcal{G}_\mu$ is nonempty, convex, and compact: every column has prescribed mass $\mu(j)$, so $0 \leq \Gamma_{ij} \leq \mu(j) \leq 1$, and every feasible plan has total mass one.

We then prove existence and uniqueness of the optimal plan.

For fixed Q, the term $\langle \Gamma,\, Q\mathbf{1}^\top - \tau C\rangle$ is linear in Γ, while $-\tau\lambda\sum_{i,j}\Gamma_{ij}\log\frac{\Gamma_{ij}}{\rho(i)\mu(j)}$ is strictly concave, because $\tau\lambda > 0$ and the corresponding entropy sum is strictly convex. Therefore $\Gamma \mapsto \mathcal{J}_Q(\Gamma)$ is strictly concave on $\mathcal{G}_\mu$. Continuity of $\mathcal{J}_Q$ and compactness of $\mathcal{G}_\mu$ give existence of a maximizer, while strict concavity gives uniqueness; denote this unique plan by $\Gamma^\star(Q)$. Lemma 1 gives the explicit formula

$$\Gamma^\star_{ij}(Q) = \mu(j)\,\frac{\rho(i)\exp\big((Q_i - \tau C_{ij})/(\tau\lambda)\big)}{Z_j(Q)}, \qquad Z_j(Q) := \sum_k \rho(k)\exp\Big(\frac{Q_k - \tau C_{kj}}{\tau\lambda}\Big).$$

Full support and entrywise-finite C imply $Z_j(Q) > 0$ and $\Gamma^\star_{ij}(Q) > 0$ for every finite Q and every i, j, so the optimizer is not only unique but also lies in the relative interior of each column simplex.

We now prove uniqueness of the maximizing policy.

Define the row marginal of the unique plan by $\pi^{\mathrm{LOT}}(Q) := \Gamma^\star(Q)\mathbf{1}$. This is a probability vector, since $\pi^{\mathrm{LOT}}(Q) \geq 0$ and $\mathbf{1}^\top\pi^{\mathrm{LOT}}(Q) = \sum_{i,j}\Gamma^\star_{ij}(Q) = \sum_j \mu(j) = 1$. Because the fixed-feasible-set reformulation above is exact and $\Gamma^\star(Q)$ attains its maximum, this row marginal attains the maximum in (8). Moreover, $\Gamma^\star(Q)$ must minimize $\mathcal{R}$ among couplings with marginals $(\pi^{\mathrm{LOT}}(Q), \mu)$: otherwise, replacing it by a lower-cost coupling with the same marginals would strictly increase $\mathcal{J}_Q$, contradicting its optimality. To prove uniqueness, suppose that $\widetilde{\pi}$ is another maximizer of (8), and let $\widetilde{\Gamma}$ be the transport minimizer defining $W^\rho_{\lambda,s}(\widetilde{\pi}, \mu)$. Then $\mathcal{J}_Q(\widetilde{\Gamma}) = \langle \widetilde{\pi},\, Q\rangle - \tau W^\rho_{\lambda,s}(\widetilde{\pi}, \mu) = F_s^{\mathrm{LOT}}(Q)$, so $\widetilde{\Gamma}$ is a maximizer of the semi-relaxed problem over $\mathcal{G}_\mu$. Uniqueness of that maximizer gives $\widetilde{\Gamma} = \Gamma^\star(Q)$, and therefore $\widetilde{\pi} = \widetilde{\Gamma}\mathbf{1} = \Gamma^\star(Q)\mathbf{1} = \pi^{\mathrm{LOT}}(Q)$. Thus $\pi^{\mathrm{LOT}}(Q)$ is the unique maximizer of (8).

We next prove convexity of the LOT aggregator.

For each fixed $\pi \in \Delta_{\mathcal{A}}$, the map $Q \mapsto \langle \pi,\, Q\rangle - \tau W^\rho_{\lambda,s}(\pi, \mu)$ is affine in Q, so $F_s^{\mathrm{LOT}}(Q) = \sup_{\pi\in\Delta_{\mathcal{A}}}\{\langle \pi,\, Q\rangle - \tau W^\rho_{\lambda,s}(\pi, \mu)\}$ is a pointwise supremum of affine functions and is therefore convex. Equivalently, $F_s^{\mathrm{LOT}} = [\tau W^\rho_{\lambda,s}(\cdot, \mu) + \iota_{\Delta_{\mathcal{A}}}]^\star$ is the convex conjugate of the transport penalty restricted to the simplex, where $\iota_{\Delta_{\mathcal{A}}}$ is zero on $\Delta_{\mathcal{A}}$ and $+\infty$ outside it. The same conclusion follows directly from the semi-relaxed formulation: for every fixed $\Gamma \in \mathcal{G}_\mu$, the map $Q \mapsto \mathcal{J}_Q(\Gamma)$ is affine, so maximizing over Γ preserves convexity.

We continue to apply Danskin's theorem.

The semi-relaxed feasible set $\mathcal{G}_\mu$ is fixed and compact, and $\mathcal{J}_Q(\Gamma)$ is continuous jointly in (Q, Γ) and differentiable in Q, with $\nabla_Q\mathcal{J}_Q(\Gamma) = \Gamma\mathbf{1}$. Because the maximizer $\Gamma^\star(Q)$ is unique, Danskin's theorem gives

$$\nabla_Q F_s^{\mathrm{LOT}}(Q) = \Gamma^\star(Q)\mathbf{1} = \pi^{\mathrm{LOT}}(Q),$$

so F_s^{LOT} is differentiable at every finite Q, and its gradient is exactly the unique policy selected by the regularized backup. Uniqueness is essential here: without it, Danskin's theorem would generally

identify a set of possible subgradients rather than a single gradient. Danskin's theorem and uniqueness establish differentiability; continuity of the gradient and the stronger smoothness assertion follow from the closed form below.

Finally, we prove continuous differentiability and smoothness.

Let $T_0 = \tau\lambda$. Lemma 1 gives $F_s^{\mathrm{LOT}}(Q) = T_0 \sum_j \mu(j) \log Z_j(Q)$ with $Z_j(Q) = \sum_i \rho(i) \exp((Q_i - \tau C_{ij})/T_0)$. Every $Z_j(Q)$ is a finite sum of strictly positive exponential functions, so $Z_j(Q) > 0$ for every $Q \in \mathbb{R}^K$, and both Z_j and $\log Z_j$ are real analytic on $\mathbb{R}^K$. Therefore $F_s^{\mathrm{LOT}} \in C^\infty(\mathbb{R}^K)$; in particular, F_s^{LOT} is continuously differentiable. Differentiating componentwise, using $\partial Z_j(Q)/\partial Q_i = \frac{1}{T_0}\rho(i)\exp((Q_i - \tau C_{ij})/T_0)$,

$$\frac{\partial F_s^{\mathrm{LOT}}(Q)}{\partial Q_i} = \sum_j \mu(j)\, \frac{\rho(i)\exp\big((Q_i - \tau C_{ij})/T_0\big)}{Z_j(Q)} = \sum_j \Gamma^\star_{ij}(Q) = \pi_i^{\mathrm{LOT}}(Q),$$

which directly confirms the gradient identity obtained from Danskin's theorem. Finally, every component $\Gamma^\star_{ij}(Q)$ in the closed-form plan above is a quotient of real-analytic functions with an everywhere-positive denominator, so $Q \mapsto \Gamma^\star(Q)$ is real analytic, and summing its columns shows that $Q \mapsto \pi^{\mathrm{LOT}}(Q) = \Gamma^\star(Q)\mathbf{1}$ is also real analytic and therefore smooth. This proves that F_s^{LOT} is convex and continuously differentiable, that $\nabla_Q F_s^{\mathrm{LOT}}(Q) = \pi^{\mathrm{LOT}}(Q)$, that $\pi^{\mathrm{LOT}}(Q)$ is the unique maximizer of (8), and that the policy-response map $Q \mapsto \pi^{\mathrm{LOT}}(Q)$ is smooth. □

G.4.2 Proof of Theorem 12

Fix a state s and suppress the state index on ρ_s, μ_s, and C_s when convenient; the cost matrix and the two distributions are held fixed while differentiating with respect to Q. Put $\beta := 1/(\tau\lambda)$ and $\mathsf{K}_{C,ij} := e^{-C_{ij}/\lambda}$. Assumption 1, together with entrywise finiteness of C, implies $\rho_s(i) > 0$, $\mu_s(j) > 0$, and $\mathsf{K}_{C,ij} > 0$ for every pair of actions i, j.

We first recall the closed-form plan and policy.

Define $u_i(Q) := \rho_s(i) e^{\beta Q_i}$ and, for each anchor j, $Z_j(Q) := (\mathsf{K}_C^\top u)_j = \sum_\ell \mathsf{K}_{C,\ell j} u_\ell$; all terms in this sum are positive, so $Z_j(Q) > 0$. By Lemma 1, $v_j(Q) = \mu_s(j)/Z_j(Q)$ and

$$\Gamma^\star_{ij}(Q) = u_i \mathsf{K}_{C,ij} v_j = \frac{u_i \mathsf{K}_{C,ij} \mu_s(j)}{Z_j(Q)}, \qquad \pi_i^{\mathrm{LOT}}(Q) = \sum_j \Gamma^\star_{ij}(Q) = u_i \sum_j \frac{\mathsf{K}_{C,ij}\mu_s(j)}{Z_j(Q)},$$

the LOT policy being the row marginal of the optimal plan. All quantities in these expressions are smooth functions of Q.

Next, we differentiate the basic factors.

Since u_i depends only on the coordinate Q_i, $\partial u_i/\partial Q_k = \beta u_i \delta_{ik}$, where δ_{ik} is the Kronecker delta. Differentiating the partition function, $\partial Z_j/\partial Q_k = \sum_\ell \mathsf{K}_{C,\ell j} \beta u_\ell \delta_{\ell k} = \beta u_k \mathsf{K}_{C,kj}$, and it follows that

$$\frac{\partial}{\partial Q_k}\Big(\frac{1}{Z_j}\Big) = -\frac{1}{Z_j^2}\frac{\partial Z_j}{\partial Q_k} = -\beta\, \frac{u_k \mathsf{K}_{C,kj}}{Z_j^2}.$$

We then differentiate the LOT policy componentwise.

Applying the product rule to the representation $\pi_i^{\mathrm{LOT}} = u_i \sum_j \mathsf{K}_{C,ij}\mu_s(j)/Z_j$ and substituting the derivatives just computed yields

$$\frac{\partial \pi_i^{\mathrm{LOT}}}{\partial Q_k} = \beta \delta_{ik} u_i \sum_j \frac{\mathsf{K}_{C,ij}\mu_s(j)}{Z_j} - \beta u_i u_k \sum_j \frac{\mathsf{K}_{C,ij}\mathsf{K}_{C,kj}\mu_s(j)}{Z_j^2},$$

and since the first term equals $\beta\delta_{ik}\pi_i^{\mathrm{LOT}}$,

$$\frac{\partial \pi_i^{\mathrm{LOT}}}{\partial Q_k} = \beta\delta_{ik}\pi_i^{\mathrm{LOT}} - \beta u_i u_k \sum_j \frac{\mathsf{K}_{C,ij}\mathsf{K}_{C,kj}\mu_s(j)}{Z_j^2}. \tag{298}$$

We now express the cross term through the optimal plan.

Since $\mu_s(j) > 0$, the closed-form plan $\Gamma^\star_{ij} = u_i \mathsf{K}_{C,ij} \mu_s(j)/Z_j$ gives

$$\frac{\Gamma^\star_{ij}\Gamma^\star_{kj}}{\mu_s(j)} = u_i u_k \, \frac{\mathsf{K}_{C,ij}\mathsf{K}_{C,kj}\mu_s(j)}{Z_j^2},$$

so (298) becomes $\partial \pi_i^{\mathrm{LOT}}/\partial Q_k = \beta\big(\delta_{ik}\pi_i^{\mathrm{LOT}} - \sum_j \Gamma^\star_{ij}\Gamma^\star_{kj}/\mu_s(j)\big)$. By Proposition 5, $\nabla_Q F_s^{\mathrm{LOT}}(Q) = \pi^{\mathrm{LOT}}(Q)$, so the policy Jacobian is the Hessian of the LOT aggregator:

$$\big[\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\big]_{ik} = \frac{1}{\tau\lambda}\Big(\delta_{ik}\pi_i^{\mathrm{LOT}} - \sum_j \frac{\Gamma^\star_{ij}\Gamma^\star_{kj}}{\mu_s(j)}\Big). \tag{299}$$

We next identify the transport co-occurrence matrix.

Define $A := \Gamma^\star \operatorname{diag}(\mu_s)^{-1}\Gamma^{\star\top}$; the inverse is well defined because μ_s has full support. Componentwise, $A_{ik} = \sum_j \Gamma^\star_{ij}\Gamma^\star_{kj}/\mu_s(j)$, which is precisely the transport co-occurrence matrix in (93), so (299) gives

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda}\big[\operatorname{diag}(\pi^{\mathrm{LOT}}) - A\big].$$

The matrix A is symmetric and entrywise nonnegative, and it has the Gram representation $A = [\Gamma^\star \operatorname{diag}(\mu_s)^{-1/2}][\Gamma^\star \operatorname{diag}(\mu_s)^{-1/2}]^\top$, hence $A \succeq 0$. This Gram representation explains the co-occurrence interpretation: A_{ik} is large when actions i and k repeatedly receive substantial mass through the same reference actions. The fact that $A \succeq 0$ alone does not yet imply that $\operatorname{diag}(A\mathbf{1}) - A$ is positive semidefinite; that property will follow from the nonnegative symmetric edge weights through the graph-energy identity below.

We continue to calculate the weighted degrees.

The optimal plan satisfies the prescribed column marginal $\sum_k \Gamma^\star_{kj} = \mu_s(j)$, so for every action i,

$$(A\mathbf{1})_i = \sum_j \frac{\Gamma^\star_{ij}}{\mu_s(j)} \sum_k \Gamma^\star_{kj} = \sum_j \frac{\Gamma^\star_{ij}}{\mu_s(j)}\,\mu_s(j) = \sum_j \Gamma^\star_{ij} = \pi_i^{\mathrm{LOT}}.$$

Hence $A\mathbf{1} = \pi^{\mathrm{LOT}}$ and $\operatorname{diag}(A\mathbf{1}) = \operatorname{diag}(\pi^{\mathrm{LOT}})$. Defining $L_A := \operatorname{diag}(A\mathbf{1}) - A$, we obtain

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda}\big(\operatorname{diag}(\pi^{\mathrm{LOT}}) - \Gamma^\star \operatorname{diag}(\mu_s)^{-1}\Gamma^{\star\top}\big) = \frac{1}{\tau\lambda} L_A,$$

which is the Hessian identity in (94).

We then prove the graph-Laplacian properties.

Because A is symmetric and entrywise nonnegative, L_A is the combinatorial Laplacian of an undirected weighted graph; the diagonal entries A_{ii} can be regarded as self-loops, which cancel in $\operatorname{diag}(A\mathbf{1}) - A$ and therefore do not affect the Laplacian quadratic form. First, $L_A\mathbf{1} = \operatorname{diag}(A\mathbf{1})\mathbf{1} - A\mathbf{1} = A\mathbf{1} - A\mathbf{1} = 0$. Second, for any $x \in \mathbb{R}^K$, $x^\top L_A x = \sum_i (\sum_k A_{ik}) x_i^2 - \sum_{i,k} A_{ik} x_i x_k$, and expanding the square below with the symmetry $A_{ik} = A_{ki}$ recovers exactly this expression, giving the graph-energy identity

$$x^\top L_A x = \frac{1}{2}\sum_{i,k} A_{ik}(x_i - x_k)^2 \geq 0,$$

the inequality holding because $A_{ik} \geq 0$. Thus $L_A \succeq 0$, and it follows from (94) and $\tau\lambda > 0$ that $\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) \succeq 0$, which independently verifies the convexity of F_s^{LOT}. Under the full-support and finite-cost assumptions, $\Gamma^\star_{ij} > 0$ for every i, j, so $A_{ik} = \sum_j \Gamma^\star_{ij}\Gamma^\star_{kj}/\mu_s(j) > 0$ for every $i \neq k$: the induced co-occurrence graph is complete, and in particular connected. Hence $\ker L_A = \operatorname{span}\{\mathbf{1}\}$ and $\operatorname{rank}(L_A) = K - 1$. The theorem itself only requires the weaker statement $L_A\mathbf{1} = 0$, but connectedness will be useful when the tangent-space inverse of L_A is considered.

We now verify the affine translation direction.

For every scalar c and every $\pi \in \Delta_{\mathcal{A}}$, $\langle \pi,\, Q + c\mathbf{1}\rangle = \langle \pi,\, Q\rangle + c\langle \pi,\, \mathbf{1}\rangle = \langle \pi,\, Q\rangle + c$, and the transport penalty is independent of Q, so $F_s^{\mathrm{LOT}}(Q+c\mathbf{1}) = c+\max_{\pi\in\Delta_{\mathcal{A}}}\{\langle \pi,\, Q\rangle - \tau W^{\rho}_{\lambda,s}(\pi,\mu_s)\} = F_s^{\mathrm{LOT}}(Q) + c$: the vector $\mathbf{1}$ is the affine translation direction of the aggregator. The same fact follows from the closed form: $u_i(Q + c\mathbf{1}) = e^{\beta c}u_i(Q)$, hence $Z_j(Q + c\mathbf{1}) = e^{\beta c}Z_j(Q)$, and the common factor cancels from each normalized local policy, so $\pi^{\mathrm{LOT}}(Q + c\mathbf{1}) = \pi^{\mathrm{LOT}}(Q)$. Differentiating the translation identity twice is consistent with $\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\mathbf{1} = \frac{1}{\tau\lambda} L_A \mathbf{1} = 0$.

We next recover the covariance representation.

Theorem 1 gives $\Gamma^{\star}_{ij} = \mu_s(j)p_i^{(j)}$ and $\pi^{\mathrm{LOT}} = \sum_j \mu_s(j)p^{(j)}$. Substituting the first identity into A gives $A_{ik} = \sum_j \mu_s(j)p_i^{(j)}p_k^{(j)}$, i.e. $A = \sum_j \mu_s(j)p^{(j)}p^{(j)\top}$, and similarly $\mathrm{diag}(\pi^{\mathrm{LOT}}) = \sum_j \mu_s(j)\,\mathrm{diag}(p^{(j)})$. It follows that

$$L_A = \mathrm{diag}(\pi^{\mathrm{LOT}}) - A = \sum_j \mu_s(j)\big[\mathrm{diag}(p^{(j)}) - p^{(j)}p^{(j)\top}\big],$$

so the graph-Laplacian representation (94) is identical to the covariance-mixture representation (13). In probabilistic terms, each matrix $\mathrm{diag}(p^{(j)}) - p^{(j)}p^{(j)\top}$ is the covariance matrix of a one-hot categorical action drawn from the local policy $p^{(j)}$, and L_A is their average over the anchors.

Finally, we prove the eigenvalue and smoothness bounds.

Since L_A is symmetric positive semidefinite, $\|L_A\|_{\mathrm{op}} = \lambda_{\max}(L_A)$, and (94) therefore gives the exact pointwise identity $\big\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\big\|_{\mathrm{op}} = \frac{1}{\tau\lambda}\lambda_{\max}(L_A(Q))$. For completeness, we now prove the dimension-free bound $\lambda_{\max}(L_A) \le 1/2$. For any probability vector p, let $\Sigma(p) := \mathrm{diag}(p) - pp^{\top}$, whose diagonal and off-diagonal entries are $\Sigma(p)_{ii} = p_i(1-p_i)$ and $\Sigma(p)_{ik} = -p_ip_k$ for $i \neq k$. The absolute row sum of row i is therefore

$$\sum_k |\Sigma(p)_{ik}| = p_i(1-p_i) + p_i\sum_{k\neq i} p_k = 2p_i(1-p_i) \le \frac{1}{2},$$

and since $\Sigma(p)$ is symmetric, $\|\Sigma(p)\|_{\mathrm{op}} \le \sqrt{\|\Sigma(p)\|_1\,\|\Sigma(p)\|_\infty} = \|\Sigma(p)\|_\infty \le \frac{1}{2}$. Using the covariance representation established above and $\sum_j \mu_s(j) = 1$,

$$\|L_A\|_{\mathrm{op}} = \Big\|\sum_j \mu_s(j)\Sigma(p^{(j)})\Big\|_{\mathrm{op}} \le \sum_j \mu_s(j)\,\Big\|\Sigma(p^{(j)})\Big\|_{\mathrm{op}} \le \frac{1}{2},$$

and since $L_A \succeq 0$, $\lambda_{\max}(L_A) = \|L_A\|_{\mathrm{op}} \le \frac{1}{2}$. Combining this with (94) gives $\big\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\big\|_{\mathrm{op}} \le \frac{1}{2\tau\lambda}$: thus $1/(2\tau\lambda)$ is a valid global Euclidean smoothness constant, while the exact pointwise curvature norm at a particular Q is $\frac{1}{\tau\lambda}\lambda_{\max}(L_A(Q))$. Finally, to pass from the pointwise Hessian bound to global Lipschitzness of the gradient, let $Q, \widehat{Q} \in \mathbb{R}^K$. The fundamental theorem of calculus gives $\nabla_Q F_s^{\mathrm{LOT}}(Q) - \nabla_Q F_s^{\mathrm{LOT}}(\widehat{Q}) = \int_0^1 \nabla_Q^2 F_s^{\mathrm{LOT}}(\widehat{Q} + t(Q-\widehat{Q}))(Q-\widehat{Q})\,dt$, and consequently

$$\begin{aligned}\Big\|\nabla_Q F_s^{\mathrm{LOT}}(Q) - \nabla_Q F_s^{\mathrm{LOT}}(\widehat{Q})\Big\|_2 &\le \int_0^1 \Big\|\nabla_Q^2 F_s^{\mathrm{LOT}}\big(\widehat{Q} + t(Q-\widehat{Q})\big)\Big\|_{\mathrm{op}}\Big\|Q-\widehat{Q}\Big\|_2\,dt \\ &\le \frac{1}{2\tau\lambda}\Big\|Q-\widehat{Q}\Big\|_2.\end{aligned}$$

This proves the Hessian identity, the graph-Laplacian properties, translation equivariance, convexity, and the stated smoothness bound. □

G.4.3 Proof of Theorem 13

Fix s and hence ρ_s, μ_s, C_s, set $T := \tau\lambda > 0$, and write $p_j(Q) := p^{(j)}(Q)$. For every anchor j, let

$$Z_j(Q) := \sum_\ell \rho_s(\ell)e^{(Q_\ell - \tau C_{s,\ell j})/T},$$

$$p_{j,i}(Q) := \frac{\rho_s(i)e^{(Q_i - \tau C_{s,ij})/T}}{Z_j(Q)}.$$

Assumption 1 and finite costs imply $Z_j(Q) > 0$, so each $p_j(Q)$ is a probability vector; the closed form also makes p_j and F_s^{LOT} smooth on $\mathbb{R}^K$.

We first compute the Jacobian of each local softmax.

For fixed j and coordinates i, k, differentiating the numerator and partition function gives

$$\partial_{Q_k}\left[\rho_s(i)e^{(Q_i-\tau C_{s,ij})/T}\right] = \frac{\rho_s(i)e^{(Q_i-\tau C_{s,ij})/T}}{T}\delta_{ik},$$
$$\partial_{Q_k} Z_j(Q) = \frac{Z_j(Q)}{T}p_{j,k}(Q).$$

Thus, by the quotient rule,

$$\begin{aligned}\partial_{Q_k} p_{j,i}(Q) &= \frac{1}{T}p_{j,i}(Q)\big(\delta_{ik} - p_{j,k}(Q)\big),\\ Dp_j(Q) &= \frac{1}{T}\big[\mathrm{diag}(p_j(Q)) - p_j(Q)p_j(Q)^\top\big].\end{aligned} \tag{300}$$

Equivalently, with the categorical covariance $\Sigma(p) := \mathrm{diag}(p) - pp^\top$ for $p \in \Delta_{\mathcal{A}}$, $Dp_j(Q) = T^{-1}\Sigma(p_j(Q))$.

Next, we establish the dimension-free covariance bound.

For $p \in \Delta_{\mathcal{A}}$, $\Sigma(p)_{ii} = p_i(1-p_i)$ and $\Sigma(p)_{ik} = -p_i p_k$ for $i \neq k$. Hence every row satisfies

$$\sum_k |\Sigma(p)_{ik}| = p_i(1-p_i) + \sum_{k\neq i} p_i p_k = 2p_i(1-p_i) \leq \frac{1}{2}.$$

Because $\Sigma(p)$ is symmetric, $\|\Sigma(p)\|_{\mathrm{op}} \leq \sqrt{\|\Sigma(p)\|_1\|\Sigma(p)\|_\infty} = \|\Sigma(p)\|_\infty \leq \frac{1}{2}$; together with (300), this gives

$$\|Dp_j(Q)\|_{\mathrm{op}} \leq \frac{1}{2T} \qquad \text{for all } j, Q. \tag{301}$$

We continue to deduce Lipschitz continuity of the local policies.

Let $\Delta := Q - \widehat{Q}$. The fundamental theorem of calculus and (301) yield

$$\begin{aligned}\|p_j(Q) - p_j(\widehat{Q})\|_2 &= \left\|\int_0^1 Dp_j(\widehat{Q} + t\Delta)\Delta\, dt\right\|_2\\ &\leq \int_0^1 \|Dp_j(\widehat{Q} + t\Delta)\|_{\mathrm{op}}\|\Delta\|_2\, dt\\ &\leq \frac{1}{2T}\|Q - \widehat{Q}\|_2.\end{aligned} \tag{302}$$

We then bound the change of one covariance matrix.

Fix j and put $p := p_j(Q)$, $\widehat{p} := p_j(\widehat{Q})$, and $d := p - \widehat{p}$. Then

$$\begin{aligned}\Sigma(p) - \Sigma(\widehat{p}) &= \mathrm{diag}(d) - (pp^\top - \widehat{p}\widehat{p}^\top),\\ \|\,\mathrm{diag}(d)\|_{\mathrm{op}} &= \|d\|_\infty \leq \|d\|_2.\end{aligned}$$

Moreover,

$$pp^\top - \widehat{p}\widehat{p}^\top = (p-\widehat{p})p^\top + \widehat{p}(p-\widehat{p})^\top,$$

so $\|xy^\top\|_{\mathrm{op}} = \|x\|_2\|y\|_2$ and $\|p\|_2, \|\widehat{p}\|_2 \leq 1$ imply

$$\|pp^\top - \widehat{p}\widehat{p}^\top\|_{\mathrm{op}} \leq (\|p\|_2 + \|\widehat{p}\|_2)\|p - \widehat{p}\|_2 \leq 2\|p - \widehat{p}\|_2.$$

Consequently, with no dependence on the action count K,

$$\|\Sigma(p) - \Sigma(\widehat{p})\|_{\mathrm{op}} \leq 3\|p - \widehat{p}\|_2. \tag{303}$$

We now prove Hessian Lipschitzness.

Write $\Sigma_j(Q) := \Sigma(p_j(Q))$. By Theorem 1, $H(Q) := \nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = T^{-1}\sum_j \mu_s(j)\Sigma_j(Q)$. Therefore, using the triangle inequality, (303), (302), and $\sum_j \mu_s(j) = 1$,

$$\begin{aligned}\|H(Q) - H(\widehat{Q})\|_{\mathrm{op}} &\le \frac{1}{T}\sum_j \mu_s(j)\|\Sigma_j(Q) - \Sigma_j(\widehat{Q})\|_{\mathrm{op}} \\ &\le \frac{3}{T}\sum_j \mu_s(j)\|p_j(Q) - p_j(\widehat{Q})\|_2 \\ &\le \frac{3}{2T^2}\|Q - \widehat{Q}\|_2.\end{aligned}$$

Since $T = \tau\lambda$,

$$\|H(Q) - H(\widehat{Q})\|_{\mathrm{op}} \le \frac{3}{2\tau^2\lambda^2}\|Q - \widehat{Q}\|_2,$$

which proves (95).

We next derive the integral second-order Taylor formula.

Set $F := F_s^{\mathrm{LOT}}$, $\Delta := Q - \widehat{Q}$, and $g(t) := F(\widehat{Q} + t\Delta)$ for $t \in [0,1]$. The chain rule gives $g'(t) = \nabla F(\widehat{Q} + t\Delta)^\top \Delta$ and $g''(t) = \Delta^\top H(\widehat{Q} + t\Delta)\Delta$. Applying the fundamental theorem of calculus twice,

$$\begin{aligned}g(1) - g(0) - g'(0) &= \int_0^1 [g'(t) - g'(0)]\,dt \\ &= \int_0^1 \int_0^t g''(u)\,du\,dt \\ &= \int_0^1 (1-u)g''(u)\,du.\end{aligned}$$

Since $\int_0^1 (1-u)\,du = 1/2$, subtracting the quadratic term yields

$$\begin{aligned}&F(Q) - F(\widehat{Q}) - \nabla F(\widehat{Q})^\top\Delta - \frac{1}{2}\Delta^\top H(\widehat{Q})\Delta \\ &\quad= \int_0^1 (1-u)\Delta^\top\big[H(\widehat{Q} + u\Delta) - H(\widehat{Q})\big]\Delta\,du. \qquad (304)\end{aligned}$$

We continue to bound the cubic Taylor remainder.

For symmetric M, $|x^\top M x| \le \|M\|_{\mathrm{op}}\|x\|_2^2$; moreover, (95) implies $\|H(\widehat{Q} + u\Delta) - H(\widehat{Q})\|_{\mathrm{op}} \le 3u\|\Delta\|_2/(2T^2)$. Taking absolute values in (304) therefore gives

$$\begin{aligned}&\left|F(Q) - F(\widehat{Q}) - \nabla F(\widehat{Q})^\top\Delta - \frac{1}{2}\Delta^\top H(\widehat{Q})\Delta\right| \\ &\quad\le \int_0^1 (1-u)\|\Delta\|_2^2\|H(\widehat{Q} + u\Delta) - H(\widehat{Q})\|_{\mathrm{op}}\,du \\ &\quad\le \frac{3\|\Delta\|_2^3}{2T^2}\int_0^1 u(1-u)\,du = \frac{\|\Delta\|_2^3}{4T^2},\end{aligned}$$

where $\int_0^1 u(1-u)\,du = 1/2 - 1/3 = 1/6$. Substituting $T = \tau\lambda$ gives

$$\left|F_s^{\mathrm{LOT}}(Q) - F_s^{\mathrm{LOT}}(\widehat{Q}) - \langle\nabla F_s^{\mathrm{LOT}}(\widehat{Q}),\,\Delta\rangle\;[-1mm] - \frac{1}{2}\Delta^\top\nabla_Q^2 F_s^{\mathrm{LOT}}(\widehat{Q})\Delta\right| \le \frac{\|\Delta\|_2^3}{4\tau^2\lambda^2}.$$

which is (96).

We then express the Hessian quadratic form as a variance.

For $p := p_j(\widehat{Q})$, direct expansion gives

$$\begin{aligned}\Delta^\top \Sigma(p)\Delta &= \Delta^\top \operatorname{diag}(p)\Delta - \Delta^\top pp^\top \Delta \\ &= \sum_i p_i \Delta_i^2 - \left(\sum_i p_i \Delta_i\right)^2 \\ &= \operatorname{Var}_{A\sim p}(\Delta_A).\end{aligned}$$

The covariance-mixture Hessian therefore satisfies

$$\begin{aligned}\Delta^\top H(\widehat{Q})\Delta &= \frac{1}{T}\sum_j \mu_s(j)\operatorname{Var}_{A\sim p_j(\widehat{Q})}(\Delta_A) \\ &= \frac{1}{T}\mathbb{E}_{J\sim\mu_s}\left[\operatorname{Var}_{A\sim p^{(J)}(\widehat{Q})}(\Delta_A)\right].\end{aligned}$$

We now convert the variance into a paired-action difference.

Conditionally on $J = j$, draw $A, A' \overset{\text{i.i.d.}}{\sim} p_j(\widehat{Q})$. Conditional independence and identical distribution give, with $\mathbb{E}_j[\cdot] := \mathbb{E}[\cdot \mid J = j]$,

$$\begin{aligned}\mathbb{E}_j[(\Delta_A - \Delta_{A'})^2] &= \mathbb{E}_j[\Delta_A^2] + \mathbb{E}_j[\Delta_{A'}^2] \\ &\quad - 2\mathbb{E}_j[\Delta_A\Delta_{A'}] \\ &= 2\left(\mathbb{E}_j[\Delta_A^2] - \mathbb{E}_j[\Delta_A]^2\right) \\ &= 2\operatorname{Var}_{p_j(\widehat{Q})}(\Delta_A),\end{aligned}$$

where $\mathbb{E}_j[\Delta_A\Delta_{A'}] = \mathbb{E}_j[\Delta_A]^2$. Averaging over J and using $T = \tau\lambda$ yields

$$\begin{aligned}\Delta^\top \nabla_Q^2 F_s^{\mathrm{LOT}}(\widehat{Q})\Delta &= \frac{1}{\tau\lambda}\mathbb{E}\left[\operatorname{Var}_{p^{(J)}(\widehat{Q})}(\Delta_A)\right] \\ &= \frac{1}{2\tau\lambda}\mathbb{E}[(\Delta_A - \Delta_{A'})^2],\end{aligned}$$

which proves (97).

Finally, we prove unbiasedness of the cross-product estimator.

Let $\mathcal{G} := \sigma(J, A, A')$, augmented if necessary by base information with respect to which $Q, \widehat{Q}$, and Δ are fixed. Conditional on $\mathcal{G}$, suppose that $\widetilde{\Delta}^{(1)}$ and $\widetilde{\Delta}^{(2)}$ are independent replicates such that, for $r \in \{1, 2\}$,

$$\mathbb{E}[\widetilde{\Delta}_A^{(r)} \mid \mathcal{G}] = \Delta_A, \qquad \mathbb{E}[\widetilde{\Delta}_{A'}^{(r)} \mid \mathcal{G}] = \Delta_{A'},$$

and assume the following products are integrable. Define $X_r := \widetilde{\Delta}_A^{(r)} - \widetilde{\Delta}_{A'}^{(r)}$. Then $\mathbb{E}[X_r \mid \mathcal{G}] = \Delta_A - \Delta_{A'}$, while conditional independence of the two replicates implies conditional independence of X_1, X_2; hence

$$\mathbb{E}[X_1X_2 \mid \mathcal{G}] = \mathbb{E}[X_1 \mid \mathcal{G}]\mathbb{E}[X_2 \mid \mathcal{G}] = (\Delta_A - \Delta_{A'})^2.$$

The tower property shows $\mathbb{E}[X_1X_2] = \mathbb{E}[(\Delta_A - \Delta_{A'})^2]$; hence the estimator in (98) satisfies

$$\begin{aligned}\mathbb{E}[\widehat{q}_2] &= \frac{1}{4\tau\lambda}\mathbb{E}[X_1X_2] \\ &= \frac{1}{4\tau\lambda}\mathbb{E}[(\Delta_A - \Delta_{A'})^2] \\ &= \frac{1}{2}\Delta^\top \nabla_Q^2 F_s^{\mathrm{LOT}}(\widehat{Q})\Delta.\end{aligned}$$

The last equality uses (97); thus $\widehat{q}_2$ is unbiased for the quadratic Taylor correction. No independence between $\widetilde{\Delta}_A^{(r)}$ and $\widetilde{\Delta}_{A'}^{(r)}$ within one replicate is needed; only the two complete replicates must be conditionally independent given $\mathcal{G}$. This is why a cross-product of two independent estimates is used: squaring one noisy estimate would generally add a variance term.

Combining the preceding arguments proves every assertion of Theorem 13. □

G.4.4 Proof of Lemma 3

Fix (s, j) and abbreviate $w_i := w_i^{(s,j)}$, $b := b_{s,j}$, and $T := T_0 > 0$. By construction, w is a probability vector: $w_i \geq 0$ and $\sum_i w_i = 1$. Full support is unnecessary; if $w_i = 0$, the corresponding Gibbs probability below is also zero. Let $\Delta := Q - \widehat{Q}$ and

$$\mathcal{R}_{s,j}(Q, \widehat{Q}) := f_{s,j}(Q) - f_{s,j}(\widehat{Q}) - \nabla f_{s,j}(\widehat{Q})^\top \Delta - \frac{1}{2}\Delta^\top \nabla^2 f_{s,j}(\widehat{Q})\Delta.$$

We first reduce the multivariate remainder to a scalar path.

By (99),

$$f_{s,j}(q) = b + T \log\left(\sum_i w_i e^{q_i/T}\right).$$

The q-independent offset b has zero derivatives and cancels from the remainder. Set $Q_u := \widehat{Q} + u\Delta$ for $u \in [0, 1]$ and $g(u) := f_{s,j}(Q_u)$. The chain rule gives

$$g'(u) = \nabla f_{s,j}(Q_u)^\top \Delta,$$
$$g''(u) = \Delta^\top \nabla^2 f_{s,j}(Q_u)\Delta.$$

Thus $g'(0) = \nabla f_{s,j}(\widehat{Q})^\top \Delta$, $g''(0) = \Delta^\top \nabla^2 f_{s,j}(\widehat{Q})\Delta$, and, because $Q_0 = \widehat{Q}$ and $Q_1 = Q$,

$$\mathcal{R}_{s,j}(Q, \widehat{Q}) = g(1) - g(0) - g'(0) - \frac{1}{2}g''(0).$$

It therefore suffices to bound g''' uniformly on $[0, 1]$.

Next, we locate the segment inside the value envelope.

Let $m_I := \inf I$, $M_I := \sup I$, and $S_I := M_I - m_I < \infty$. Since $Q_i, \widehat{Q}_i \in I$, $(Q_u)_i = (1-u)\widehat{Q}_i + uQ_i \in [m_I, M_I]$. This remains true when I is not an interval: Q_u lies in $\mathrm{conv}(I)$, which is contained in $[m_I, M_I]$ and has the same span. Hence

$$m_I \leq (Q_u)_i \leq M_I \qquad \text{for all } i \text{ and } u \in [0, 1]. \tag{305}$$

We next compare the Gibbs law with the base weights.

Along the segment, define

$$Z(u) := \sum_k w_k e^{(Q_u)_k/T}, \qquad p_u(i) := \frac{w_i e^{(Q_u)_i/T}}{Z(u)}.$$

Some w_k is positive, so positivity of the exponential gives $Z(u) > 0$. By (305),

$$e^{(Q_u)_i/T} \leq e^{M_I/T},$$
$$Z(u) \geq \sum_k w_k e^{m_I/T} = e^{m_I/T}.$$

Consequently,

$$\begin{aligned} p_u(i) &\leq \frac{w_i e^{M_I/T}}{e^{m_I/T}} = e^{(M_I - m_I)/T} w_i \\ &= e^{S_I/T} w_i \qquad (i,\ u \in [0, 1]). \end{aligned} \tag{306}$$

If $w_i = 0$, both sides vanish; no positive lower bound on individual weights is needed.

We continue to differentiate the Gibbs probabilities.

Since $(Q_u)_i = \widehat{Q}_i + u\Delta_i$,

$$\frac{d}{du} e^{(Q_u)_i/T} = \frac{\Delta_i}{T} e^{(Q_u)_i/T},$$
$$Z'(u) = \frac{1}{T}\sum_i w_i e^{(Q_u)_i/T}\Delta_i$$
$$= \frac{Z(u)}{T}\sum_i p_u(i)\Delta_i.$$

Define $m_u := \mathbb{E}_{A\sim p_u}[\Delta_A] = \sum_i p_u(i)\Delta_i$; then $Z'(u)/Z(u) = m_u/T$, and the quotient rule gives

$$\begin{aligned} p_u'(i) &= \frac{\Delta_i}{T} p_u(i) - p_u(i)\frac{Z'(u)}{Z(u)} \\ &= \frac{1}{T} p_u(i)(\Delta_i - m_u). \end{aligned}$$

Therefore, for every fixed function h on the finite action set,

$$\begin{aligned} \frac{d}{du}\mathbb{E}_{p_u}[h(\Delta_A)] &= \sum_i p_u'(i) h(\Delta_i) \\ &= \frac{1}{T}\mathbb{E}_{p_u}\left[(\Delta_A - m_u)h(\Delta_A)\right] \\ &= \frac{1}{T}\operatorname{Cov}_{p_u}\big(h(\Delta_A), \Delta_A\big). \end{aligned} \tag{307}$$

We then compute the first and second derivatives.

Because $g(u) = b + T\log Z(u)$, the preceding identities and (307) with $h(x) = x$ give

$$\begin{aligned} g'(u) &= T\frac{Z'(u)}{Z(u)} = m_u = \mathbb{E}_{p_u}[\Delta_A], \\ g''(u) &= m_u' = \frac{1}{T}\mathbb{E}_{p_u}[(\Delta_A - m_u)\Delta_A] \\ &= \frac{1}{T}\big(\mathbb{E}_{p_u}[\Delta_A^2] - m_u^2\big) = \frac{1}{T}\operatorname{Var}_{p_u}(\Delta_A). \end{aligned}$$

We now compute the third derivative.

Put $r_u := \mathbb{E}_{p_u}[\Delta_A^2]$. Applying (307) with $h(x) = x^2$ yields

$$r_u' = \frac{1}{T}\mathbb{E}_{p_u}[(\Delta_A - m_u)\Delta_A^2] = \frac{1}{T}\big(\mathbb{E}_{p_u}[\Delta_A^3] - m_u r_u\big).$$

Since $g''(u) = T^{-1}(r_u - m_u^2)$ and $m_u' = T^{-1}(r_u - m_u^2)$,

$$\begin{aligned} g'''(u) &= \frac{1}{T}(r_u' - 2m_u m_u') \\ &= \frac{1}{T^2}\left(\mathbb{E}_{p_u}[\Delta_A^3] - 3m_u\mathbb{E}_{p_u}[\Delta_A^2] + 2m_u^3\right). \end{aligned}$$

Moreover, expanding the cube and using $\mathbb{E}_{p_u}[\Delta_A] = m_u$ gives

$$\begin{aligned} \mathbb{E}_{p_u}[(\Delta_A - m_u)^3] &= \mathbb{E}_{p_u}[\Delta_A^3] - 3m_u\mathbb{E}_{p_u}[\Delta_A^2] + 3m_u^2\mathbb{E}_{p_u}[\Delta_A] - m_u^3 \\ &= \mathbb{E}_{p_u}[\Delta_A^3] - 3m_u\mathbb{E}_{p_u}[\Delta_A^2] + 2m_u^3. \end{aligned}$$

Hence

$$g'''(u) = \frac{1}{T^2}\mathbb{E}_{A\sim p_u}\left[(\Delta_A - \mathbb{E}_{p_u}\Delta_A)^3\right]. \tag{308}$$

We next bound the centered third moment.

For $a, b \geq 0$, convexity of x^3 gives $(a+b)^3 = 8((a+b)/2)^3 \leq 4(a^3 + b^3)$. Thus, whenever $\mathbb{E}|X|^3 < \infty$,

$$|X - \mathbb{E}X|^3 \leq 4\big(|X|^3 + |\mathbb{E}X|^3\big).$$

Taking expectations and using Jensen, $|\mathbb{E}X|^3 \leq (\mathbb{E}|X|)^3 \leq \mathbb{E}|X|^3$, gives

$$\mathbb{E}|X - \mathbb{E}X|^3 \leq 8\mathbb{E}|X|^3. \tag{309}$$

Since $|\mathbb{E}[(X - \mathbb{E}X)^3]| \leq \mathbb{E}|X - \mathbb{E}X|^3$, applying this to $X = \Delta_A$, $A \sim p_u$, and using (308) yields

$$|g'''(u)| \leq \frac{1}{T^2}\mathbb{E}_{p_u}|\Delta_A - m_u|^3 \leq \frac{8}{T^2}\sum_i p_u(i)|\Delta_i|^3.$$

We continue to transfer the moment bound to the base weights.

By (306), $\sum_i p_u(i)|\Delta_i|^3 \le e^{S_I/T} \sum_i w_i|\Delta_i|^3$. Hence, uniformly for $u \in [0,1]$,

$$|g'''(u)| \le \frac{8e^{S_I/T}}{T^2} \sum_i w_i|\Delta_i|^3. \tag{310}$$

We then derive the integral Taylor remainder.

Taylor's formula with integral remainder, equivalently the fundamental theorem of calculus applied three times, gives

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(0) + \frac{1}{2}\int_0^1 (1-u)^2 g'''(u)\,du,$$

$$g(1) - g(0) - g'(0) - \frac{1}{2}g''(0) = \int_0^1 \int_0^t \int_0^v g'''(u)\,du\,dv\,dt = \frac{1}{2}\int_0^1 (1-u)^2 g'''(u)\,du.$$

Consequently,

$$\mathcal{R}_{s,j}(Q, \widehat{Q}) = \frac{1}{2}\int_0^1 (1-u)^2 g'''(u)\,du. \tag{311}$$

Finally, we combine the bounds.

Put $M_3 := \sum_i w_i|\Delta_i|^3$. Using (310) in (311),

$$\begin{aligned} |\mathcal{R}_{s,j}(Q, \widehat{Q})| &\le \frac{1}{2}\int_0^1 (1-u)^2 |g'''(u)|\,du \\ &\le \frac{8e^{S_I/T}}{T^2} M_3 \frac{1}{2}\int_0^1 (1-u)^2\,du \\ &= \frac{4e^{S_I/T}}{3T^2} M_3, \end{aligned}$$

because $\frac{1}{2}\int_0^1 (1-u)^2\,du = \frac{1}{2}\cdot\frac{1}{3} = \frac{1}{6}$. Restoring $T = T_0$, $w_i = w_i^{(s,j)}$, and $\Delta_i = Q_i - \widehat{Q}_i$ gives

$$|\mathcal{R}_{s,j}(Q, \widehat{Q})| \le \frac{4e^{S_I/T_0}}{3T_0^2} \sum_i w_i^{(s,j)} |Q_i - \widehat{Q}_i|^3,$$

which is exactly (101). There is no additional factor depending on the total number of actions K. □

G.5 Fixed deterministic action pruning

G.5.1 Proof of Lemma 4

Fix (s,j) and write $I := I_{s,j}$, $w_i := w_i^{(s,j)}$, $b := b_{s,j}$, and $\alpha := \alpha_{s,j} = w^{(s,j)}(I^c)$. Since $w^{(s,j)} \in \Delta_{\mathcal{A}}$ and $\alpha < 1$,

$$w^{(s,j)}(I) = \sum_{i \in I} w_i = 1 - \alpha > 0.$$

Thus the retained set has positive total weight even if some retained weights vanish.

We first define the full and retained partition functions.

For $q \in \mathbb{R}^K$, define

$$Z(q) := \sum_i w_i e^{q_i/T_0}, \qquad Z_I(q) := \sum_{i \in I} w_i e^{q_i/T_0}.$$

All exponentials are positive and $w(I) > 0$, so $Z(q) \ge Z_I(q) > 0$. By (99) and (102),

$$f_{s,j}(q) = b + T_0 \log Z(q),$$
$$f_{s,j}^{\mathcal{I}}(q) = b + T_0 \log Z_I(q).$$

and the common offset b cancels. Define the nonnegative ratio

$$r_{s,j}(q) := \frac{Z(q) - Z_I(q)}{Z_I(q)} = \frac{\sum_{i \notin I} w_i e^{q_i/T_0}}{\sum_{i \in I} w_i e^{q_i/T_0}}. \tag{312}$$

Then $Z(q)/Z_I(q) = 1 + r_{s,j}(q)$.

Next, we bound the ratio uniformly over the value class.

Let $q_{\max} := \max_i q_i$ and $q_{\min} := \min_i q_i$. Then

$$\begin{aligned} Z(q) - Z_I(q) &= \sum_{i \notin I} w_i e^{q_i/T_0} \\ &\le e^{q_{\max}/T_0} \sum_{i \notin I} w_i \\ &= e^{q_{\max}/T_0} \alpha, \\ Z_I(q) &= \sum_{i \in I} w_i e^{q_i/T_0} \\ &\ge e^{q_{\min}/T_0} \sum_{i \in I} w_i \\ &= e^{q_{\min}/T_0} (1 - \alpha). \end{aligned}$$

Substitution into (312) gives

$$r_{s,j}(q) \le e^{(q_{\max} - q_{\min})/T_0} \frac{\alpha}{1 - \alpha} = e^{\operatorname{osc}(q)/T_0} \frac{\alpha}{1 - \alpha}.$$

Hence, if $\operatorname{osc}(q) \le S$,

$$0 \le r_{s,j}(q) \le e^{S/T_0} \frac{\alpha_{s,j}}{1 - \alpha_{s,j}} = \zeta_{s,j}(S). \tag{313}$$

This geometry-dependent bound is uniform over the stated value class. If $\alpha_{s,j} = 0$, the ratio and every pruning error below are exactly zero.

We now convert the ratio into the local value gap.

The ratio identity gives

$$f_{s,j}(q) - f^{\mathcal{I}}_{s,j}(q) = T_0 \log \frac{Z(q)}{Z_I(q)} = T_0 \log\bigl(1 + r_{s,j}(q)\bigr).$$

Since $r_{s,j}(q) \ge 0$ and $\log(1 + r)$ is increasing on $[0, \infty)$, (313) yields

$$0 \le f_{s,j}(q) - f^{\mathcal{I}}_{s,j}(q) \le T_0 \log\bigl(1 + \zeta_{s,j}(S)\bigr),$$

proving (104).

We next average the columnwise value bounds.

By definition,

$$\begin{aligned} \mathcal{F}_s(q) &= \sum_j \mu_s(j) f_{s,j}(q), \\ F^{\mathcal{I}}_s(q) &= \sum_j \mu_s(j) f^{\mathcal{I}}_{s,j}(q). \end{aligned}$$

Because $\mu_s(j) \geq 0$, averaging the nonnegative gaps gives

$$\begin{aligned} 0 &\leq \mathcal{F}_s(q) - F_s^{\mathcal{I}}(q) \\ &= \sum_j \mu_s(j)\big[f_{s,j}(q) - f_{s,j}^{\mathcal{I}}(q)\big] \\ &\leq T_0 \sum_j \mu_s(j) \log\big(1 + \zeta_{s,j}(S)\big) \\ &\leq T_0 \sup_{s'} \sum_j \mu_{s'}(j) \log\big(1 + \zeta_{s',j}(S)\big) \\ &= \Delta_{\mathcal{I}}(S). \end{aligned}$$

This is the state-aggregator bound associated with (105).

We continue to identify the retained policy as a conditional law.

The full and zero-extended retained policies are

$$p_i^{(s,j)}(q) = \frac{w_i e^{q_i/T_0}}{Z(q)},$$
$$p_i^{\mathcal{I},(s,j)}(q) = \frac{\mathbf{1}\{i \in I\} w_i e^{q_i/T_0}}{Z_I(q)}.$$

Moreover,

$$\begin{aligned} p^{(s,j)}(q)(I) &= \sum_{i \in I} \frac{w_i e^{q_i/T_0}}{Z(q)} = \frac{Z_I(q)}{Z(q)} \\ &= \frac{1}{1 + r_{s,j}(q)} > 0. \end{aligned}$$

Thus, for $i \in I$, $p_i^{\mathcal{I},(s,j)}(q) = p_i^{(s,j)}(q)/p^{(s,j)}(q)(I)$, while the retained policy is zero for $i \notin I$. Therefore

$$p^{\mathcal{I},(s,j)}(q) = p^{(s,j)}(q)(\,\cdot \mid I). \tag{314}$$

The displayed positivity makes this conditional law well defined.

We then derive the exact omitted-mass and policy-distance identities.

Let $m(q) := p^{(s,j)}(q)(I^c)$. Then

$$m(q) = 1 - \frac{1}{1 + r_{s,j}(q)} = \frac{r_{s,j}(q)}{1 + r_{s,j}(q)} \leq \frac{\zeta_{s,j}(S)}{1 + \zeta_{s,j}(S)},$$

where the inequality uses monotonicity of $r/(1+r)$ on $[0, \infty)$. On I^c, the retained policy is zero, so

$$\sum_{i \notin I} \Big| p_i^{(s,j)}(q) - p_i^{\mathcal{I},(s,j)}(q) \Big| = m(q).$$

On I, conditioning increases every coordinate:

$$p_i^{\mathcal{I},(s,j)}(q) = \frac{p_i^{(s,j)}(q)}{p^{(s,j)}(q)(I)} \geq p_i^{(s,j)}(q),$$

and hence

$$\sum_{i \in I} \Big| p_i^{(s,j)}(q) - p_i^{\mathcal{I},(s,j)}(q) \Big| = 1 - p^{(s,j)}(q)(I) = m(q).$$

Adding the two contributions,

$$\begin{aligned} \Big\| p^{(s,j)}(q) - p^{\mathcal{I},(s,j)}(q) \Big\|_1 &= 2m(q) = \frac{2r_{s,j}(q)}{1 + r_{s,j}(q)} \\ &\leq \frac{2\zeta_{s,j}(S)}{1 + \zeta_{s,j}(S)}. \end{aligned}$$

Under $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$, this is equivalently

$$\mathrm{TV}\left(p^{(s,j)}(q), p^{\mathcal{I},(s,j)}(q)\right) = p^{(s,j)}(q)(I^c) \leq \frac{\zeta_{s,j}(S)}{1+\zeta_{s,j}(S)}.$$

We now specialize the partition ratio to ordinary LOT.

By (100),

$$w_i^{(s,j)} = \frac{\rho_s(i)e^{-C_{s,ij}/\lambda}}{G_{s,j}},$$
$$G_{s,j} := \sum_k \rho_s(k)e^{-C_{s,kj}/\lambda}.$$

Fix $\zeta > 0$ and take the closed cost ball $I = I_{s,j}(R) := \{i : C_{s,ij} \leq R\}$. Assume $C_{s,jj} = 0$ and $R \geq 0$. Then $j \in I$, so the ball is nonempty and contains positive prior mass. Introduce

$$\widehat{Z}(q) := \sum_i \rho_s(i)e^{(q_i - \tau C_{s,ij})/T_0},$$
$$\widehat{Z}_I(q) := \sum_{i\in I} \rho_s(i)e^{(q_i - \tau C_{s,ij})/T_0}.$$

Since $T_0 = \tau\lambda$, $e^{(q_i-\tau C_{s,ij})/T_0} = e^{q_i/T_0}e^{-C_{s,ij}/\lambda}$, and therefore

$$Z(q) = \frac{\widehat{Z}(q)}{G_{s,j}}, \qquad Z_I(q) = \frac{\widehat{Z}_I(q)}{G_{s,j}}.$$

The common normalizer cancels:

$$r_{s,j}(q) = \frac{\widehat{Z}(q) - \widehat{Z}_I(q)}{\widehat{Z}_I(q)}. \tag{315}$$

We next lower-bound the retained raw partition.

Because $j \in I$ and $C_{s,jj} = 0$, the anchor term gives

$$\begin{aligned}\widehat{Z}_I(q) &\geq \rho_s(j)e^{(q_j - \tau C_{s,jj})/T_0} \\ &= \rho_s(j)e^{q_j/T_0} \geq \rho_s(j)e^{q_{\min}/T_0} \\ &\geq \rho_{\min}e^{q_{\min}/T_0}.\end{aligned}$$

We continue to upper-bound the omitted raw partition.

For $i \notin I$, closedness of the cost ball gives $C_{s,ij} > R$, hence $e^{-C_{s,ij}/\lambda} < e^{-R/\lambda}$. Using a weak inequality,

$$\begin{aligned}\widehat{Z}(q) - \widehat{Z}_I(q) &= \sum_{i\notin I} \rho_s(i)e^{q_i/T_0}e^{-C_{s,ij}/\lambda} \\ &\leq e^{q_{\max}/T_0 - R/\lambda}\sum_{i\notin I}\rho_s(i) \\ &\leq e^{q_{\max}/T_0 - R/\lambda}.\end{aligned}$$

Combining this with (315),

$$\begin{aligned}r_{s,j}(q) &\leq \frac{e^{q_{\max}/T_0 - R/\lambda}}{\rho_s(j)e^{q_{\min}/T_0}} \\ &= \rho_s(j)^{-1}e^{(q_{\max}-q_{\min})/T_0 - R/\lambda} \\ &\leq \rho_{\min}^{-1}e^{S/T_0 - R/\lambda},\end{aligned}$$

using $\mathrm{osc}(q) \le S$ and $\rho_s(j) \ge \rho_{\min}$.

Finally, we solve for a sufficient radius.

It suffices for $r_{s,j}(q) \le \zeta$ that

$$\rho_{\min}^{-1} e^{S/T_0 - R/\lambda} \le \zeta.$$

Taking logarithms and using $T_0 = \tau\lambda$,

$$\frac{R}{\lambda} \ge \frac{S}{T_0} + \log \frac{1}{\rho_{\min}\zeta}, \qquad \lambda \frac{S}{T_0} = \frac{S}{\tau},$$

so it is enough that

$$R \ge \frac{S}{\tau} + \lambda \log \frac{1}{\rho_{\min}\zeta}.$$

Combining this with $R \ge 0$, which retains the zero-cost anchor, gives

$$R \ge R(S,\zeta) := \max\left\{0, \frac{S}{\tau} + \lambda \log \frac{1}{\rho_{\min}\zeta}\right\}. \tag{316}$$

Equality is allowed: because the ball is closed, every omitted action still satisfies $C_{s,ij} > R$, so no slack is needed. Under (316), $r_{s,j}(q) \le \zeta$ uniformly for all q with $\mathrm{osc}(q) \le S$.

If the prior is uniform, $\rho_s(i) = 1/K$ and $\rho_{\min} = 1/K$, reducing the condition to

$$R \ge \max\left\{0, \frac{S}{\tau} + \lambda \log \frac{K}{\zeta}\right\}.$$

This proves the local-value, state-aggregator, policy-mass, total-variation, and ordinary-LOT cost-radius claims. □

G.5.2 Proof of Lemma 2

We proceed in three parts: derive the exact worst-case column and policy errors, invert the formula to obtain the smallest deterministic support, and construct a one-state MDP attaining the Bellman amplification.

We first define the retained and omitted partition functions.

Fix $T_0 > 0$, $S \ge 0$, $w \in \Delta_{\mathcal{A}}$, and a nonempty $I \subseteq \mathcal{A}$ with $\alpha := w(I^c) < 1$. Then $w(I) = 1-\alpha > 0$, even if some retained weights vanish. For vector $Q \in \mathbb{R}^K$, set

$$A(Q) := \sum_{i \in I} w_i e^{Q_i/T_0}, \qquad B(Q) := \sum_{i \notin I} w_i e^{Q_i/T_0}.$$

Here Q_i is the i^{th} action.

Thus $A(Q) > 0$, $B(Q) \ge 0$, and the full partition is $A(Q) + B(Q)$. The common offset in f and f^I cancels, so

$$f(Q) - f^I(Q) = T_0 \log\left(1 + \frac{B(Q)}{A(Q)}\right). \tag{317}$$

Crucially, (102) uses the raw retained sum $A(Q)$, not $A(Q)/(1-\alpha)$; only the restricted policy is normalized.

Next, we upper-bound the ratio over the bounded-span class.

Let $m := \min_i Q_i$ and $M := \max_i Q_i$. If $\mathrm{osc}(Q) = M - m \le S$, then $m \le Q_i \le m + S$, and

$$A(Q) \ge e^{m/T_0} \sum_{i \in I} w_i = (1-\alpha) e^{m/T_0},$$

$$B(Q) \le e^{(m+S)/T_0} \sum_{i \notin I} w_i = \alpha e^{(m+S)/T_0}.$$

Therefore

$$0 \le \frac{B(Q)}{A(Q)} \le e^{S/T_0} \frac{\alpha}{1-\alpha}. \tag{318}$$

We then show that the upper bound is attained.

For arbitrary $m \in \mathbb{R}$, define

$$Q_i^{\mathrm{worst}} := \begin{cases} m, & i \in I, \\ m + S, & i \notin I. \end{cases} \tag{319}$$

Its oscillation is S when $I, I^c \neq \varnothing$ and $S > 0$, and is still at most S if $S = 0$ or $I^c = \varnothing$. Moreover,

$$A(Q^{\mathrm{worst}}) = (1-\alpha)e^{m/T_0},$$
$$B(Q^{\mathrm{worst}}) = \alpha e^{(m+S)/T_0}.$$

so

$$\frac{B(Q^{\mathrm{worst}})}{A(Q^{\mathrm{worst}})} = e^{S/T_0}\frac{\alpha}{1-\alpha}. \tag{320}$$

This also holds for $\alpha = 0$, including when I^c contains only zero-weight actions. Since $T_0 \log(1+x)$ is increasing, (317)–(320) imply

$$\sup_{\mathrm{osc}(Q)\le S}\{f(Q) - f^I(Q)\} = T_0 \log\left(1 + e^{S/T_0}\frac{\alpha}{1-\alpha}\right),$$

The supremum is a maximum, attained by (319); this proves (17). In particular, for $S = 0$ the exact gap is $T_0 \log(1/(1-\alpha))$, which need not vanish because the retained partition is unnormalized.

We now derive the exact omitted policy mass.

The full local softmax probability at action the i^{th} is $p_i(Q) = w_i e^{Q_i/T_0}/(A(Q) + B(Q))$, hence the total probability of omitted actions

$$p(Q)(I^c) = \frac{B(Q)}{A(Q)+B(Q)} = \frac{B(Q)/A(Q)}{1 + B(Q)/A(Q)}. \tag{321}$$

Since $x/(1+x)$ increases on $[0,\infty)$, the same two-level vector is worst. Using (321) and (320) gives

$$\sup_{\mathrm{osc}(Q)\le S} p(Q)(I^c) = \frac{e^{S/T_0}\alpha}{1-\alpha+e^{S/T_0}\alpha}. \tag{322}$$

We next derive the exact policy-distance formula.

The zero-extended restricted policy is

$$p_i^I(Q) = \frac{\mathbf{1}\{i \in I\} w_i e^{Q_i/T_0}}{A(Q)}.$$

Because $p(Q)(I) = A(Q)/(A(Q)+B(Q)) > 0$ and $p_i^I(Q) = p_i(Q)/p(Q)(I)$ on I, one has $p^I(Q) = p(Q)(\,\cdot \mid I)$. Conditioning gives the exact identity

$$\left\|p(Q) - p^I(Q)\right\|_1 = 2p(Q)(I^c) = \frac{2B(Q)/A(Q)}{1 + B(Q)/A(Q)}. \tag{323}$$

Indeed, I^c contributes $p(Q)(I^c)$ because p^I vanishes there, and conditioning redistributes exactly the same mass on I. Combining (322) and (323) therefore gives

$$\sup_{\mathrm{osc}(Q)\le S}\left\|p(Q) - p^I(Q)\right\|_1 = \frac{2e^{S/T_0}\alpha}{1-\alpha+e^{S/T_0}\alpha}.$$

Under $\mathrm{TV}(p, p^I) = \frac{1}{2}\left\|p - p^I\right\|_1$, the sharp error is

$$\sup_{\mathrm{osc}(Q)\le S}\mathrm{TV}(p(Q), p^I(Q)) = \frac{e^{S/T_0}\alpha}{1-\alpha+e^{S/T_0}\alpha}.$$

We continue to invert the exact value-error formula.

Fix $\xi > 0$ and write $z := e^{\xi/T_0} - 1 > 0$ and $a := e^{S/T_0} \geq 1$. By (17), the uniform condition $f(Q) - f^I(Q) \leq \xi$ for all $\mathrm{osc}(Q) \leq S$ is equivalent to

$$T_0 \log\left(1 + a\frac{\alpha}{1-\alpha}\right) \leq \xi \quad \Longleftrightarrow \quad a\frac{\alpha}{1-\alpha} \leq z.$$

Here we used monotonicity of the logarithm and exponential. Since $1 - \alpha > 0$,

$$a\alpha \leq z(1-\alpha) \Longleftrightarrow (a+z)\alpha \leq z \Longleftrightarrow \alpha \leq \frac{z}{a+z}.$$

Substitution gives

$$\frac{z}{a+z} = \frac{e^{\xi/T_0} - 1}{e^{S/T_0} + e^{\xi/T_0} - 1} = \alpha^\star(S, \xi).$$

Because $a, z > 0$, $0 < \alpha^\star(S, \xi) < 1$, within the domain of Definition 1. Consequently,

$$\begin{aligned} & f(Q) - f^I(Q) \leq \xi \quad \forall Q : \mathrm{osc}(Q) \leq S \\ \Longleftrightarrow \quad & w(I^c) = \alpha \leq \alpha^\star(S, \xi) \\ \Longleftrightarrow \quad & w(I) \geq 1 - \alpha^\star(S, \xi), \end{aligned}$$

which proves (19).

We then characterize the smallest set for one column.

For $\beta \in (0, 1)$, define

$$m_w(\beta) := \min\{|J| : J \subseteq \mathcal{A},\ w(J) \geq 1 - \beta\}.$$

Finiteness of the action set ensures attainment. Order actions so that $w_{(1)} \geq \cdots \geq w_{(K)}$, breaking ties by a fixed deterministic order, and let

$$\begin{aligned} k_\beta &:= \min\left\{k : \sum_{\ell=1}^{k} w_{(\ell)} \geq 1 - \beta\right\}, \\ J_\beta &:= \{(1), \ldots, (k_\beta)\}. \end{aligned}$$

The prefix has sufficient mass, and no smaller set can: among all k-element subsets, the k largest weights maximize total mass. Therefore $m_w(\beta) = k_\beta$, attained by a top-weight prefix.

We now prove the exact worst-anchor fixed-set complexity.

Set $\beta := \alpha^\star(S, \xi)$ and

$$m_{s,j}(\beta) := \min\{|I| : I \subseteq \mathcal{A},\ w^{(s,j)}(I) \geq 1 - \beta\}.$$

The inverted value-error formula forces every admissible family $\mathcal{I} = \{I_{s,j}\}$ in $d_{\mathrm{fixed}}(S, \xi)$ to satisfy $|I_{s,j}| \geq m_{s,j}(\beta)$, whence

$$d_{\mathrm{fixed}}(S, \xi) \geq \sup_{s,j} m_{s,j}(\beta). \tag{324}$$

Conversely, fix a global deterministic tie-breaking order and choose for each (s, j) the shortest decreasing-weight prefix $I^\star_{s,j}$ with mass at least $1 - \beta$. It depends only on the fixed geometry, not on q; the one-column optimality result gives $|I^\star_{s,j}| = m_{s,j}(\beta)$, and the inverted value-error formula gives uniform error at most ξ. Thus

$$d_{\mathrm{fixed}}(S, \xi) \leq \sup_{s,j} m_{s,j}(\beta). \tag{325}$$

Combining,

$$d_{\mathrm{fixed}}(S, \xi) = \sup_{s,j} m_{s,j}(\beta) = d_{\mathrm{ker}}(\beta) = d_{\mathrm{ker}}(\alpha^\star(S, \xi)),$$

which proves (20). This equality is exact only within the stated fixed-set class: deterministic, Q-independent neighborhoods with the same uniform per-column tolerance. It is not a support lower bound for directly averaged anchor errors, Q-adaptive neighborhoods, randomized action sampling, or multilevel Monte Carlo.

We next impose the cascade mass condition.

The cascade condition $w^{(s,j)}(I_{s,j}) \geq 1/2$ is equivalent to tail mass at most $1/2$, while the exact value-error condition requires tail mass at most $\alpha^\star(S,\xi)$. Both therefore hold exactly when

$$w^{(s,j)}(I_{s,j}^c) \leq \min\left\{\frac{1}{2}, \alpha^\star(S,\xi)\right\}.$$

Repeating the preceding worst-anchor argument with this smaller tail budget gives the exact support complexity with $\alpha^\star(S,\xi)$ replaced by the displayed minimum. For $S = S_Q$ and $\xi := (1-\gamma)\varepsilon/2$, let

$$z_\varepsilon := e^{\xi/T_0} - 1 = \exp\left(\frac{(1-\gamma)\varepsilon}{2T_0}\right) - 1.$$

Then

$$\min\left\{\frac{1}{2}, \alpha^\star(S_Q,\xi)\right\} = \min\left\{\frac{1}{2}, \frac{z_\varepsilon}{e^{S_Q/T_0} + z_\varepsilon}\right\} = \alpha_\varepsilon,$$

as defined in (2).

We continue to construct an MDP attaining Bellman amplification.

Take a one-state MDP with actions $\mathcal{A}$ and deterministic self-loops $P(s \mid s,i) = 1$. Define

$$r_i := S\mathbf{1}\{i \notin I\}, \tag{326}$$

so $r_i = 0$ on I, $r_i = S$ on I^c, and $\operatorname{osc}(\mathbf{r}) \leq S$. Give every reference column the same w, I, b. Since $\sum_j \mu_s(j) = 1$, the mixtures reduce to $\mathcal{F}(Q) = f(Q)$ and $F^{\mathcal{I}}(Q) = f^I(Q)$.

Both maps are translation equivariant:

$$\mathcal{F}(Q + c\mathbf{1}) = \mathcal{F}(Q) + c, \qquad F^{\mathcal{I}}(Q + c\mathbf{1}) = F^{\mathcal{I}}(Q) + c. \tag{327}$$

Indeed, multiplying every partition term by e^{c/T_0} adds c to its scaled log partition.

Finally, we solve the two one-state Bellman fixed points.

Let $V_{\mathcal{F}}^\star, V_{\mathcal{I}}^\star$ be the scalar values. Self-loops imply

$$Q_{\mathcal{F}}^\star = \mathbf{r} + \gamma V_{\mathcal{F}}^\star \mathbf{1}, \qquad Q_{\mathcal{I}}^\star = \mathbf{r} + \gamma V_{\mathcal{I}}^\star \mathbf{1}.$$

Translation equivariance gives

$$V_{\mathcal{F}}^\star = \mathcal{F}(\mathbf{r}) + \gamma V_{\mathcal{F}}^\star,$$
$$V_{\mathcal{I}}^\star = F^{\mathcal{I}}(\mathbf{r}) + \gamma V_{\mathcal{I}}^\star,$$

and hence

$$V_{\mathcal{F}}^\star = \frac{\mathcal{F}(\mathbf{r})}{1-\gamma}, \qquad V_{\mathcal{I}}^\star = \frac{F^{\mathcal{I}}(\mathbf{r})}{1-\gamma}.$$

Therefore

$$V_{\mathcal{F}}^\star - V_{\mathcal{I}}^\star = \frac{\mathcal{F}(\mathbf{r}) - F^{\mathcal{I}}(\mathbf{r})}{1-\gamma}. \tag{328}$$

Every cloned column sees the maximizing vector (319) with $m = 0$, so

$$\mathcal{F}(\mathbf{r}) - F^{\mathcal{I}}(\mathbf{r}) = T_0 \log\left(1 + e^{S/T_0}\frac{\alpha}{1-\alpha}\right),$$

and substitution into (328) yields

$$V_{\mathcal{F}}^\star - V_{\mathcal{I}}^\star = \frac{T_0}{1-\gamma}\log\left(1 + e^{S/T_0}\frac{\alpha}{1-\alpha}\right).$$

Because both fixed points use the same rewards,

$$Q_{\mathcal{F}}^\star - Q_{\mathcal{I}}^\star = \gamma\left(V_{\mathcal{F}}^\star - V_{\mathcal{I}}^\star\right)\mathbf{1} = \frac{\gamma T_0}{1-\gamma}\log\left(1 + e^{S/T_0}\frac{\alpha}{1-\alpha}\right)\mathbf{1}.$$

Finally,

$$\operatorname{osc}(Q_{\mathcal{F}}^\star) = \operatorname{osc}(Q_{\mathcal{I}}^\star) = \operatorname{osc}(\mathbf{r}) \leq S,$$

because continuation is action-independent. The construction therefore stays in the same span class and exactly attains the value and action-value amplification factors in the subsequent Bellman perturbation bound. This proves Lemma 2. □

G.5.3 Proof of Theorem 14

Define the full and restricted Bellman operators by

$$(\mathcal{T}_{\mathcal{F}}Q)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\mathcal{F}_{s'}(Q(s',\cdot))\right],$$
$$(\mathcal{T}^{\mathcal{I}}Q)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[F^{\mathcal{I}}_{s'}(Q(s',\cdot))\right].$$

They share rewards and transitions and differ only in the state aggregator. The retained sets are fixed independently of q and satisfy

$$w^{(s,j)}(I_{s,j}) > 0 \quad \Longleftrightarrow \quad \alpha_{s,j} := w^{(s,j)}(I^c_{s,j}) < 1.$$

This is essential when the kernel has zeros: nonempty $I_{s,j}$ alone does not ensure a positive restricted partition.

We first establish the properties of the restricted aggregator.

For fixed (s,j),

$$f^{\mathcal{I}}_{s,j}(q) = b_{s,j} + T_0 \log\left(\sum_{i\in I_{s,j}} w_i^{(s,j)} e^{q_i/T_0}\right).$$

The retained weights sum to $w^{(s,j)}(I_{s,j}) > 0$, so the partition is positive and the local value finite for every finite q.

If $q \le \widehat{q}$ coordinatewise, then $w_i^{(s,j)} e^{q_i/T_0} \le w_i^{(s,j)} e^{\widehat{q}_i/T_0}$ on $I_{s,j}$; summation and the increasing logarithm show that $f^{\mathcal{I}}_{s,j}$ is monotone. For every scalar c,

$$\begin{aligned} f^{\mathcal{I}}_{s,j}(q + c\mathbf{1}) &= b_{s,j} + T_0 \log\left(e^{c/T_0}\sum_{i\in I_{s,j}} w_i^{(s,j)} e^{q_i/T_0}\right) \\ &= f^{\mathcal{I}}_{s,j}(q) + c, \end{aligned}$$

so it is translation equivariant. Finally, with $d := \|q - \widehat{q}\|_\infty$, $\widehat{q} - d\mathbf{1} \le q \le \widehat{q} + d\mathbf{1}$; monotonicity and translation equivariance give

$$f^{\mathcal{I}}_{s,j}(\widehat{q}) - d \le f^{\mathcal{I}}_{s,j}(q) \le f^{\mathcal{I}}_{s,j}(\widehat{q}) + d,$$

hence

$$\left|f^{\mathcal{I}}_{s,j}(q) - f^{\mathcal{I}}_{s,j}(\widehat{q})\right| \le \|q - \widehat{q}\|_\infty. \tag{329}$$

Since $F^{\mathcal{I}}_s = \sum_j \mu_s(j) f^{\mathcal{I}}_{s,j}$ is a convex combination, it is also monotone, translation equivariant, and

$$\left|F^{\mathcal{I}}_s(q) - F^{\mathcal{I}}_s(\widehat{q})\right| \le \|q - \widehat{q}\|_\infty. \tag{330}$$

The full aggregator $\mathcal{F}_s$ has the same three properties.

Next, we prove Bellman monotonicity and contraction.

If $Q \le \widehat{Q}$, monotonicity gives $F^{\mathcal{I}}_{s'}(Q(s',\cdot)) \le F^{\mathcal{I}}_{s'}(\widehat{Q}(s',\cdot))$ for every s'; taking conditional expectations and adding the common reward proves $\mathcal{T}^{\mathcal{I}}Q \le \mathcal{T}^{\mathcal{I}}\widehat{Q}$. For bounded $Q, \widehat{Q}$,

$$\begin{aligned} &|(\mathcal{T}^{\mathcal{I}}Q)(s,a) - (\mathcal{T}^{\mathcal{I}}\widehat{Q})(s,a)| \\ &\quad \le \gamma \mathbb{E}_{s'|s,a}\left[\left|F^{\mathcal{I}}_{s'}(Q(s',\cdot)) - F^{\mathcal{I}}_{s'}(\widehat{Q}(s',\cdot))\right|\right] \\ &\quad \le \gamma \mathbb{E}_{s'|s,a}\left[\|Q(s',\cdot) - \widehat{Q}(s',\cdot)\|_\infty\right] \le \gamma \|Q - \widehat{Q}\|_\infty. \end{aligned}$$

Thus

$$\|\mathcal{T}^{\mathcal{I}}Q - \mathcal{T}^{\mathcal{I}}\widehat{Q}\|_\infty \le \gamma \|Q - \widehat{Q}\|_\infty. \tag{331}$$

The same argument makes $\mathcal{T}_{\mathcal{F}}$ a monotone γ-contraction; hence both operators have unique bounded fixed points.

We then establish pointwise ordering of the two operators.

For every (s, j, q),

$$\sum_{i \in I_{s,j}} w_i^{(s,j)} e^{q_i/T_0} \le \sum_i w_i^{(s,j)} e^{q_i/T_0}.$$

The offsets agree, so $f_{s,j}^{\mathcal{I}}(q) \le f_{s,j}(q)$. Averaging over $j \sim \mu_s$ yields

$$F_s^{\mathcal{I}}(q) \le \mathcal{F}_s(q) \qquad \text{for all } s, q, \tag{332}$$

and consequently

$$\mathcal{T}^{\mathcal{I}} Q \le \mathcal{T}_{\mathcal{F}} Q \qquad \text{for every bounded } Q. \tag{333}$$

We now prove componentwise ordering of the fixed points.

By (333), at the full fixed point,

$$\mathcal{T}^{\mathcal{I}} Q_{\mathcal{F}}^\star \le \mathcal{T}_{\mathcal{F}} Q_{\mathcal{F}}^\star = Q_{\mathcal{F}}^\star.$$

Set $Q^{(0)} := Q_{\mathcal{F}}^\star$ and $Q^{(n+1)} := \mathcal{T}^{\mathcal{I}} Q^{(n)}$. Then $Q^{(1)} \le Q^{(0)}$, and monotonicity inductively gives $Q^{(n+1)} \le Q^{(n)}$. Contraction implies

$$Q^{(n)} = (\mathcal{T}^{\mathcal{I}})^n Q_{\mathcal{F}}^\star \longrightarrow Q_{\mathcal{I}}^\star \quad \text{in } \|\cdot\|_\infty.$$

Passing to the componentwise limit,

$$Q_{\mathcal{I}}^\star \le Q_{\mathcal{F}}^\star. \tag{334}$$

This proves the first componentwise assertion in (106).

We next bound the operator perturbation at the full fixed point.

By assumption, $\operatorname{osc}(Q_{\mathcal{F}}^\star(s,\cdot)) \le S_Q$ for every s. Lemma 4 therefore gives

$$0 \le \mathcal{F}_s(Q_{\mathcal{F}}^\star(s,\cdot)) - F_s^{\mathcal{I}}(Q_{\mathcal{F}}^\star(s,\cdot)) \le \Delta_{\mathcal{I}}(S_Q). \tag{335}$$

Consequently, for every (s, a),

$$\begin{aligned} 0 &\le (\mathcal{T}_{\mathcal{F}} Q_{\mathcal{F}}^\star)(s,a) - (\mathcal{T}^{\mathcal{I}} Q_{\mathcal{F}}^\star)(s,a) \\ &= \gamma \mathbb{E}_{s'|s,a}\left[\mathcal{F}_{s'}(Q_{\mathcal{F}}^\star(s',\cdot)) - F_{s'}^{\mathcal{I}}(Q_{\mathcal{F}}^\star(s',\cdot))\right] \\ &\le \gamma \Delta_{\mathcal{I}}(S_Q). \end{aligned}$$

Taking the supremum gives

$$\|\mathcal{T}_{\mathcal{F}} Q_{\mathcal{F}}^\star - \mathcal{T}^{\mathcal{I}} Q_{\mathcal{F}}^\star\|_\infty \le \gamma \Delta_{\mathcal{I}}(S_Q). \tag{336}$$

Only the span of $Q_{\mathcal{F}}^\star$ is used; no span assumption on $Q_{\mathcal{I}}^\star$ is required.

We continue to derive the action-value fixed-point error.

Let $D_Q := \|Q_{\mathcal{F}}^\star - Q_{\mathcal{I}}^\star\|_\infty$. Using both fixed-point identities, adding and subtracting $\mathcal{T}^{\mathcal{I}} Q_{\mathcal{F}}^\star$, and applying (336) and (331),

$$\begin{aligned} D_Q &= \|\mathcal{T}_{\mathcal{F}} Q_{\mathcal{F}}^\star - \mathcal{T}^{\mathcal{I}} Q_{\mathcal{I}}^\star\|_\infty \\ &\le \|\mathcal{T}_{\mathcal{F}} Q_{\mathcal{F}}^\star - \mathcal{T}^{\mathcal{I}} Q_{\mathcal{F}}^\star\|_\infty \\ &\quad + \|\mathcal{T}^{\mathcal{I}} Q_{\mathcal{F}}^\star - \mathcal{T}^{\mathcal{I}} Q_{\mathcal{I}}^\star\|_\infty \\ &\le \gamma \Delta_{\mathcal{I}}(S_Q) + \gamma D_Q. \end{aligned}$$

Thus $(1-\gamma) D_Q \le \gamma \Delta_{\mathcal{I}}(S_Q)$ and

$$\|Q_{\mathcal{F}}^\star - Q_{\mathcal{I}}^\star\|_\infty \le \frac{\gamma \Delta_{\mathcal{I}}(S_Q)}{1-\gamma}. \tag{337}$$

Together with (334), this gives $0 \le Q_{\mathcal{F}}^\star - Q_{\mathcal{I}}^\star$ componentwise and the displayed norm bound, proving both action-value assertions in (106).

We then prove the ordering of the state values.

By definition,

$$V_{\mathcal{F}}^\star(s) = \mathcal{F}_s(Q_{\mathcal{F}}^\star(s,\cdot)), \qquad V_{\mathcal{I}}^\star(s) = F_s^{\mathcal{I}}(Q_{\mathcal{I}}^\star(s,\cdot)).$$

The fixed-point order, monotonicity of $F_s^{\mathcal{I}}$, and aggregator order imply

$$F_s^{\mathcal{I}}(Q_{\mathcal{I}}^\star(s,\cdot)) \le F_s^{\mathcal{I}}(Q_{\mathcal{F}}^\star(s,\cdot)) \le \mathcal{F}_s(Q_{\mathcal{F}}^\star(s,\cdot)).$$

Therefore

$$V_{\mathcal{I}}^\star(s) \le V_{\mathcal{F}}^\star(s). \tag{338}$$

We now derive the state-value error.

Adding and subtracting the restricted aggregator at the full fixed point, then using (335), the 1-Lipschitz property, and (337),

$$\begin{aligned} V_{\mathcal{F}}^\star(s) - V_{\mathcal{I}}^\star(s) &= \big[\mathcal{F}_s(Q_{\mathcal{F}}^\star) - F_s^{\mathcal{I}}(Q_{\mathcal{F}}^\star)\big] \\ &\quad + \big[F_s^{\mathcal{I}}(Q_{\mathcal{F}}^\star) - F_s^{\mathcal{I}}(Q_{\mathcal{I}}^\star)\big] \\ &\le \Delta_{\mathcal{I}}(S_Q) + \|Q_{\mathcal{F}}^\star - Q_{\mathcal{I}}^\star\|_\infty \\ &\le \frac{\Delta_{\mathcal{I}}(S_Q)}{1-\gamma}, \end{aligned}$$

where the state argument $(s,\cdot)$ is suppressed inside the brackets. Together with (338),

$$0 \le V_{\mathcal{F}}^\star(s) - V_{\mathcal{I}}^\star(s) \le \frac{\Delta_{\mathcal{I}}(S_Q)}{1-\gamma}. \tag{339}$$

The constants differ because the Q-operator sees only the discounted next-state discrepancy $\gamma\Delta_{\mathcal{I}}(S_Q)$, whereas the state value also has the immediate discrepancy $\Delta_{\mathcal{I}}(S_Q)$:

$$\Delta_{\mathcal{I}}(S_Q) + \frac{\gamma\Delta_{\mathcal{I}}(S_Q)}{1-\gamma} = \frac{\Delta_{\mathcal{I}}(S_Q)}{1-\gamma}.$$

Finally, we verify the conservative choice $S_Q = 2\bar{B}_Q$.

Under Theorem 15, define

$$\begin{aligned} B_F &:= \sup_s |\mathcal{F}_s(0)| < \infty, \quad \bar{B}_F := B_F + T_0 \log 2, \\ B_Q &:= \frac{R_{\max} + \gamma B_F}{1-\gamma}, \qquad \bar{B}_Q := \frac{R_{\max} + \gamma \bar{B}_F}{1-\gamma}. \end{aligned}$$

The 1-Lipschitz property gives $|\mathcal{F}_s(q)| \le |\mathcal{F}_s(0)| + |\mathcal{F}_s(q) - \mathcal{F}_s(0)| \le B_F + \|q\|_\infty$. At the full fixed point, using $|r(s,a)| \le R_{\max}$,

$$\begin{aligned} \|Q_{\mathcal{F}}^\star\|_\infty &= \|\mathcal{T}_{\mathcal{F}} Q_{\mathcal{F}}^\star\|_\infty \\ &\le R_{\max} + \gamma\big(B_F + \|Q_{\mathcal{F}}^\star\|_\infty\big). \end{aligned}$$

Thus $(1-\gamma)\|Q_{\mathcal{F}}^\star\|_\infty \le R_{\max} + \gamma B_F$, so

$$\|Q_{\mathcal{F}}^\star\|_\infty \le B_Q \le \bar{B}_Q. \tag{340}$$

Finally, for every state,

$$\begin{aligned} \operatorname{osc}(Q_{\mathcal{F}}^\star(s,\cdot)) &\le 2\|Q_{\mathcal{F}}^\star(s,\cdot)\|_\infty \\ &\le 2\|Q_{\mathcal{F}}^\star\|_\infty \le 2B_Q \le 2\bar{B}_Q. \end{aligned}$$

Thus $S_Q = 2\bar{B}_Q$ satisfies the span hypothesis. Combining The preceding arguments prove Theorem 14. □

G.6 Exact heat support and locally computable truncated heat

G.6.1 Proof of Theorem 6

Fix a finite connected undirected action graph G with nonnegative edge weights and combinatorial Laplacian L. Recall that $T_0 = \tau\lambda > 0$, and set $H_t := e^{-tL}$ for $t > 0$. We first identify its

columns with the normalized-kernel backup, then convert the heat-tail and volume estimates into an effective-dimension bound, and finally derive the Gaussian specialization and planning rate.

We first prove that H_t is a strictly positive stochastic kernel.

Since G is undirected, L is symmetric, and $L\mathbf{1} = 0$. Hence

$$H_t\mathbf{1} = e^{-tL}\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top H_t = (H_t\mathbf{1})^\top = \mathbf{1}^\top,$$

so every row and column of H_t sums to one. To verify entrywise positivity, choose $\nu_{\mathrm{u}} > d_{\max}$ and set $P_G := I - L/\nu_{\mathrm{u}}$. Because $L = D - W$, P_G has off-diagonal entries $W_{ik}/\nu_{\mathrm{u}} \geq 0$, diagonal entries $1 - d_i/\nu_{\mathrm{u}} > 0$, and $P_G\mathbf{1} = \mathbf{1}$; thus it is a nonnegative stochastic matrix. Since $L = \nu_{\mathrm{u}}(I - P_G)$,

$$H_t = e^{-\nu_{\mathrm{u}}t}e^{\nu_{\mathrm{u}}tP_G} = e^{-\nu_{\mathrm{u}}t}\sum_{m=0}^{\infty}\frac{(\nu_{\mathrm{u}}t)^m}{m!}P_G^m. \tag{341}$$

Every term is entrywise nonnegative. For $i \neq j$, connectedness gives a positive-weight path from j to i of some length m, hence $(P_G^m)_{ij} > 0$, whose coefficient in (341) is positive because $t > 0$; for $i = j$, the $m = 0$ term suffices. Therefore

$$H_t(i,j) > 0 \quad \text{for all } i, j.$$

Each column $H_t(:,j)$ is thus a strictly positive probability vector, and $0 < H_t(i,j) \leq 1$, so the logarithmic heat cost below is finite and nonnegative.

Next, we identify the heat columns with normalized LOT columns.

Take the uniform prior $\rho_s(i) = 1/K$ and define $C_{ij}^{\mathrm{HK}} := -\lambda \log H_t(i,j)$. Then $e^{-C_{ij}^{\mathrm{HK}}/\lambda} = H_t(i,j)$, and column stochasticity gives

$$G_{s,j} = \sum_i \rho_s(i)e^{-C_{ij}^{\mathrm{HK}}/\lambda} = \frac{1}{K}\sum_i H_t(i,j) = \frac{1}{K}.$$

Thus the normalization in (100) yields

$$\begin{aligned} w_i^{(s,j)} &= \frac{\rho_s(i)e^{-C_{ij}^{\mathrm{HK}}/\lambda}}{G_{s,j}} = H_t(i,j), \\ b_{s,j} &= T_0 \log G_{s,j} = -T_0 \log K. \end{aligned} \tag{342}$$

Consequently, the uncentered uniform-prior LOT aggregator satisfies

$$\begin{aligned} F_s^{\log\mathrm{H}}(q) &= \sum_j \mu_s(j)\Big[-T_0 \log K \\ &\qquad + T_0 \log \sum_i H_t(i,j)e^{q_i/T_0}\Big] \\ &= F_s^{\mathrm{HK}}(q) - T_0 \log K, \\ F_s^{\mathrm{HK}}(q) &:= T_0 \sum_j \mu_s(j) \log \sum_i H_t(i,j)e^{q_i/T_0}, \end{aligned}$$

where F_s^{HK} is the centered direct-heat aggregator in (57). Equivalently, centering keeps $w^{(s,j)} = H_t(:,j)$ and replaces the offset $-T_0 \log K$ by zero. With $c := T_0 \log K$,

$$F_s^{\mathrm{HK}}(q) = F_s^{\log\mathrm{H}}(q) + c. \tag{343}$$

Thus centering changes only the backup's absolute value, not the local softmax probabilities, because the common column normalization cancels.

We next derive the corresponding Bellman fixed-point shifts.

Both aggregators are translation equivariant: $F_s(q + d\mathbf{1}) = F_s(q) + d$ for every scalar d. If $\mathcal{T}_{\mathrm{HK}}$ and $\mathcal{T}_{\log\mathrm{H}}$ are their Bellman operators, (343) implies

$$\mathcal{T}_{\mathrm{HK}}Q = \mathcal{T}_{\log\mathrm{H}}Q + \gamma c\mathbf{1}, \qquad \mathcal{T}_{\log\mathrm{H}}(Q + d\mathbf{1}) = \mathcal{T}_{\log\mathrm{H}}Q + \gamma d\mathbf{1}. \tag{344}$$

Let $Q^\star_{\mathrm{logH}}$ be the uncentered fixed point and set $d_Q := \gamma c/(1-\gamma)$, so $d_Q = \gamma(d_Q + c)$, and write $\widetilde{Q} := Q^\star_{\mathrm{logH}} + d_Q \mathbf{1}$. Then

$$\begin{aligned}\mathcal{T}_{\mathrm{HK}}\widetilde{Q} &= \mathcal{T}_{\mathrm{logH}}\widetilde{Q} + \gamma c \mathbf{1} \\ &= Q^\star_{\mathrm{logH}} + \gamma(d_Q + c)\mathbf{1} = Q^\star_{\mathrm{logH}} + d_Q \mathbf{1}.\end{aligned}$$

Uniqueness of the discounted Bellman fixed point and translation equivariance of the aggregators therefore give, with $q^\star_s := Q^\star_{\mathrm{logH}}(s, \cdot)$,

$$\begin{aligned}Q^\star_{\mathrm{HK}} &= Q^\star_{\mathrm{logH}} + \frac{\gamma c}{1-\gamma}\mathbf{1} \\ &= Q^\star_{\mathrm{logH}} + \frac{\gamma T_0 \log K}{1-\gamma}\mathbf{1}, \\ V^\star_{\mathrm{HK}}(s) &= F_s^{\mathrm{logH}}(q^\star_s + d_Q \mathbf{1}) + c \\ &= V^\star_{\mathrm{logH}}(s) + d_Q + c \\ &= V^\star_{\mathrm{logH}}(s) + \frac{T_0 \log K}{1-\gamma}.\end{aligned}$$

This proves (58). At the corresponding translated action values, the policies agree; their derivatives also agree because a common logit shift leaves a softmax unchanged. Hence centering leaves the policies, Hessians, pruning differences, and local Taylor estimators unchanged.

We continue to verify the centered value envelopes.

At $q = 0$, column stochasticity yields

$$\begin{aligned}F_s^{\mathrm{HK}}(0) &= T_0 \sum_j \mu_s(j) \log \sum_i H_t(i,j) = 0, \\ B_F &:= \sup_s |F_s^{\mathrm{HK}}(0)| = 0.\end{aligned}$$

Suppose each retained set contains at least half of its heat column: $H_t(I_{s,j}, j) := \sum_{i \in I_{s,j}} H_t(i,j) \geq 1/2$ for every (s,j). Then

$$\begin{aligned}F_s^{\mathcal{I}}(0) &= T_0 \sum_j \mu_s(j) \log H_t(I_{s,j}, j), \\ -\log 2 &\leq \log H_t(I_{s,j}, j) \leq 0.\end{aligned}$$

Since μ_s is a probability distribution, $-T_0 \log 2 \leq F_s^{\mathcal{I}}(0) \leq 0$, and therefore

$$B_F^{\mathcal{I}} := \sup_s |F_s^{\mathcal{I}}(0)| \leq T_0 \log 2, \tag{345}$$

uniformly in the ambient action count K.

We then convert the heat-tail bound into a mass-retaining set.

Fix $\alpha \in (0,1)$. Since $\Psi_t(r) \to 0$ as $r \to \infty$, the set $\{m \in \mathbb{N}_0 : \Psi_t(m) \leq \alpha\}$ is nonempty, so

$$r_\alpha := \min\{m \in \mathbb{N}_0 : \Psi_t(m) \leq \alpha\}$$

is well defined and attained. For each anchor j, choose $I_j := B_G(j, r_\alpha) = \{i : d_G(i,j) \leq r_\alpha\}$. The tail assumption and unit column mass give

$$\begin{aligned}H_t(I_j^c, j) &= \sum_{i : d_G(i,j) > r_\alpha} H_t(i,j) \leq \Psi_t(r_\alpha) \leq \alpha, \\ H_t(I_j, j) &= 1 - H_t(I_j^c, j) \geq 1 - \alpha.\end{aligned} \tag{346}$$

Thus $B_G(j, r_\alpha)$ is an admissible mass-retaining set for the heat column centered at j.

We now use graph-volume growth to bound the effective dimension.

For the exact-heat columns, define

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) := \sup_j \min\left\{ |I| : I \subseteq \mathcal{A},\ \sum_{i\in I} H_t(i,j) \geq 1-\alpha \right\}.$$

If the heat graph varies with the state, the supremum also ranges over s; the uniform bound below is unchanged. By (346), every $B_G(j, r_\alpha)$ is feasible, so the volume-growth estimate in (53) gives

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) \leq \sup_j |B_G(j, r_\alpha)| \leq v_0(1+r_\alpha)^{d_g},$$

proving (54). This is an effective-support, or mass-concentration, statement; it does not assert $H_t(i,j) = 0$ outside $B_G(j, r_\alpha)$. Indeed, strict positivity of exact heat shows that exact H_t has full support for connected G and $t > 0$.

We next derive the Gaussian-radius specialization.

Suppose that, throughout a stated diffusive range $0 \leq r \leq r_{\mathrm{diff}}$,

$$\Psi_t(r) \leq c_H \exp\left(-\frac{r^2}{c_t t}\right), \qquad c_H, c_t \geq 1.$$

Assume $c_H/\alpha \geq 1$ and define

$$R_\alpha := \sqrt{c_t t \log(c_H/\alpha)}, \qquad m_\alpha := \lceil R_\alpha \rceil, \qquad m_\alpha \leq r_{\mathrm{diff}},$$

where the last condition ensures that the estimate holds at the rounded integer radius. Since $m_\alpha \geq R_\alpha$,

$$\begin{aligned}\Psi_t(m_\alpha) &\leq c_H e^{-m_\alpha^2/(c_t t)} \\ &\leq c_H e^{-R_\alpha^2/(c_t t)} \\ &= c_H e^{-\log(c_H/\alpha)} = \alpha.\end{aligned}$$

Minimality of r_α thus gives $r_\alpha \leq m_\alpha = \lceil \sqrt{c_t t \log(c_H/\alpha)} \rceil$, and (54) yields

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) \leq v_0 \left(1 + \left\lceil \sqrt{c_t t \log(c_H/\alpha)} \right\rceil\right)^{d_g}, \tag{347}$$

which is (55). For fixed t, c_H, c_t, v_0, d_g, as $\alpha \downarrow 0$ within the Gaussian regime,

$$1 + \left\lceil \sqrt{c_t t \log(c_H/\alpha)} \right\rceil = O\left(\sqrt{\log(1/\alpha)}\right),$$

and hence

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha) = O\left((\log(1/\alpha))^{d_g/2}\right). \tag{348}$$

This conclusion is restricted to graph-family and accuracy sequences for which the required radius stays in the Gaussian regime and before finite-graph neighborhood saturation. Neither Gaussian decay nor polynomial volume growth follows from bounded degree alone.

We continue to relate the planning tail budget to the target accuracy.

For the centered heat branch, the centered-envelope calculation above gives $B_F = 0$, so

$$\begin{aligned}\bar{B}_F &= T_0 \log 2, \\ \bar{B}_Q &= \frac{R_{\max} + \gamma T_0 \log 2}{1-\gamma}, \\ S_Q &= 2\bar{B}_Q.\end{aligned}$$

Thus S_Q is independent of K when $(R_{\max}, \gamma, T_0)$ are fixed. From (2),

$$\begin{aligned}z_\varepsilon &= \exp\left(\frac{(1-\gamma)\varepsilon}{2T_0}\right) - 1, \\ \alpha_\varepsilon &= \min\left\{\frac{1}{2}, \frac{z_\varepsilon}{e^{S_Q/T_0} + z_\varepsilon}\right\}.\end{aligned}$$

For fixed $T_0 > 0$, $\gamma < 1$, and $S_Q < \infty$,

$$\begin{aligned} z_\varepsilon &= \frac{1-\gamma}{2T_0}\varepsilon + O(\varepsilon^2) = \Theta(\varepsilon), \\ e^{S_Q/T_0} + z_\varepsilon &= e^{S_Q/T_0} + O(\varepsilon), \\ \frac{z_\varepsilon}{e^{S_Q/T_0} + z_\varepsilon} &= \frac{1-\gamma}{2T_0} e^{-S_Q/T_0}\varepsilon + O(\varepsilon^2) = \Theta(\varepsilon). \end{aligned}$$

For sufficiently small ε the last quantity is below $1/2$, so the minimum is inactive and

$$\alpha_\varepsilon = \Theta(\varepsilon). \tag{349}$$

Therefore $\log(1/\alpha_\varepsilon) = \log(1/\varepsilon) + O(1)$, and (348) implies

$$d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha_\varepsilon) = O\left((\log(1/\varepsilon))^{d_g/2}\right). \tag{350}$$

Finally, we substitute the dimension into the planning bound.

The root-value bound (23) is

$$n(\varepsilon, \delta) \le \frac{c_2}{\varepsilon^3}\big(c_2 d_{\mathrm{ker}}(\alpha_\varepsilon)\big)^{c_3[1+\log\log(e+\bar{B}_Q/\varepsilon)]} \log\frac{2}{\delta}.$$

Set $L_\varepsilon := \log(1/\varepsilon)$. By (350), for some constant C independent of ε,

$$\begin{aligned} d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha_\varepsilon) &\le C L_\varepsilon^{d_g/2}, \\ \log\big(c_2 d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha_\varepsilon)\big) &= O(\log L_\varepsilon), \\ 1 + \log\log(e + \bar{B}_Q/\varepsilon) &= O(\log L_\varepsilon). \end{aligned}$$

where $\bar{B}_Q$ is fixed. Hence the logarithm of the geometry-dependent factor is

$$\begin{aligned} &c_3[1 + \log\log(e + \bar{B}_Q/\varepsilon)] \\ &\qquad \times \log\big(c_2 d_{\mathrm{ker}}^{\mathrm{HK}}(\alpha_\varepsilon)\big) \\ &= O\left((\log L_\varepsilon)^2\right) \\ &= O\left((\log\log(1/\varepsilon))^2\right). \end{aligned}$$

Exponentiating and substituting gives

$$\begin{aligned} n(\varepsilon, \delta) &\le \varepsilon^{-3} \exp\left(O((\log\log(1/\varepsilon))^2)\right) \log\frac{2}{\delta} \\ &= (1/\varepsilon)^{3+o(1)} \log\frac{2}{\delta}. \end{aligned} \tag{351}$$

Indeed, $(\log\log(1/\varepsilon))^2 / \log(1/\varepsilon) \to 0$, so the exponential factor is $(1/\varepsilon)^{o(1)}$. This proves (59).

The final asymptotic statement is understood along graph-family and accuracy sequences for which the rounded Gaussian radius remains in the stated diffusive range and before graph balls saturate at all K actions. Exact H_t also remains dense: the theorem shows that small sets retain most of its mass, but computing the corresponding exact heat weights in sub-K time additionally requires Assumption 2.

Combining these arguments proves Theorem 6 and the corresponding clauses of Theorem 7. □

G.6.2 Proof of Proposition 1

At $C = 0$ and uniform ρ, every base weight is $1/K$, so retaining $1 - \alpha$ mass requires exactly $\lceil (1-\alpha)K \rceil$ actions. □

G.6.3 Proof of Proposition 2

Fix an anchor j and suppress the state index.

We first lower-bound every Gibbs weight.

Under the uniform prior, $\rho(i) = 1/K$, so

$$w_i^{(j)} = \frac{e^{-C_{ij}/\lambda}}{\sum_{k=1}^{K} e^{-C_{kj}/\lambda}}.$$

Since $0 \le C_{ij} \le C_{\max}$,

$$e^{-C_{ij}/\lambda} \ge e^{-C_{\max}/\lambda}, \qquad \sum_{k=1}^{K} e^{-C_{kj}/\lambda} \le K.$$

Hence

$$w_i^{(j)} \ge \frac{e^{-C_{\max}/\lambda}}{K} \qquad \text{for every } i, j. \tag{352}$$

Thus, for fixed $C_{\max}/\lambda$, every action has mass of order $1/K$.

Next, we lower-bound the required cardinality.

For any retained set $I \subseteq \mathcal{A}$ with $d := |I|$,

$$w^{(j)}(I^c) = \sum_{i \in I^c} w_i^{(j)} \ge \frac{K-d}{K} e^{-C_{\max}/\lambda}.$$

Therefore, $w^{(j)}(I) \ge 1-\alpha$, equivalently $w^{(j)}(I^c) \le \alpha$, implies

$$d \ge K\left(1 - \alpha e^{C_{\max}/\lambda}\right). \tag{353}$$

Since $d \ge 0$, uniformly

$$d \ge K\left(1 - \alpha e^{C_{\max}/\lambda}\right)_+, \qquad (x)_+ := \max\{x, 0\}.$$

The positive part is needed only when $\alpha e^{C_{\max}/\lambda} \ge 1$, where the preceding bound is nonpositive.

We then pass to the kernel effective dimension.

The cardinality bound above holds for every anchor j, so

$$\min_{\substack{I \subseteq \mathcal{A} \\ w^{(j)}(I) \ge 1-\alpha}} |I| \ge K\left(1 - \alpha e^{C_{\max}/\lambda}\right)_+.$$

If the geometry is state-dependent and the same $C_{\max}$ holds uniformly over states, this also holds for every s. Taking the supremum over state–anchor pairs in Definition 1 gives

$$d_{\ker}(\alpha) \ge K\left(1 - \alpha e^{C_{\max}/\lambda}\right)_+, \tag{354}$$

which is (56). In particular, if $C_{\max}/\lambda$ is fixed and $\alpha \downarrow 0$, then

$$\alpha e^{C_{\max}/\lambda} = o(1), \qquad d_{\ker}(\alpha) \ge K(1 - o(1)),$$

so the deterministic kernel-mass pruning rule must retain almost all actions at high accuracy.

We now express squared diffusion distance spectrally.

Let $C = C_t^L$ be generated by a finite connected undirected graph, with

$$L = \sum_{k=1}^{K} \lambda_k \phi_k \phi_k^\top, \qquad 0 = \lambda_1 < \lambda_2 \le \cdots \le \lambda_K, \qquad \phi_1 = \frac{1}{\sqrt{K}}\mathbf{1}.$$

The squared diffusion distance is

$$(C_t^L)_{ij} = \sum_{k=1}^{K} e^{-2t\lambda_k} \big(\phi_k(i) - \phi_k(j)\big)^2.$$

Since ϕ_1 is constant, its contribution vanishes, yielding

$$(C_t^L)_{ij} = \sum_{k=2}^{K} e^{-2t\lambda_k} \big(\phi_k(i) - \phi_k(j)\big)^2. \tag{355}$$

Finally, we bound the diffusion cost.

For $k \geq 2$, $\lambda_k \geq \lambda_2$; hence, for $t \geq 0$,

$$(C_t^L)_{ij} \leq e^{-2t\lambda_2} \sum_{k=2}^{K} \big(\phi_k(i) - \phi_k(j)\big)^2.$$

Orthonormality and Parseval's identity give

$$\sum_{k=2}^{K} \big(\phi_k(i) - \phi_k(j)\big)^2 = \|e_i - e_j\|_2^2,$$

because the $k = 1$ term is zero. This norm is zero for $i = j$ and equals $1 + 1 - 0 = 2$ otherwise. Therefore

$$0 \leq (C_t^L)_{ij} \leq 2e^{-2t\lambda_2} \qquad \text{for all } i, j. \tag{356}$$

Thus $C_{\max} \leq 2e^{-2t\lambda_2}$, and (354) gives

$$d_{\ker}(\alpha) \geq K \left(1 - \alpha \exp\left\{\frac{2e^{-2t\lambda_2}}{\lambda}\right\}\right)_+.$$

This concerns deterministic, value-independent kernel-mass neighborhoods. It neither asserts that the final LOT policy is dense for every action-value vector nor lower-bounds every randomized action-integration or planning method. □

G.6.4 Proof of Theorem 7

The centering, support, Gaussian-range, and planning-rate claims were established in the proof of Theorem 6. The two density comparisons are Propositions 1 and 2. Together these results prove every clause of the expanded statement. □

G.6.5 Proof of Theorem 3

Fix $r \in \mathbb{N}_0$. Recall that $G = (\mathcal{A}, E, W)$ is a finite connected undirected weighted graph, $L = D - W$ with $D_{ii} = \sum_k W_{ik}$, and

$$\nu_{\mathrm{u}} \geq d_{\max} := \max_i D_{ii}, \qquad \nu_{\mathrm{u}} > 0, \qquad \theta := \nu_{\mathrm{u}} t > 0.$$

We prove the stochasticity, support, computation, approximation, decision, and recursive-planning claims in turn.

We first verify that P_G is symmetric and stochastic.

Since

$$P_G := I - \frac{L}{\nu_{\mathrm{u}}} = I - \frac{D}{\nu_{\mathrm{u}}} + \frac{W}{\nu_{\mathrm{u}}},$$

its off-diagonal entries are $W_{ik}/\nu_{\mathrm{u}} \geq 0$, while

$$(P_G)_{ii} = 1 - \frac{D_{ii}}{\nu_{\mathrm{u}}} + \frac{W_{ii}}{\nu_{\mathrm{u}}} \geq 1 - \frac{D_{ii}}{\nu_{\mathrm{u}}} \geq 0.$$

Moreover,

$$\sum_k (P_G)_{ik} = 1 - \frac{D_{ii}}{\nu_{\mathrm{u}}} + \frac{1}{\nu_{\mathrm{u}}} \sum_k W_{ik} = 1,$$

so P_G is nonnegative and row stochastic. Since $W = W^\top$ and D is diagonal, $P_G = P_G^\top$, whence $\mathbf{1}^\top P_G = (P_G \mathbf{1})^\top = \mathbf{1}^\top$. Thus P_G, and every power P_G^k, is symmetric, nonnegative, and doubly stochastic.

Next, we represent exact heat by Poisson uniformization.

Because $L = \nu_{\mathrm{u}}(I - P_G)$ and I commutes with P_G,

$$H_t = e^{-tL} = e^{-\theta} e^{\theta P_G} = \sum_{k=0}^{\infty} a_k P_G^k, \qquad a_k := e^{-\theta} \frac{\theta^k}{k!}.$$

The $a_k > 0$ sum to one and form the pmf of $\mathrm{Pois}(\theta)$. Define

$$s_r := \sum_{k=0}^{r} a_k > 0,$$

$$\beta_r := 1 - s_r = \sum_{k=r+1}^{\infty} a_k,$$

$$H_{t,r} := \frac{1}{s_r} \sum_{k=0}^{r} a_k P_G^k.$$

Hence $H_{t,r}$ is a convex combination of symmetric, nonnegative, doubly stochastic matrices and has probability-distribution columns.

We next prove the exact finite-radius support identity.

For an anchor j, $(P_G^k)_{ij}$ is a sum over length-k sequences $j = x_0, \ldots, x_k = i$ of products of the corresponding entries of P_G. A positive off-diagonal factor traverses a positive-weight edge and a diagonal factor stays at the current vertex; thus

$$(P_G^k)_{ij} > 0 \implies d_G(i,j) \le k,$$
$$\mathrm{supp}(P_G^k e_j) \subseteq B_G(j,k).$$

Nonnegativity then gives

$$\mathrm{supp}(H_{t,r}(:,j)) \subseteq \bigcup_{k=0}^{r} B_G(j,k) = B_G(j,r).$$

Conversely, fix $i \in B_G(j,r)$ and put $\ell := d_G(i,j) \le r$. If $i \ne j$, a shortest positive-weight path $j = x_0, \ldots, x_\ell = i$ satisfies

$$(P_G^\ell)_{ij} \ge \prod_{m=0}^{\ell-1} \frac{W_{x_{m+1},x_m}}{\nu_{\mathrm{u}}} > 0, \qquad a_\ell = e^{-\theta} \frac{\theta^\ell}{\ell!} > 0,$$

and therefore $H_{t,r}(i,j) \ge a_\ell (P_G^\ell)_{ij} / s_r > 0$. If $i = j$, the $k = 0$ term gives $H_{t,r}(j,j) \ge a_0 / s_r > 0$. Consequently,

$$\mathrm{supp}(H_{t,r}(:,j)) = B_G(j,r), \qquad H_{t,r}(j,j) > 0. \tag{357}$$

The reverse inclusion needs no positive holding probability: a vertex at distance ℓ is reached by the $k = \ell$ term.

We continue to prove the local recurrence and arithmetic cost.

Starting with $v_0 = e_j$ and $v_{k+1} = P_G v_k$ gives $v_k = P_G^k e_j$, so

$$y_r := \sum_{k=0}^{r} a_k v_k, \qquad H_{t,r} e_j = \frac{y_r}{s_r}.$$

Thus the recurrence computes the truncated column exactly. Since $\mathrm{supp}(v_k) \subseteq B_G(j,k)$, computing v_{k+1} only processes the current support and its adjacency lists and never visits outside $B_G(j,k+1)$; the full recurrence stays within $B_G(j,r)$.

Let $d_r := \max_j |B_G(j,r)|$. At most d_r vertices are visited, each with at most Δ_G positive-weight neighbors. Each sparse multiplication costs $O((\Delta_G+1)d_r)$; there are r such multiplications, and accumulating $r+1$ weighted vectors and returning the result costs $O((r+1)d_r)$. Uniformly for $r \geq 0$,

$$O\left((r+1)(\Delta_G+1)d_r\right). \tag{358}$$

For $r \geq 1$ this is $O(r(\Delta_G+1)d_r)$. If $r=0$, the column is e_j but returning it still costs $O(1)$; hence a theorem allowing $r=0$ must use $r+1$ or treat that case separately. Two recurrence vectors, the running sum, and stored discovered adjacency lists use $O((\Delta_G+1)d_r)$ working memory. This proves (64), subject to the $r=0$ clarification.

Once constructed, the column induces

$$p_i^{(j,r)}(q) = \frac{H_{t,r}(i,j)e^{q_i/T_0}}{\sum_{k\in B_G(j,r)} H_{t,r}(k,j)e^{q_k/T_0}}, \qquad i \in B_G(j,r).$$

One pass over at most d_r entries computes the denominator and all probabilities in $O(d_r)$ time. An alias table also costs $O(d_r)$ to build and then yields $O(1)$ time per sample.

We then decompose exact heat and bound the kernel error.

Define $R_r := \sum_{k=r+1}^{\infty} a_k P_G^k$. Splitting the uniformization series gives

$$H_t = \sum_{k=0}^{r} a_k P_G^k + R_r = s_r H_{t,r} + R_r. \tag{359}$$

R_r is nonnegative and symmetric, and every row and column sums to $\sum_{k=r+1}^{\infty} a_k = \beta_r$. Since $s_r = 1-\beta_r$,

$$H_t - H_{t,r} = R_r - \beta_r H_{t,r}. \tag{360}$$

For every row i,

$$\begin{aligned}\sum_j |H_t(i,j) - H_{t,r}(i,j)| &\leq \sum_j R_r(i,j) + \beta_r \sum_j H_{t,r}(i,j)\\ &= 2\beta_r.\end{aligned}$$

so $\|H_t - H_{t,r}\|_\infty \leq 2\beta_r$ in the induced maximum-row-sum norm. Symmetry gives

$$\max_j \sum_i |H_t(i,j) - H_{t,r}(i,j)| \leq 2\beta_r. \tag{361}$$

For $i \notin B_G(j,r)$, $(P_G^k)_{ij} = 0$ for $0 \leq k \leq r$, so $H_{t,r}(i,j) = 0$ and

$$\sum_{i\notin B_G(j,r)} H_t(i,j) = \sum_{i\notin B_G(j,r)} R_r(i,j) \leq \sum_i R_r(i,j) = \beta_r.$$

These bounds prove (66) and (27). They also show why the recurrence does not compute exact H_t restricted to the ball: for $i \in B_G(j,r)$, $H_t(i,j) = s_r H_{t,r}(i,j) + R_r(i,j)$, and $R_r(i,j)$ may be positive because walks longer than r can leave and return or otherwise end inside the ball.

We now express exact heat as a mixture of stochastic kernels.

If $\beta_r > 0$, set $\bar{R}_r := R_r/\beta_r$. It is symmetric and doubly stochastic, and (359) becomes

$$H_t = (1-\beta_r)H_{t,r} + \beta_r \bar{R}_r. \tag{362}$$

Fix $q \in \mathbb{R}^K$ with $\operatorname{osc}(q) \le S$ and define

$$\begin{aligned} u_i &:= e^{q_i/T_0}, & u_{\min} &:= \min_i u_i, \\ u_{\max} &:= \max_i u_i, \\ \kappa &:= \frac{u_{\max}}{u_{\min}} \\ &= \exp\left(\frac{\max_i q_i - \min_i q_i}{T_0}\right) \\ &\le e^{S/T_0}. \end{aligned}$$

where $\kappa \ge 1$. For each anchor j, let

$$\begin{aligned} z_{r,j} &:= \sum_i H_{t,r}(i,j)u_i = \langle H_{t,r}(:,j), u\rangle, \\ \bar{z}_j &:= \sum_i \bar{R}_r(i,j)u_i = \langle \bar{R}_r(:,j), u\rangle, \\ z_j &:= \sum_i H_t(i,j)u_i = \langle H_t(:,j), u\rangle \\ &= (1-\beta_r)z_{r,j} + \beta_r \bar{z}_j. \end{aligned}$$

Since both columns are probability distributions, $u_{\min} \le z_{r,j}, \bar{z}_j \le u_{\max}$.

We next derive the uniform log-partition error.

Dividing the preceding mixture by $z_{r,j} > 0$ and using the interval bounds gives

$$1 - \beta_r + \frac{\beta_r}{\kappa} \le \frac{z_j}{z_{r,j}} \le 1 + \beta_r(\kappa - 1). \tag{363}$$

Set $L_\kappa := 1 - \beta_r(1 - \kappa^{-1})$ and $U_\kappa := 1 + \beta_r(\kappa - 1)$. Then

$$\begin{aligned} U_\kappa L_\kappa &= 1 + \frac{\beta_r(1-\beta_r)(\kappa-1)^2}{\kappa} \ge 1, \\ -\log L_\kappa &\le \log U_\kappa. \end{aligned}$$

Thus the upper logarithmic deviation dominates the lower, and (363) yields

$$\begin{aligned} |T_0 \log z_j - T_0 \log z_{r,j}| &\le T_0 \log U_\kappa \\ &= T_0 \log(1 + \beta_r(\kappa - 1)) \\ &\le T_0 \log(1 + \beta_r(e^{S/T_0} - 1)) \\ &= \delta_r(S). \end{aligned}$$

If $\beta_r = 0$, then $H_t = H_{t,r}$ and the same statement holds with zero error. Hence, for every anchor,

$$\left| T_0 \log \sum_i H_t(i,j) e^{q_i/T_0} - T_0 \log \sum_i H_{t,r}(i,j) e^{q_i/T_0} \right| \le \delta_r(S).$$

Averaging with μ_s and applying the triangle inequality gives

$$|F^{\mathrm{HK}}_{s,t}(q) - F^{\mathrm{HK}}_{s,t,r}(q)| \le \delta_r(S),$$

uniformly in s and all q with $\operatorname{osc}(q) \le S$. This proves (69) and (28).

We continue to establish the Bellman contractions.

Under the standing assumptions $\gamma \in (0,1)$ and bounded rewards, both Bellman maps act on the complete space of bounded action-value functions. Their aggregators are mixtures of maps $q \mapsto T_0 \log \sum_i w_i e^{q_i/T_0}$ with $w \in \Delta$. Such maps are monotone and translation equivariant:

$$T_0 \log \sum_i w_i e^{(q_i+c)/T_0} = c + T_0 \log \sum_i w_i e^{q_i/T_0}.$$

If $d := \|q - q'\|_\infty$, then $q' - d\mathbf{1} \le q \le q' + d\mathbf{1}$, so monotonicity and translation equivariance imply $|F_s(q) - F_s(q')| \le d$. Thus both aggregators are 1-Lipschitz, and

$$\|\mathcal{T}_t Q - \mathcal{T}_t Q'\|_\infty \le \gamma \|Q - Q'\|_\infty,$$
$$\|\mathcal{T}_{t,r} Q - \mathcal{T}_{t,r} Q'\|_\infty \le \gamma \|Q - Q'\|_\infty.$$

Both operators are therefore γ-contractions with unique bounded fixed points.

We then propagate the error to the action-value fixed points.

Assume $\operatorname{osc}(Q_t^\star(s,\cdot)) \le S$ for every s; in the expanded theorem, $S = S_Q = 2\bar{B}_Q$. The uniform log-partition bound at the exact fixed point gives

$$|F_{s,t}^{\mathrm{HK}}(Q_t^\star(s,\cdot)) - F_{s,t,r}^{\mathrm{HK}}(Q_t^\star(s,\cdot))| \le \delta_r(S),$$

and hence $\|\mathcal{T}_t Q_t^\star - \mathcal{T}_{t,r} Q_t^\star\|_\infty \le \gamma \delta_r(S)$. Using the fixed-point identities, adding and subtracting $\mathcal{T}_{t,r} Q_t^\star$, and applying the contraction,

$$\begin{aligned}\|Q_{t,r}^\star - Q_t^\star\|_\infty &\le \|\mathcal{T}_{t,r} Q_{t,r}^\star - \mathcal{T}_{t,r} Q_t^\star\|_\infty \\ &\quad + \|\mathcal{T}_{t,r} Q_t^\star - \mathcal{T}_t Q_t^\star\|_\infty \\ &\le \gamma \|Q_{t,r}^\star - Q_t^\star\|_\infty + \gamma \delta_r(S).\end{aligned}$$

Rearranging gives

$$\|Q_{t,r}^\star - Q_t^\star\|_\infty \le \frac{\gamma}{1-\gamma} \delta_r(S). \tag{364}$$

Only the exact-heat fixed point needs the span assumption; no span bound on $Q_{t,r}^\star$ is used.

We now propagate the error to the state values.

By definition, $V_t^\star(s) = F_{s,t}^{\mathrm{HK}}(Q_t^\star(s,\cdot))$ and $V_{t,r}^\star(s) = F_{s,t,r}^{\mathrm{HK}}(Q_{t,r}^\star(s,\cdot))$. Adding and subtracting the truncated aggregator at $Q_t^\star$ gives

$$\begin{aligned}|V_{t,r}^\star(s) - V_t^\star(s)| &\le \|Q_{t,r}^\star - Q_t^\star\|_\infty + \delta_r(S) \\ &\le \frac{\gamma \delta_r(S)}{1-\gamma} + \delta_r(S) = \frac{\delta_r(S)}{1-\gamma}.\end{aligned}$$

Taking the supremum over states,

$$\|V_{t,r}^\star - V_t^\star\|_\infty \le \frac{\delta_r(S)}{1-\gamma}. \tag{365}$$

Together, (364) and (365) prove the decision-error claims; setting $S = S_Q$ gives (70).

We next identify the normalized-kernel form.

Since every column of $H_{t,r}$ is a probability distribution,

$$F_{s,t,r}^{\mathrm{HK}}(q) = T_0 \sum_j \mu_s(j) \log \sum_i H_{t,r}(i,j) e^{q_i/T_0}$$

has the normalized-kernel form (99) with $w^{(s,j)} = H_{t,r}(:,j)$ and $b_{s,j} = 0$. By the finite-radius support identity,

$$I_{s,j} = B_G(j,r), \quad w^{(s,j)}(I_{s,j}) = 1,$$
$$|I_{s,j}| \le d_r, \qquad d_{\mathcal{I}} = d_r.$$

Moreover, $F_{s,t,r}^{\mathrm{HK}}(0) = T_0 \sum_j \mu_s(j) \log \sum_i H_{t,r}(i,j) = 0$, so $B_F = 0$. Thus the cascade retains mass one, rather than merely one half.

We continue to compose with the recursive estimator.

Under Theorem 15's assumptions—an independent generative model, $|R| \le R_{\max}$, $\gamma \in (0,1)$, fixed geometry, and $0 < \varepsilon_{\mathrm{stat}} \le \varepsilon_0$—apply it with $I_{s,j} = B_G(j,r)$, $d_{\mathcal{I}} = d_r$, and $B_F = 0$. Its median-of-means outer estimator estimates $V^\star_{t,r}(s_0)$ within $\varepsilon_{\mathrm{stat}}$ with probability at least $1-\delta$, using

$$n_r(\varepsilon_{\mathrm{stat}}, \delta) \le \frac{C_{\mathrm{u}}}{\varepsilon_{\mathrm{stat}}^3} (C_{\mathrm{u}} d_r)^{c_{\mathrm{u}}[1+\log\log(e+\bar{B}_Q/\varepsilon_{\mathrm{stat}})]} \log \frac{2}{\delta},$$

which is (71).

After sampling anchor J, the local policy is supported exactly on $B_G(J,r)$, so the recursive estimator evaluates no action-value coordinate and makes no transition-oracle call outside that ball. The local recurrence also constructs the kernel column without a full K-scan unless the ball has saturated at all K actions. The constants from Theorem 15 depend on the reward bound, discount, temperature, and fixed cascade choices, but not additionally on

$$K, \quad r, \quad d_r, \quad \varepsilon_{\mathrm{stat}}, \quad \text{or} \quad \delta.$$

The graph-column arithmetic $O((r+1)(\Delta_G+1)d_r)$ is separate from the transition-oracle complexity.

Finally, we clarify the approximation target.

The recurrence constructs the normalized finite Poisson head $H_{t,r}$ exactly, not exact H_t restricted to $B_G(j,r)$: exact in-ball heat also contains $R_r(:,j)$ from walks longer than r. Accordingly, the recursive estimator directly targets $V^\star_{t,r}$, while

$$\|V^\star_{t,r} - V^\star_t\|_\infty \le \frac{\delta_r(S)}{1-\gamma}$$

controls its deterministic discrepancy from the exact centered-heat target. Thus local computation and approximation to exact heat are distinct parts of the guarantee.

Combining these arguments proves Theorem 3 and every clause of Theorem 8. □

G.6.6 Proof of Theorem 8

Every clause of the expanded statement was established in the preceding proof of Theorem 3. □

G.6.7 Proof of Corollaries 1 and 2

Fix $0 < \delta < 1$ and $0 < \varepsilon \le \varepsilon_0$, and set

$$\theta := \nu_{\mathrm{u}} t, \quad X \sim \mathrm{Pois}(\theta), \quad M_\varepsilon := \max\left\{1, 2e\theta, \log_2 \frac{1}{b_\varepsilon}\right\}.$$

By (30), $r_\varepsilon + 1 = \lceil M_\varepsilon \rceil$. We first make the exact–truncated heat bias at most $\varepsilon/2$, then estimate the truncated-heat value to accuracy $\varepsilon/2$. The standing assumptions imply $\theta > 0$.

We first verify the action-value span used in the heat error bound.

The centered exact-heat aggregator is

$$F^{\mathrm{HK}}_{s,t}(q) = T_0 \sum_j \mu_s(j) \log \sum_i H_t(i,j) e^{q_i/T_0}.$$

Because every heat column is a probability vector, $F^{\mathrm{HK}}_{s,t}(0) = 0$. If $d = \|q - q'\|_\infty$, then $q' - d\mathbf{1} \le q \le q' + d\mathbf{1}$; monotonicity and translation equivariance give

$$F^{\mathrm{HK}}_{s,t}(q') - d \le F^{\mathrm{HK}}_{s,t}(q) \le F^{\mathrm{HK}}_{s,t}(q') + d.$$

Thus $F^{\mathrm{HK}}_{s,t}$ is one-Lipschitz in $\|\cdot\|_\infty$ and $|F^{\mathrm{HK}}_{s,t}(q)| \le \|q\|_\infty$.

For the exact centered-heat fixed point $Q^\star_t$, bounded rewards imply

$$\begin{aligned}
\|Q^\star_t\|_\infty &= \|\mathcal{T}^{\mathrm{HK}}_t Q^\star_t\|_\infty \\
&\le R_{\max} + \gamma \sup_s |F^{\mathrm{HK}}_{s,t}(Q^\star_t(s,\cdot))| \\
&\le R_{\max} + \gamma \|Q^\star_t\|_\infty,
\end{aligned}$$

so $\|Q_t^\star\|_\infty \le R_{\max}/(1-\gamma)$ and, for every s,

$$\operatorname{osc}(Q_t^\star(s,\cdot)) \le \frac{2R_{\max}}{1-\gamma} \le S_{\mathrm{h}},$$
$$S_{\mathrm{h}} := \frac{2(R_{\max} + \gamma T_0 \log 2)}{1-\gamma}.$$

This verifies the span hypothesis in (29). The extra $T_0 \log 2$ is the retained-mass allowance of the generic localized cascade and is independent of K.

Next, we identify the Poisson tail controlled by the radius.

By (26),

$$\beta_r = 1 - \sum_{k=0}^{r} e^{-\theta}\frac{\theta^k}{k!} = \sum_{k=r+1}^{\infty} e^{-\theta}\frac{\theta^k}{k!}$$
$$= \Pr\{X \ge r+1\}.$$

Thus β_r is exactly the uniformization mass omitted after walk length r. For $u > 0$, Markov's inequality yields

$$\Pr\{X \ge r+1\} \le e^{-u(r+1)}\mathbb{E}[e^{uX}]$$
$$= \exp\{\theta(e^u - 1) - u(r+1)\}.$$

If $r+1 > \theta$, this is minimized at $u_\star = \log((r+1)/\theta) > 0$, giving

$$\beta_r \le \exp\left\{r + 1 - \theta - (r+1)\log\frac{r+1}{\theta}\right\}$$
$$= e^{-\theta}\left(\frac{e\theta}{r+1}\right)^{r+1}.$$

We then prove that r_ε enforces $\beta_{r_\varepsilon} \le b_\varepsilon$.

Construction gives $r_\varepsilon + 1 \ge 2e\theta$ and $r_\varepsilon + 1 > \theta$, so the preceding bound applies and

$$\beta_{r_\varepsilon} \le e^{-\theta}\left(\frac{e\theta}{r_\varepsilon+1}\right)^{r_\varepsilon+1} \le 2^{-(r_\varepsilon+1)}.$$

Also $r_\varepsilon + 1 \ge \log_2(1/b_\varepsilon)$, whence

$$2^{-(r_\varepsilon+1)} \le 2^{-\log_2(1/b_\varepsilon)} = b_\varepsilon, \qquad \boxed{\beta_{r_\varepsilon} \le b_\varepsilon}.$$

The term 1 in M_ε ensures $r_\varepsilon \ge 0$ even when both other thresholds are below one.

We now convert the Poisson-tail bound into an exact-heat fixed-point bound.

The standing assumptions $T_0 > 0$ and $\gamma \in (0,1)$ imply $S_{\mathrm{h}} > 0$, hence $e^{S_{\mathrm{h}}/T_0} - 1 > 0$ and the definition in (30) is well defined. It gives

$$b_\varepsilon \le \frac{e^{(1-\gamma)\varepsilon/(2T_0)} - 1}{e^{S_{\mathrm{h}}/T_0} - 1}.$$

Together with the Poisson-tail bound above,

$$1 + \beta_{r_\varepsilon}(e^{S_{\mathrm{h}}/T_0} - 1) \le e^{(1-\gamma)\varepsilon/(2T_0)}.$$

Applying Theorem 3 with $S = S_{\mathrm{h}}$,

$$\|V_{t,r_\varepsilon}^\star - V_t^\star\|_\infty \le \frac{T_0}{1-\gamma}\log\left(1 + \beta_{r_\varepsilon}(e^{S_{\mathrm{h}}/T_0} - 1)\right)$$
$$\le \frac{\varepsilon}{2}.$$

In particular, $|V_{t,r_\varepsilon}^\star(s_0) - V_t^\star(s_0)| \le \varepsilon/2$; this is a deterministic bias bound between the truncated- and exact-heat targets.

We next verify that the truncated heat kernel satisfies the cascade hypotheses.

Theorem 3 gives, for every j, a probability vector satisfying

$$\operatorname{supp}(H_{t,r_\varepsilon}(:,j)) = B_G(j, r_\varepsilon).$$

Choose the deterministic, value-independent set $I_{s,j} := B_G(j, r_\varepsilon)$. It contains the entire truncated column, so $H_{t,r_\varepsilon}(I_{s,j}, j) = 1 \geq 1/2$; hence Theorem 15 applies with no pruning error relative to the truncated operator. Moreover,

$$d_{\mathcal{I}} = \sup_{s,j} |I_{s,j}| = \max_j |B_G(j, r_\varepsilon)| = d_\varepsilon.$$

Normalized columns and zero offset give $F^{\mathrm{HK}}_{s,t,r_\varepsilon}(0) = 0$, hence $B_F = 0$, while

$$\bar{B}_Q = \frac{R_{\max} + \gamma T_0 \log 2}{1-\gamma} = \frac{S_{\mathrm{h}}}{2}.$$

Thus all cascade constants depend only on $(R_{\max}, \gamma, T_0)$ and the fixed schedules, with no further dependence on K, r_ε, or d_ε.

We continue to estimate the truncated-heat value statistically.

Run the cascade and outer median-of-means estimator with $\varepsilon_{\mathrm{stat}} = \varepsilon/2$. Choose ε_0 small enough that this lies in the cascade's admissible range whenever $0 < \varepsilon \leq \varepsilon_0$. Its estimate $\widehat{V}_{t,r_\varepsilon}(s_0)$ then satisfies, with probability at least $1-\delta$,

$$|\widehat{V}_{t,r_\varepsilon}(s_0) - V^\star_{t,r_\varepsilon}(s_0)| \leq \frac{\varepsilon}{2}.$$

On this event, the deterministic fixed-point bound and the triangle inequality give

$$\begin{aligned} |\widehat{V}_{t,r_\varepsilon}(s_0) - V^\star_t(s_0)| &\leq |\widehat{V}_{t,r_\varepsilon}(s_0) - V^\star_{t,r_\varepsilon}(s_0)| \\ &\quad + |V^\star_{t,r_\varepsilon}(s_0) - V^\star_t(s_0)| \leq \varepsilon. \end{aligned}$$

This proves (31); no confidence split is needed because the second term is deterministic.

We then derive the transition-oracle complexity.

The median-of-means consequence of Theorem 15, equivalently (71), yields

$$\begin{aligned} N_{\mathrm{heat}} &\leq \frac{C_{\mathrm{u}}}{(\varepsilon/2)^3} (C_{\mathrm{u}} d_\varepsilon)^{c_{\mathrm{u}}[1+\log\log(e+\bar{B}_Q/(\varepsilon/2))]} \log\frac{2}{\delta} \\ &= \frac{8C_{\mathrm{u}}}{\varepsilon^3} (C_{\mathrm{u}} d_\varepsilon)^{c_{\mathrm{u}}[1+\log\log(e+2\bar{B}_Q/\varepsilon)]} \log\frac{2}{\delta}. \end{aligned}$$

Here $\bar{B}_Q$ depends only on $(R_{\max}, \gamma, T_0)$. For $M_B := \max\{1, 2\bar{B}_Q\}$,

$$e + \frac{2\bar{B}_Q}{\varepsilon} \leq M_B \left(e + \frac{1}{\varepsilon}\right).$$

Since $\log(e + 1/\varepsilon) \geq 1$,

$$\log\log\left(e + \frac{2\bar{B}_Q}{\varepsilon}\right) \leq \log(1 + \log M_B) + \log\log(e + 1/\varepsilon).$$

Absorbing the resulting structural exponent shift, 8, and $(C_{\mathrm{u}}, c_{\mathrm{u}})$ into $C, c > 0$ gives

$$N_{\mathrm{heat}} \leq \frac{C}{\varepsilon^3} (C d_\varepsilon)^{c[1+\log\log(e+1/\varepsilon)]} \log\frac{2}{\delta}.$$

This is (32); C, c depend only on $(R_{\max}, \gamma, T_0)$ and the fixed cascade choices.

We now bound the selected radius.

Let

$$a := \frac{1-\gamma}{2T_0}, \qquad d_{\mathrm{h}} := e^{S_{\mathrm{h}}/T_0} - 1.$$

These are positive and independent of ε. Since $e^{a\varepsilon} - 1 = a\varepsilon + O(\varepsilon^2)$, for sufficiently small ε the second argument in the minimum defining b_ε is below $1/2$, and

$$b_\varepsilon = \frac{e^{a\varepsilon} - 1}{d_{\mathrm{h}}} = \frac{1-\gamma}{2T_0(e^{S_{\mathrm{h}}/T_0} - 1)}\varepsilon + O(\varepsilon^2) = \Theta(\varepsilon).$$

Therefore $\log_2(1/b_\varepsilon) = \log_2(1/\varepsilon) + O(1)$ and

$$\boxed{r_\varepsilon = O\left(1 + \nu_{\mathrm{u}} t + \log\frac{1}{\varepsilon}\right)}.$$

We next use polynomial graph-volume growth.

If $|B_G(j,m)| \le C_{\mathrm{vol}}(m+1)^{d_g}$ for every j, m, then

$$\boxed{d_\varepsilon \le \min\left\{K, C_{\mathrm{vol}}(r_\varepsilon + 1)^{d_g}\right\}}.$$

Indeed, this follows by applying the volume bound and the trivial bound K to each ball and maximizing over j. If $\nu_{\mathrm{u}} t = O(1)$ and all other structural parameters are fixed, then

$$r_\varepsilon = O(\log(1/\varepsilon)), \qquad d_\varepsilon = O((\log(1/\varepsilon))^{d_g}),$$

the latter being sub-K until the balls saturate the graph.

We continue to simplify the action-dependent complexity factor.

Write $L_\varepsilon := \log(1/\varepsilon)$ and $\ell_\varepsilon := \log L_\varepsilon$. Under the polynomial-growth bound above, $d_\varepsilon \le C_d L_\varepsilon^{d_g}$ for a structural C_d, so

$$\log(Cd_\varepsilon) = O(1 + d_g \ell_\varepsilon),$$
$$1 + \log\log(e + 1/\varepsilon) = O(1 + \ell_\varepsilon).$$

Consequently, the logarithm of the action-dependent factor in (32) is

$$c[1 + \log\log(e + 1/\varepsilon)]\log(Cd_\varepsilon) = O((d_g + 1)\ell_\varepsilon^2).$$

For the conventional $d_g \ge 1$, this is $O(d_g[\log\log(1/\varepsilon)]^2)$; hence the factor equals

$$\exp\left(O(d_g[\log\log(1/\varepsilon)]^2)\right).$$

For fixed $d_g \ge 1$, $d_g[\log\log(1/\varepsilon)]^2/\log(1/\varepsilon) \to 0$, so this is $(1/\varepsilon)^{o(1)}$. Substitution in (32) proves

$$\boxed{N_{\mathrm{heat}} = (1/\varepsilon)^{3+o(1)} \log\frac{2}{\delta}}.$$

If the degenerate $d_g = 0$ is allowed, retain the intermediate $O((d_g + 1)[\log\log(1/\varepsilon)]^2)$ bound; the near-cubic conclusion is unchanged.

Finally, we record local computation, memory, and scope.

For each sampled anchor j, the recurrence $v_{k+1} = P_G v_k$ constructs the column of H_{t,r_ε} while visiting only $B_G(j, r_\varepsilon)$. Uniformly for $r_\varepsilon \ge 0$, its arithmetic and working-memory costs are

$$O((r_\varepsilon + 1)(\Delta_G + 1)d_\varepsilon), \qquad O((\Delta_G + 1)d_\varepsilon),$$

respectively; for $r_\varepsilon \ge 1$, the first is equivalently $O(r_\varepsilon(\Delta_G + 1)d_\varepsilon)$. Under fixed $\nu_{\mathrm{u}} t$ and polynomial growth, these become

$$O((\Delta_G + 1)\log^{d_g+1}(1/\varepsilon)), \qquad O((\Delta_G + 1)\log^{d_g}(1/\varepsilon)),$$

respectively. These graph-computation costs are separate from the transition-oracle count in (32).

Finally, the algorithm computes the normalized finite Poisson head H_{t,r_ε}, not the entries of exact H_t inside the ball: walks longer than r_ε may return to it. The Poisson-tail bound pays for this difference, making the guarantee valid for the exact centered-heat target. The proved target is the scalar regularized root value, not root-action identification, simple regret, or a learned policy. □

G.7 Full recursive estimator: value bounds, bias, moments, and cost

For completeness, this subsection records the constants suppressed in the main statement of Theorem 15. The proof has six parts: recursive construction, a finite-depth base case, baseline-error control, bias control through Taylor cancellation, third-moment control, and cost accounting with outer confidence amplification. Define

$$\begin{aligned} B_F^{\mathcal{I}} &= \sup_s |F_s^{\mathcal{I}}(0)|, \\ B_Q &= \frac{R_{\max} + \gamma B_F}{1-\gamma}, \quad B_Q^{\mathcal{I}} = \frac{R_{\max} + \gamma B_F^{\mathcal{I}}}{1-\gamma}, \\ B_V^{\mathcal{I}} &= B_F^{\mathcal{I}} + B_Q^{\mathcal{I}}. \end{aligned} \tag{366}$$

The retained-mass condition gives

$$\begin{aligned} B_F^{\mathcal{I}} &\le \bar{B}_F := B_F + T_0 \log 2, \\ B_Q^{\mathcal{I}} &\le \bar{B}_Q := \frac{R_{\max} + \gamma \bar{B}_F}{1-\gamma}. \end{aligned} \tag{367}$$

Indeed, if $m_{s,j} = w^{(s,j)}(I_{s,j}) \ge 1/2$, then $f_{s,j}^{\mathcal{I}}(0) = f_{s,j}(0) + T_0 \log m_{s,j}$. Averaging over j gives $|F_s^{\mathcal{I}}(0)| \le |\mathcal{F}_s(0)| + T_0 \log 2$, which is the first bound in (367); the second follows by solving the usual discounted value envelope. There are constants $c_b, c_q, c_N > 0$, an integer $m_2 < \infty$, and $a \in (1, 1/\gamma)$, all fixed independently of $(K, d_{\mathcal{I}}, \eta, \varepsilon, \delta)$ and uniformly over admissible active-set families, such that only coordinate baselines are projected to $[-B_Q^{\mathcal{I}}, B_Q^{\mathcal{I}}]$ and the returned sample is unclipped. The deterministic worst-case call count obeys

$$\begin{aligned} C(\eta) \le\ & c d_{\mathcal{I}} \eta^{-2/3} \big[1 + C(c_b \eta^{1/3})\big] + 1 + C(a\eta) \\ & + 4 m_2 \big[1 + C(c_q \eta^{2/3})\big] + c, \end{aligned} \tag{368}$$

which resolves to (108). The proof below constructs the coarse routine and verifies the bias and third-moment interfaces rather than assuming them.

Complete planner interface. The theorem proves the existence of fixed structural constants $\eta_0, C_{\mathrm{b}}, C_3, c_b, c_q, c_N, m_2$, rather than universal numerical values for every problem class. Algorithm 2 therefore treats these analysis-supplied constants as offline inputs. This keeps the pseudocode faithful to the theorem and makes every online step, including confidence amplification, explicit. All cascade trees drawn in different calls are independent.

Table 9: Planner-specific notation used by Algorithms 2–9.

Symbol	Meaning
$T_0 = \tau\lambda$	effective soft-backup temperature
S_Q	valid upper bound on the action-value span
ξ, α_ε	one-step pruning budget and allowed omitted kernel mass
$I_{s,j}, \mathcal{I}$	retained actions for anchor j and the full fixed active-set family
$d_{\mathcal{I}}$	largest retained-set size, $\sup_{s,j} \lvert I_{s,j}\rvert$
u, η	outer statistical tolerance and one-cascade bias tolerance
$Y_\eta^{\mathcal{I}}(s)$	one recursive localized value sample at state s
$\eta_0, C_{\mathrm{b}}, C_3$	coarse threshold, bias constant, and uniform third-moment bound
r_η, N_η	fine-scale radius $\eta^{1/3}$ and baseline batch size
$\eta_b, \eta_\ell, \eta_q$	child tolerances for baseline, linear, and quadratic terms
m_2	fixed batch size for each paired quadratic correction
$B_Q^{\mathcal{I}}, B_V^{\mathcal{I}}$	action-value and state-value projection envelopes
$H_0, m_\star$	coarse-tree horizon and explicit common coarse batch size

The choice of η makes the cascade bias at most $u/2$. Since $\mathbb{E}|Y_\eta|^3 \le C_3$, its variance is at most $C_3^{2/3}$; the displayed block size makes a block mean miss $\mathbb{E}Y_\eta$ by more than $u/2$ with probability at most $1/4$. The odd number of blocks then gives failure probability at most δ by the usual median-of-means bound. Pruning uses the other $\varepsilon/2$ of the error budget.

G.7.1 Proof of Theorem 15

We first specify the recursion and schedules. Fix $\mathcal{I}$, suppress it on Y_η, write $Q^\star = Q^\star_{\mathcal{I}}$ and $V^\star = V^\star_{\mathcal{I}}$. Recall the restricted weights, recursive schedules, and centered local contribution from (43), (45), and (44). Each oracle call at (s,i) returns an independent (R,S') with $|R| \le R_{\max}$, $\mathbb{E}R = r(s,i)$, and $S' \sim P(\cdot \mid s,i)$; all calls and recursive subtrees are fresh. Constants c_b, c_q, η_0 are chosen below; $\eta_0 < \min\{c_b^{3/2}, c_q^3\}$ makes every child tolerance larger than η below η_0, proving termination.

For deterministic q, $\mathbb{E}_{J\sim\mu_s} \bar{f}^{\mathcal{I}}_{s,J}(q) = F^{\mathcal{I}}_s(q)$, and monotonicity plus translation equivariance gives $\min_i q_i \le f^{\mathcal{I}}_{s,j}(q) - f^{\mathcal{I}}_{s,j}(0) \le \max_i q_i$; hence $|\bar{f}^{\mathcal{I}}_{s,j}(q)| \le B^{\mathcal{I}}_F + B^{\mathcal{I}}_Q = B^{\mathcal{I}}_V$ on $[-B^{\mathcal{I}}_Q, B^{\mathcal{I}}_Q]^{I_{s,j}}$. The zero terms are known and precomputable: both vanish for centered dense heat-kernel LOT, while a retained column has $f^{\mathcal{I}}_{s,j}(0) = T_0 \log w^{(s,j)}(I_{s,j}) \in [-T_0 \log 2, 0]$. Unlike the dense cascade of Dam (2026), Algorithm 3 is unclipped, uses asymmetric child tolerances and predetermined retained sets, and proves bias, moments, termination, and cost jointly rather than assuming a recursive interface.

Given $(J, \widehat{Q}, A, A')$, the two batches are conditionally independent; primes denote independent subtrees, and only $\widehat{Q}$ is projected.

Next, we establish the envelopes and a genuine base case. Translation equivariance and 1-Lipschitzness imply $|F^{\mathcal{I}}_s(q)| \le B^{\mathcal{I}}_F + \|q\|_\infty$, so the fixed point gives $\|Q^\star\|_\infty \le R_{\max} + \gamma(B^{\mathcal{I}}_F + \|Q^\star\|_\infty)$, hence $\|Q^\star\|_\infty \le B^{\mathcal{I}}_Q$ and $|V^\star(s)| \le B^{\mathcal{I}}_V$.

Construct the base routine as follows. Put $\bar{B}_V = \bar{B}_F + \bar{B}_Q$, choose H_0 with $\gamma^{H_0}\bar{B}_V \le \eta_0/8$, and, from terminal value zero, perform depth-H_0 Monte Carlo backups. At each node average anchors $J \sim \mu_s$; for each anchor query only $I_{s,J}$ using fresh transitions/subtrees, evaluate $\bar{f}^{\mathcal{I}}_{s,J}$, and project the state estimate to $[-B^{\mathcal{I}}_V, B^{\mathcal{I}}_V]$.

Set $\kappa_0 = (1-\gamma)\eta_0/16$ and $\zeta_0 = \eta_0/[16(\bar{B}_V+1)]$. Hoeffding and a union bound over at most $(1 + m_J d_{\mathcal{I}} m_P)^{H_0}$ nodes (all sampled quantities are bounded) give finite batches with every mean within κ_0 of its conditional mean and failure probability at most ζ_0; e.g., $m_J = m_P = \lceil (c_\star(1+d_{\mathcal{I}}))^{2H_0} \rceil$ works for large structural $c_\star$. Since H_0, η_0 and the envelopes are structural, the cost is at most $(c(1+d_{\mathcal{I}}))^{p_0}$ for fixed c, p_0. The explicit batch choice and finite-depth estimator are given in (376) and Algorithm 9. The batch condition has a finite solution because its exponential term dominates its polynomial term; set $m_J = m_P = m_\star$. This routine is $\mathsf{Coarse}^{\mathcal{I}}_{\eta_0}(s)$ in Algorithm 3. On the good event, backward induction and contraction give error $\gamma^{H_0}\bar{B}_V + 2\kappa_0/(1-\gamma) \le \eta_0/4$ after the fixed rescaling; on failure, projection bounds it by $2\bar{B}_V$. Thus $\mathbb{E}|\mathsf{Coarse}^{\mathcal{I}}_{\eta_0}(s) - V^\star(s)| \le \eta_0/4 + 2\bar{B}_V\zeta_0 \le 3\eta_0/8 < \eta_0$, and Jensen gives

$$|\mathbb{E}\mathsf{Coarse}^{\mathcal{I}}_{\eta_0}(s) - V^\star(s)| \le \eta_0, \qquad |\mathsf{Coarse}^{\mathcal{I}}_{\eta_0}(s)| \le B^{\mathcal{I}}_V. \tag{369}$$

Its deterministic cost is at most $(c(1+d_{\mathcal{I}}))^{p_0}$; any $C_{\mathrm{b}} \ge 1$ covers its rescaled bias and $B^{\mathcal{I}}_V$ its third moment.

We then bound the baseline error. For child tolerance θ, define $X_\theta = R + \gamma Y_\theta(S')$, $e_\theta(s,i) = \mathbb{E}X_\theta(s,i) - Q^\star(s,i)$, $b(\theta) = \sup_s |\mathbb{E}Y_\theta(s) - V^\star(s)|$, and $x(\theta) = \sup_s \|Y_\theta(s)\|_3$; then $\|e_\theta\|_\infty \le \gamma b(\theta)$. Assuming at larger child tolerances that $b(\theta) \le C_{\mathrm{b}}\theta$ and $x(\theta) \le S_3$, the X_{η_b} have uniformly bounded third moments. The $p=3$ Rosenthal inequality and projection nonexpansiveness around $Q^\star_i \in [-B^{\mathcal{I}}_Q, B^{\mathcal{I}}_Q]$ give, for large c_N and small c_b,

$$\|\widehat{Q}_i - Q^\star_i\|_3 \le c_X N_\eta^{-1/2} + \gamma C_{\mathrm{b}}\eta_b \le c_\Delta r_\eta, \tag{370}$$

hence, without a good event or recursive confidence allocation,

$$\sup_{s,j}\sup_{i\in I_{s,j}} \mathbb{E}|\widehat{Q}_i - Q^\star_i|^3 \le c_\Delta^3 \eta. \tag{371}$$

We now carry out the bias induction. Let $\Delta_i = Q^\star(s,i) - \widehat{Q}_i$ for $i \in I_{s,J}$. Lemma 3, (43), (371), and $\sum_i w_i^{\mathcal{I},(s,j)} = 1$ bound the expected Taylor remainder by $c_R\eta$. Also, (306) and Hölder imply, after averaging over $\widehat{Q}$,

$$\mathbb{E}[|\Delta_A| \mid J] \le e^{2B^{\mathcal{I}}_Q/T_0} \sum_i w_i^{\mathcal{I},(s,J)} \mathbb{E}|\Delta_i| \le c_A r_\eta, \tag{372}$$

for $A \sim p^{\mathcal{I},(s,J)}(\widehat{Q})$. Given $(J, \widehat{Q})$, the linear correction has mean $p^{\mathcal{I},(s,J)}(\widehat{Q})^\top(\Delta + e_{\eta_\ell})$, differing from the ideal term by at most $\gamma b(a\eta)$.

Conditional batch independence makes the quadratic mean $(\Delta_A - \Delta_{A'} + e_{\eta_q}(A) - e_{\eta_q}(A'))^2/(4T_0)$. With $d = \Delta_A - \Delta_{A'}$ and $u = e_{\eta_q}(A) - e_{\eta_q}(A')$, the ideal mean is $d^2/(4T_0)$, while (372) gives $\mathbb{E}|d| \le 2c_A r_\eta$ and $|u| \le 2\gamma b(c_q\eta^{2/3})$. Therefore

$$\begin{aligned}\mathbb{E}\left|\frac{(d+u)^2 - d^2}{4T_0}\right| &\le \frac{\mathbb{E}|du|}{2T_0} + \frac{\mathbb{E}u^2}{4T_0} \\ &\le \frac{2c_A\gamma}{T_0} r_\eta b(c_q\eta^{2/3}) + \frac{\gamma^2}{T_0} b(c_q\eta^{2/3})^2.\end{aligned}$$

Writing the columnwise Taylor expansion as

$$f^{\mathcal{I}}_{s,J}(Q^\star) = f^{\mathcal{I}}_{s,J}(\widehat{Q}) + p^{\mathcal{I},(s,J)}(\widehat{Q})^\top\Delta + \tfrac{1}{2}\Delta^\top\nabla^2 f^{\mathcal{I}}_{s,J}(\widehat{Q})\Delta + \mathcal{R}_{s,J},$$

the centered ideal estimator satisfies, conditionally on $(J, \widehat{Q})$,

$$\begin{aligned}&\bar{f}^{\mathcal{I}}_{s,J}(\widehat{Q}) + p^{\mathcal{I},(s,J)}(\widehat{Q})^\top\Delta + \mathbb{E}[d^2/(4T_0) \mid J, \widehat{Q}] \\ &\quad = F^{\mathcal{I}}_s(0) + f^{\mathcal{I}}_{s,J}(Q^\star) - f^{\mathcal{I}}_{s,J}(0) - \mathcal{R}_{s,J}.\end{aligned}$$

Averaging $J \sim \mu_s$ leaves $F^{\mathcal{I}}_s(Q^\star) = V^\star(s)$ minus the remainder. Hence

$$b(\eta) \le c_0\eta + \gamma b(a\eta) + c_1 r_\eta b(c_q\eta^{2/3}) + c_2 b(c_q\eta^{2/3})^2. \tag{373}$$

With $q_b = \gamma a < 1$, induction in (373) yields $b(\eta) \le [c_0 + q_b C_{\mathrm{b}} + c_1 c_q C_{\mathrm{b}} + c_2 c_q^2 C_{\mathrm{b}}^2 \eta_0^{1/3}]\eta$.

For a noncircular choice, fix the provisional S_3 below, choose c_N, set $\bar{c}_\Delta = c_X c_N^{-1/2} + 1$, and let $\bar{c}_0$ be the remainder constant from Lemma 3 at $c_\Delta = \bar{c}_\Delta$. Take $C_{\mathrm{b}} \ge \max\{1, 4\bar{c}_0/(1-q_b)\}$ and $c_b \le (\gamma C_{\mathrm{b}})^{-1}$; then $c_\Delta \le \bar{c}_\Delta$ and $c_0 \le (1-q_b)C_{\mathrm{b}}/4$. Choose c_q so that $c_1 c_q \le (1-q_b)/4$, then reduce η_0 until $c_2 c_q^2 C_{\mathrm{b}} \eta_0^{1/3} \le (1-q_b)/4$, the conditions after (45) hold, and fine-regime child mean biases are at most one. The coarse routine remains available at fixed η_0, and $b(\eta) \le C_{\mathrm{b}}\eta$ closes.

We next establish a uniform third-moment bound. The local term is bounded by $B^{\mathcal{I}}_V$ and $\|G_{\mathrm{lin}}\|_3 \le R_{\max} + B^{\mathcal{I}}_Q + \gamma x(a\eta)$. For $\mathcal{H} = \sigma(J, \widehat{Q}, A, A')$, each $D_{\ell h}$ has bounded conditional mean D_0 because $Q^\star, \widehat{Q}$ and child biases are bounded. Conditional Rosenthal gives

$$\|\overline{D}_\ell\|_{3|\mathcal{H}} \le D_0 + \frac{c_D}{\sqrt{m_2}}(1 + \gamma x(c_q\eta^{2/3})). \tag{374}$$

Writing $M_\ell = \mathbb{E}[|\overline{D}_\ell|^3 \mid \mathcal{H}]$, batch independence gives $\mathbb{E}[|\overline{D}_1\overline{D}_2|^3 \mid \mathcal{H}] = M_1 M_2$, avoiding sixth moments. Thus

$$\begin{aligned}x(\eta) \le A_0 + \gamma x(a\eta) + \frac{A_1}{\sqrt{m_2}}(1 + x(c_q\eta^{2/3})) \\ + \frac{A_2}{m_2}(1 + x(c_q\eta^{2/3}))^2.\end{aligned} \tag{375}$$

Choose $S_3 \ge 4A_0/(1-\gamma)$, then fixed m_2 so the last two terms in (375) are at most $(1-\gamma)S_3/2$ for child norms $\le S_3$. Simultaneous induction gives $\sup_{\eta,s}\|Y_\eta(s)\|_3 \le S_3$ and the second claim of (107) with $C_3 = S_3^3$. Compatible choices are a, A_0; then S_3; then $c_N, C_{\mathrm{b}}, c_b, c_q, \eta_0$; finally m_2, so the provisional radius is no extra assumption.

Finally, we bound the cost and confidence amplification. All loop sizes are bounded by $d_{\mathcal{I}}$ before sampling, so Algorithm 3 has worst-case deterministic cost

$$C(\eta) \le d_{\mathcal{I}} N_\eta[1 + C(\eta_b)] + 1 + C(\eta_\ell) + 4m_2[1 + C(\eta_q)] + c,$$

which is (368). Take a nonincreasing integer ℓ_c, zero in the base regime and, below η_0, satisfying $\ell_c(c_b\eta^{1/3}) \le \ell_c(\eta) - 1$, $\ell_c(a\eta) \le \ell_c(\eta)$, and $\ell_c(c_q\eta^{2/3}) \le \ell_c(\eta)$. Iterating $x \mapsto c_b x^{1/3}$ gives

$$\ell_c(\eta) \le c_\ell + c'_\ell \log\left(2 + \log_+\frac{\eta_0}{\eta}\right), \qquad \log_+ x = \max\{0, \log x\}.$$

Compare with $G(\eta) = A\eta^{-1}D^{\ell_c(\eta)+p_0}$, $D = d_0(1+d_{\mathcal{I}})$, where p_0 absorbs the coarse-base polynomial. The baseline, linear, and quadratic recursive ratios are at most $cd_{\mathcal{I}}/(c_bD)$, $1/a$, and $(4m_2/c_q)\eta^{1/3}$. Since $a > 1$, enlarge d_0 and then reduce fixed η_0 until their sum is below one; enlarge A for ceilings, nonrecursive calls, and the compact boundary band. Strong induction gives $C(\eta) \le G(\eta)$, i.e. (108).

Finally $C_3 < \infty$ gives uniform variance. For tolerance u, take $\eta = u/(2C_{\mathrm{b}})$ and a median of block means from $O(u^{-2}\log(2/\delta))$ independent $Y_\eta^{\mathcal{I}}(s_0)$ samples. With probability $1-\delta$, stochastic error is at most $u/2$, and bias is at most $u/2$. This proves the localized result. No internal confidence budget depends on δ: deterministic cost, bias and moments control the recursion, and the single outer amplification yields one $\log(1/\delta)$ factor rather than recursive confidence splitting. This proves Theorem 15. □

G.7.2 Proof of Theorem 2

The cascade theorem assumes $0 < \eta \le \eta_0$. Set

$$\varepsilon_0^{\mathrm{cor}} := \min\{1, 4C_{\mathrm{b}}\eta_0\}, \qquad 0 < \varepsilon \le \varepsilon_0^{\mathrm{cor}}, \qquad \delta \in (0,1),$$

and recall

$$\bar{B}_F = B_F + T_0\log 2, \quad \bar{B}_Q = \frac{R_{\max} + \gamma\bar{B}_F}{1-\gamma}, \quad S_Q = 2\bar{B}_Q.$$

We first select deterministic retained sets and bound their fixed-point error, then estimate the restricted fixed point using independent cascade samples and one outer median-of-means construction.

We first verify that $S_Q = 2\bar{B}_Q$ is a valid span envelope. One-Lipschitzness of the normalized-kernel aggregator and bounded rewards give

$$\begin{aligned}
|\mathcal{F}_s(q)| &\le |\mathcal{F}_s(0)| + \|q\|_\infty \le B_F + \|q\|_\infty,\\
\|Q^\star_{\mathcal{F}}\|_\infty &\le R_{\max} + \gamma(B_F + \|Q^\star_{\mathcal{F}}\|_\infty)\\
\|Q^\star_{\mathcal{F}}\|_\infty &\le \frac{R_{\max} + \gamma B_F}{1-\gamma} \le \bar{B}_Q,\\
\operatorname{osc}(Q^\star_{\mathcal{F}}(s,\cdot)) &\le 2\|Q^\star_{\mathcal{F}}\|_\infty \le S_Q \quad (\forall s).
\end{aligned}$$

Thus the span hypothesis of Theorem 14 holds. The $T_0\log 2$ allowance covers active sets retaining at least half of each base column.

Next, we record the accuracy-dependent tail budget. As in (2), define

$$\begin{aligned}
z_\varepsilon &:= \exp\left(\frac{(1-\gamma)\varepsilon}{2T_0}\right) - 1,\\
\alpha_\varepsilon &:= \min\left\{\frac{1}{2}, \frac{z_\varepsilon}{e^{S_Q/T_0} + z_\varepsilon}\right\}.
\end{aligned}$$

Since $T_0 > 0$, $\gamma < 1$, and $\varepsilon > 0$, we have $z_\varepsilon > 0$, $0 < z_\varepsilon/(e^{S_Q/T_0} + z_\varepsilon) < 1$, and $0 < \alpha_\varepsilon \le 1/2$. Thus retaining $1-\alpha_\varepsilon$ mass retains at least half of the column, as required by Theorem 15.

We continue to choose deterministic minimum-cardinality active sets. For each (s,j), order actions by decreasing $w_i^{(s,j)}$ with a fixed deterministic tie rule, and let $I_{s,j}$ be the shortest prefix with $w^{(s,j)}(I_{s,j}) \ge 1-\alpha_\varepsilon$. These sets depend only on the known geometry, are fixed before transition sampling, and never depend on sampled or estimated values. A top-weight prefix has minimum cardinality; hence Definition 1 yields

$$\begin{aligned}
|I_{s,j}| &= \min\{|I| : w^{(s,j)}(I) \ge 1-\alpha_\varepsilon\}\\
&\le d_{\ker}(\alpha_\varepsilon),\\
d_{\mathcal{I}} := \sup_{s,j}|I_{s,j}| &\le d_{\ker}(\alpha_\varepsilon).
\end{aligned}$$

Writing $\alpha_{s,j} := w^{(s,j)}(I_{s,j}^c)$ gives

$$0 \le \alpha_{s,j} \le \alpha_\varepsilon \le \tfrac{1}{2} < 1.$$

Thus every restricted log partition is finite, even if the original column has exact zero entries.

We then invert the pruning-error formula. Let $E_Q := e^{S_Q/T_0}$ and $\alpha^\star := z_\varepsilon/(E_Q + z_\varepsilon)$. Since $\alpha_\varepsilon \le \alpha^\star$ and $(x/(1-x))' = (1-x)^{-2} > 0$ on $[0,1)$,

$$E_Q \frac{\alpha_\varepsilon}{1-\alpha_\varepsilon} \le E_Q \frac{\alpha^\star}{1-\alpha^\star} = E_Q \frac{z_\varepsilon/(E_Q+z_\varepsilon)}{E_Q/(E_Q+z_\varepsilon)} = z_\varepsilon.$$

Equivalently,

$$\boxed{e^{S_Q/T_0} \frac{\alpha_\varepsilon}{1-\alpha_\varepsilon} \le z_\varepsilon},$$

$$\alpha^\star = \frac{e^{\xi/T_0}-1}{e^{S_Q/T_0}+e^{\xi/T_0}-1}, \qquad \xi = \frac{(1-\gamma)\varepsilon}{2}.$$

This is the omitted-mass condition in (19); the cap $1/2$ enforces the cascade retained-mass condition.

We now bound every local pruning error. Lemma 4 defines

$$\zeta_{s,j}(S_Q) := e^{S_Q/T_0} \frac{\alpha_{s,j}}{1-\alpha_{s,j}} \le e^{S_Q/T_0} \frac{\alpha_\varepsilon}{1-\alpha_\varepsilon} \le z_\varepsilon.$$

Thus, uniformly for $\mathrm{osc}(q) \le S_Q$,

$$0 \le f_{s,j}(q) - f^{\mathcal{I}}_{s,j}(q) \le T_0 \log(1+\zeta_{s,j}(S_Q)) \le T_0 \log(1+z_\varepsilon).$$

Here $f^{\mathcal{I}}_{s,j}$ uses the raw retained partition in (102). If retained weights are normalized for policy sampling, its value must include $T_0 \log w^{(s,j)}(I_{s,j})$; omitting this centering term changes the operator and invalidates the uniform pruning guarantee.

We next bound the state-aggregator gap. By (105), the local pruning bound above, and $\sum_j \mu_s(j) = 1$,

$$\Delta_{\mathcal{I}}(S_Q) = T_0 \sup_s \sum_j \mu_s(j) \log(1+\zeta_{s,j}(S_Q))$$

$$\le T_0 \log(1+z_\varepsilon) = \frac{(1-\gamma)\varepsilon}{2}.$$

We continue to propagate pruning through the Bellman fixed point. By the span-envelope bound above and Theorem 14,

$$0 \le V^\star_{\mathcal{F}}(s_0) - V^\star_{\mathcal{I}}(s_0) \le \frac{\Delta_{\mathcal{I}}(S_Q)}{1-\gamma} \le \frac{\varepsilon}{2}.$$

Thus pruning uses at most half the root-value error budget. The analogous Q-fixed-point bound has the smaller factor $\gamma/(1-\gamma)$, whereas the state value targeted here has factor $1/(1-\gamma)$.

We then choose the inner cascade accuracy. Let $u := \varepsilon/2$, $Y_\eta := Y^{\mathcal{I}}_\eta(s_0)$, and $m_\eta := \mathbb{E}Y_\eta$. For $0 < \eta \le \eta_0$, (107) gives

$$|m_\eta - V^\star_{\mathcal{I}}(s_0)| \le C_{\mathrm{b}}\eta, \qquad \mathbb{E}|Y_\eta|^3 \le C_3.$$

Choose $\eta := u/(2C_{\mathrm{b}}) = \varepsilon/(4C_{\mathrm{b}})$. The definition of $\varepsilon_0^{\mathrm{cor}}$ ensures $\eta \le \eta_0$, and hence $|m_\eta - V^\star_{\mathcal{I}}(s_0)| \le u/2$.

We now derive a variance bound from the third moment. Lyapunov's inequality yields

$$\mathrm{Var}(Y_\eta) \le \mathbb{E}[Y_\eta^2] \le (\mathbb{E}|Y_\eta|^3)^{2/3} \le C_3^{2/3}.$$

For each block b, take independent copies $Y_{b,1}, \dots, Y_{b,M}$ and set

$$\overline{Y}_b := M^{-1} \sum_{m=1}^{M} Y_{b,m}, \qquad M := \left\lceil \frac{16 C_3^{2/3}}{u^2} \right\rceil.$$

Then $\mathrm{Var}(\overline{Y}_b) \le C_3^{2/3}/M$, so Chebyshev gives

$$\Pr\{|\overline{Y}_b - m_\eta| > u/2\} \le \frac{4C_3^{2/3}}{Mu^2} \le \frac{1}{4}.$$

Each block is therefore good with probability at least $3/4$.

We next amplify confidence using one outer median. Take the odd number of independent blocks

$$B := 2\left\lceil 4\log\frac{2}{\delta}\right\rceil + 1, \qquad \widehat{V}_{\mathcal{I}}(s_0) := \operatorname{median}\{\overline{Y}_1, \dots, \overline{Y}_B\}.$$

If G_b indicates a bad block, then the G_b are independent, $\mathbb{E}G_b \le 1/4$, and median failure requires at least $B/2$ bad blocks. Hoeffding's inequality gives

$$\begin{aligned}\Pr\left\{\sum_{b=1}^{B} G_b \ge \frac{B}{2}\right\} &\le \exp\left[-2B\left(\frac{1}{2}-\frac{1}{4}\right)^2\right] \\ &= e^{-B/8} \le \frac{\delta}{2} \le \delta.\end{aligned}$$

Thus, with probability at least $1-\delta$, $|\widehat{V}_{\mathcal{I}}(s_0) - m_\eta| \le u/2$, and

$$\begin{aligned}|\widehat{V}_{\mathcal{I}}(s_0) - V^\star_{\mathcal{I}}(s_0)| &\le |\widehat{V}_{\mathcal{I}}(s_0) - m_\eta| \\ &\quad + |m_\eta - V^\star_{\mathcal{I}}(s_0)| \le u/2 + u/2 = u.\end{aligned}$$

All confidence amplification occurs in this one outer step; no confidence budget is divided recursively across the cascade tree.

We continue to combine statistical and pruning errors. On the preceding probability-$1-\delta$ event,

$$\begin{aligned}|\widehat{V}_{\mathcal{I}}(s_0) - V^\star_{\mathcal{F}}(s_0)| &\le |\widehat{V}_{\mathcal{I}}(s_0) - V^\star_{\mathcal{I}}(s_0)| \\ &\quad + |V^\star_{\mathcal{I}}(s_0) - V^\star_{\mathcal{F}}(s_0)| \\ &\le \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

Hence the output estimates the *unpruned* regularized root value with error at most ε and probability at least $1-\delta$.

We then bound the number of outer cascade samples. The number of independent recursive samples is MB. Since $\varepsilon \le \varepsilon_0^{\mathrm{cor}} \le 1$ and $u \le 1$, the ceiling bounds give, for constants depending only on C_3,

$$\begin{aligned}M &\le c_M u^{-2}, & B &\le c_B \log\frac{2}{\delta}, \\ MB &\le c_{\mathrm{out}} u^{-2}\log\frac{2}{\delta} \le \frac{c'_{\mathrm{out}}}{\varepsilon^2}\log\frac{2}{\delta}.\end{aligned}$$

Each sample uses a fresh recursion tree and at most $C(\eta)$ transition-oracle calls.

We now substitute the one-sample cascade cost. Equation (108) states

$$C(\eta) \le \frac{c_0}{\eta}(c_0 d_{\mathcal{I}})^{c_1[1+\log\log(e+\bar{B}_Q/\eta)]}.$$

Using $\eta = \varepsilon/(4C_{\mathrm{b}})$ and $d_{\mathcal{I}} \le d_{\mathrm{ker}}(\alpha_\varepsilon)$,

$$C(\eta) \le \frac{4c_0 C_{\mathrm{b}}}{\varepsilon}\big(c_0 d_{\mathrm{ker}}(\alpha_\varepsilon)\big)^{c_1[1+\log\log(e+4C_{\mathrm{b}}\bar{B}_Q/\varepsilon)]}.$$

For explicit constant absorption, put $M_{\mathrm{b}} := \max\{1, 4C_{\mathrm{b}}\}$. Then

$$e + \frac{4C_{\mathrm{b}}\bar{B}_Q}{\varepsilon} \le M_{\mathrm{b}}\left(e + \frac{\bar{B}_Q}{\varepsilon}\right),$$

and, since $\log(e + \bar{B}_Q/\varepsilon) \ge 1$,

$$\begin{aligned}\log\log\left(e + \frac{4C_{\mathrm{b}}\bar{B}_Q}{\varepsilon}\right) &\le \log\left(\log M_{\mathrm{b}} + \log(e + \bar{B}_Q/\varepsilon)\right) \\ &\le \log(1 + \log M_{\mathrm{b}}) \\ &\quad + \log\log(e + \bar{B}_Q/\varepsilon).\end{aligned}$$

Absorbing fixed factors into structural constants and increasing the base constant to at least one gives $c_2, c_3 > 0$ such that

$$C(\eta) \le \frac{c_2}{\varepsilon}\big(c_2 d_{\ker}(\alpha_\varepsilon)\big)^{c_3[1+\log\log(e+\bar{B}_Q/\varepsilon)]}.$$

Multiplication by MB and one further increase of c_2, c_3 yields

$$n(\varepsilon,\delta) \le \frac{c_2}{\varepsilon^3}\big(c_2 d_{\ker}(\alpha_\varepsilon)\big)^{c_3[1+\log\log(e+\bar{B}_Q/\varepsilon)]}\log\frac{2}{\delta},$$

which is (23).

Finally, we cover locality and the bounded-dimension case. Every recursive transition query uses an action in the corresponding fixed $I_{s,j}$, so none lies outside the retained neighborhood of its sampled anchor; the union over anchors may nevertheless be the full action set. Transition-oracle locality alone does not imply sub-K arithmetic: finding top-weight sets and evaluating their mass without scanning all actions also requires Assumption 2.

If $d_{\ker}(\alpha_\varepsilon)$ stays uniformly bounded as $\varepsilon \downarrow 0$, the action-dependent factor in (23) is

$$\begin{aligned}&\big(c_2 d_{\ker}(\alpha_\varepsilon)\big)^{c_3[1+\log\log(e+\bar{B}_Q/\varepsilon)]}\\ &\quad= \exp(O(\log\log(1/\varepsilon))) = (1/\varepsilon)^{o(1)}.\end{aligned}$$

so $n(\varepsilon,\delta) = (1/\varepsilon)^{3+o(1)}\log(2/\delta)$. Equivalently this is $\varepsilon^{-3+o(1)}\log(1/\delta)$ after redefining the signed $o(1)$ exponent. The guarantee concerns only the scalar regularized root value of the chosen unpruned normalized-kernel Bellman operator: it neither identifies a root action, proves simple regret, bounds a learned policy's return, nor removes the separate regularization bias to the hard-max Bellman target. □

G.8 Complete specification of LocalLOTPlan

Fix the active-set family $\mathcal{I} = \{I_{s,j}\}_{s,j}$ from Theorem 2. Use the restricted weights and centered local contribution in (43) and (44). Let

$$B_F^{\mathcal{I}} = \sup_s |F_s^{\mathcal{I}}(0)|, \quad B_Q^{\mathcal{I}} = \frac{R_{\max} + \gamma B_F^{\mathcal{I}}}{1-\gamma}, \quad B_V^{\mathcal{I}} = B_F^{\mathcal{I}} + B_Q^{\mathcal{I}}.$$

The retained-mass condition implies $B_F^{\mathcal{I}} \le \bar{B}_F$ and $B_Q^{\mathcal{I}} \le \bar{B}_Q$. The proof fixes structural constants $\eta_0, C_{\mathrm{b}}, C_3, c_b, c_q, c_N > 0$, an integer m_2, and $a \in (1, 1/\gamma)$, independently of $(K, d_{\mathcal{I}}, \varepsilon, \delta)$. These analysis-supplied constants are offline inputs to Algorithms 3 and 9.

For a fine-scale call, use the schedules in (45). All oracle calls and recursive subtrees below are mutually independent. Projection is applied only to the coordinate baselines $\widehat{Q}_i$; the returned recursive sample is not clipped.

For the finite-depth base case, put $\bar{B}_V = \bar{B}_F + \bar{B}_Q$ and choose H_0 so that $\gamma^{H_0}\bar{B}_V \le \eta_0/8$. Let

$$\kappa_0 = \frac{(1-\gamma)\eta_0}{16}, \qquad \zeta_0 = \frac{\eta_0}{16(\bar{B}_V+1)}, \qquad B_0 = \max\{1, B_V^{\mathcal{I}}, R_{\max} + \gamma B_V^{\mathcal{I}}\},$$

and let $m_\star$ be the smallest positive integer satisfying

$$2(1+m_\star^2 d_{\mathcal{I}})^{H_0+1}\exp\left(-\frac{m_\star \kappa_0^2}{2B_0^2}\right) \le \zeta_0. \tag{376}$$

The base-case call in Algorithm 3 is $\mathsf{Coarse}_{\eta_0}^{\mathcal{I}}(s) := \mathsf{CoarseNode}^{\mathcal{I}}(s, 0)$.

G.9 Local policy metric, MaxEnt limit, and transport locality

G.9.1 Proof of Corollary 3

Fix s and suppress its index. By Assumption 1, $\rho = \rho_s$ and $\mu = \mu_s$ have full support; moreover, $\tau, \lambda > 0$. Set $T_0 := \tau\lambda$ and specialize ordinary finite-cost LOT to $C_s = 0$. We reduce the

Algorithm 9 $\mathsf{CoarseNode}^{\mathcal{I}}(s,h)$: finite-depth base case

Input: state s, depth h, active sets $\mathcal{I}$, horizon H_0, batch size $m_\star$
if $h = H_0$ **then**
  return 0
end if
for $r = 1, \dots, m_\star$ **do**
  Draw $J_r \sim \mu_s$ and set $I_r \leftarrow I_{s,J_r}$
  for $i \in I_r$ **do**
    for $p = 1, \dots, m_\star$ **do**
      Draw (R_{rip}, S'_{rip}) at (s,i)
      $X_{rip} \leftarrow R_{rip} + \gamma \mathsf{CoarseNode}^{\mathcal{I}}(S'_{rip}, h+1)$
    end for
    $\widehat{Q}_{ri} \leftarrow \Pi_{[-B_Q^{\mathcal{I}}, B_Q^{\mathcal{I}}]} \left(m_\star^{-1} \sum_{p=1}^{m_\star} X_{rip} \right)$
  end for
  $U_r \leftarrow \bar{f}^{\mathcal{I}}_{s,J_r}(\widehat{Q}_r)$
end for
return $\Pi_{[-B_V^{\mathcal{I}}, B_V^{\mathcal{I}}]} \left(m_\star^{-1} \sum_{r=1}^{m_\star} U_r \right)$

regularizer to policy KL, recover the policy and coupling from the closed form, and then simplify the co-occurrence graph and Hessian.

We first reduce the zero-cost transport regularizer to policy KL.

For fixed $\pi \in \Delta_{\mathcal{A}}$,

$$W^{\rho}_{\lambda,s}(\pi,\mu) = \lambda \min_{\Gamma \in \Pi(\pi,\mu)} \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho(i)\mu(j)}.$$

Every feasible Γ satisfies

$$\begin{aligned} \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho(i)\mu(j)} &= \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\pi(i)\mu(j)} \quad + \sum_{i,j} \Gamma_{ij} \log \frac{\pi(i)}{\rho(i)} \\ &= D_{\mathrm{KL}}(\Gamma \| \pi\mu^\top) + D_{\mathrm{KL}}(\pi\|\rho). \end{aligned}$$

The identity remains valid when $\pi(i) = 0$, under the standard extended-KL convention (in particular, $0 \log 0 = 0$), because feasibility then forces row i of Γ to vanish. The first divergence is nonnegative and vanishes iff $\Gamma = \pi\mu^\top$, which is feasible since $(\pi\mu^\top)\mathbf{1} = \pi$ and $(\pi\mu^\top)^\top \mathbf{1} = \mu$. Hence

$$W^{\rho}_{\lambda,s}(\pi,\mu) = \lambda D_{\mathrm{KL}}(\pi\|\rho), \qquad \Gamma^\star_\pi = \pi\mu^\top,$$

with a unique minimizing coupling. Thus μ disappears from the zero-cost policy objective but remains the second factor of the optimal product coupling.

Next, we reduce the LOT backup to prior-weighted MaxEnt.

Substitution gives

$$F^{\mathrm{LOT}}_s(Q) = \max_{\pi \in \Delta_{\mathcal{A}}} \left\{ \langle \pi, Q \rangle - T_0 D_{\mathrm{KL}}(\pi\|\rho) \right\}, \qquad T_0 = \tau\lambda.$$

Define

$$Z_\rho(Q) := \sum_i \rho(i) e^{Q_i/T_0}, \qquad \pi^Q_i := \frac{\rho(i) e^{Q_i/T_0}}{Z_\rho(Q)}.$$

Full support of ρ and finiteness of Q imply $Z_\rho(Q) > 0$ and $\pi^Q_i > 0$. For every $\pi \in \Delta_{\mathcal{A}}$,

$$\begin{aligned} D_{\mathrm{KL}}(\pi\|\pi^Q) &= D_{\mathrm{KL}}(\pi\|\rho) - T_0^{-1}\langle \pi, Q\rangle + \log Z_\rho(Q), \\ \Rightarrow \langle \pi, Q\rangle - T_0 D_{\mathrm{KL}}(\pi\|\rho) &= T_0 \log Z_\rho(Q) - T_0 D_{\mathrm{KL}}(\pi\|\pi^Q). \end{aligned}$$

Thus KL nonnegativity yields the unique optimizer and value

$$\pi_i^{\mathrm{LOT}}(Q) = \frac{\rho_s(i)e^{Q_i/(\tau\lambda)}}{\sum_k \rho_s(k)e^{Q_k/(\tau\lambda)}}, F_s^{\mathrm{LOT}}(Q) = \tau\lambda \log \sum_i \rho_s(i)e^{Q_i/(\tau\lambda)}.$$

Hence zero-cost LOT is exactly KL-regularized MaxEnt with prior ρ_s and temperature $T_0 = \tau\lambda$. Below abbreviate $Z := Z_\rho(Q)$ and $p := \pi^{\mathrm{LOT}}(Q)$.

We next recover the result from the closed-form transport scaling.

At zero cost, $\mathsf{K}_{C,ij} = e^{-C_{s,ij}/\lambda} = 1$, so $\mathsf{K}_C = \mathbf{1}\mathbf{1}^\top$. With $u_i := \rho_s(i)e^{Q_i/T_0}$ as in Lemma 1, the column scaling and unique optimal plan are

$$(\mathsf{K}_C^\top u)_j = \sum_i \rho_s(i)e^{Q_i/T_0} = Z, \quad v_j = \frac{\mu_s(j)}{Z},$$
$$\Gamma^\star_{ij} = u_i \mathsf{K}_{C,ij} v_j = p_i \mu_s(j), \quad \Gamma^\star = p\mu_s^\top.$$

The marginal constraints hold directly:

$$\sum_i \Gamma^\star_{ij} = \mu_s(j) \sum_i p_i = \mu_s(j),$$
$$\sum_j \Gamma^\star_{ij} = p_i \sum_j \mu_s(j) = p_i.$$

Thus p is exactly the free row marginal. The value formula (9) agrees because every anchor has the same partition:

$$Z_j(Q) = \sum_i \rho_s(i) \exp\left(\frac{Q_i - \tau C_{s,ij}}{\tau\lambda}\right) = Z,$$
$$F_s^{\mathrm{LOT}}(Q) = \tau\lambda \sum_j \mu_s(j) \log Z_j(Q) = \tau\lambda \log Z,$$

where $\sum_j \mu_s(j) = 1$.

We continue to simplify the transport co-occurrence graph.

Assumption 1 gives $\mu_s(j) > 0$, so (93) is well defined. The product form yields

$$A_{ik} = \sum_j \frac{\Gamma^\star_{ij}\Gamma^\star_{kj}}{\mu_s(j)} = p_i p_k \sum_j \mu_s(j) = p_i p_k,$$
$$A = pp^\top, \qquad A\mathbf{1} = p,$$
$$L_A = \mathrm{diag}(A\mathbf{1}) - A = \mathrm{diag}(p) - pp^\top.$$

Thus L_A is the covariance matrix of a one-hot categorical action drawn from p.

We then recover the prior-weighted softmax Jacobian.

Substitution into (94) gives

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda}\left[\mathrm{diag}(p) - pp^\top\right].$$

Direct differentiation gives, for every i, k,

$$\frac{\partial p_i}{\partial Q_k} = \frac{1}{T_0}\frac{\rho_s(i)e^{Q_i/T_0}}{Z}\left(\delta_{ik} - \frac{\rho_s(k)e^{Q_k/T_0}}{Z}\right) = \frac{1}{\tau\lambda}p_i(\delta_{ik} - p_k),$$
$$D_Q p = \nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda}\left[\mathrm{diag}(p) - pp^\top\right].$$

This is the multinomial-covariance Jacobian of the prior-weighted softmax. For every $x \in \mathbb{R}^K$,

$$x^\top\left[\mathrm{diag}(p) - pp^\top\right]x = \mathrm{Var}_{A\sim p}(x_A) \geq 0,$$

so the Hessian is positive semidefinite. It annihilates $\mathbf{1}$, in agreement with $F_s^{\mathrm{LOT}}(Q + c\mathbf{1}) = F_s^{\mathrm{LOT}}(Q) + c$ and softmax shift invariance. Full support gives nullspace $\mathrm{span}\{\mathbf{1}\}$ and rank $K - 1$ for $K \geq 2$; for $K = 1$ the Hessian is zero and the same identities remain valid.

Finally, we record the uniform-prior normalization.

If ρ_s is uniform, then

$$\begin{aligned}\pi_i^{\mathrm{LOT}}(Q) &= \frac{e^{Q_i/T_0}}{\sum_k e^{Q_k/T_0}},\\ F_s^{\mathrm{LOT}}(Q) &= T_0 \log\left(\frac{1}{K}\sum_i e^{Q_i/T_0}\right)\\ &= T_0 \log \sum_i e^{Q_i/T_0} - T_0 \log K.\end{aligned}$$

Thus the policy is the usual softmax, while the prior-normalized log-sum-exp value is lower by $T_0 \log K$ than the common Shannon-entropy-bonus convention. This constant affects neither policy nor Hessian, but must be retained when comparing Bellman fixed points or centered and uncentered backups.

We have proved

$$\Gamma^\star = \pi^{\mathrm{LOT}}\mu_s^\top, \qquad A = \pi^{\mathrm{LOT}}\big(\pi^{\mathrm{LOT}}\big)^\top,$$

and shown that the general transport-graph curvature in (94) reduces exactly to the prior-weighted softmax covariance Jacobian. General LOT replaces this categorical covariance by the Laplacian of the value-dependent transport co-occurrence graph. □

G.9.2 Proof of Theorem 5

Fix s and suppress it when unambiguous. Set

$$g_s = \Omega_s + \iota_{\Delta_\mathcal{A}}, \qquad \Omega_s(\pi) = \tau W^\rho_{\lambda,s}(\pi, \mu_s),$$

where Ω_s is defined on $\Delta_\mathcal{A}$ and g_s is its extended-real extension to $\mathbb{R}^K$. Also let

$$\begin{aligned}&H = \{\pi : \mathbf{1}^\top \pi = 1\}, \quad T = \{v : \mathbf{1}^\top v = 0\},\\ &P_T = I - K^{-1}\mathbf{1}\mathbf{1}^\top.\end{aligned}$$

Thus $H = \mathrm{aff}(\Delta_\mathcal{A})$, T is its tangent space, and P_T is the orthogonal projection onto T. All derivatives of Ω_s are relative to $\mathrm{ri}(\Delta_\mathcal{A}) \subset H$.

We first establish convexity, closedness, and Fenchel conjugacy.

For fixed π, $\Pi(\pi, \mu_s)$ is convex, and

$$(\Gamma, \pi) \mapsto \langle \Gamma, C\rangle + \lambda \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho_s(i)\mu_s(j)}$$

is jointly convex on its feasible domain. Partial minimization under the affine marginal constraints therefore makes $W^\rho_{\lambda,s}(\pi, \mu_s)$ convex in π. Since the action set is finite, the coupling polytopes are compact, the bases $\rho_s(i)\mu_s(j)$ are positive, and C is finite, the minimum is finite and continuous on $\Delta_\mathcal{A}$ by the standard finite-dimensional entropic-OT/Berge continuity argument. Hence g_s is proper, closed, and convex. By (8),

$$\begin{aligned}F_s^{\mathrm{LOT}}(Q) &= \sup_{\pi \in \Delta_\mathcal{A}} \{\langle Q, \pi\rangle - \Omega_s(\pi)\}\\ &= g_s^\star(Q) = (\Omega_s + \iota_{\Delta_\mathcal{A}})^\star(Q),\end{aligned}$$

which is the first identity in (39).

Next, we identify the unique Fenchel response.

The semi-relaxed formulation optimizes over $\{\Gamma \geq 0 : \Gamma^\top \mathbf{1} = \mu_s\}$. Its reward–cost term plus negative relative entropy is strictly concave on this compact convex set, so it has a unique optimizer Γ_Q and hence a unique row marginal

$$\pi_Q = \Gamma_Q \mathbf{1} = \nabla_Q F_s^{\mathrm{LOT}}(Q),$$

where the last equality also follows from the closed form and Danskin's theorem. Fenchel–Young equality gives $F_s^{\mathrm{LOT}}(Q) + g_s(\pi_Q) = \langle Q, \pi_Q \rangle$, equivalently

$$Q \in \partial g_s(\pi_Q), \qquad \pi_Q \in \partial g_s^\star(Q). \tag{377}$$

Since $g_s^\star = F_s^{\mathrm{LOT}}$ is differentiable, the second relation again gives $\pi_Q = \nabla F_s^{\mathrm{LOT}}(Q)$.

We now prove strict positivity of the optimal coupling and connectivity of the induced graph.

Lemma 1 yields

$$(\Gamma_Q)_{ij} = u_i \mathsf{K}_{C,ij} v_j, \qquad u_i = \rho_s(i) e^{Q_i/(\tau\lambda)},$$
$$\mathsf{K}_{C,ij} = e^{-C_{ij}/\lambda}, \qquad v_j = \frac{\mu_s(j)}{(\mathsf{K}_C^\top u)_j}.$$

Assumption 1 and finite C make every factor positive (the denominator is a finite positive sum), whence

$$(\Gamma_Q)_{ij} > 0 \quad \text{for all } i, j. \tag{378}$$

Thus $(\pi_Q)_i = \sum_j (\Gamma_Q)_{ij} > 0$: every finite-$Q$ response is interior under the theorem's assumptions, and "matched interior" merely specifies where relative policy derivatives are taken.

Define A_Q, L_Q as in (38). Then

$$(A_Q)_{ik} = \sum_j \frac{(\Gamma_Q)_{ij}(\Gamma_Q)_{kj}}{\mu_s(j)}$$

is symmetric and nonnegative, and (378) implies

$$(A_Q)_{ik} > 0 \quad \text{for all } i, k. \tag{379}$$

Ignoring self-loops, it is therefore a complete, connected weighted undirected graph. Using $\sum_k (\Gamma_Q)_{kj} = \mu_s(j)$ gives

$$A_Q \mathbf{1} = \pi_Q, \qquad L_Q = \mathrm{diag}(\pi_Q) - A_Q, \qquad L_Q \mathbf{1} = 0. \tag{380}$$

Moreover,

$$x^\top L_Q x = \frac{1}{2} \sum_{i,k} (A_Q)_{ik} (x_i - x_k)^2.$$

Hence $L_Q \succeq 0$, and strict positivity of every off-diagonal weight makes equality equivalent to x being constant. Consequently,

$$\ker L_Q = \mathrm{span}\{\mathbf{1}\}, \qquad \mathrm{rank}(L_Q) = K - 1,$$
$$x^\top L_Q x > 0 \quad (0 \neq x \in T). \tag{381}$$

We next identify the backup Hessian and its quotient geometry.

Theorem 12, at $\Gamma^\star = \Gamma_Q$, gives

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda} \left[\mathrm{diag}(\pi_Q) - \Gamma_Q \, \mathrm{diag}(\mu_s)^{-1} \Gamma_Q^\top \right]$$
$$= \frac{1}{\tau\lambda} L_Q,$$

the first equality in (40). Translation equivariance gives

$$F_s^{\mathrm{LOT}}(Q + c\mathbf{1}) = F_s^{\mathrm{LOT}}(Q) + c,$$
$$\nabla F_s^{\mathrm{LOT}}(Q + c\mathbf{1}) = \nabla F_s^{\mathrm{LOT}}(Q). \tag{382}$$

Thus the response depends only on $[Q] = Q + \mathrm{span}\{\mathbf{1}\}$, whose canonical representative is $q = P_T Q = Q - K^{-1}(\mathbf{1}^\top Q)\mathbf{1}$. Represent the quotient map by

$$\Phi : T \to \mathrm{ri}(\Delta_{\mathcal{A}}), \qquad \Phi(q) = \nabla F_s^{\mathrm{LOT}}(q).$$

The translation invariance above also gives $L_{P_T Q} = L_Q$. Because $\mathbf{1}^\top \Phi(q) = 1$, its derivative maps T into T; by the Hessian identity and (381),

$$D\Phi(q)|_T = \frac{1}{\tau\lambda} L_q|_T : T \to T, \quad \text{positive definite and invertible.} \tag{383}$$

It is also globally injective on T: by (381), for $q_1 \neq q_0$, put $d = q_1 - q_0 \in T \setminus \{0\}$; then

$$d^\top[\Phi(q_1) - \Phi(q_0)] = \frac{1}{\tau\lambda} \int_0^1 d^\top L_{q_0 + td} d \, dt > 0.$$

This is strict convexity after quotienting out translations.

We continue to compute the simplex normal cone and the relative gradient.

For $\pi_Q = \Phi(P_T Q)$, interiority leaves only the affine constraint, so

$$N_{\Delta_{\mathcal{A}}}(\pi_Q) = T^\perp = \mathrm{span}\{\mathbf{1}\}. \tag{384}$$

Indeed, both signs of every sufficiently small $d \in T$ are feasible, so a normal n must satisfy $n^\top d = 0$ for all $d \in T$. From (377), for every $\pi \in \Delta_{\mathcal{A}}$,

$$\begin{aligned} \Omega_s(\pi) &\geq \Omega_s(\pi_Q) + Q^\top(\pi - \pi_Q) \\ &= \Omega_s(\pi_Q) + (P_T Q)^\top(\pi - \pi_Q), \end{aligned}$$

so $P_T Q$ is a relative subgradient. If another relative subgradient $z \in T$ existed, extending by $g_s = +\infty$ off $\Delta_{\mathcal{A}}$ would give $z \in \partial g_s(\pi_Q)$, hence by Fenchel reciprocity and differentiability,

$$\Phi(z) = \nabla F_s^{\mathrm{LOT}}(z) = \pi_Q = \Phi(P_T Q).$$

Injectivity on T forces $z = P_T Q$. A finite convex function with singleton relative subdifferential is relatively differentiable; therefore

$$\boxed{\nabla_T \Omega_s(\pi_Q) = P_T Q.} \tag{385}$$

This is the second identity in (39). Together with (384) it also yields

$$\partial g_s(\pi_Q) = \nabla_T \Omega_s(\pi_Q) + N_{\Delta_{\mathcal{A}}}(\pi_Q) = P_T Q + \mathrm{span}\{\mathbf{1}\}. \tag{386}$$

Thus Q is determined only up to an additive constant, as translation equivariance requires.

We then construct the local inverse on the tangent space.

Let $q = P_T Q$. By (383) and the finite-dimensional inverse-function theorem, there is an open neighborhood $U \subset T$ of q such that $V := \Phi(U)$ is open in H, contains π_Q, lies in $\mathrm{ri}(\Delta_{\mathcal{A}})$, and $\Phi : U \to V$ is a C^1 diffeomorphism. Let $\Psi : V \to U$ be its local inverse. Every $\pi \in V$ is matched, $\pi = \Phi(\Psi(\pi))$, so the normal-cone calculation above gives $\nabla_T \Omega_s(\pi) = \Psi(\pi)$. Hence Ω_s is twice relatively differentiable on V, and

$$\nabla_T^2 \Omega_s(\pi_Q) = D\Psi(\pi_Q) = \left(D\Phi(q)|_T\right)^{-1}. \tag{387}$$

Only local surjectivity near the matched policy is used; this does not say that every interior policy lies in the global response range.

We now identify the inverse with the Laplacian pseudoinverse.

The Hessian identity above gives $D\Phi(q)|_T = (\tau\lambda)^{-1} L_Q|_T$. Since L_Q is symmetric, positive definite on T, and has kernel $\mathrm{span}\{\mathbf{1}\}$, its Moore–Penrose inverse satisfies

$$\begin{gathered} L_Q^\dagger \mathbf{1} = 0, \qquad L_Q^\dagger|_T = (L_Q|_T)^{-1}, \\ L_Q L_Q^\dagger = L_Q^\dagger L_Q = P_T. \end{gathered}$$

Therefore, on T,

$$(D\Phi(q)|_T)^{-1} = \tau\lambda(L_Q|_T)^{-1} = \tau\lambda L_Q^\dagger|_T.$$

Using the canonical ambient representative that vanishes on $T^\perp = \mathrm{span}\{\mathbf{1}\}$ gives

$$\boxed{\nabla_T^2 \Omega_s(\pi_Q) = \tau\lambda L_Q^\dagger.} \tag{388}$$

Together with the Hessian identity above this proves (40). On T,

$$\left[\nabla_T^2 \Omega_s(\pi_Q) \nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\right]\big|_T = L_Q^\dagger L_Q\big|_T = P_T|_T = I_T.$$

Thus the Hessians are inverse on probability-preserving directions, but neither is invertible on $\mathbb{R}^K$ because common shifts are null.

We next read the inverse curvature as effective resistance.

By (379), the graph has positive off-diagonal conductances and is connected, so its effective resistance is finite and equals

$$R_Q(i,k) = (e_i - e_k)^\top L_Q^\dagger (e_i - e_k).$$

Since $d_{ik} = e_i - e_k \in T$, (388) implies

$$d_{ik}^\top \nabla_T^2 \Omega_s(\pi_Q) d_{ik} = \tau\lambda R_Q(i,k),$$

which is (41). More explicitly, whenever $\pi_Q + h(e_i - e_k) \in \Delta_{\mathcal{A}}$, its local second-order quadratic term is

$$\frac{h^2}{2}(e_i - e_k)^\top \nabla_T^2 \Omega_s(\pi_Q)(e_i - e_k) = \frac{\tau\lambda h^2}{2} R_Q(i,k).$$

Thus resistance is the pointwise quadratic cost of transferring infinitesimal policy mass from k to i.

Finally, we record the scope of the identity.

Γ_Q, A_Q, L_Q depend on Q; hence $\tau\lambda L_Q^\dagger$ is a pointwise Riemannian-type metric tensor on the matched response manifold, not a fixed global distance between arbitrary policies. Its graph is the induced transport co-occurrence graph, not necessarily the input graph used to form C. Strict positivity, connectedness, and rank $K-1$ require full-support (ρ_s, μ_s) and entrywise-finite C. A finite-support kernel with exact zeros may instead produce a disconnected co-occurrence graph and extra null directions, requiring a further restriction of the inverse-function argument or an additional rank condition. Finally, the Hessian identity alone proves neither natural-gradient convergence, policy smoothness on the input graph, nor a global policy distance.

This completes the proof of Theorem 5.

□

G.9.3 Proof of Corollary 4

Fix s and impose Assumption 1. Set $T_0 = \tau\lambda$ with $\tau, \lambda > 0$, and let $C = 0$. For the matched LOT response $\pi_Q = \nabla F_s^{\mathrm{LOT}}(Q)$, positivity of the prior and finiteness of Q make every coordinate strictly positive, so $\pi_Q \in \mathrm{ri}(\Delta_{\mathcal{A}})$. Hence relative derivatives are well defined on

$$T = \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\}.$$

We reduce zero-cost transport to KL, compute its conjugate and tangent Hessian, and identify that Hessian with the inverse categorical Laplacian.

We first reduce the zero-cost transport penalty to a policy KL.

At $C = 0$,

$$W_{\lambda,s}^{\rho}(\pi, \mu_s) = \lambda \min_{\Gamma \in \Pi(\pi, \mu_s)} \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho_s(i)\mu_s(j)}.$$

For each positive row put $r_i(j) = \Gamma_{ij}/\pi_i$; zero rows contribute nothing. The KL chain rule yields

$$\sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\rho_s(i)\mu_s(j)} = \mathrm{KL}(\pi\|\rho_s) + \sum_i \pi_i \mathrm{KL}(r_i\|\mu_s).$$

The second term is nonnegative and vanishes for the feasible independent coupling $\Gamma = \pi\mu_s^\top$. Therefore

$$W^\rho_{\lambda,s}(\pi, \mu_s) = \lambda \mathrm{KL}(\pi\|\rho_s). \tag{389}$$

Multiplication by the outer weight τ gives

$$\boxed{\Omega_s(\pi) = \tau W^\rho_{\lambda,s}(\pi, \mu_s) = T_0 \mathrm{KL}(\pi\|\rho_s).} \tag{390}$$

This proves the first assertion. The factor λ in (389) is essential because the transport objective contains λ times the coupling KL.

Next, we compute the KL conjugate and its matched response.

Define

$$Z_\rho(Q) = \sum_i \rho_s(i) e^{Q_i/T_0}, \qquad p_{Q,i} = \frac{\rho_s(i) e^{Q_i/T_0}}{Z_\rho(Q)}.$$

For every $\pi \in \Delta_{\mathcal{A}}$,

$$\mathrm{KL}(\pi\|p_Q) = \mathrm{KL}(\pi\|\rho_s) - T_0^{-1}\langle \pi, Q\rangle + \log Z_\rho(Q).$$

Equivalently,

$$\langle \pi, Q\rangle \quad - \quad T_0 \mathrm{KL}(\pi\|\rho_s) \quad = \quad T_0 \log Z_\rho(Q) \quad - \quad T_0 \mathrm{KL}(\pi\|p_Q).$$

Nonnegativity of KL, with equality only at $\pi = p_Q$, implies

$$(\Omega_s + \iota_{\Delta_{\mathcal{A}}})^\star(Q) = T_0 \log \sum_i \rho_s(i) e^{Q_i/T_0}, \qquad \pi_Q = p_Q. \tag{391}$$

Thus the response is the unique, strictly positive ρ_s-weighted softmax, recovering the zero-cost Fenchel relation in Theorem 5.

We continue to compute the zero-cost transport graph.

All anchors have the same local softmax, so the unique optimal coupling is $\Gamma_Q = \pi_Q \mu_s^\top$. Full support makes $\mathrm{diag}(\mu_s)^{-1}$ well defined, and (38) gives

$$\begin{aligned} A_Q &= \Gamma_Q \,\mathrm{diag}(\mu_s)^{-1} \Gamma_Q^\top \\ &= \pi_Q \big[\mu_s^\top \,\mathrm{diag}(\mu_s)^{-1} \mu_s\big] \pi_Q^\top = \pi_Q \pi_Q^\top, \end{aligned}$$

since the bracket equals $\sum_j \mu_s(j) = 1$. As $A_Q \mathbf{1} = \pi_Q$,

$$\boxed{L_Q = \mathrm{diag}(\pi_Q) - \pi_Q \pi_Q^\top.} \tag{392}$$

This proves the second assertion. The matrix in (392) is the covariance of a one-hot categorical draw. Full support gives

$$\ker L_Q = \mathrm{span}\{\mathbf{1}\}, \qquad \mathrm{rank}(L_Q) = K - 1,$$

so L_Q is invertible on T.

We then compute the relative KL Hessian directly.

On $\mathrm{ri}(\Delta_{\mathcal{A}})$,

$$\begin{aligned} \Omega_s(\pi) &= T_0 \sum_i \pi_i \log \frac{\pi_i}{\rho_s(i)}, \\ \partial_i \Omega_s &= T_0 \left(\log \frac{\pi_i}{\rho_s(i)} + 1 \right), \qquad \partial^2_{ik} \Omega_s = T_0 \frac{\mathbf{1}\{i = k\}}{\pi_i}. \end{aligned}$$

These are the derivatives of its ambient smooth extension near an interior point. On the simplex the ambient gradient is defined modulo a multiple of $\mathbf{1}$, but this ambiguity vanishes against tangent directions. Thus, for $v, w \in T$,

$$v^\top \nabla^2_T \Omega_s(\pi) w = T_0 \sum_i \frac{v_i w_i}{\pi_i}. \tag{393}$$

In particular,

$$v^\top \nabla_T^2 \Omega_s(\pi_Q) v = T_0 \sum_i \frac{v_i^2}{\pi_{Q,i}} = \tau\lambda \sum_i \frac{v_i^2}{\pi_{Q,i}}, \qquad v \in T, \tag{394}$$

which is (109).

We now identify the tangent inverse and pseudoinverse explicitly.

To verify the inverse relation (40), write

$$D_Q = \mathrm{diag}(\pi_Q), \qquad P_T = I - \frac{\mathbf{1}\mathbf{1}^\top}{K},$$
$$B_Q = P_T D_Q^{-1} P_T.$$

Then B_Q is symmetric and kills $\mathbf{1}$. For $x \in T$, let

$$a_x = K^{-1}\mathbf{1}^\top D_Q^{-1} x, \qquad y = B_Q x = D_Q^{-1} x - a_x \mathbf{1} \in T.$$

Since

$$D_Q y = x - a_x \pi_Q, \qquad \pi_Q^\top y = \mathbf{1}^\top x - a_x = -a_x,$$

we have $L_Q y = (D_Q - \pi_Q \pi_Q^\top) y = x$. Hence $B_Q|_T = (L_Q|_T)^{-1}$. Both matrices are symmetric and vanish on $T^\perp = \mathrm{span}\{\mathbf{1}\}$, so

$$\boxed{L_Q^\dagger = P_T D_Q^{-1} P_T.} \tag{395}$$

Equivalently, $L_Q L_Q^\dagger = L_Q^\dagger L_Q = P_T$. For $v, w \in T$, the projections in (395) disappear, giving

$$v^\top L_Q^\dagger w = v^\top D_Q^{-1} w = \sum_i \frac{v_i w_i}{\pi_{Q,i}}.$$

Consequently,

$$\nabla_T^2 \Omega_s(\pi_Q) = T_0 L_Q^\dagger = \tau\lambda L_Q^\dagger \quad \text{on } T. \tag{396}$$

The equality holds as operators and bilinear forms and is precisely Theorem 5. An ordinary inverse on $\mathbb{R}^K$ is invalid: the simplex and softmax share the null direction $\mathbf{1}$.

We next identify the Fisher–Rao metric.

The categorical Fisher–Rao metric is

$$g_\pi(v, w) = \sum_i \frac{v_i w_i}{\pi_i}, \qquad v, w \in T.$$

Indeed, for $v \in T$ and sufficiently small h such that $\pi + hv > 0$,

$$\begin{aligned} \mathrm{KL}(\pi + hv \| \pi) &= \sum_i (\pi_i + hv_i) \log\left(1 + h\frac{v_i}{\pi_i}\right) \\ &= \frac{h^2}{2} \sum_i \frac{v_i^2}{\pi_i} + o(h^2), \end{aligned}$$

the linear term vanishing because $\sum_i v_i = 0$. Thus (393) induces the metric $\tau\lambda g_{\pi_Q}$. The prior determines π_Q, but conditional on that policy the metric is the standard categorical Fisher–Rao metric.

Finally, we state the proximal and natural-gradient qualification.

The preceding identity is pointwise; an algorithmic claim additionally requires a policy objective, its differential, and an update rule. Given a chosen local policy-objective gradient $g \in \mathbb{R}^K$ at an interior π, the Bregman divergence generated by Ω_s is

$$\begin{aligned} D_{\Omega_s}(\pi' \| \pi) &= \Omega_s(\pi') - \Omega_s(\pi) - \langle \nabla_T \Omega_s(\pi), \pi' - \pi \rangle \\ &= T_0 \mathrm{KL}(\pi' \| \pi), \end{aligned}$$

so the prior cancels. An explicitly defined KL-proximal (mirror) step is

$$\pi^+ \in \arg\max_{\pi' \in \Delta_{\mathcal{A}}} \left\{ \langle g, \pi' - \pi \rangle - \eta^{-1} D_{\Omega_s}(\pi' \| \pi) \right\}, \qquad \eta > 0. \tag{397}$$

Its stationarity equations give

$$\pi_i^+ = \frac{\pi_i \exp(\eta g_i / T_0)}{\sum_k \pi_k \exp(\eta g_k / T_0)}. \tag{398}$$

For small η, (398) yields

$$\pi^+ - \pi = \frac{\eta}{T_0} \big[\mathrm{diag}(\pi) - \pi\pi^\top \big] g + O(\eta^2).$$

This is the Fisher–Rao natural-gradient direction in probability coordinates: it is the unique $d \in T$ satisfying

$$T_0 \sum_i \frac{d_i v_i}{\pi_i} = g^\top v \qquad \text{for every } v \in T.$$

Thus the KL-proximal step has this direction only in its infinitesimal limit; the finite exponentiated step generally differs from a finite additive natural-gradient step. Moreover, Corollary 4 specifies neither g, a step-size schedule, nor an iterative actor, and therefore proves neither convergence of a natural-policy-gradient method nor RL objective improvement. Such claims require a separately defined proximal or natural-gradient algorithm and convergence analysis.

Combining (390), (392), and (394) proves Corollary 4.

□

G.9.4 Proof of Proposition 6

Fix a state s and a matched response, assume $K \geq 2$ as is implicit in the use of $\lambda_2(L_A)$, and write $\pi^{\mathrm{LOT}} = \nabla F_s^{\mathrm{LOT}}(Q)$, $A = A_Q$, and $L_A = L_Q$. Under the hypotheses of Theorem 5 (full-support prior and reference distribution, entrywise-finite cost matrix, and $\tau, \lambda > 0$), the optimal coupling is strictly positive. Hence all off-diagonal conductances of A are positive, so the co-occurrence graph is connected and

$$L_A \succeq 0, \qquad \ker L_A = \mathrm{span}\{\mathbf{1}\}, \qquad \mathrm{rank}(L_A) = K - 1,$$

in particular its algebraic connectivity $\lambda_2(L_A)$ is positive.

We first diagonalize the transport Laplacian on the tangent space.

Choose an orthonormal eigenbasis $u_1, \ldots, u_K$ of the real symmetric matrix L_A, where $u_1 = \mathbf{1}/\sqrt{K}$ and

$$0 = \lambda_1(L_A) < \lambda_2(L_A) \leq \cdots \leq \lambda_K(L_A) = \lambda_{\max}(L_A).$$

Then

$$L_A = \sum_{\ell=2}^{K} \lambda_\ell(L_A) u_\ell u_\ell^\top. \tag{399}$$

Thus

$$T = \{v : \mathbf{1}^\top v = 0\} = \mathrm{span}\{u_2, \ldots, u_K\}$$

is the simplex tangent space, and

$$\mathrm{spec}(L_A|_T) = \{\lambda_2(L_A), \ldots, \lambda_K(L_A)\}.$$

In particular, $L_A|_T$ is positive definite.

Next, we compute the pseudoinverse spectrum.

The Moore–Penrose pseudoinverse has the same eigenvectors, with zero on $\mathbf{1}$ and reciprocal positive eigenvalues:

$$\begin{aligned} L_A^\dagger &= \sum_{\ell=2}^{K} \frac{u_\ell u_\ell^\top}{\lambda_\ell(L_A)}, \\ L_A^\dagger \mathbf{1} = 0, &\qquad L_A^\dagger|_T = (L_A|_T)^{-1}. \end{aligned} \tag{400}$$

By (40),

$$H_\Omega := \nabla_T^2 \Omega_s(\pi^{\mathrm{LOT}}) = \tau\lambda L_A^\dagger = \sum_{\ell=2}^{K} \frac{\tau\lambda}{\lambda_\ell(L_A)} u_\ell u_\ell^\top.$$

Its tangent-space spectrum is

$$\mathrm{spec}(H_\Omega|_T) = \left\{ \frac{\tau\lambda}{\lambda_\ell(L_A)} : 2 \le \ell \le K \right\},$$

in reverse order to that of $L_A|_T$. Therefore

$$\lambda_{\min}(H_\Omega|_T) = \frac{\tau\lambda}{\lambda_{\max}(L_A)}, \qquad \lambda_{\max}(H_\Omega|_T) = \frac{\tau\lambda}{\lambda_2(L_A)}. \tag{401}$$

We continue to derive the exact pointwise quadratic moduli.

For $v = \sum_{\ell=2}^{K} a_\ell u_\ell \in T$, orthonormality and (400) give

$$\|v\|_2^2 = \sum_{\ell=2}^{K} a_\ell^2,$$

$$v^\top H_\Omega v = \tau\lambda v^\top L_A^\dagger v = \sum_{\ell=2}^{K} \frac{\tau\lambda}{\lambda_\ell(L_A)} a_\ell^2.$$

Since

$$\frac{\tau\lambda}{\lambda_{\max}(L_A)} \le \frac{\tau\lambda}{\lambda_\ell(L_A)} \le \frac{\tau\lambda}{\lambda_2(L_A)},$$

we obtain

$$\sigma\|v\|_2^2 \le v^\top \nabla_T^2 \Omega_s(\pi^{\mathrm{LOT}}) v \le \beta\|v\|_2^2, \qquad v \in T, \tag{402}$$

where

$$\sigma = \frac{\tau\lambda}{\lambda_{\max}(L_A)}, \qquad \beta = \frac{\tau\lambda}{\lambda_2(L_A)}. \tag{403}$$

These constants are pointwise sharp: equality on the left holds for an eigenvector of $\lambda_{\max}(L_A)$ and equality on the right for a Fiedler vector. Thus (403) proves the two spectral moduli in (110).

Here "local strong-convexity and smoothness moduli" means the Hessian at this policy. Indeed, for $v \in T$ and sufficiently small h with $\pi^{\mathrm{LOT}} + hv$ in the simplex,

$$\Omega_s(\pi^{\mathrm{LOT}} + hv) = \Omega_s(\pi^{\mathrm{LOT}}) + h\langle \nabla_T \Omega_s(\pi^{\mathrm{LOT}}), v\rangle + \frac{h^2}{2} v^\top \nabla_T^2 \Omega_s(\pi^{\mathrm{LOT}}) v + o(h^2).$$

Thus (402) bounds the quadratic coefficient by $\sigma\|v\|_2^2/2$ and $\beta\|v\|_2^2/2$; without additional uniformity it is not a finite-neighborhood or global strong-convexity/smoothness inequality.

We then prove the universal one-half upper bound for $\lambda_{\max}(L_A)$.

The mixture representation in Theorem 1 gives

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda} \sum_j \mu_s(j) \big[\mathrm{diag}(p^{(j)}) - p^{(j)} p^{(j)\top} \big]. \tag{404}$$

For any probability vector p, let $\Sigma_p = \mathrm{diag}(p) - pp^\top$. For every unit x, let $I \sim p$. Then x_I is supported on an interval of length $\max_i x_i - \min_i x_i$, and such a variable has variance at most one quarter of the squared interval length. Hence

$$x^\top \Sigma_p x = \mathrm{Var}_{I\sim p}(x_I) \le \frac{1}{4}(\max_i x_i - \min_i x_i)^2 \le \frac{1}{2},$$

because, for maximizing and minimizing indices i_+, i_-,

$$\max_i x_i - \min_i x_i = \langle x, e_{i_+} - e_{i_-} \rangle \le \|x\|_2 \|e_{i_+} - e_{i_-}\|_2 \le \sqrt{2}.$$

Consequently,

$$\|\Sigma_p\|_{\mathrm{op}} = \lambda_{\max}(\Sigma_p) \le \frac{1}{2}. \tag{405}$$

Since $\mu_s(j) \ge 0$ and $\sum_j \mu_s(j) = 1$, the operator-norm triangle inequality in (404), followed by (405), yields

$$\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\|_{\mathrm{op}} \le \frac{1}{\tau\lambda} \sum_j \mu_s(j) \|\Sigma_{p^{(j)}}\|_{\mathrm{op}} \le \frac{1}{2\tau\lambda}.$$

On the other hand, Theorem 12 and $L_A \succeq 0$ give

$$\nabla_Q^2 F_s^{\mathrm{LOT}}(Q) = \frac{1}{\tau\lambda} L_A.$$

Hence

$$\|\nabla_Q^2 F_s^{\mathrm{LOT}}(Q)\|_{\mathrm{op}} = \frac{\lambda_{\max}(L_A)}{\tau\lambda}.$$

Therefore

$$\boxed{\lambda_{\max}(L_A) \le \frac{1}{2}.} \tag{406}$$

This is precisely the dependence on Theorem 1 stated in the proposition, now supplied with its covariance-norm proof. Substitution of (406) into $\sigma = \tau\lambda/\lambda_{\max}(L_A)$ gives

$$\sigma \ge 2\tau\lambda,$$

which is the inequality in (110).

We now compute the pointwise Hessian condition number.

Since $H_\Omega|_T \succ 0$, (401) gives

$$\begin{aligned} \mathrm{cond}_2(H_\Omega|_T) &= \frac{\lambda_{\max}(H_\Omega|_T)}{\lambda_{\min}(H_\Omega|_T)} = \frac{\tau\lambda/\lambda_2(L_A)}{\tau\lambda/\lambda_{\max}(L_A)} \\ &= \frac{\lambda_{\max}(L_A)}{\lambda_2(L_A)} = \frac{\beta}{\sigma} = \kappa_A. \end{aligned}$$

This proves (111): reciprocal curvature reverses the extreme eigenvalues but preserves the condition number of $L_A|_T$.

Finally, we explain why the result is pointwise rather than an iterative rate.

The objects Γ_Q, A_Q, and $L_A = L_Q$ depend on the matched Q and thus on the current policy response. Full support and finite cost imply $\lambda_2(L_A) > 0$ at each fixed finite Q, but not a uniform positive lower bound along arbitrary policy sequences. The co-occurrence graph may approach a bottleneck and $\lambda_2(L_A)$ may approach zero, making β arbitrarily large; if $\lambda_2(L_A)/\lambda_{\max}(L_A) \to 0$, then $\kappa_A \to \infty$. Although $\lambda_{\max}(L_A) \le 1/2$ uniformly lower bounds σ, it does not prevent this loss of connectivity.

Thus the proposition gives exact local quadratic conditioning at one matched policy. A policy-mirror-descent or natural-gradient convergence theorem would additionally require an explicit update and objective, an admissible interior trajectory, uniform lower and upper curvature bounds along it (in particular, a uniform positive lower bound on $\lambda_2(L_A)$), and the descent, smoothness, or stochastic-error conditions of the chosen algorithm. No iterative conclusion follows from the pointwise spectral identity alone. Finally, L_A is the Laplacian of the value-dependent transport co-occurrence graph, not necessarily that of the input action graph defining the cost.

This completes the proof of Proposition 6.

□

G.9.5 Proof of Theorem 16

Fix the state s and action-value vector Q from the theorem, and set $T_0 := \tau\lambda > 0$. Full support of ρ_s and finiteness of the relevant costs make $\log \rho_s(i)$ and all scores $S_{ij}(Q)$ finite. For $E \subseteq \mathcal{A}$, write

$$p^{(j)}(E) := \sum_{i \in E} p_i^{(j)}(Q), \qquad \pi^{\mathrm{LOT}}(E) := \sum_{i \in E} \pi_i^{\mathrm{LOT}}(Q).$$

The nonempty sets B, R have well-defined maxima, and their disjointness prevents double counting below. We retain the theorem's cost notation C^L; no spectral property is used, so the proof applies verbatim to any finite LOT cost C_s after replacing C^L by C_s. In the barrier specialization, $\eta > 0$ is the stated margin.

We first rewrite the local policy as an exact Gibbs law over the scores.

By (10),

$$p_i^{(j)}(Q) = \frac{\rho_s(i) e^{(Q_i - \tau C_{ij}^L)/(\tau\lambda)}}{\sum_k \rho_s(k) e^{(Q_k - \tau C_{kj}^L)/(\tau\lambda)}}.$$

Since $T_0 = \tau\lambda$, (112) gives

$$\begin{aligned}
\rho_s(i) e^{(Q_i - \tau C_{ij}^L)/T_0} &= \exp\left(\log \rho_s(i) + \frac{Q_i - \tau C_{ij}^L}{T_0}\right) \\
&= \exp\left(\frac{Q_i + T_0 \log \rho_s(i) - \tau C_{ij}^L}{T_0}\right) \\
&= e^{S_{ij}(Q)/T_0}.
\end{aligned}$$

Thus

$$p_i^{(j)}(Q) = \frac{e^{S_{ij}(Q)/T_0}}{\sum_{k \in \mathcal{A}} e^{S_{kj}(Q)/T_0}}. \tag{407}$$

Hence value, prior mass, and transport cost enter through one local score.

Next, we isolate the Gibbs masses of the target and comparison sets.

For the fixed anchor j, let

$$\begin{aligned}
N_B &:= \sum_{i \in B} e^{S_{ij}/T_0}, \qquad & N_R &:= \sum_{k \in R} e^{S_{kj}/T_0}, \\
N_O &:= \sum_{h \in \mathcal{A} \setminus (B \cup R)} e^{S_{hj}/T_0}.
\end{aligned}$$

Here the suppressed argument is Q. Finiteness and nonemptiness give $N_B, N_R > 0$, while $N_O \geq 0$. Since $B \cap R = \varnothing$, the partition is $N_B + N_R + N_O$; hence (407) yields

$$p^{(j)}(B) = \frac{N_B}{N_B + N_R + N_O} \leq \frac{N_B}{N_B + N_R}.$$

Thus mass outside $B \cup R$ can only decrease this two-set upper bound.

We next bound the two set-partition terms by their best scores.

Set

$$M_B := \max_{i \in B} S_{ij}(Q), \qquad M_R := \max_{k \in R} S_{kj}(Q).$$

Bounding every target summand by its maximum and retaining a maximizer in R gives

$$N_B \leq |B| e^{M_B/T_0}, \tag{408}$$

and

$$N_R \geq e^{M_R/T_0}. \tag{409}$$

Because $x/(x+y)$ increases in $x > 0$ and decreases in $y > 0$, (408)– (409) imply

$$p^{(j)}(B) \leq \frac{|B| e^{M_B/T_0}}{|B| e^{M_B/T_0} + e^{M_R/T_0}} = \frac{|B| e^{-(M_R - M_B)/T_0}}{1 + |B| e^{-(M_R - M_B)/T_0}}.$$

By (113), $M_R - M_B = \Delta_j(B, R; Q)$. Defining

$$a_j(B, R; Q) := |B| \exp\left[-\frac{\Delta_j(B, R; Q)}{\tau\lambda}\right]$$

therefore gives

$$p^{(j)}(B) \le \frac{a_j(B,R;Q)}{1+a_j(B,R;Q)}. \tag{410}$$

Since $a/(1+a) \le a$ for $a > 0$, this proves both inequalities in (114). The logistic bound is the sharper one and is always at most one. When $\Delta_j < 0$, the comparison set has a lower best score than B; the simpler bound a_j can then exceed one and be uninformative, whereas positive margins give exponential leakage decay.

We continue to pass from anchor-local policies to the LOT mixture.

By (11) and finiteness of the sums,

$$\pi^{\mathrm{LOT}}(Q) = \sum_{\ell} \mu_s(\ell) p^{(\ell)}(Q),$$
$$\pi^{\mathrm{LOT}}(B) = \sum_{i\in B} \sum_{\ell} \mu_s(\ell) p_i^{(\ell)}(Q) = \sum_{\ell} \mu_s(\ell) p^{(\ell)}(B).$$

For each ℓ, choose any nonempty $R_\ell \subseteq \mathcal{A} \setminus B$. Applying (410) anchorwise yields

$$\pi^{\mathrm{LOT}}(B) \le \sum_{\ell} \mu_s(\ell) \frac{a_\ell(B,R_\ell;Q)}{1+a_\ell(B,R_\ell;Q)},$$

which is (115). The comparison sets may differ across anchors; only the pointwise bounds and the fact that μ_s is a probability distribution are used.

We then separate prior-adjusted value from transport cost.

Define $U_i(Q) := Q_i + T_0 \log \rho_s(i)$. Then

$$S_{ij}(Q) = U_i(Q) - \tau C^L_{ij}. \tag{411}$$

Accordingly,

$$\Delta_j(B,R;Q) = \max_{k\in R}(U_k - \tau C^L_{kj}) - \max_{i\in B}(U_i - \tau C^L_{ij}).$$

Let $k_U \in R$ maximize U_k. Evaluating the first maximum at k_U gives

$$\max_{k\in R}(U_k - \tau C^L_{kj}) \ge \max_{k\in R} U_k - \tau \max_{k\in R} C^L_{kj},$$

while, for every $i \in B$,

$$U_i \le \max_{h\in B} U_h, \qquad -\tau C^L_{ij} \le -\tau \min_{h\in B} C^L_{hj},$$

and hence

$$\max_{i\in B}(U_i - \tau C^L_{ij}) \le \max_{i\in B} U_i - \tau \min_{i\in B} C^L_{ij}.$$

With

$$b_j(B,R) := \min_{i\in B} C^L_{ij} - \max_{k\in R} C^L_{kj},$$

subtracting yields

$$\Delta_j(B,R;Q) \ge \tau b_j(B,R) - \left(\max_{i\in B} U_i(Q) - \max_{k\in R} U_k(Q)\right). \tag{412}$$

We now apply the barrier condition.

Suppose $b_j(B,R) > 0$ and (116) holds, namely

$$\max_{i\in B} U_i(Q) - \max_{k\in R} U_k(Q) \le \tau b_j(B,R) - \eta.$$

Substitution into (412) gives

$$\Delta_j(B,R;Q) \ge \tau b_j(B,R) - [\tau b_j(B,R) - \eta] = \eta.$$

Since $x \mapsto e^{-x/T_0}$ is decreasing,

$$a_j(B, R; Q) \le |B| e^{-\eta/T_0} = |B| e^{-\eta/(\tau\lambda)}.$$

The second inequality in (114) now gives

$$\boxed{p^{(j)}(B) \le |B| e^{-\eta/(\tau\lambda)}.} \tag{413}$$

The sharper conclusion

$$p^{(j)}(B) \le \frac{|B| e^{-\eta/(\tau\lambda)}}{1 + |B| e^{-\eta/(\tau\lambda)}}$$

follows from the logistic bound.

Finally, we record edge cases and scope.

The condition $b_j(B, R) > 0$ encodes a genuine conservative min–max transport barrier: every action in B is costlier relative to j than the most costly action permitted in R. Algebraically, however, the implication from (116) to (413) remains valid for $b_j(B, R) \le 0$ whenever the displayed condition holds.

The choice $\eta > 0$ gives exponential decay. At $\eta = 0$ the bound is valid but may be no better than $|B|$, while for $\eta < 0$ it remains algebraically valid but is generally uninformative. The mixture claim requires a nonempty R_ℓ disjoint from B for each anchor, which is impossible for $B = \mathcal{A}$; it is thus intended for proper target subsets. Full support makes all adjusted scores finite. With zero prior masses one could assign score $-\infty$ to unsupported actions, but that extension lies outside the assumptions. Disjointness is essential to the set-mass decomposition above because overlap would double count $N_B + N_R$.

Finally, this is a state-local, one-backup probability bound. It implies neither global smoothness on the input action graph, persistence of locality under arbitrary Bellman recursion, nor convergence of an iterative policy algorithm. It needs no connectedness or effective-resistance hypothesis: only the exact local Gibbs law and score comparison enter.

This completes the proof of Theorem 16.

□

G.9.6 Proof of Proposition 4

Fix s. The standing assumptions give $\tau, \lambda > 0$; set $T_0 = \tau\lambda$ and $Z_\rho(Q) = \sum_i \rho_s(i) e^{Q_i/T_0}$. Let the normalized prior ρ_s have full support and μ_s be a probability distribution over reference actions. For all cost claims assume, for every s, i, j, $0 \le C_{s,ij} \le \bar{C}$, where $\bar{C} = \sup_s \|C_s\|_\infty < \infty$. For fixed-point claims the operators share a discounted MDP—bounded reward, transition kernel, and $0 \le \gamma < 1$—and differ only in state aggregation.

We first identify the zero-cost regularizer, policy, and value.

For $C_s = 0$ and every $\Gamma \in \Pi(\pi, \mu_s)$, the KL chain rule gives

$$\mathrm{KL}(\Gamma \| \rho_s \otimes \mu_s) = \mathrm{KL}(\pi \| \rho_s) + \sum_i \pi_i \mathrm{KL}(\Gamma(\cdot \mid i) \| \mu_s).$$

For $\pi_i = 0$, the weighted conditional-KL term is defined as zero and $\Gamma(\cdot \mid i)$ may be chosen arbitrarily. The last term is nonnegative, and the independent coupling $\Gamma = \pi \mu_s^\top$ makes it zero. Thus

$$W^\rho_{\lambda,s}(\pi, \mu_s)\big|_{C_s = 0} = \lambda \mathrm{KL}(\pi \| \rho_s),$$

and the aggregator is

$$F_s^{C=0}(Q) = \max_{\pi \in \Delta_\mathcal{A}} \{\langle \pi, Q \rangle - T_0 \mathrm{KL}(\pi \| \rho_s)\}.$$

Defining p_Q and expanding $\mathrm{KL}(\pi \| p_Q)$ gives

$$\begin{aligned} p_{Q,i} &= \frac{\rho_s(i) e^{Q_i/T_0}}{Z_\rho(Q)}, \\ \langle \pi, Q \rangle - T_0 \mathrm{KL}(\pi \| \rho_s) &= T_0 \log Z_\rho(Q) - T_0 \mathrm{KL}(\pi \| p_Q) \\ &\le T_0 \log Z_\rho(Q), \end{aligned}$$

with equality iff $\pi = p_Q$. Therefore

$$\boxed{\begin{aligned} F_s^{C=0}(Q) = F_s^{\rho}(Q) = T_0 \log \sum_i \rho_s(i) e^{Q_i/T_0}, \\ \pi_i^{\rho}(Q) = \frac{\rho_s(i) e^{Q_i/T_0}}{Z_\rho(Q)}. \end{aligned}} \tag{414}$$

This is prior-weighted, prior-normalized MaxEnt at matched temperature $T_0 = \tau\lambda$. Every zero-cost anchor has the same local problem, so μ_s disappears from value and policy, remaining only in the independent optimal plan $\Gamma^\star = \pi^\rho \mu_s^\top$.

Next, we compare each cost-based local partition with the zero-cost partition.

For a general nonnegative cost,

$$\begin{aligned} F_s^C(Q) &= T_0 \sum_j \mu_s(j) \log Z_{s,j}^C(Q), \\ Z_{s,j}^C(Q) &= \sum_i \rho_s(i) \exp\left(\frac{Q_i - \tau C_{s,ij}}{T_0}\right) \\ &= \sum_i \rho_s(i) e^{Q_i/T_0} e^{-C_{s,ij}/\lambda}, \end{aligned}$$

using $T_0 = \tau\lambda$. The cost bounds imply $e^{-\bar{C}/\lambda} \le e^{-C_{s,ij}/\lambda} \le 1$. Multiplying by $\rho_s(i) e^{Q_i/T_0} \ge 0$ and summing gives, for every j,

$$e^{-\bar{C}/\lambda} Z_\rho(Q) \le Z_{s,j}^C(Q) \le Z_\rho(Q). \tag{415}$$

Positivity permits logs and multiplication by T_0, giving $T_0 \log Z_\rho(Q) - T_0\bar{C}/\lambda \le T_0 \log Z_{s,j}^C(Q) \le T_0 \log Z_\rho(Q)$. Since $T_0/\lambda = \tau$,

$$F_s^\rho(Q) - \tau\bar{C} \le T_0 \log Z_{s,j}^C(Q) \le F_s^\rho(Q). \tag{416}$$

We next average over anchors and prove the uniform value gap.

As $\mu_s(j) \ge 0$ and $\sum_j \mu_s(j) = 1$, averaging (416) gives

$$F_s^\rho(Q) - \tau\bar{C} \le F_s^C(Q) \le F_s^\rho(Q). \tag{417}$$

Equivalently, uniformly in s, Q,

$$0 \le F_s^\rho(Q) - F_s^C(Q) \le \tau\bar{C}. \tag{418}$$

This proves (91) and the first inequality of (90). Normalization and $Q_i \le \max_k Q_k$ give $Z_\rho(Q) \le e^{\max_k Q_k/T_0} \sum_i \rho_s(i) = e^{\max_k Q_k/T_0}$, and hence

$$F_s^\rho(Q) \le \max_i Q_i. \tag{419}$$

Combining (417) and (419) proves the full ordering (90).

We continue to define and order the three Bellman operators.

Put $q_{S'} = Q(S', \cdot)$. The cost-based LOT, matched-prior MaxEnt, and hard-max Bellman operators are

$$\begin{aligned} (\mathcal{T}_C Q)(s,a) &= r(s,a) + \gamma \mathbb{E}_{S' \sim P(\cdot|s,a)}[F_{S'}^C(q_{S'})], \\ (\mathcal{T}_\rho Q)(s,a) &= r(s,a) + \gamma \mathbb{E}_{S' \sim P(\cdot|s,a)}[F_{S'}^\rho(q_{S'})], \\ (\mathcal{T} Q)(s,a) &= r(s,a) + \gamma \mathbb{E}_{S' \sim P(\cdot|s,a)}[\max_i Q(S', i)]. \end{aligned}$$

Expectation, $\gamma \ge 0$ (including $\gamma = 0$), and the common reward preserve the uniform value order above; hence every bounded Q satisfies

$$\mathcal{T}_C Q \le \mathcal{T}_\rho Q \le \mathcal{T} Q \quad \text{coordinatewise.} \tag{420}$$

Each aggregator is monotone and sup-norm one-Lipschitz: for $d = \|q - q'\|_\infty$, the bounds $q' - d\mathbf{1} \le q \le q' + d\mathbf{1}$ and translation equivariance imply $|F_s(q) - F_s(q')| \le d$; likewise for the hard maximum. Thus all operators are monotone γ-contractions, with unique fixed points $Q^\star_C$, $Q^\star_\rho$, and $Q^\star$.

We then prove the coordinatewise fixed-point ordering.

From $Q^\star_C = \mathcal{T}_C Q^\star_C$ and the first relation in (420), $Q^\star_C \le \mathcal{T}_\rho Q^\star_C$. For $q_0 = Q^\star_C$, $q_{n+1} = \mathcal{T}_\rho q_n$, monotonicity gives $q_0 \le q_1 \le q_2 \le \cdots$. Contraction iterates from every bounded point converge in sup norm to $Q^\star_\rho$, so $Q^\star_C \le Q^\star_\rho$ coordinatewise.

Likewise, $Q^\star_\rho = \mathcal{T}_\rho Q^\star_\rho \le \mathcal{T} Q^\star_\rho$; monotone hard-max iteration and contraction give $Q^\star_\rho \le Q^\star$. Therefore

$$\boxed{Q^\star_C \le Q^\star_\rho \le Q^\star \quad \text{coordinatewise.}} \tag{421}$$

This proves the ordering part of (92).

We now propagate the uniform backup gap to the fixed points.

Equation (418) implies, for bounded Q, $\|\mathcal{T}_\rho Q - \mathcal{T}_C Q\|_\infty \le \gamma\tau\bar{C}$. Adding and subtracting $\mathcal{T}_\rho Q^\star_C$ between the fixed-point equations gives

$$\begin{aligned}
\|Q^\star_\rho - Q^\star_C\|_\infty &= \|\mathcal{T}_\rho Q^\star_\rho - \mathcal{T}_C Q^\star_C\|_\infty \\
&\le \|\mathcal{T}_\rho Q^\star_\rho - \mathcal{T}_\rho Q^\star_C\|_\infty \\
&\quad + \|\mathcal{T}_\rho Q^\star_C - \mathcal{T}_C Q^\star_C\|_\infty \\
&\le \gamma \|Q^\star_\rho - Q^\star_C\|_\infty + \gamma\tau\bar{C}.
\end{aligned}$$

Rearranging,

$$\boxed{\|Q^\star_\rho - Q^\star_C\|_\infty \le \frac{\gamma\tau\bar{C}}{1-\gamma}.} \tag{422}$$

Together with (421), this proves all of (92).

We next explain the prior normalization.

If the prior is uniform, $\rho_s(i) = 1/K$, then $F^\rho_s(Q) = T_0 \log(\frac{1}{K}\sum_i e^{Q_i/T_0}) = T_0 \log \sum_i e^{Q_i/T_0} - T_0 \log K$, which is the claimed uniform-prior normalization, although the policy is the usual softmax $\pi^\rho_i(Q) = e^{Q_i/T_0} / \sum_k e^{Q_k/T_0}$. The Shannon-entropy-bonus value $T_0 \log \sum_i e^{Q_i/T_0}$ is larger by $T_0 \log K$. This offset ensures $F^\rho_s(Q) \le \max_i Q_i$; the unshifted value need not obey that numerical order. It leaves the one-step policy unchanged but changes absolute Bellman values and discounted fixed points. The proposition therefore requires the same prior-normalized convention and effective temperature on both sides.

Finally, we record the scope of the ordering.

The upper partition bound uses $C_s \ge 0$; the uniform lower bound uses $\|C_s\|_\infty \le \bar{C}$. Negative entries can make $F^C_s(Q) > F^\rho_s(Q)$, while unbounded costs give no displayed finite perturbation constant.

The order $Q^\star_C \le Q^\star_\rho \le Q^\star$ concerns regularized numerical fixed points only. It does not order the three induced policies' unregularized returns, prove smaller simple regret, or show that adding geometry reduces regularization bias. With matched prior and temperature, nonnegative uncentered transport cost only lowers the regularized backup from MaxEnt.

The argument applies directly to ordinary uncentered cost-based LOT, whose value is F^{logH}_s. Centered direct-heat LOT has a known additive translation from the corresponding uniform-prior log-heat cost representation; remove it before applying the order, or one compares value normalizations rather than nonnegative transport cost. For $c = T_0 \log K$, $F^{\mathrm{HK}}_s = F^{\mathrm{logH}}_s + c$. Restoring c leaves the pointwise policy unchanged but shifts the action-value fixed point by $\gamma c/(1-\gamma)$ and the state-value fixed point by $c/(1-\gamma)$.

Finally, the coordinate order through $Q^\star$ alone gives no uniform numerical bound on $\|Q^\star - Q^\star_\rho\|_\infty$ or $\|Q^\star - Q^\star_C\|_\infty$; such hard-max regularization-bias bounds require the separate argument in Proposition 3.

This completes the proof of Proposition 4.

□

G.9.7 Proof of Theorem 17

Let G be the finite connected undirected weighted graph from the theorem, with symmetric combinatorial Laplacian, and $K \geq 2$; $K = 1$ is trivial but has no λ_2 notation. For heat semigroups, $t \geq 0$. Set

$$T_0 := \tau\lambda, \qquad \delta_t := 2e^{-2t\lambda_2}, \qquad d_t := \delta_t/\lambda,$$

where $\tau, \lambda > 0$. Connectedness makes the zero eigenvalue simple, $\lambda_2 > 0$, and permits the corresponding normalized eigenvector $\phi_1 = K^{-1/2}\mathbf{1}$, so $\phi_1(i) - \phi_1(j) = 0$. For the fixed-point claim, operators share the bounded reward, transition kernel, and discount $0 \leq \gamma < 1$; only their state aggregators differ.

We first bound every diffusion-distance entry by the spectral gap. For actions i, j, nonnegativity of the summands gives $(C_t^L)_{ij} \geq 0$. Since $t \geq 0$ and $\lambda_k \geq \lambda_2$ for $k \geq 2$, $e^{-2t\lambda_k} \leq e^{-2t\lambda_2}$, whence

$$\begin{aligned}(C_t^L)_{ij} &= \sum_{k=2}^{K} e^{-2t\lambda_k}(\phi_k(i) - \phi_k(j))^2 \\ &\leq e^{-2t\lambda_2} \sum_{k=2}^{K} (\phi_k(i) - \phi_k(j))^2.\end{aligned}$$

For $v_{ij} := e_i - e_j$, orthonormality and completeness of the eigenvectors give Parseval's identity:

$$\sum_{k=1}^{K} (\phi_k(i) - \phi_k(j))^2 = \sum_{k=1}^{K} (\phi_k^\top v_{ij})^2 = \|v_{ij}\|_2^2.$$

The $k = 1$ term vanishes; hence

$$\sum_{k=2}^{K} (\phi_k(i) - \phi_k(j))^2 = \|e_i - e_j\|_2^2 = \begin{cases} 0, & i = j, \\ 2, & i \neq j. \end{cases}$$

Thus, for all i, j,

$$0 \leq (C_t^L)_{ij} \leq 2e^{-2t\lambda_2} = \delta_t.$$

Taking the largest absolute entry proves (118), whose cost-matrix ℓ_∞ norm is entrywise maximum; moreover $(C_t^L)_{ii} = 0$ exactly.

Next, we compare each local partition function with its zero-cost counterpart. Fix s and $Q \in \mathbb{R}^K$, and define the strictly positive unnormalized MaxEnt weights, their partition, and the zero-cost prior-weighted backup by

$$q_i := \rho_s(i) e^{Q_i/T_0} > 0, \qquad Z_0 := \sum_i q_i > 0,$$

$$F_s^\rho(Q) = T_0 \log Z_0,$$

where positivity follows from full support of ρ_s. For anchor j, let

$$Z_{j,t} := \sum_i q_i e^{-(C_t^L)_{ij}/\lambda}, \qquad f_{j,t}(Q) := T_0 \log Z_{j,t}.$$

This is the local log-partition in (12), since

$$q_i e^{-(C_t^L)_{ij}/\lambda} = \rho_s(i) \exp\left(\frac{Q_i - \tau(C_t^L)_{ij}}{\tau\lambda}\right).$$

The spectral-gap cost bound yields

$$e^{-\delta_t/\lambda} \leq e^{-(C_t^L)_{ij}/\lambda} \leq 1.$$

Multiplying by $q_i > 0$ and summing gives

$$e^{-d_t} Z_0 \leq Z_{j,t} \leq Z_0.$$

Monotonicity of log and $T_0 = \tau\lambda$ then give

$$\log Z_0 - d_t \le \log Z_{j,t} \le \log Z_0,$$

$$F_s^\rho(Q) - \tau\delta_t \le f_{j,t}(Q) \le F_s^\rho(Q),$$
$$0 \le F_s^\rho(Q) - f_{j,t}(Q) \le \tau\delta_t.$$

Because $\mu_{s,t}$ is a probability distribution,

$$F_{s,t}^{\mathrm{LOT}}(Q) := \sum_j \mu_{s,t}(j) f_{j,t}(Q),$$
$$0 \le F_s^\rho(Q) - F_{s,t}^{\mathrm{LOT}}(Q) \le \tau\delta_t.$$

This uniform-in-$s, Q, \mu_{s,t}$ one-sided result is stronger than, and proves, the absolute-value claim (119).

We continue to derive the exact anchor-local likelihood ratio. At the fixed s, Q, define the full-support geometry-free softmax and the anchor-j local LOT policy

$$p_{0,i} := \pi^{\mathrm{soft}}_{\rho_s,i}(Q) = \frac{q_i}{Z_0} > 0, \qquad p_{t,i}^{(j)} = \frac{q_i e^{-(C_t^L)_{ij}/\lambda}}{Z_{j,t}}.$$

The local normalizing ratio is the p_0-expectation of a quantity in $[e^{-d_t}, 1]$:

$$m_{j,t} := \frac{Z_{j,t}}{Z_0} = \sum_k p_{0,k} e^{-(C_t^L)_{kj}/\lambda},$$
$$e^{-d_t} \le m_{j,t} \le 1.$$

Consequently the exact componentwise ratio is

$$r_{ij,t} := \frac{p_{t,i}^{(j)}}{p_{0,i}} = \frac{e^{-(C_t^L)_{ij}/\lambda}}{m_{j,t}},$$
$$\log r_{ij,t} = -\frac{(C_t^L)_{ij}}{\lambda} - \log m_{j,t}.$$

The two terms lie respectively in $[-d_t, 0]$ and $[0, d_t]$, so

$$-d_t \le \log r_{ij,t} \le d_t, \qquad e^{-d_t} \le r_{ij,t} \le e^{d_t}.$$

Thus the bound includes both the direct geometry factor and the changed local partition function.

We then bound the local KL divergence and local ℓ_1 error. The ratio gives the exact identity

$$D_{\mathrm{KL}}(p_t^{(j)} \| p_0) = \sum_i p_{t,i}^{(j)} \log r_{ij,t}$$
$$= -\frac{1}{\lambda} \sum_i p_{t,i}^{(j)} (C_t^L)_{ij} - \log m_{j,t}.$$

The cost term is nonpositive and $-\log m_{j,t} \le d_t$, hence

$$0 \le D_{\mathrm{KL}}(p_t^{(j)} \| p_0) \le d_t;$$

the same follows directly from $\log r_{ij,t} \le d_t$. Since $p_{t,i}^{(j)} = p_{0,i} r_{ij,t}$,

$$\|p_t^{(j)} - p_0\|_1 = \sum_i p_{0,i} |r_{ij,t} - 1|$$
$$\le \max\{e^{d_t} - 1, 1 - e^{-d_t}\} \sum_i p_{0,i}$$
$$= e^{d_t} - 1,$$

because $d_t \geq 0$ implies $1 - e^{-d_t} \leq e^{d_t} - 1$.

We now pass from the local policies to the LOT mixture. Because the same full-support p_0 serves every anchor, (11) gives

$$\pi_{s,t}^{\mathrm{LOT}}(Q) = \sum_j \mu_{s,t}(j) p_t^{(j)},$$

$$\frac{\pi_{s,t,i}^{\mathrm{LOT}}(Q)}{p_{0,i}} = \sum_j \mu_{s,t}(j) \frac{p_{t,i}^{(j)}}{p_{0,i}} = \sum_j \mu_{s,t}(j) r_{ij,t}.$$

The latter is a convex combination in $[e^{-d_t}, e^{d_t}]$, so, for every i,

$$e^{-d_t} \leq \frac{\pi_{s,t,i}^{\mathrm{LOT}}(Q)}{p_{0,i}} \leq e^{d_t}.$$

Therefore

$$D_{\mathrm{KL}}(\pi_{s,t}^{\mathrm{LOT}}(Q) \| p_0) = \sum_i \pi_{s,t,i}^{\mathrm{LOT}}(Q) \log \frac{\pi_{s,t,i}^{\mathrm{LOT}}(Q)}{p_{0,i}} \leq d_t \sum_i \pi_{s,t,i}^{\mathrm{LOT}}(Q) = d_t.$$

Substitution of $d_t = \delta_t / \lambda$ proves (121). Equivalently, convexity gives

$$D_{\mathrm{KL}}\left(\sum_j \mu_{s,t}(j) p_t^{(j)} \middle\| p_0 \right) \leq \sum_j \mu_{s,t}(j) D_{\mathrm{KL}}(p_t^{(j)} \| p_0) \leq d_t.$$

Likewise, by the mixed ratio or norm convexity,

$$\|\pi_{s,t}^{\mathrm{LOT}}(Q) - p_0\|_1 \leq \sum_j \mu_{s,t}(j) \|p_t^{(j)} - p_0\|_1 \leq e^{d_t} - 1,$$

which is (122). Since any two distributions are at ℓ_1 distance at most 2, this may be sharpened to $\min\{2, e^{d_t} - 1\}$, though the theorem states the simpler ratio bound. Despite the label "tv," the display uses ℓ_1; under $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$, halve its right-hand side.

We next verify contraction and propagate the one-step value error to the fixed points. Both aggregators are sup-norm one-Lipschitz: if $a = \|Q - Q'\|_\infty$, every local exponential weight changes by a factor in $[e^{-a/T_0}, e^{a/T_0}]$, so every local log-partition, its anchor average, and F_s^ρ change by at most a. Their Bellman operators are therefore γ-contractions with unique fixed points. Since their rewards and kernels coincide, the local partition comparison gives, for every (s, a),

$$\begin{aligned} &|(\mathcal{T}_t^{\mathrm{LOT}} Q)(s, a) - (\mathcal{T}_\rho Q)(s, a)| \\ &= \gamma \Big| \mathbb{E}_{s' \sim P(\cdot|s,a)} \Big[F_{s',t}^{\mathrm{LOT}}(Q(s', \cdot)) \\ &\qquad - F_{s'}^\rho(Q(s', \cdot)) \Big] \Big| \\ &\leq \gamma \tau \delta_t, \end{aligned}$$

and thus, uniformly in Q,

$$\|\mathcal{T}_t^{\mathrm{LOT}} Q - \mathcal{T}_\rho Q\|_\infty \leq \gamma \tau \delta_t.$$

The same partition comparison also gives $\mathcal{T}_t^{\mathrm{LOT}} Q \leq \mathcal{T}_\rho Q$ coordinatewise, so monotonicity and contraction yield the stronger order $Q_t^\star \leq Q_\rho^\star$, although the theorem only requires the norm bound

below. At the fixed points,

$$\begin{aligned}\|Q_t^\star - Q_\rho^\star\|_\infty &= \|\mathcal{T}_t^{\mathrm{LOT}} Q_t^\star - \mathcal{T}_\rho Q_\rho^\star\|_\infty \\ &\le \|\mathcal{T}_t^{\mathrm{LOT}} Q_t^\star - \mathcal{T}_t^{\mathrm{LOT}} Q_\rho^\star\|_\infty + \|\mathcal{T}_t^{\mathrm{LOT}} Q_\rho^\star - \mathcal{T}_\rho Q_\rho^\star\|_\infty \\ &\le \gamma \|Q_t^\star - Q_\rho^\star\|_\infty + \gamma\tau\delta_t.\end{aligned}$$

Rearranging,

$$\|Q_t^\star - Q_\rho^\star\|_\infty \le \frac{\gamma\tau\delta_t}{1-\gamma} = \frac{2\gamma\tau}{1-\gamma} e^{-2t\lambda_2},$$

which proves (120).

Finally, we record edge cases and scope of the comparison. At $t = 0$ the proof remains valid with $\delta_0 = 2$, not necessarily a small estimate. As $t \to \infty$, connectedness gives $\lambda_2 > 0$, so $\delta_t \to 0$ exponentially and all three asserted discrepancies vanish. The marginal $\mu_{s,t}$ may vary with t: the bounds are uniform in it because every anchor has the same zero-cost local policy. Full support of ρ_s makes all displayed likelihood ratios defined; full support of $\mu_{s,t}$ is a standing LOT assumption, although this proof needs only nonnegativity and unit mass.

The envelope does not assert that $t \mapsto F_{s,t}^{\mathrm{LOT}}(Q)$ is monotone when $\mu_{s,t}$ varies, and is not an online tracking theorem for a reference distribution changing within one Bellman iteration. The policy bounds compare policies at the same fixed Q; by themselves, they do not compare policies evaluated at $Q_t^\star$ and $Q_\rho^\star$. This argument concerns Gibbs similarities from diffusion-distance cost and implies neither finite support nor a local algorithm for constructing the generally dense Gibbs columns. Finally, connectedness and $t \ge 0$ are essential for exponential decay: disconnected graphs retain additional zero modes, whereas for $t < 0$ the inequality $e^{-2t\lambda_k} \le e^{-2t\lambda_2}$ may fail. This completes the proof. □

G.10 Learned geometry and single-trajectory estimates

G.10.1 Proof of Theorem 18

Set

$$T_0 := \tau\lambda, \qquad d_C := \varepsilon_C/\lambda.$$

Here $\tau, \lambda > 0$; ρ_s, μ_s are common full-support action and reference distributions; and

$$|\widehat{C}_{s,ij} - C_{s,ij}| \le \varepsilon_C \quad \text{for all } s, i, j.$$

Bounded, entrywise-finite costs make every local partition finite and positive. Full support of ρ_s makes every local and mixed LOT policy strictly positive.

We first compare the local partition functions multiplicatively. Fix s, $Q \in \mathbb{R}^K$, and anchor j; suppress s and set

$$\begin{aligned}q_i &:= \rho_s(i) e^{Q_i/T_0} > 0, \\ Z_j^C &:= \sum_i q_i e^{-C_{ij}/\lambda}, \\ Z_j^{\widehat{C}} &:= \sum_i q_i e^{-\widehat{C}_{ij}/\lambda}.\end{aligned}$$

The first row below is the cost bound; dividing it by $\lambda > 0$, changing sign, and exponentiating gives the second:

$$\begin{aligned}-\varepsilon_C \le \widehat{C}_{ij} - C_{ij} &\le \varepsilon_C, \\ e^{-d_C} \le e^{-(\widehat{C}_{ij} - C_{ij})/\lambda} &\le e^{d_C}.\end{aligned}$$

Multiplication by $q_i e^{-C_{ij}/\lambda} > 0$ gives

$$\begin{aligned}e^{-d_C} q_i e^{-C_{ij}/\lambda} &\le q_i e^{-\widehat{C}_{ij}/\lambda} \\ &\le e^{d_C} q_i e^{-C_{ij}/\lambda}.\end{aligned}$$

Summation in i yields the exact multiplicative comparison

$$e^{-d_C} Z_j^C \le Z_j^{\widehat{C}} \le e^{d_C} Z_j^C.$$

Positivity then gives

$$e^{-d_C} \le \frac{Z_j^{\widehat{C}}}{Z_j^C} \le e^{d_C}, \qquad |\log Z_j^{\widehat{C}} - \log Z_j^C| \le d_C.$$

Next, we convert the partition comparison into the uniform aggregator bound. The multiplicative partition bound and the local-softmax formula give

$$\begin{aligned} f_j^C(Q) &= T_0 \log Z_j^C, \\ f_j^{\widehat{C}}(Q) &= T_0 \log Z_j^{\widehat{C}}, \\ |f_j^{\widehat{C}}(Q) - f_j^C(Q)| &= T_0 |\log Z_j^{\widehat{C}} - \log Z_j^C| \\ &\le T_0 d_C = \tau \varepsilon_C. \end{aligned}$$

For common μ_s,

$$\begin{aligned} F_s^C(Q) &= \sum_j \mu_s(j) f_j^C(Q), \\ F_s^{\widehat{C}}(Q) &= \sum_j \mu_s(j) f_j^{\widehat{C}}(Q), \end{aligned}$$

so the triangle inequality and $\sum_j \mu_s(j) = 1$ yield

$$\begin{aligned} &|F_s^{\widehat{C}}(Q) - F_s^C(Q)| \\ &= \left| \sum_j \mu_s(j) (f_j^{\widehat{C}}(Q) - f_j^C(Q)) \right| \\ &\le \sum_j \mu_s(j) |f_j^{\widehat{C}}(Q) - f_j^C(Q)| \\ &\le \tau \varepsilon_C. \end{aligned}$$

Thus (123) holds uniformly in s, Q. No sign order follows because $\widehat{C} - C$ may have either sign.

We continue to propagate the one-step error through the Bellman fixed point. Both aggregators are sup-norm one-Lipschitz: for $a = \|Q - Q'\|_\infty$ and every i,

$$e^{-a/T_0} e^{Q_i/T_0} \le e^{Q_i'/T_0} \le e^{a/T_0} e^{Q_i/T_0}.$$

Every local partition obeys these bounds, so its T_0-scaled log, and then its anchor average, changes by at most a. Thus $\mathcal{T}^C, \mathcal{T}^{\widehat{C}}$ are γ-contractions with unique fixed points.

Common rewards and transitions and (123) give, for every bounded Q and (s, a),

$$\begin{aligned} &|(\mathcal{T}^{\widehat{C}} Q)(s,a) - (\mathcal{T}^C Q)(s,a)| \\ &= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[F_{s'}^{\widehat{C}}(Q(s', \cdot)) - F_{s'}^C(Q(s', \cdot)) \right] \right| \\ &\le \gamma \tau \varepsilon_C, \end{aligned}$$

whence

$$\|\mathcal{T}^{\widehat{C}} Q - \mathcal{T}^C Q\|_\infty \le \gamma \tau \varepsilon_C \qquad \text{for every } Q.$$

The fixed-point equations, with $\mathcal{T}^{\widehat{C}} Q_C^\star$ inserted, give

$$\begin{aligned} \|Q_{\widehat{C}}^\star - Q_C^\star\|_\infty = \|\mathcal{T}^{\widehat{C}} Q_{\widehat{C}}^\star - \mathcal{T}^C Q_C^\star\|_\infty &\le \|\mathcal{T}^{\widehat{C}} Q_{\widehat{C}}^\star - \mathcal{T}^{\widehat{C}} Q_C^\star\|_\infty + \|\mathcal{T}^{\widehat{C}} Q_C^\star - \mathcal{T}^C Q_C^\star\|_\infty \\ &\le \gamma \|Q_{\widehat{C}}^\star - Q_C^\star\|_\infty + \gamma \tau \varepsilon_C. \end{aligned}$$

Solving with $1 - \gamma > 0$ gives

$$\|Q_{\widehat{C}}^\star - Q_C^\star\|_\infty \le \frac{\gamma \tau}{1 - \gamma} \varepsilon_C,$$

i.e., (124).

We then derive the exact local policy likelihood ratio. For the fixed s, Q, j,

$$p_i^{C,(j)} = \frac{q_i e^{-C_{ij}/\lambda}}{Z_j^C}, \qquad p_i^{\widehat{C},(j)} = \frac{q_i e^{-\widehat{C}_{ij}/\lambda}}{Z_j^{\widehat{C}}}.$$

Their exact likelihood ratio is

$$R_{ij} := \frac{p_i^{\widehat{C},(j)}}{p_i^{C,(j)}} = e^{-(\widehat{C}_{ij}-C_{ij})/\lambda} \frac{Z_j^C}{Z_j^{\widehat{C}}},$$
$$\log R_{ij} = -\frac{\widehat{C}_{ij} - C_{ij}}{\lambda} + \log \frac{Z_j^C}{Z_j^{\widehat{C}}}.$$

Both factors lie in $[e^{-d_C}, e^{d_C}]$: the cost bound controls the first and the multiplicative comparison above gives the inverse partition ratio. Hence

$$e^{-2d_C} \le R_{ij} \le e^{2d_C}, \qquad |\log R_{ij}| \le 2d_C,$$
$$e^{-2d_C} \le \frac{p_i^{C,(j)}}{p_i^{\widehat{C},(j)}} \le e^{2d_C}.$$

The factor 2 accounts for one e^{d_C} each from the unnormalized weight and its anchor-local normalizer.

We now establish both local KL directions and the local ℓ_1 bound. The ratio formula gives the exact identities

$$D_{\mathrm{KL}}\left(p^{\widehat{C},(j)} \| p^{C,(j)}\right) = -\frac{1}{\lambda}\mathbb{E}_{i\sim p^{\widehat{C},(j)}}[\widehat{C}_{ij} - C_{ij}] + \log \frac{Z_j^C}{Z_j^{\widehat{C}}},$$
$$D_{\mathrm{KL}}\left(p^{C,(j)} \| p^{\widehat{C},(j)}\right) = \frac{1}{\lambda}\mathbb{E}_{i\sim p^{C,(j)}}[\widehat{C}_{ij} - C_{ij}] + \log \frac{Z_j^{\widehat{C}}}{Z_j^C}.$$

Each cost contribution and log-partition ratio has absolute value at most d_C; hence

$$D_{\mathrm{KL}}\left(p^{\widehat{C},(j)} \| p^{C,(j)}\right) \le 2d_C,$$
$$D_{\mathrm{KL}}\left(p^{C,(j)} \| p^{\widehat{C},(j)}\right) \le 2d_C.$$

The pointwise log-likelihood bounds above give the same conclusion. Also,

$$\begin{aligned}\|p^{\widehat{C},(j)} - p^{C,(j)}\|_1 &= \sum_i p_i^{C,(j)} |R_{ij} - 1| \\ &\le \max\{e^{2d_C} - 1, 1 - e^{-2d_C}\} \\ &\quad \times \sum_i p_i^{C,(j)} \\ &= e^{2d_C} - 1,\end{aligned}$$

because $d_C \ge 0$ makes $1 - e^{-2d_C} \le e^{2d_C} - 1$.

We next show that the componentwise ratio bounds survive anchor mixing. Common μ_s gives the LOT policies

$$\pi_s^C(Q) = \sum_j \mu_s(j) p^{C,(j)},$$
$$\pi_s^{\widehat{C}}(Q) = \sum_j \mu_s(j) p^{\widehat{C},(j)}.$$

For every i, j, the pointwise ratio bound above gives the first line below; multiplying it by $\mu_s(j) \geq 0$ and summing over j gives the second:

$$e^{-2d_C} p_i^{C,(j)} \leq p_i^{\widehat{C},(j)} \leq e^{2d_C} p_i^{C,(j)},$$
$$e^{-2d_C} \pi_{s,i}^{C}(Q) \leq \pi_{s,i}^{\widehat{C}}(Q) \leq e^{2d_C} \pi_{s,i}^{C}(Q).$$

Strict positivity permits division:

$$e^{-2d_C} \leq \frac{\pi_{s,i}^{\widehat{C}}(Q)}{\pi_{s,i}^{C}(Q)} \leq e^{2d_C} \qquad \text{for every } i.$$

Its explicit convex-combination form is

$$\frac{\pi_{s,i}^{\widehat{C}}(Q)}{\pi_{s,i}^{C}(Q)} = \sum_j \omega_{ij} R_{ij},$$
$$\omega_{ij} := \frac{\mu_s(j) p_i^{C,(j)}}{\sum_\ell \mu_s(\ell) p_i^{C,(\ell)}},$$
$$\omega_{ij} \geq 0, \qquad \sum_j \omega_{ij} = 1.$$

This explains why the anchor distribution must be common. Reciprocation gives the same interval for $\pi_{s,i}^{C}(Q)/\pi_{s,i}^{\widehat{C}}(Q)$.

We continue by deriving the mixed-policy KL and ℓ_1 bounds. The mixed componentwise ratios imply

$$D_{\mathrm{KL}}\left(\pi_s^{\widehat{C}}(Q) \| \pi_s^{C}(Q)\right)$$
$$= \sum_i \pi_{s,i}^{\widehat{C}}(Q) \log \frac{\pi_{s,i}^{\widehat{C}}(Q)}{\pi_{s,i}^{C}(Q)}$$
$$\leq 2d_C \sum_i \pi_{s,i}^{\widehat{C}}(Q) = 2d_C,$$
$$D_{\mathrm{KL}}\left(\pi_s^{C}(Q) \| \pi_s^{\widehat{C}}(Q)\right)$$
$$= \sum_i \pi_{s,i}^{C}(Q) \log \frac{\pi_{s,i}^{C}(Q)}{\pi_{s,i}^{\widehat{C}}(Q)} \leq 2d_C.$$

Substitution of $d_C = \varepsilon_C/\lambda$ proves (125); alternatively, joint convexity of KL and The local bounds above give both directions.

With $R_i = \pi_{s,i}^{\widehat{C}}(Q)/\pi_{s,i}^{C}(Q)$,

$$\|\pi_s^{\widehat{C}}(Q) - \pi_s^{C}(Q)\|_1 = \sum_i \pi_{s,i}^{C}(Q)|R_i - 1|$$
$$\leq e^{2d_C} - 1 = \exp\left(\frac{2\varepsilon_C}{\lambda}\right) - 1.$$

This is (126). Since probability-vector ℓ_1 distance is at most 2, it sharpens to $\min\{2, \exp(2\varepsilon_C/\lambda) - 1\}$; the theorem states the direct likelihood-ratio consequence.

Finally, we record edge cases and scope. If $\varepsilon_C = 0$, costs coincide entrywise, all partition and likelihood ratios equal one, and every displayed error bound vanishes exactly. For $\varepsilon_C > 0$, value perturbation is linear in $\tau\varepsilon_C$ and lacks $1/\lambda$, because the log partition's $1/\lambda$ cancels with $T_0 = \tau\lambda$. Policy sensitivity retains $1/\lambda$, so its bound weakens as entropic smoothing vanishes.

The fixed-point result needs state-uniform cost error and common reward and transition model. The policy claims compare policies at a common state and a common Q; they do not by themselves compare readouts at the two different fixed points. Common ρ_s, μ_s, and temperatures are essential: changing any adds error uncontrolled by ε_C. Finally, this stability upper bound asserts neither statistical consistency of the learned cost nor an unregularized policy-return guarantee from the regularized fixed-point bound. This completes the proof. □

G.10.2 Proof of Corollary 5

For cost perturbations use the entrywise matrix sup norm

$$\|D\|_\infty := \max_{i,j} |D_{ij}|,$$

and for linear maps the Euclidean operator norm $\|\cdot\|_2$. All feature and spectral-filter bounds are uniform over states when the geometry is state dependent (automatic for a shared geometry). Both LOT constructions have the same full-support action prior ρ_s, anchor distribution μ_s, and temperatures $\tau, \lambda > 0$, as required by Theorem 18.

We first interpret the M-seminorm and record its basic inequalities. Since $M \succeq 0$ has a positive-semidefinite square root,

$$\|x\|_M = \sqrt{x^\top M x} = \|M^{1/2} x\|_2.$$

Thus $\|\cdot\|_M$ is a seminorm (a norm if $M \succ 0$), and the Euclidean triangle and Cauchy–Schwarz inequalities applied after $M^{1/2}$ give

$$\|x+y\|_M \le \|x\|_M + \|y\|_M, \qquad |x^\top M y| \le \|x\|_M \|y\|_M.$$

These remain valid for singular M, so invertibility is unnecessary.

Next, we bound every feature-induced cost entry. For fixed i, j, let

$$d_{ij} := \Phi_i - \Phi_j, \quad u_i := \widehat{\Phi}_i - \Phi_i, \quad e_{ij} := u_i - u_j,$$

so $\widehat{\Phi}_i - \widehat{\Phi}_j = d_{ij} + e_{ij}$. The assumptions and triangle inequality yield

$$\|d_{ij}\|_M \le \|\Phi_i\|_M + \|\Phi_j\|_M \le 2B,$$
$$\|e_{ij}\|_M \le \|u_i\|_M + \|u_j\|_M \le 2\eta.$$

Consequently,

$$\begin{aligned} C^{\widehat{\Phi}}_{ij} - C^{\Phi}_{ij} &= \|d_{ij} + e_{ij}\|_M^2 - \|d_{ij}\|_M^2 \\ &= 2 d_{ij}^\top M e_{ij} + \|e_{ij}\|_M^2, \\ |C^{\widehat{\Phi}}_{ij} - C^{\Phi}_{ij}| &\le 2|d_{ij}^\top M e_{ij}| + \|e_{ij}\|_M^2 \\ &\le 2\|d_{ij}\|_M \|e_{ij}\|_M + \|e_{ij}\|_M^2 \\ &\le 2(2B)(2\eta) + (2\eta)^2 \\ &= 8B\eta + 4\eta^2. \end{aligned}$$

The bound is independent of i, j, hence

$$\|C^{\widehat{\Phi}} - C^{\Phi}\|_\infty \le 8B\eta + 4\eta^2.$$

For completeness, the simultaneous bound $\|\widehat{\Phi}_i\|_M \le B$ also gives the optional sharpening

$$\begin{aligned} |C^{\widehat{\Phi}}_{ij} - C^{\Phi}_{ij}| &= \big| \|\widehat{\Phi}_i - \widehat{\Phi}_j\|_M - \|\Phi_i - \Phi_j\|_M \big| \\ &\quad \times \big(\|\widehat{\Phi}_i - \widehat{\Phi}_j\|_M + \|\Phi_i - \Phi_j\|_M \big) \\ &\le \|e_{ij}\|_M (2B + 2B) \le 8B\eta. \end{aligned}$$

The corollary instead uses the direct expansion bound $8B\eta + 4\eta^2$, which does not exploit this sharpening.

We then apply learned-geometry stability to the feature costs. Set $\varepsilon_\Phi := 8B\eta + 4\eta^2$. The feature-cost bound above verifies the cost-error hypothesis of Theorem 18 with $\varepsilon_C = \varepsilon_\Phi$. Thus, for every state s and common action-value vector Q,

$$|F_s^{C^{\widehat{\Phi}}}(Q) - F_s^{C^{\Phi}}(Q)| \le \tau \varepsilon_\Phi,$$
$$\|\mathcal{T}^{C^{\widehat{\Phi}}} Q - \mathcal{T}^{C^{\Phi}} Q\|_\infty \le \gamma \tau \varepsilon_\Phi.$$

Its fixed-point clause gives

$$\|Q^\star_{\hat{\Phi}} - Q^\star_{\Phi}\|_\infty \le \frac{\gamma\tau}{1-\gamma}\varepsilon_\Phi = \frac{\gamma\tau}{1-\gamma}(8B\eta + 4\eta^2),$$

the feature-embedding fixed-point claim. Although the corollary displays only this consequence, the policy clauses of Theorem 18 also give, at any common s, Q,

$$\max\left\{\begin{array}{l} D_{\mathrm{KL}}\left(\pi_s^{C^\Phi}(Q) \,\|\, \pi_s^{C^{\hat{\Phi}}}(Q)\right), \\ D_{\mathrm{KL}}\left(\pi_s^{C^{\hat{\Phi}}}(Q) \,\|\, \pi_s^{C^\Phi}(Q)\right) \end{array}\right\} \le \frac{2\varepsilon_\Phi}{\lambda},$$

and

$$\|\pi_s^{C^{\hat{\Phi}}}(Q) - \pi_s^{C^\Phi}(Q)\|_1 \le \exp\left(\frac{2\varepsilon_\Phi}{\lambda}\right) - 1.$$

These are inherited consequences, not additional assumptions.

We now bound every spectral-filter cost entry. Let

$$E_g := g(\widehat{L}) - g(L), \qquad v_{ij} := e_i - e_j.$$

Because $L, \widehat{L}$ are symmetric and g is defined by spectral functional calculus, $g(L), g(\widehat{L})$, and E_g are symmetric; by assumption, $\|E_g\|_2 \le \varepsilon_g$. Hence

$$C^{\widehat{L}}_{ij} - C^L_{ij} = v_{ij}^\top g(\widehat{L}) v_{ij} - v_{ij}^\top g(L) v_{ij} = v_{ij}^\top E_g v_{ij},$$
$$|C^{\widehat{L}}_{ij} - C^L_{ij}| \le \|E_g\|_2 \|v_{ij}\|_2^2 \le \varepsilon_g \|e_i - e_j\|_2^2 \le 2\varepsilon_g,$$

where we used $|v^\top E v| \le \|E\|_2 \|v\|_2^2$ for symmetric E, and $\|e_i - e_j\|_2^2 = 2$ for $i \ne j$ (zero for $i = j$). Taking the maximum over all entries gives

$$\|C^{\widehat{L}} - C^L\|_\infty \le 2\varepsilon_g.$$

We next apply learned-geometry stability to the spectral-filter costs. Set $\varepsilon_L := 2\varepsilon_g$. Theorem 18, with $\varepsilon_C = \varepsilon_L$, gives for every s, Q

$$|F_s^{C^{\hat{L}}}(Q) - F_s^{C^L}(Q)| \le 2\tau\varepsilon_g,$$
$$\|\mathcal{T}^{C^{\hat{L}}} Q - \mathcal{T}^{C^L} Q\|_\infty \le 2\gamma\tau\varepsilon_g,$$

and

$$\|Q^\star_{\widehat{L}} - Q^\star_L\|_\infty \le \frac{\gamma\tau}{1-\gamma}\varepsilon_L = \frac{2\gamma\tau}{1-\gamma}\varepsilon_g,$$

which is the spectral-filter claim. At any common s, Q, its policy clauses also give

$$\max\left\{\begin{array}{l} D_{\mathrm{KL}}\left(\pi_s^{C^L}(Q) \,\|\, \pi_s^{C^{\hat{L}}}(Q)\right), \\ D_{\mathrm{KL}}\left(\pi_s^{C^{\hat{L}}}(Q) \,\|\, \pi_s^{C^L}(Q)\right) \end{array}\right\} \le \frac{4\varepsilon_g}{\lambda},$$

and

$$\|\pi_s^{C^{\hat{L}}}(Q) - \pi_s^{C^L}(Q)\|_1 \le \exp\left(\frac{4\varepsilon_g}{\lambda}\right) - 1.$$

Finally, we record edge cases and scope. If $\eta = 0$, every feature perturbation lies in the null space of the M-seminorm, so the feature-induced costs agree exactly. If $\varepsilon_g = 0$, the filtered matrices agree in operator norm and the spectral costs agree exactly even when $L \ne \widehat{L}$. All corresponding deterministic stability bounds then vanish.

The spectral argument directly controls $g(\widehat{L})$ and $g(L)$; it does not claim that an arbitrary filter is operator-Lipschitz in its Laplacian, so deriving ε_g from $\|\widehat{L}-L\|_2$ needs a separate estimate. Likewise, the deterministic feature argument supplies no statistical rate for η.

Finally, fixed-point conclusions require the displayed geometry errors uniformly over every state entering the MDP aggregator. A one-state bound gives only the corresponding one-step aggregator and policy bounds there, not a global fixed-point bound without further control. The inherited policy inequalities compare policies at the same Q; comparing readouts at their respective fixed points also requires controlling the change in Q. This completes the proof. □

G.10.3 Proof of Corollary 6

Fix (s,a) and index its successive visits. The observed next-state one-hot vectors, centered by $P(\cdot \mid s,a)$, form a bounded martingale-difference sequence. Since $N_t(s,a) \to \infty$, the martingale strong law gives $\widehat{P}_t(\cdot \mid s,a) \to P(\cdot \mid s,a)$ almost surely. Finiteness of the tables makes the convergence uniform over all rows; continuity of $P \mapsto W_s(P)$ and of the Laplacian construction therefore yields $\widehat{L}_{s,t} \to L_s$ in operator norm.

For positive-semidefinite Laplacians, Duhamel's identity gives

$$e^{-2h\widehat{L}_{s,t}} - e^{-2hL_s} = -2h \int_0^1 e^{-2h(1-u)\widehat{L}_{s,t}} \times (\widehat{L}_{s,t} - L_s) e^{-2huL_s}\, du.$$

Both exponential factors are operator-norm contractions, hence

$$\left\| e^{-2h\widehat{L}_{s,t}} - e^{-2hL_s} \right\|_2 \le 2h \left\| \widehat{L}_{s,t} - L_s \right\|_2 .$$

Since $\|e_a - e_b\|_2^2 = 2$, this proves (131). Moreover, $0 \le (\widehat{C}_{s,t}^{,h})_{ab} \le 2$, because the heat filter is positive semidefinite with operator norm at most one. Thus the predictable estimator satisfies all hypotheses of Theorem 9. □

G.10.4 Proof of Theorem 19

We first obtain simultaneous visitation coverage. We separate visitation coverage from transition-row estimation. For $x \in \mathcal{S} \times \mathcal{A}$, the corrected Bernstein inequality for stationary Markov chains (Paulin, 2015), applied to $f_x(z) = \mathbf{1}\{z = x\}$, gives

$$\Pr\left\{ N_n(x) < \frac{n\nu_\beta(x)}{2} \right\} \le 2\exp\left\{ -\frac{n\nu_\beta(x)\gamma_{\mathrm{ps}}}{104} \right\}. \tag{423}$$

A union bound and (133) therefore imply

$$\Pr\left\{ \min_x N_n(x) < \frac{n\underline{\nu}}{2} \right\} \le \frac{\delta}{2}. \tag{424}$$

On the complementary event, every first-m successor in (132) occurs by time n.

Next, we concentrate the stopped transition-row samples. For rigor, define the first-m variables on the conceptual infinite trajectory and concentrate them without conditioning on the coverage event. Fix $x = (s,a)$, set $p_x = P(\cdot \mid s,a)$, and write

$$Y_{x,k} = e_{s_{T_{x,k}+1}} - p_x.$$

The strong Markov property gives $\mathbb{E}[Y_{x,k} \mid \mathcal{F}_{T_{x,k}}] = 0$. More explicitly, with $\mathcal{G}_{x,0}$ trivial and $\mathcal{G}_{x,k} = \mathcal{F}_{T_{x,k}+1}$ for $k \ge 1$, one has $\mathcal{G}_{x,k-1} \subseteq \mathcal{F}_{T_{x,k}}$ and therefore $\mathbb{E}[Y_{x,k} \mid \mathcal{G}_{x,k-1}] = 0$. Thus these vectors form a martingale-difference sequence under the stopped filtration. Introduce the self-adjoint dilation

$$\mathcal{D}(y) = \begin{pmatrix} 0 & y^\top \\ y & 0_{S_0 \times S_0} \end{pmatrix}.$$

It satisfies $\|\mathcal{D}(y)\|_2 = \|y\|_2$. Since $Y_{x,k}$ is the difference between a one-hot vector and a probability vector, $\|Y_{x,k}\|_2 \le \sqrt{2}$. Moreover,

$$\mathcal{D}(Y_{x,k})^2 = \begin{pmatrix} \|Y_{x,k}\|_2^2 & 0 \\ 0 & Y_{x,k}Y_{x,k}^\top \end{pmatrix},$$

while conditionally on the past

$$\mathbb{E}\,\|Y_{x,k}\|_2^2 = 1 - \|p_x\|_2^2 \le 1, \qquad \mathbb{E}[Y_{x,k}Y_{x,k}^\top] = \operatorname{diag}(p_x) - p_x p_x^\top \preceq I.$$

Thus the predictable matrix-variance process of the first m dilations has operator norm at most m. Matrix Freedman's inequality (Tropp, 2011) yields, for every $u > 0$,

$$\Pr\left\{ \|\textstyle\sum_{k=1}^m Y_{x,k}\|_2 > \sqrt{2mu} + \dfrac{2\sqrt{2}}{3}u \right\} \le 2(S_0+1)e^{-u}. \tag{425}$$

Take $u = \ell_\delta$ from (134) and union bound over the M_0 rows. The resulting estimation failure probability is at most $\delta/2$. Intersecting this event with (424) proves (134). This separate-event argument avoids the selection bias that would result from conditioning the row samples on finite-horizon coverage.

Finally, we propagate transition error to geometry and value. Finally, (135) and the learned-geometry fixed-point perturbation bound (124) give (136). □

G.10.5 Proof of Corollary 7

For symmetric positive-semidefinite Laplacians L and $\widetilde{L}$, Duhamel's identity gives

$$e^{-2h\widetilde{L}} - e^{-2hL} = -2h \int_0^1 e^{-2h(1-u)\widetilde{L}}(\widetilde{L} - L)e^{-2huL}\,du.$$

Both semigroups contract in operator norm, so

$$\left\| e^{-2h\widetilde{L}} - e^{-2hL} \right\|_2 \le 2h \left\| \widetilde{L} - L \right\|_2.$$

Because $\|e_a - e_b\|_2^2 = 2$, combining this inequality with (137) proves (138); (124) then proves (139).

For the fixed-mask kernel construction, put $\epsilon = \left\| \widetilde{P} - P \right\|_{2,\infty}$. On every permitted edge, the reverse triangle inequality and Lipschitzness of κ give

$$|\widetilde{W}_{s,ab} - W_{s,ab}| \le L_\kappa \left\| (\widetilde{P}_{s,a} - P_{s,a}) - (\widetilde{P}_{s,b} - P_{s,b}) \right\|_2 \le 2L_\kappa\epsilon.$$

The absolute row sum of the Laplacian difference is at most twice the sum of the incident edge-weight errors. Symmetry and maximum degree D therefore give

$$\left\| \widetilde{L}_s - L_s \right\|_2 \le \left\| \widetilde{L}_s - L_s \right\|_\infty \le 4DL_\kappa\epsilon.$$

Thus $L_{\mathcal{L}} = 4DL_\kappa$ in (137), proving the stated constants. □

G.11 Spectral approximation and moving references

G.11.1 Proof of Theorem 20

For any coupling Γ, $|\langle \Gamma,\, C^L - \widehat{C^L} \rangle| \le \left\| C^L - \widehat{C^L} \right\|_\infty \sum_{ij} \Gamma_{ij} = \left\| C^L - \widehat{C^L} \right\|_\infty \le \varepsilon_d$, since the plan has unit mass. The KL term is cost-independent, so $W^\rho_{\lambda,s}$ is 1-Lipschitz in the cost; hence F_s^{LOT} is also 1-Lipschitz in the cost (a maximum of such terms), giving $|F_s^{\mathrm{LOT}} - \widehat{F}_s^{\mathrm{LOT}}| \le \tau\varepsilon_d$. Both operators are γ-contractions (Prop. 3); a uniform $\tau\varepsilon_d$ perturbation of the backup propagates to a fixed-point gap $\le \gamma\tau\varepsilon_d/(1-\gamma)$. The cost-perturbation bound $\left\| C^L - \widehat{C^L} \right\|_\infty \le \varepsilon_d$ is immediate from (6). □

G.11.2 Proof of Theorem 21

Because Π_d is κ_d-Lipschitz in $\|\cdot\|_\infty$ and $\mathcal{T}^{C_d}$ is a γ-contraction, the projected operator $\widehat{\mathcal{T}}_d = \Pi_d \mathcal{T}^{C_d}$ is a $\gamma\kappa_d$-contraction. Since $\gamma\kappa_d < 1$, it has a unique fixed point. Let $Q_C^\star = \mathcal{T}^C Q_C^\star$. Then

$$\begin{aligned}\left\|\widehat{Q}_d^\star - Q_C^\star\right\|_\infty &= \left\|\widehat{\mathcal{T}}_d \widehat{Q}_d^\star - \mathcal{T}^C Q_C^\star\right\|_\infty \\ &\le \left\|\widehat{\mathcal{T}}_d \widehat{Q}_d^\star - \widehat{\mathcal{T}}_d Q_C^\star\right\|_\infty + \left\|\widehat{\mathcal{T}}_d Q_C^\star - \mathcal{T}^C Q_C^\star\right\|_\infty \\ &\le \gamma\kappa_d \left\|\widehat{Q}_d^\star - Q_C^\star\right\|_\infty + \left\|\Pi_d \mathcal{T}^{C_d} Q_C^\star - \mathcal{T}^C Q_C^\star\right\|_\infty .\end{aligned}$$

For the second term,

$$\begin{aligned}\left\|\Pi_d \mathcal{T}^{C_d} Q_C^\star - \mathcal{T}^C Q_C^\star\right\|_\infty &\le \left\|\Pi_d \mathcal{T}^{C_d} Q_C^\star - \Pi_d \mathcal{T}^C Q_C^\star\right\|_\infty + \left\|\Pi_d \mathcal{T}^C Q_C^\star - \mathcal{T}^C Q_C^\star\right\|_\infty \\ &\le \kappa_d \left\|\mathcal{T}^{C_d} Q_C^\star - \mathcal{T}^C Q_C^\star\right\|_\infty + \varepsilon_V^\star(d) \\ &\le \kappa_d \gamma\tau \varepsilon_C(d) + \varepsilon_V^\star(d),\end{aligned}$$

where the last step is the cost-perturbation bound from Thm. 18. Rearranging gives (144). If the uniform residual (145) holds on a bounded invariant set containing $Q_C^\star$, then $\varepsilon_V^\star(d) \le \varepsilon_V(d)$. Finally, the spectral truncation theorem gives $\varepsilon_C(d) = \varepsilon_d$ for the rank-d cost, and setting $\kappa_d = 1$ yields (146). □

G.11.3 Proof of Theorem 22

We first bound the aggregator and fixed-point perturbations. Using the local-softmax value form, write

$$f_{s,j}(Q) = \tau\lambda \log \sum_i \rho_s(i) \exp\left(\frac{Q_i - \tau C_{s,ij}}{\tau\lambda}\right).$$

Then $F_s^\mu(Q) = \sum_j \mu_s(j) f_{s,j}(Q)$ and $F_s^\nu(Q) = \sum_j \nu_s(j) f_{s,j}(Q)$. For any reference columns j, k, the definition of Δ_C implies

$$\max_i |C_{s,ij} - C_{s,ik}| \le \Delta_C.$$

Writing $q_i = \rho_s(i) e^{Q_i/(\tau\lambda)}$, we have

$$e^{-\Delta_C/\lambda} \sum_i q_i e^{-C_{s,ik}/\lambda} \le \sum_i q_i e^{-C_{s,ij}/\lambda} \le e^{\Delta_C/\lambda} \sum_i q_i e^{-C_{s,ik}/\lambda}.$$

Taking $\tau\lambda\log(\cdot)$ gives $|f_{s,j}(Q) - f_{s,k}(Q)| \le \tau\Delta_C$, hence $\operatorname{span}_j f_{s,j} \le \tau\Delta_C$. For $d_j = \mu_s(j) - \nu_s(j)$, $\sum_j d_j = 0$, so subtracting any midpoint of the range of $f_{s,j}$ yields

$$|F_s^\mu(Q) - F_s^\nu(Q)| = \left|\sum_j d_j f_{s,j}(Q)\right| \le \frac{\operatorname{span}_j f_{s,j}}{2} \|\mu_s - \nu_s\|_1 \le \frac{\tau\Delta_C}{2} \|\mu_s - \nu_s\|_1 ,$$

which proves (149). Therefore

$$\|\mathcal{T}^\mu Q - \mathcal{T}^\nu Q\|_\infty \le \frac{\gamma\tau\Delta_C}{2}\delta_\mu$$

uniformly in Q, and the fixed-point perturbation bound for two γ-contractions gives (150).

Next, we bound the policy change at a fixed value vector. At fixed Q, the local policies $p_s^{(j)}(Q)$ do not depend on the mixing reference, so

$$\pi_s^\mu(Q) - \pi_s^\nu(Q) = \sum_j (\mu_s(j) - \nu_s(j)) p_s^{(j)}(Q),$$

and (151) follows because each $p_s^{(j)}$ has ℓ_1 norm one. Applying (150) to consecutive references gives

$$\left\|Q_{k+1}^\star - Q_k^\star\right\|_\infty \le B_\mu \delta_k.$$

We continue to unroll the moving-target tracking recursion. For the tracking recursion,

$$\begin{aligned} e_{k+1} &= \left\| \mathcal{T}^{\mu^k} Q_k - Q^\star_{k+1} \right\|_\infty \\ &\le \left\| \mathcal{T}^{\mu^k} Q_k - \mathcal{T}^{\mu^k} Q^\star_k \right\|_\infty + \left\| Q^\star_k - Q^\star_{k+1} \right\|_\infty \\ &\le \gamma e_k + B_\mu \delta_k, \end{aligned}$$

which is (153). Unrolling gives (154), and the limsup bound (155) follows from the stability of the scalar recursion. If $\delta_k \to 0$, then the geometrically weighted convolution in (154) tends to zero, so $e_k \to 0$.

Finally, we specialize to heat-diffused references. For heat-diffused references, H_{t_μ} is a Markov kernel and therefore nonexpansive in ℓ_1. Thus

$$\left\| \mu_s^{k+1} - \mu_s^k \right\|_1 = (1-\xi) \left\| H_{t_\mu}(\pi_s^{k+1} - \pi_s^k) \right\|_1 \le (1-\xi) \left\| \pi_s^{k+1} - \pi_s^k \right\|_1,$$

which proves (156). □

G.11.4 Proof of Corollary 8

Insert the intermediate operator $\mathcal{T}^{C,\widehat{\mu}_t}$. The learned cost bound gives

$$\sup_Q \left\| \mathcal{T}^{\widehat{C}_t,\widehat{\mu}_t} Q - \mathcal{T}^{C,\widehat{\mu}_t} Q \right\|_\infty \le \gamma\tau\varepsilon_t,$$

and Theorem 22 gives

$$\sup_Q \left\| \mathcal{T}^{C,\widehat{\mu}_t} Q - \mathcal{T}^{C,\mu} Q \right\|_\infty \le \frac{\gamma\tau\Delta_C}{2} \varepsilon_{\mu,t}.$$

Their sum proves (157), so the vanishing-drift argument of Theorem 9 applies unchanged. Actor convergence combines (126), (151), and continuity in Q. □

G.12 Continuous actions, finite grids, and the Gaussian example

G.12.1 Proof of Theorem 23

We first disintegrate the coupling and apply the Gibbs variational identity. Because any feasible plan in (158) with finite KL is absolutely continuous with respect to $\rho_s \otimes \mu_s$ and has second marginal μ_s, it admits the disintegration $\Gamma(da, db) = \nu_b(da)\mu_s(db)$ with $\nu_b \ll \rho_s$ for μ_s-a.e. b. The chain rule for relative entropy gives

$$\mathrm{KL}(\Gamma \| \rho_s \otimes \mu_s) = \int \mathrm{KL}(\nu_b \| \rho_s)\, d\mu_s(b).$$

Thus the supremum separates over reference atoms b:

$$\mathcal{F}_s^c(Q) = \int \sup_{\nu \in \mathcal{P}(\mathcal{A})} \left\{ \int [Q(a) - \tau c_s(a,b)] d\nu(a) - \tau\lambda \mathrm{KL}(\nu \| \rho_s) \right\} d\mu_s(b).$$

For fixed b, let $\psi_b(a) = (Q(a) - \tau c_s(a,b))/(\tau\lambda)$ and $Z_b = \int e^{\psi_b(a)} d\rho_s(a)$. The Gibbs variational identity follows by writing, for any $\nu \ll \rho_s$,

$$\int \psi_b \, d\nu - \mathrm{KL}(\nu \| \rho_s) = \log Z_b - \mathrm{KL}(\nu \| p^Q_{s,b}),$$

where $p^Q_{s,b}(da) = e^{\psi_b(a)} d\rho_s(a)/Z_b$. The unique maximizer is $p^Q_{s,b}$, giving (161)–(163). Convexity follows because (163) is an integral of log-partition functions. Monotonicity and translation equivariance are immediate from the same formula; if $\|Q - Q'\|_\infty \le \epsilon$, then $e^{-\epsilon/(\tau\lambda)} Z_{s,b}(Q') \le Z_{s,b}(Q) \le e^{\epsilon/(\tau\lambda)} Z_{s,b}(Q')$, so $|\mathcal{F}_s^c(Q) - \mathcal{F}_s^c(Q')| \le \epsilon$.

Next, we differentiate the log partition. For the derivative, differentiate the log partition under the integral. Since $\mathcal{A}$ is compact and Q, c_s, h are bounded and continuous, dominated convergence applies:

$$D \log Z_{s,b}(Q)[h] = \frac{1}{\tau\lambda} \int h(a)\, dp^Q_{s,b}(a).$$

Multiplication by $\tau\lambda$ and integration over $b \sim \mu_s$ prove (164). A second differentiation gives the covariance identity (165). Positivity is the nonnegativity of variance, and the bound follows from Popoviciu's inequality, $\mathrm{Var}_p(h) \le \mathrm{osc}(h)^2/4$.

Finally, we recover special cases and the Bellman fixed point. If $c_s \equiv 0$, $Z_{s,b}$ and $p^Q_{s,b}$ do not depend on b, which yields (166). If ρ_s and μ_s are atomic, all integrals reduce to sums, giving the finite kernel-relative LOT formulas. Under the statewise kernel and joint-measurability assumptions in the theorem, standard parameterized-integration results make the partition function, its μ_s-integral, and hence $s \mapsto \mathcal{F}^c_s(Q(s,\cdot))$ measurable for bounded jointly measurable Q. Thus $\mathcal{T}^c$ maps the bounded measurable functions into themselves. Finally, the Lipschitz property implies

$$\|\mathcal{T}^c Q - \mathcal{T}^c Q'\|_\infty \le \gamma \|Q - Q'\|_\infty,$$

so Banach's fixed-point theorem gives the unique fixed point of the continuous-action Bellman operator. For (168), write $S_b(a) = Q(a) - \tau c_s(a,b)$; the numerator of $p^Q_{s,b}(B)$ is at most $\rho_s(B) e^{\sup_B S_b/(\tau\lambda)}$, while the denominator is at least $\rho_s(R) e^{\inf_R S_b/(\tau\lambda)}$. □

G.12.2 Proof of Corollary 9

We first prove uniform convergence of the value backup. Using the pushforward definitions, the finite-grid backup can be written as

$$\mathcal{F}^{\mathrm{LOT}}_{s,m}(Q_m) = \tau\lambda \int \log \int \exp\left(\frac{Q(r_m(a)) - \tau c_s(r_m(a), r_m(b))}{\tau\lambda}\right) \rho_s(da)\, \mu_s(db).$$

For every (a,b), the exponent numerator differs from $Q(a) - \tau c_s(a,b)$ by at most $\Delta_m = \omega_Q(\eta_m) + \tau\omega_c(\eta_m)$. Hence each inner partition function is within a multiplicative factor $[e^{-\Delta_m/(\tau\lambda)}, e^{\Delta_m/(\tau\lambda)}]$ of the continuous partition function. Taking $\tau\lambda$ times the logarithm and integrating over b proves (169). Since compactness gives uniform continuity, $\Delta_m \to 0$ whenever $\eta_m \to 0$.

Finally, we prove weak convergence of the policy measures. For weak policy convergence, test against any bounded continuous h and set $\omega_h(\eta) = \sup_{d(a,a')\le\eta} |h(a) - h(a')|$. When the finite-grid policy is viewed as a measure on $\mathcal{A}$, its test integral is

$$\int h\, d\pi^{\mathrm{LOT}}_{s,m} = \int \frac{\int h(r_m(a)) e^{[Q(r_m(a)) - \tau c_s(r_m(a), r_m(b))]/(\tau\lambda)}\, \rho_s(da)}{\int e^{[Q(r_m(a)) - \tau c_s(r_m(a), r_m(b))]/(\tau\lambda)}\, \rho_s(da)}\, \mu_s(db).$$

Thus it contains $h(r_m(a))$, not $h(a)$. Uniform continuity supplies the comparison

$$\sup_a |h(r_m(a)) - h(a)| \le \omega_h(\eta_m) \longrightarrow 0.$$

The exponential weights and their partition functions converge uniformly by the bound with Δ_m above. Applying these two estimates to the tilted numerator and denominator, and then integrating over $b \sim \mu_s$, gives $\int h\, d\pi^{\mathrm{LOT}}_{s,m} \to \int h\, d\pi^{Q,c}_s$. This is exactly weak convergence of the finite-grid LOT policies. □

G.12.3 Proof of Proposition 7

The density of $p^Q_{s,b}$ with respect to Lebesgue measure is proportional to

$$\exp\left(-\tfrac{1}{2}(a - m_0)^\top \Sigma_0^{-1}(a - m_0) + \frac{q^\top a - \frac{1}{2} a^\top H a}{\tau\lambda} - \frac{\frac{1}{2}(a-b)^\top M (a-b)}{\lambda}\right).$$

Collecting quadratic and linear terms in a gives precision $\Lambda = \Sigma_0^{-1} + H/(\tau\lambda) + M/\lambda$ and natural parameter $\eta_b = \Sigma_0^{-1} m_0 + q/(\tau\lambda) + Mb/\lambda$. Completing the square yields (170). □

H Experiment Details

This section gives the data construction and comparison protocol behind Figures 1, 2, 3, 4, 5, and 6. The first two experiments isolate two steps of the proposed mechanism: action geometry must

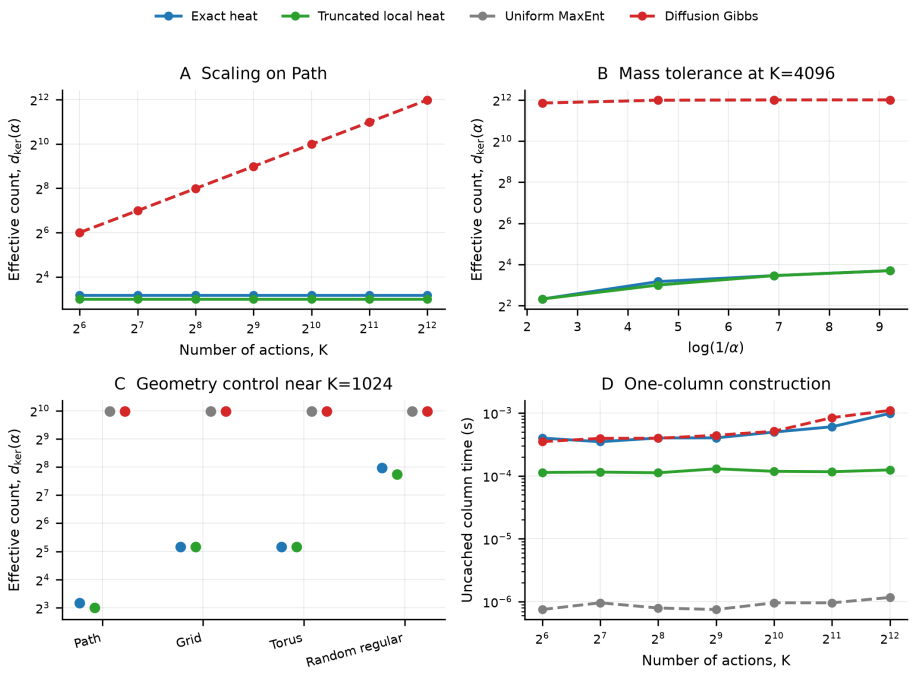


Figure 1: Kernel design map on synthetic action graphs. Panels A and B compare the number of actions needed to retain a prescribed kernel mass as the action count and mass tolerance vary. Panel C changes the action graph at fixed size and diffusion scale; Panel D reports cold construction time for one kernel column. Exact heat and normalized finite-walk heat are compared with uniform MaxEnt and diffusion-distance Gibbs, which define different kernel targets.

first concentrate kernel mass in a neighborhood that can be constructed locally, and restricting computation to that neighborhood must then have a controlled effect on a Bellman backup and its fixed point. Both studies use synthetic, fully specified inputs so that geometry and approximation error can be examined without confounding from a learned model. Ring planning then tests whether locality reduces candidate-action work in a complete control calculation at comparable root-value accuracy. The geometry-learning experiment asks whether the action geometry itself can be estimated from reward-free transitions and reused when the reward goal changes. Finally, we visualize the effective-resistance geometry induced by the optimal LOT coupling and check its policy-curvature identity.

H.1 Kernel concentration

Data and protocol. Figure 1 reports results on undirected action graphs, with no rewards, MDP transitions, or sampled trajectories. We use paths, square grids, square tori, and seeded random 4-regular graphs to vary the growth of graph neighborhoods. Requested action counts are $K = 2^m$ for $m = 6, \dots, 12$; the square graphs use the nearest square size, and analysis uses their realized vertex counts. We use one seeded random-regular realization per requested size. For each graph we evaluate Poisson diffusion scales $\theta \in \{0.5, 2, 8\}$ and omitted-mass tolerances $\alpha \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. These are generated graph and kernel data, rather than a training set.

For a kernel column $w(\cdot \mid j)$, the primary statistic is the smallest number of candidate actions that retain at least $1 - \alpha$ of its mass, maximized over anchors:

$$d_{\mathrm{ker}}(\alpha) = \max_j \min_{I \subseteq \mathcal{A}} \left\{ |I| : \sum_{i \in I} w(i \mid j) \geq 1 - \alpha \right\}. \tag{426}$$

We also record the smallest high-mass graph-ball radius, construction time for one uncached column, and local graph accesses. The comparison includes uniform MaxEnt, exact heat, normalized finite-walk heat, and diffusion-distance Gibbs with cost scale $\lambda = 1$. These kernels define different regularization targets: MaxEnt and diffusion-distance Gibbs are controls for concentration, not estimators of the exact-heat backup. The finite-walk kernel is constructed from adjacency access with radius r chosen so that its omitted Poisson tail $\beta_r = \Pr\{\mathrm{Pois}(\theta) > r\} \le \alpha$; it has a separate, certified comparison to exact heat.

Comparison shown in the figure. Panels A and B of Figure 1 fix $\theta = 2$ on a path. At $\alpha = 10^{-2}$, increasing K from 64 to 4096 leaves the exact-heat effective count at nine actions, while the truncated kernel needs eight; uniform MaxEnt and the chosen Gibbs kernel require approximately $0.99K$ actions. Tightening the retained-mass requirement in Panel B enlarges the heat neighborhood gradually rather than forcing it to cover the full path. These observations concern mass concentration, not value-function accuracy; in particular, the smaller effective count of truncated heat does not make it a more accurate kernel.

Panel C holds $K = 1024$, $\alpha = 10^{-2}$, and $\theta = 2$ fixed while changing only the action graph. Exact heat requires nine actions on the path, 36 on the grid and torus, and 252 on the random 4-regular graph to retain 99% of a column's mass. Thus a short diffusion radius translates into a small candidate set when graph balls grow slowly; rapid neighborhood growth weakens the locality advantage. Panel D tests whether concentration can be used algorithmically: on the path, cold construction of one finite-walk column stays nearly constant as K grows, whereas construction of an exact-heat column becomes more expensive. These timings are medians of three single-threaded repetitions and describe kernel construction only, not a complete planning run. Taken together, the panels identify the geometric regime in which a dense exact-heat prior has a small, locally accessible high-mass core.

H.2 Pruning and heat-truncation error

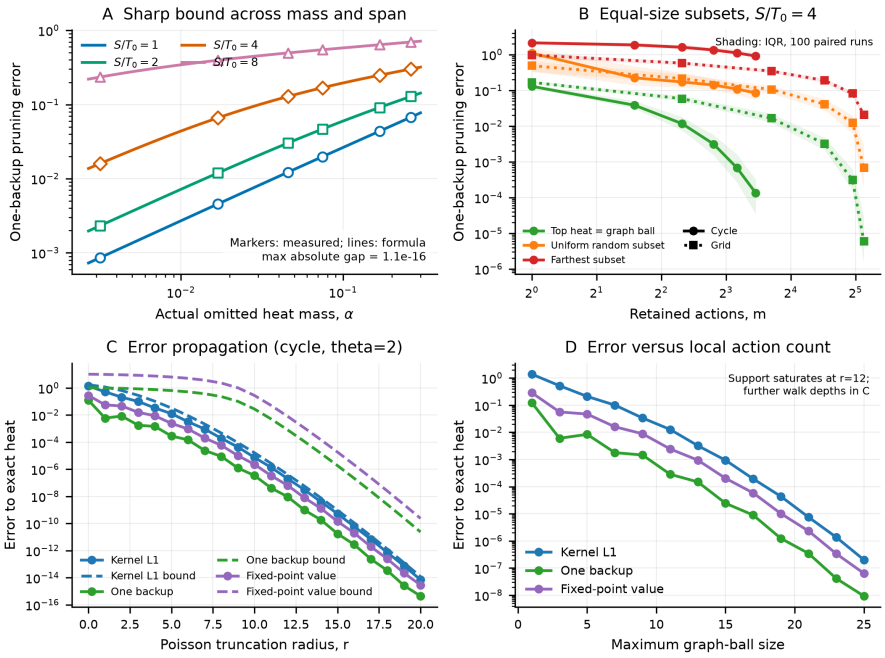
Data and controlled comparisons. Figure 2 uses synthetic action values and a small tabular control problem. Panels A and B prune entries of an *exact* heat column without renormalizing the retained weights, so they measure approximation to the same exact-heat backup. Panel A uses a 41-action cycle, six nominal mass budgets, and action-value spans $S/T_0 \in \{1, 2, 4, 8\}$ at $T_0 = 0.1$. Panel B uses a 25-action cycle and a 6×6 grid with 100 paired smooth-value fields per setting: randomly phase-shifted sinusoids on the cycle and heat-smoothed Gaussian fields on the grid. At each graph-ball size, top-heat-weight, graph-ball, uniform-random, and farthest-action subsets retain exactly the same number of actions. Pairing reuses the same value field across selection methods, so differences reflect the retained set rather than a different realization of Q.

Exact-heat pruning: Panels A–B. Let α denote the *actual* mass omitted by a retained set and $S = \max_i Q_i - \min_i Q_i$. The signed one-backup pruning error is bounded by

$$0 \le F_H(Q) - F_{H,I}(Q) \le T_0 \log\left(1 + e^{S/T_0} \frac{\alpha}{1-\alpha}\right). \tag{427}$$

Panel A places the lowest action values on retained actions and the highest on omitted actions. Its measured errors coincide with the expression above, showing that the worst-case certificate is sharp. Panel B then tests a less adversarial setting. On the displayed cycle and centered grid, graph balls select the same actions as the largest exact heat weights and have much smaller median errors than equally sized random or farthest subsets. This comparison attributes the gain to *which* actions are evaluated: a local set is useful when it retains the heat mass relevant to the backup.

Normalized heat truncation: Panels C–D. Here the operation changes. We replace exact heat by the normalized Poisson head $H_{t,r}$, supported inside the r-hop graph ball; it defines a truncated-heat target, and every error is measured against FullExactHeat. The omitted Poisson probability β_r is distinct from the exact-kernel pruning mass α, and the kernel certificate is $\|H_t(:, j) - H_{t,r}(:, j)\|_1 \le 2\beta_r$. We sweep $r = 0, \ldots, 20$ on 25-action path, cycle, and 5×5 grid graphs for $\theta \in \{0.5, 2, 8\}$. Backup and fixed-point errors are evaluated on a five-state tabular MDP with discount $\gamma = 0.9$. Action a selects the nominal displacement $u(a) = \lfloor 5a/25 \rfloor$; transition noise leaves it unchanged with probability 0.8 or shifts it by one state in either direction with probability 0.1 each. Rewards



A-B prune unnormalized exact-heat weights (same target). C-D compare the truncated-heat fixed point with FullExactHeat; beta_r is distinct from the pruning mass alpha.

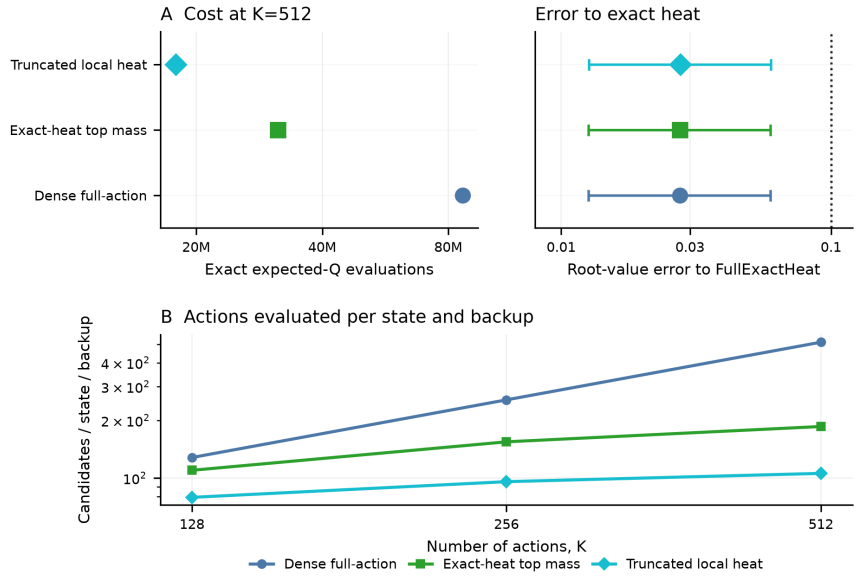
Figure 2: Sharp pruning and propagation of local-heat error. Panel A compares measured worst-case Exact Heat pruning error with its sharp formula. Panel B compares equal-size action subsets over paired smooth value fields. Panels C and D show kernel, backup, and fixed-point errors against Poisson radius and local action count, respectively, with the theoretical bounds in C. Panels A–B prune unnormalized Exact Heat weights; Panels C–D compare normalized truncated heat with FullExactHeat.

penalize distance from the state goal and from a state-dependent anchor action. Exact and truncated heat use the same rewards and transition dynamics.

Panels C and D display the cycle at $\theta = 2$, first against walk radius and then against the maximum number of actions in a local graph ball. As the Poisson tail shrinks, observed kernel, backup, and fixed-point errors fall overall and remain below their respective certificates across all 189 graph–scale–radius cases. At $r = 6$, the local support contains at most 13 of 25 actions and the maximum fixed-point value error across states is below 10^{-3} relative to FullExactHeat. Panel D stops when the graph ball first covers all actions: additional walk depths can still improve the kernel, but no longer reduce the number of candidate actions. Figure 2 therefore links the concentration seen in Figure 1 to a measured local-action budget and controlled Bellman error, while keeping deterministic pruning and normalized heat truncation as separate approximations.

H.3 Ring-control planning

Control problem and reference. We use the stochastic ring MDP with states and actions in $\mathbb{Z}_K$, transition $S' = (S + a + \xi) \bmod K$, and noise probabilities $\Pr(\xi = 0) = 0.8$ and $\Pr(\xi = \pm 1) = 0.1$ each. The goal is $g = 0$; reward penalizes squared cyclic distance from the goal and a 0.05-weighted squared cyclic action displacement, both normalized by $(K/2)^2$. We report the absolute value error at root state $s = K/2$. The action graph is a cycle, separate from the MDP transition kernel P. Its heat prior uses Poisson mean $\theta = 2$ and regularization temperature $T_0 = 0.1$; the discount is $\gamma = 0.95$. For each K, we solve the full exact-heat Bellman fixed point with the population anchor prior to obtain the reference $V^*_{\mathrm{FullExactHeat}}$. Every displayed root-value error is measured against this same K-specific reference.



A: Median and IQR over paired seeds; dotted line marks requested ϵ=0.1. Errors use FullExactHeat as reference; local heat has a truncated-heat target.

Figure 3: Ring-control planning with hyperparameters fixed across $K \in \{128, 256, 512\}$. Panel A compares the median number of exact expected-Q evaluations and root-value error at $K = 512$; whiskers show interquartile ranges over 30 paired seeds, and the dotted line marks the requested error threshold $\epsilon = 0.1$. Panel B shows candidate actions evaluated per state and backup as K increases. All methods use the same weighted stratified anchor draws. Local finite-walk heat defines a truncated-heat target; all plotted errors are measured against the FullExactHeat fixed point.

Paired estimators and cost accounting. The anchor prior is $\mu_s = 0.95\delta_{j(s)} + 0.05\,\mathrm{Unif}(\mathbb{Z}_K)$, where $j(s)$ is the nominal goal-directed action. For every state and seed, we include $j(s)$ with weight 0.95 and draw eight anchors from the uniform component, each with weight $0.05/8$. The resulting anchor table is shared by all methods in a paired run. This stratification ensures that the small uniform component is represented; it is applied equally to local heat, exact-heat top-mass pruning, and dense full-action evaluation. The dense method evaluates every candidate action for the sampled anchors. Top-mass retains actions under the configured, certified exact-heat mass budget. The local method constructs normalized finite-walk heat columns from adjacency access and evaluates their graph-ball candidates. Top-mass and dense use exact heat; local heat has the truncated-heat target, while its error is still reported against FullExactHeat.

The transition kernel P is known, and each queried action value uses the exact expectation $Q(s, a) = r(s, a) + \gamma \mathbb{E}_P[V(S')]$. Thus the cost axis counts exact expected-Q evaluations, not sampled environment transitions. For local heat, the plotted count charges the union of candidates across all nine weighted anchors, rather than treating one heat column as the entire backup cost.

Validation across action counts. After fixing the cycle graph, heat scale, anchor rule, and requested $\epsilon = 0.1$ using development runs, we evaluate $K \in \{128, 256, 512\}$ with 30 new paired seeds at each size. In Figure 3, local heat has median root-value errors 0.024, 0.024, and 0.028, respectively. The median paired ratio of dense to local exact expected-Q evaluations grows from 1.61 to 2.67 to 4.84; at $K = 512$ its bootstrap 95% interval is $[4.83, 4.85]$. At that size, local, top-mass, and dense use 17.86, 31.37, and 86.51 million median evaluations, respectively, while their median errors are 0.02762, 0.02755, and 0.02755; the local error has bootstrap 95% interval $[0.01862, 0.05374]$. The paired median absolute root-value difference between local and dense is only 7.7×10^{-5}. The local heat radius is six, giving 13 candidates per heat column; Panel B shows that candidate evaluations per state and backup grow more slowly for local heat than for dense evaluation as K increases. The

deterministic exact-heat versus truncated-heat fixed-point gap at $K = 512$ is about 3×10^{-6}, much smaller than the observed root-value error.

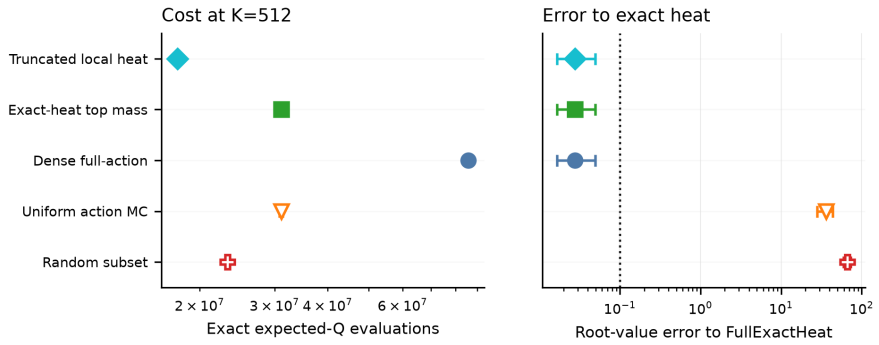


Figure 4: Additional action-selection comparisons in a separate $K = 512$ paired-seed run. Markers and whiskers show medians and interquartile ranges for exact expected-Q evaluations and root-value error to FullExactHeat. Local finite-walk heat, exact-heat top-mass pruning, and dense full-action evaluation share stratified anchor draws with uniform action Monte Carlo and an equal-size random subset. The dotted line marks $\epsilon = 0.1$; the error axis is logarithmic to display all five methods. Local heat defines a truncated-heat target, whereas the other four methods approximate the exact-heat target.

Random-action controls and interpretation. Figure 4 uses a separate set of 30 paired seeds at $K = 512$ and adds two exact-heat-target randomized action-selection baselines. With the same stratified anchors, uniform-action Monte Carlo and an equal-size random subset have median root errors 36.28 and 66.75, respectively, whereas local heat has error 0.02760. Their action samples can miss the narrow high-mass neighborhood of an anchor; the comparison shows why selecting candidates through action geometry matters at these allocated budgets. The top-mass comparator uses a conservative certified mass budget, so its evaluation count is specific to that rule. Across the validation sizes, local heat meets the requested root-error threshold in $29/30$, $30/30$, and $29/30$ runs. These are empirical frequencies, rather than a verified $(\epsilon, \delta) = (0.1, 0.05)$ coverage guarantee. On this cycle-action problem, local heat matches the observed dense root-value accuracy while using fewer exact expected-Q evaluations.

H.4 Reward-free geometry learning

Reward-free data and geometry. We use the $K = 16$ stochastic ring-control system with 16 states and actions. For each action a, we draw $n \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$ transitions (S, S'), with S sampled from the fixed, full-support probe $\nu = 0.9\delta_0 + 0.1\,\mathrm{Unif}(\mathbb{Z}_{16})$ and $S' \sim P(\cdot \mid S, a)$. The 30 paired seeds use nested prefixes of the same reward-free transition batches across sample sizes. From these samples, we estimate each signed transition signature $z_a = P_a^\top \nu - \nu$ and a shared state Laplacian L_ν; their empirical counterparts define the learned action cost

$$\widehat{C}_t(a, b) = \kappa_C \left\| e^{-t\widehat{L}_\nu} (\widehat{z}_a - \widehat{z}_b) \right\|_2^2, \qquad t = 1, \quad \kappa_C = 1. \tag{428}$$

The oracle cost C_t uses the corresponding population quantities. The probe is chosen independently of the subsequent reward goals. A uniform-probe ablation checks this choice: on the translation-invariant ring, a uniform probe is stationary under every action, making all population signatures zero and hence providing no informative action geometry.

Reward transfer and comparison target. We reuse each learned cost matrix for five goal locations, $g \in \{0, 3, 6, 10, 13\}$, without collecting new transitions for a new goal. Every downstream Bellman solve uses the *true* MDP transition kernel; only the action cost changes from C_t to $\widehat{C}_t$. Thus the primary comparison measures geometry-learning error rather than transition-model error. We report $\|\widehat{C}_t - C_t\|_{\max}$ and $\|Q^*_{\widehat{C}_t} - Q^*_{C_t}\|_\infty$, where the latter compares the regularized fixed points

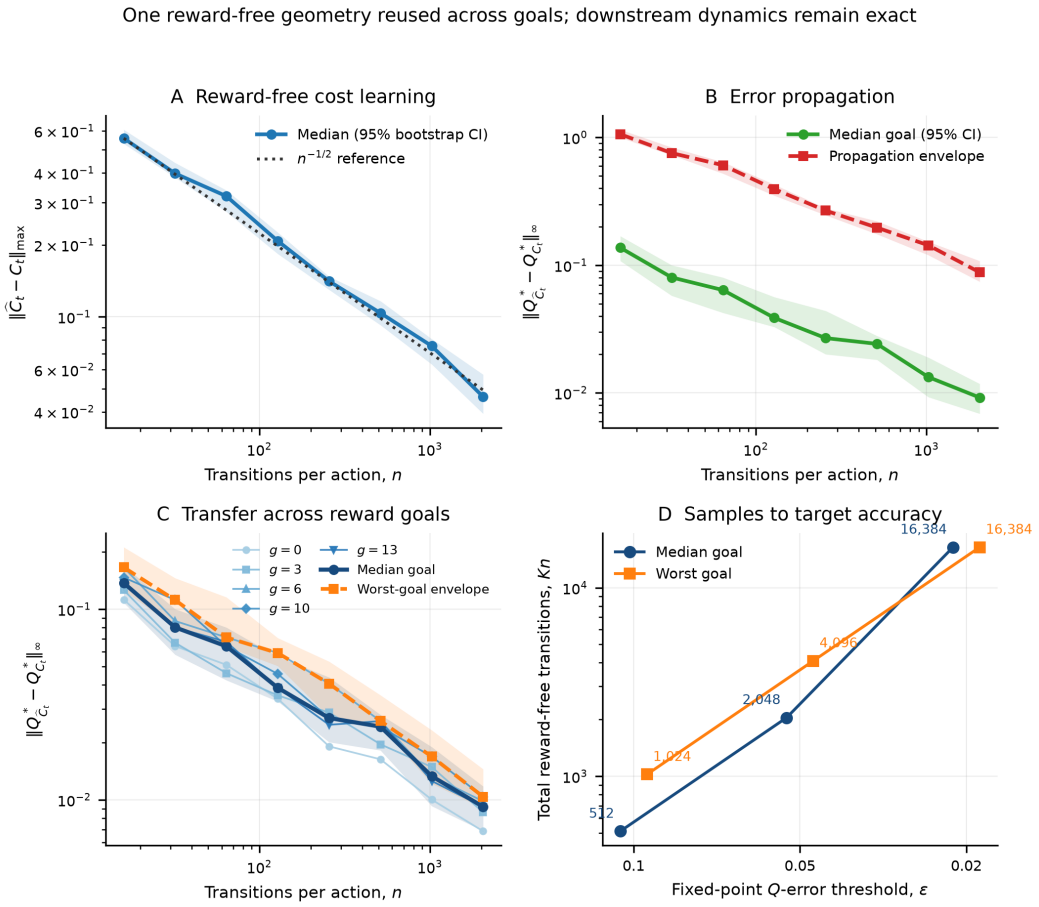


Figure 5: Learning transition-induced action geometry without rewards on a 16-state ring. Panel A shows the maximum cost-matrix error against the population diffusion cost, with an $n^{-1/2}$ reference slope. Panel B compares median-goal fixed-point Q error with its deterministic propagation bound. Panel C shows errors for five reward goals using the same learned geometry, and Panel D gives the total reward-free transition samples Kn needed to reach each error threshold. Shaded bands are pointwise paired-bootstrap 95% confidence intervals over 30 seeds.

for the same reward task and population-cost target. The deterministic propagation envelope is $\gamma\tau\|\widehat{C}_t - C_t\|_{\max}/(1-\gamma)$, with $\gamma = 0.95$ and $\tau = 0.1$. No-diffusion, raw-coordinate, permuted-geometry, and uniform MaxEnt controls define different action-cost or regularization targets; they are diagnostic references, not estimators of C_t.

Results. Figure 5 shows that the reward-free estimate becomes more accurate with additional transition data: the median maximum cost error decreases from 0.56 at $n = 16$ to 0.046 at $n = 2048$, tracking the displayed $n^{-1/2}$ reference trend. This improvement carries through the Bellman fixed point. At $n = 2048$, the median across the five goal-specific median Q errors is 0.0092, and the largest of those five medians is 0.0104; all observed primary runs satisfy the propagation envelope. Panel D makes the shared data budget explicit: based on goal-specific seed medians, the median-goal error first falls below 0.05 at $Kn = 2048$ reward-free transitions, while the largest goal-specific median first falls below 0.05 at $Kn = 4096$. These are one-time geometry samples reused across all five goals. The experiment therefore supports the paper's reward-transfer claim in a controlled setting: an action cost learned without reward labels can approach its population-cost planning target across several rewards, without relearning the dynamics or the geometry for each task.

H.5 EFFECTIVE-RESISTANCE CURVATURE

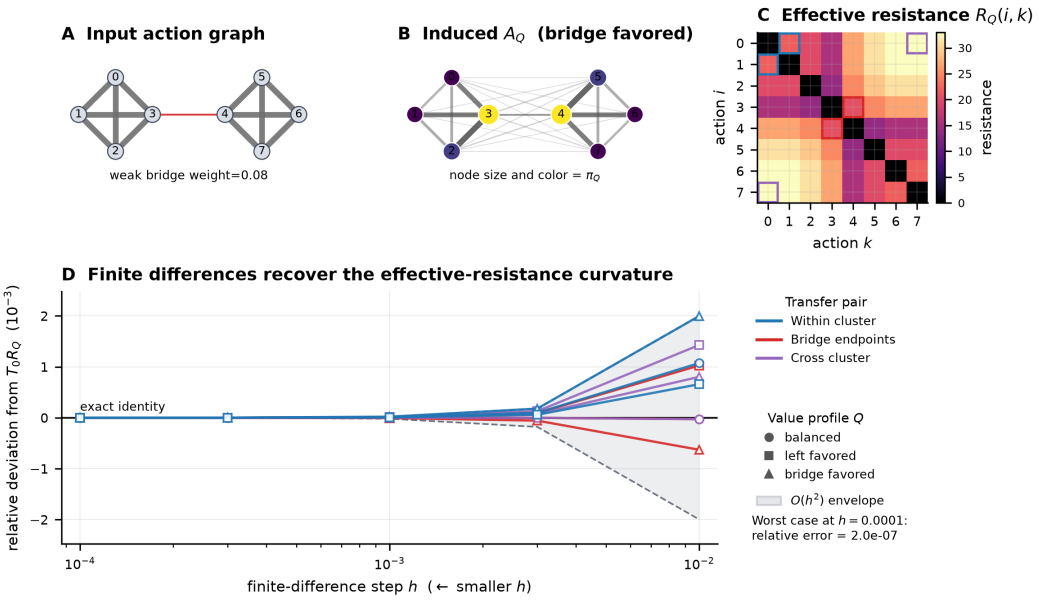


Figure 6: Effective-resistance interpretation of LOT policy curvature. Panel A shows the input action graph used to define the transport cost. For the bridge-favored action values, Panel B shows the coupling-induced co-occurrence graph, with node size and color proportional to policy mass, and Panel C shows its effective-resistance matrix. Colored boxes mark the three action pairs tested in Panel D. Panel D compares central finite-difference curvature with $T_0 R_Q(i,k)$ across three action-value profiles and five step sizes.

Graph, costs, and action values. Figure 6 uses a deterministic, eight-action construction rather than sampled trajectories. The input graph consists of two four-action cliques, with edge weight 1 inside each clique and a single weak edge of weight 0.08 between actions 3 and 4 (Panel A). We set the transport cost to 0.2 times the effective resistance of this *input* graph, and use uniform candidate and anchor distributions. The regularization parameters are $\tau = 0.1$ and $\lambda = 1$, so $T_0 = \tau\lambda = 0.1$. We evaluate a constant Q profile, a profile that favors the left cluster, and a profile that favors the bridge endpoints. These controlled changes to Q let us examine how the induced policy geometry responds while holding the input graph and cost matrix fixed.

Geometry induced by the coupling. For each Q, we calculate its optimal semi-relaxed LOT coupling Γ_Q and policy $\pi_Q = \Gamma_Q \mathbf{1}$. With anchor distribution μ, the co-occurrence weights are $A_Q = \Gamma_Q \operatorname{diag}(\mu)^{-1} \Gamma_Q^\top$; the corresponding Laplacian $L_Q = \operatorname{diag}(A_Q \mathbf{1}) - A_Q$ gives $R_Q(i,k) = (e_i - e_k)^\top L_Q^\dagger (e_i - e_k)$. Panel B depicts A_Q for the bridge-favored profile and Panel C displays its resistance matrix. The co-occurrence graph connects actions according to shared anchor mass and is therefore distinct from the sparse graph in Panel A. This distinction is visible in the pairwise values: under constant Q, the within-cluster pair $(0,1)$ has resistance 16.04, compared with 26.85 for bridge endpoints $(3,4)$ and 27.33 for cross-cluster pair $(0,7)$. Favoring the bridge endpoints changes these values to 21.86, 20.54, and 33.00, respectively. In particular, the bridge pair becomes less resistant while the distant cross-cluster pair becomes more resistant. The curvature geometry thus depends on the current action values, rather than only on graph topology.

Finite-difference verification. Panel D tests the local identity along policy transfers $d_{ik} = e_i - e_k$ for all three Q profiles and the three highlighted pairs. At each interior policy π_Q, we evaluate the

regularizer Ω with both policy and anchor marginals fixed and compute

$$\begin{aligned}\Delta_h^{ik}\Omega(\pi_Q) &= \frac{1}{h^2}\big[\Omega(\pi_Q + h d_{ik}) \\ &\qquad - 2\Omega(\pi_Q) + \Omega(\pi_Q - h d_{ik})\big], \\ \Delta_h^{ik}\Omega(\pi_Q) &\longrightarrow T_0 R_Q(i,k).\end{aligned} \tag{429}$$

We use $h \in \{10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}\}$. Across the nine profile–pair combinations, the maximum relative deviation falls from 2.0×10^{-3} at the coarsest step to 2.1×10^{-7} at the finest step; all nine are below 10^{-6} at the finest step. The decay follows the expected second-order trend for central differences. Together, Panels B–D give a direct numerical interpretation of the paper's curvature identity: moving policy mass between two actions is locally more costly when those actions have larger effective resistance in the Q-dependent co-occurrence graph.