SAMPLE-EFFICIENT RL FOR POLICY IMPROVEMENT VIA FLOW-REVERSED ACTION ANCHORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalist robot policies offer strong behavioral priors, but achieving deployment-level reliability and throughput remains difficult. Reinforcement learning (RL) can improve these policies, yet its effectiveness is constrained by the cost of online interaction. Existing methods adapt frozen flow-matching policies by treating noise space as a continuous action space. We show that this behaviorally unstructured space makes exploration inefficient, often preventing these methods from extracting meaningful learning signals on challenging tasks. We introduce ANCHORQ, which enables coarse-to-fine policy improvement through a structured, refinable latent action representation. At each decision step, ANCHORQ uses flow reversal to map reference behaviors to state-conditioned latent anchors, samples local particles around those anchors, and applies value-guided sampling to explore promising refinements. Our analysis shows that anchors reduce identification complexity at the cost of support bias; particle refinement trades this bias against action-class complexity; and critic accuracy determines the reliability of KL-regularized, value-guided selection. Across simulated and real-world manipulation tasks, ANCHORQ outperforms state-of-the-art post-training baselines in sample efficiency, delivering a **37%** improvement in real-world success rate and **97%** higher simulated task throughput under the same interaction budget. Additional visualizations are available on our anonymous website [1].

1 INTRODUCTION

Generalist vision-language-action (VLA) models provide a powerful starting point for robotic manipulation. Through extensive pretraining, these policies can follow language instructions and transfer to objects, scenes, and tasks beyond the training set (Black et al., 2025b;a; Kim et al., 2025; TRI LBM Team et al., 2025). Yet deployment demands more than broad competence: robots must achieve high success rates and throughput in their target environments, while adaptation itself must remain inexpensive. Pretrained VLAs remain imperfect under changes in geometry, clutter, object placement, and task execution, while collecting demonstrations that densely cover these variations is costly. Reinforcement-learning (RL) post-training offers a natural route to deployment-specific improvement, but fine-tuning a large VLA is computationally expensive and can degrade useful pretrained behaviors. An appealing alternative is therefore to keep the generalist policy frozen and learn how to select and refine its actions.

Prior methods adapts pretrained VLA policies by learning lightweight action corrections, steering latent inputs, or selecting among candidate actions (Yuan et al., 2025; Wagenmaker et al., 2025; Tang et al., 2026; Nakamoto et al., 2024). Flow-matching policies provide an additional adaptation interface through the noise that initializes action generation. For a fixed state s and numerical sampler, an initial latent z is deterministically mapped to an action chunk, $a = \pi_0(s, z)$. The latent can therefore serve directly as the action for RL: changing how z is selected changes the executed behavior without modifying the VLA. For example, diffusion steering via reinforcement learning (DSRL) learns directly in the latent noise space of a frozen policy (Wagenmaker et al., 2025).

This raises a more fundamental question: *what action representation should be used for RL to adapt pretrained VLA policies?* At one extreme, a handful of behaviorally meaningful actions, such as

[1] https://anonymous.4open.science/r/AnchorQ/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

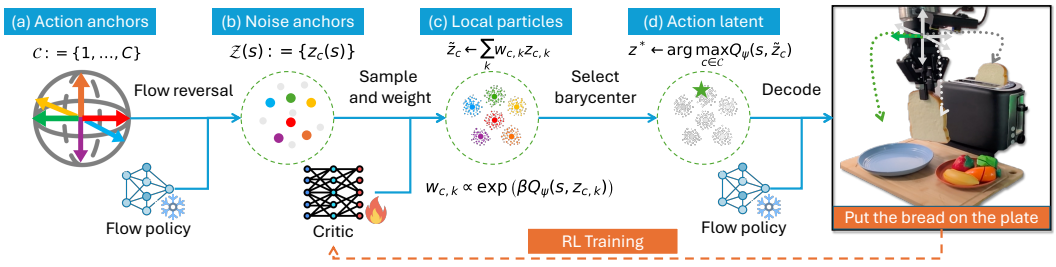

Figure 1: **Overview of ANCHORQ.** (a) A small set of reference action anchors defines coarse behaviors. (b) Flow reversal maps these references into state-conditioned noise anchors. (c) ANCHORQ samples local particles around every anchor and reweights them using the learned critic to form refined barycenters. (d) It selects the highest-value barycenter, decodes it through the frozen flow policy, and executes the resulting action. The critic is trained online using Bellman updates.

moving left or right, can make online learning much easier, but only by restricting the behaviors the policy can express. At the other extreme, existing approaches (Wagenmaker et al., 2025; Yuan et al., 2025; Xiao et al., 2025) search expressive but behaviorally unstructured continuous action or latent spaces. The learner must discover which action corrections or latent vectors produce useful behaviors in each state, without an explicit set of behavioral alternatives to guide this search. This makes efficient exploration extremely difficult: as we demonstrate in our experiments, DSRL and action-space residual RL baselines fail to learn effective policies on challenging tasks within the interaction budget. What is missing is an ***online action representation that is both structured and refinable***: compact when coarse behaviors suffice, but able to expose richer behavior when needed.

To address this gap, we construct a structured and refinable action representation around state-conditioned noise-space anchors, which we generate at every decision by applying flow reversal (Tang et al., 2026) to a small library of semantically meaningful reference actions. Building on this representation, we introduce ANCHORQ, a coarse-to-fine RL method for adapting frozen flow-matching policies directly in their latent noise space. ANCHORQ first samples local particles around each anchor and applies exponential value tilting to bias sampling toward high-value particles. It then selects an anchor and samples a refined noise latent from its tilted particle distribution, which the frozen flow policy decodes into an action. We train the critic Q using temporal-difference (TD) learning from online experience to guide both particle tilting and anchor selection.

Contributions. The main contributions are summarized as follows:

- **A structured and refinable action representation for online RL.** We propose a new online action representation and a coarse-to-fine RL method that turn continuous-space exploration into selection and refinement of behaviorally meaningful actions.
- **Guarantees for coarse-to-fine latent search.** We show that the flow-reversed noise anchors reduces identification complexity (Theorem 1), while particle refinement trades support bias against the complexity of the local action class (Theorem 2). We further show that value tilting reliably amplifies high-value refinements under critic error (Theorem 3), while epistemic uncertainty adaptation calibrates exploitation to critic confidence (Theorem 9, Appendix I).
- **Consistent empirical gains in simulation and on real robots.** Across three benchmarks in LIBERO (Liu et al., 2023a) and RoboTwin (Chen et al., 2025), ANCHORQ outperforms the state-of-the-art steering baselines, which include continuous noise space RL and residual-based RL, with 97% higher throughput than the closest baseline. We also deploy ANCHORQ to solve three challenging real robot manipulation tasks, where it outperforms both the frozen base policy and the RL baseline under the same interaction budget.

Related Work. The most relevant related work is presented here; Appendix A is a broader review.

RL and steering in frozen generative policies. Several methods adapt frozen policies via their latent noise space. Value-guided methods rerank or steer sampled action chunks (Nakamoto et al., 2024; Wang et al., 2025), while DSRL (Wagenmaker et al., 2025) treats the noise itself as a continuous action and learns a policy over the full latent space. FRS (Tang et al., 2026) instead inverts

the frozen flow to recover latents corresponding to reference motions, using this primitive for zero-shot steering and to initialize continuous noise-space RL. UniSteer (Lu et al., 2026) also combines flow inversion with noise-space RL, yet uses online human corrections and learns a continuous latent actor. On the other hand, ANCHORQ uses inversion to map a fixed library of coarse reference behaviors into latent anchors that are refined locally with continuous particles, and ranked by a critic.

Reduced and coarse-to-fine action spaces. Reducing the effective action space can substantially simplify RL, through discretization (Tang & Agrawal, 2020), learned primitives (Shafiullah et al., 2022; Lee et al., 2024), or hierarchical refinement (Seo et al., 2024; Seo & Abbeel, 2024). These methods typically construct abstractions directly in the action coordinates. In contrast, ANCHORQ maps a finite set of coarse reference actions to state-conditioned latent anchors through flow inversion, then refines these anchors locally using sampled continuous perturbations.

2 ANCHORQ: VALUE LEARNING OVER ONLINE FLOW-REVERSED LATENTS

ANCHORQ adapts a frozen flow-matching VLA by learning to select actions in its latent noise space. At each decision, it flow-reverses a small library of coarse reference behaviors into state-dependent latent anchors, samples local refinements around those anchors, and uses a learned critic to select a latent for execution through the frozen VLA.

We first introduce the flow-reversed anchors, then describe local value-tilted refinement, and finally present action selection and critic training. Figure 1 and Algorithm 1 summarize the method.

2.1 ONLINE FLOW-REVERSED ANCHORS

Given a state $s \in \mathcal{S}$ and latent noise $z \in \mathbb{R}^{d_z}$, the frozen flow-matching VLA produces an action chunk $a = \pi_{\text{flow}}(s, z) \in \mathbb{R}^{d_z}$ such that $d_z := h_a \cdot d_a$ where h_a is the chunk horizon and d_a the dimension of each per-timestep action. With the ODE solver for the flow policy fixed, this mapping is deterministic, and we therefore treat z as the RL action.

ANCHORQ constructs a compact, state-dependent set of such latent actions from a fixed library of coarse reference behaviors (e.g. cardinal Cartesian motions) indexed by $\mathcal{C} := \{1, \dots, C\}$. At state $s \in \mathcal{S}$, a reference map φ specifies an *action anchor* $a_c^{\text{ref}}(s) := \varphi(s, c)$, a reference action chunk in the environment action space. Flow reversal (cf. (Rout et al., 2024; Wang et al., 2024; Kim et al., 2022)), recently adopted in robotics (Tang et al., 2026), maps each action anchor backward through the flow-matching policy, conditioned on s, to obtain a corresponding *noise anchor*:

$$z_c(s) := \textsc{RevFlow}\big(s, a_c^{\text{ref}}(s)\big) \in \mathbb{R}^{d_z}.$$

The resulting noise-anchor set is $\mathcal{Z}(s) := \{z_c(s) : c \in \mathcal{C}\}$. ANCHORQ reconstructs $\mathcal{Z}(s)$ at every decision using the current state, and parameterizes the critic as $Q_\psi(s, z)$ so it evaluates noise anchors and their continuous refinements.

2.2 LOCAL PARTICLES AND VALUE-TILTED REFINEMENT

The anchor-only variant of ANCHORQ restricts action selection to $\mathcal{Z}(s)$; the reference behaviors underlying the noise anchors may be too coarse for fine-grained control. We therefore refine each noise anchor using local particles weighted by the learned critic.

For each $c \in \mathcal{C}$, we retain the noise anchor as $z_{c,0}(s) := z_c(s)$ and sample M additional particles:

$$z_{c,k}(s) := z_c(s) + \sigma_p \xi_{c,k}, \qquad \xi_{c,k} \sim \mathcal{N}(0, I), \qquad k = 1, \dots, M,$$

where σ_p controls the perturbation scale. Let $\mathcal{B}_c(s) := (z_{c,0}(s), \dots, z_{c,M}(s))$ denote the resulting local particle cloud. Within each cloud, value tilting favors particles with higher critic values. Indeed, denote the probability simplex $\Delta_{M+1} := \{p \in \mathbb{R}_+^{M+1} : \sum_{k=0}^{M} p_k = 1\}$, and let $u \in \Delta_{M+1}$ be the uniform element. Then, for a tilting strength $\beta > 0$, define weights

$$w_{c,\beta}(\cdot \mid s) := \arg\max_{p \in \Delta_{M+1}} \left\{ \sum_{k=0}^{M} p_k Q_\psi(s, z_{c,k}(s)) - \tfrac{1}{\beta} D_{\text{KL}}(p \| u) \right\}.$$

The unique solution is $\exp(\beta Q_\psi(s, z_{c,k}(s))) / \sum_{\ell=0}^{M} \exp(\beta Q_\psi(s, z_{c,\ell}(s)))$, yielding an exponential *value tilting* of the local particles toward higher-value refinements. These weights define a refined

latent via $\tilde{z}_{c,\beta}(s) := \sum_{k=0}^{M} w_{c,\beta}(k \mid s)\, z_{c,k}(s)$. Since the weights are nonnegative and sum to one, $\tilde{z}_{c,\beta}(s)$ lies in the convex hull of $\mathcal{B}_c(s)$, so that the refined latent remains in a neighborhood of $z_c(s)$. ANCHORQ constructs a refined latent for all noise anchors before selecting among them.

σ_p controls the reach of local exploration, while β controls how strongly refinement favors high-value particles. As $\sigma_p \to 0$, the method recovers exact-anchor ANCHORQ. As $\beta \to 0$, the weights become uniform; as $\beta \to \infty$, the refined latent converges to the highest-valued particle when the maximizer is unique. In practice, we adapt both parameters using the procedures in Appendix C.

2.3 ACTION SELECTION AND CRITIC TRAINING

During RL training, ANCHORQ explores with ϵ-greedy selection among the refined latents. With probability ϵ, it samples $c_t \sim \mathrm{Unif}(\mathcal{C})$; otherwise, $c_t = \arg\max_{c \in \mathcal{C}} Q_\psi\big(s_t, \tilde{z}_{c,\beta}(s_t)\big)$. We execute the action chunk $a_t = \pi_{\text{flow}}(s_t, z_t)$ with $z_t := \tilde{z}_{c_t,\beta}(s_t)$, observe reward r_t, next state s_{t+1}, and terminal indicator d_t, and store $(s_t, z_t, r_t, s_{t+1}, d_t)$ in the replay buffer $\mathcal{D}$. We cache the flow-reversed anchors with each replay observation and reuse them during critic updates, while drawing fresh local particles. During evaluation, we use the same refinement procedure with greedy selection.

For critic updates, we sample batched transitions (s, z, r, s', d) from $\mathcal{D}$ and use temporal difference (TD) learning. Each transition spans one latent-action decision, with reward r from the executed action chunk and discount γ per decision. We minimize

$$\mathcal{L}(\psi) := \mathbb{E}_{(s,z,r,s',d)\sim\mathcal{D}}\left[\big(Q_\psi(s,z) - r - \gamma(1-d)V^\star(s')\big)^2\right],$$

where $V^\star(s') := Q_{\bar{\psi}}(s', \tilde{z}_{c^\star(s'),\beta}(s'))$ is the target critic's value of the refined latent selected by the online critic, with $c^\star(s') := \arg\max_{c\in\mathcal{C}} Q_\psi(s', \tilde{z}_{c,\beta}(s'))$. For nonterminal transitions, all noise anchors are retrieved from cache and refined at s' before selection; when $d = 1$, the bootstrap term vanishes. The VLA remains frozen throughout training, and latent actions are obtained directly through critic-guided refinement and selection, without requiring a separately trained latent actor. Algorithm 1 gives the detailed procedure, with implementation details in Appendix C.

Algorithm 1 ANCHORQ with local value-tilted refinement

Input: frozen flow policy π_{flow} and reverse map REVFLOW; reference map φ; critic Q_ψ and target critic $Q_{\bar{\psi}}$; replay $\mathcal{D}$; particles M, scale σ_p, tilt β, discount γ, exploration ϵ

for each latent-action decision t **do**
  Construct anchors: $z_c(s_t) \leftarrow \text{REVFLOW}(s_t, \varphi(s_t, c))$ for all $c \in \mathcal{C}$
  Sample particles: $z_{c,k} \leftarrow z_c(s_t) + \sigma_p \xi_{c,k}$, $\forall c \in \mathcal{C}$, $k \in [M]$, with $\xi_{c,0} = 0$, $\xi_{c,k} \sim \mathcal{N}(0, I)$
  Tilt locally: $w_{c,k} \propto \exp\left(\beta Q_\psi(s_t, z_{c,k})\right)$ for each $c \in \mathcal{C}$
  Refine: $\tilde{z}_c \leftarrow \sum_{k=0}^{M} w_{c,k} z_{c,k}$ for each $c \in \mathcal{C}$
  Select refined action: w.p. ϵ, $c_t \sim \mathrm{Unif}(\mathcal{C})$; otherwise $c_t \leftarrow \arg\max_{c\in\mathcal{C}} Q_\psi(s_t, \tilde{z}_c)$
  Execute $\pi_{\text{flow}}(s_t, \tilde{z}_{c_t})$, observe (r_t, s_{t+1}, d_t), and store $(s_t, \tilde{z}_{c_t}, r_t, s_{t+1}, d_t)$ in $\mathcal{D}$
  for each critic update **do**
    Sample $(s, z, r, s', d) \sim \mathcal{D}$
    Retrieve cached anchors $\{z_c(s')\}_{c\in\mathcal{C}}$
    Sample $\{z^\star_{c,k}\}_{k=0}^{M}$ around $z_c(s')$ for all $c \in \mathcal{C}$; set $w^\star_{c,k} \propto \exp(\beta Q_\psi(s', z^\star_{c,k}))$
    Set $\tilde{z}^\star_c \leftarrow \sum_k w^\star_{c,k} z^\star_{c,k}$ for all $c \in \mathcal{C}$ and set $c^\star \leftarrow \arg\max_c Q_\psi(s', \tilde{z}^\star_c)$
    Set $V^\star(s') \leftarrow Q_{\bar{\psi}}(s', \tilde{z}^\star_{c^\star})$ and set $y \leftarrow r + \gamma(1-d)V^\star(s')$
    Update Q_ψ from $\big(Q_\psi(s,z) - y\big)^2$ and update $Q_{\bar{\psi}}$
  end for
end for

2.4 ONLINE NOISE ANCHORS IMPROVE SAMPLE EFFICIENCY

ANCHORQ uses state-dependent noise anchors and local particle refinements to construct and select actions online. Our results isolate three consequences of this algorithmic structure:

- **Sample efficiency.** Restricting online search to the actions represented by the noise anchors can reduce the number of latent alternatives that must be identified (Theorem 1).

- **A bias–complexity frontier.** Restricting the action class introduces support bias; local particles reduce this bias by enlarging the class, at the cost of increased statistical complexity (Theorem 2).
- **Risk-sensitive local refinement.** Exponential value tilting provides an entropic, upper-tail-seeking reweighting of particles around each anchor. With a sufficiently accurate critic, it amplifies genuinely high-value refinements and, under local smoothness, yields a corresponding value guarantee for the weighted latent (Theorem 3).

Below we provide the key results in informal theorems; full statements are given in the appendix.

Sample Efficiency: Online noise anchors reduce identification complexity. To isolate the statistical effect of restricting the online action class, we consider a one-decision MDP with N behaviorally distinct latent modes. Flow reversal exposes a subset of $C < N$ modes, whereas an unrestricted learner may select any of the N.

Theorem 1 (Online noise anchors reduce complexity)**.** Fix $\varepsilon \in (0, 1/8]$ and $\alpha \in (0, 1/4]$. If the best noise anchor is $\varepsilon/2$-optimal, then restricting online search to the C noise anchors identifies an ε-optimal action with probability at least $1-\alpha$ using $\mathcal{O}(\frac{C}{\varepsilon^2}\log\frac{C}{\alpha})$ episodes. Conversely, any learner that is (ε, α)-correct over all N actions requires, on some instance whose optimal action lies in the noise-anchor set, $\Omega(\frac{N-C}{\varepsilon^2}\log\frac{1}{\alpha})$ expected samples from actions outside the noise-anchor set.

The upper bound shows that when the C noise anchors contain a near-optimal action, searching only among them is sufficient to identify an ε-optimal action. For the lower bound construction, each non-flow-reversed action is suboptimal in the base instance but optimal in a corresponding alternative instance. Hence, a learner searching all N latent modes must still collect enough samples to rule out all $N - C$ outside modes, even when a noise anchor is globally optimal. That is, unrestricted residual RL, even with FRS-based prefill (e.g., Tang et al. (2026)), cannot eliminate this exploration cost. The full result is in Theorem 4, Appendix E.

Bias–Complexity Frontier: Trading support bias for learning complexity. Local particles interpolate between selection among exact noise anchors and broader latent search. For the more general sequential MDP, we view the frozen VLA as a control problem whose action is the latent z: choosing z at state s executes $\pi_{\text{flow}}(s, z)$ and induces the corresponding environment reward and transition. Let $V^\star$ and $Q^\star$ denote the optimal value and action-value functions when any latent $z \in \mathbb{R}^{d_z}$ may be selected.

To isolate the bias–complexity tradeoff, we analyze an idealized local action class. Indeed, following Section 2.2, let $\rho \geq 0$ denote an *idealized local search radius*, where $\rho = 0$ allows only the exact noise anchors, while larger ρ allows any latent in $B_2(z_c(s), \rho) := \{\tilde{z} \in \mathbb{R}^{d_z} : \|\tilde{z} - z_c(s)\|_2 \leq \rho\}$. Define $\mathcal{Z}_\rho(s) := \bigcup_{c\in\mathcal{C}} B_2\left(z_c(s), \rho\right)$, and define the corresponding one-step support gap $\Delta_\rho(s) := V^\star(s) - \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s, z)$ and set $\bar{\Delta}_\rho := \sup_s \Delta_\rho(s)$. Thus $\Delta_\rho(s)$ measures how much value is unavailable as action selection remains local to the noise anchors.

Theorem 2 (Bias–complexity frontier)**.** Increasing the local search radius can only reduce the support bias: $0 \leq \rho_1 \leq \rho_2$ implies $\bar{\Delta}_{\rho_2} \leq \bar{\Delta}_{\rho_1}$. Moreover, suppose a finite particle representation of the ρ-local action class incurs one-step representation error at most ϵ_{rep}, and its learned selector incurs one-step selection error at most $\epsilon_{\text{sel}}(n, \rho, \alpha)$ after n interactions. Let $\pi_{\rho,n}$ denote the resulting learned policy and $V^{\pi_{\rho,n}}$ its discounted value function. With probability at least $1-\alpha$, the bound holds: $\|V^\star - V^{\pi_{\rho,n}}\|_\infty \leq (\bar{\Delta}_\rho + \epsilon_{\text{rep}} + \epsilon_{\text{sel}}(n, \rho, \alpha))/(1-\gamma)$.

This result exposes a Pareto-type frontier: *larger local search implies smaller support bias but a larger action class to learn*. Exact-anchor ANCHORQ gives the strongest restriction, while local particles progressively enlarge the available action class. Formal support and sample-complexity results are given in Theorems 5 and 6 (Appendices F and G), respectively. Theorem 2 is stated in terms of an idealized ρ, whereas the algorithm realizes local search via Gaussian particles. Proposition 9 (Appendix G.7) connects these two views: sufficiently many Gaussian particles around the anchors cover the ρ-local action class to a prescribed resolution with high probability. Thus (σ_p, M) control the stochastic coverage of the idealized local class rather than setting ρ directly.

Risk-Sensitive Local Refinement: Value tilting exploits particle diversity. The weight vector $w_{c,\beta}$ defined in Section 2.2 is exactly the Gibbs change of measure solving the entropic variational problem $\frac{1}{\beta}\log(\frac{1}{M+1}\sum_{k=0}^{M} e^{\beta Q_\psi(s, z_{c,k}(s))}) = \max_{p\in\Delta_{M+1}}\{\sum_{k=0}^{M} p_k Q_\psi(s, z_{c,k}(s)) -$

$\frac{1}{\beta} D_{\mathrm{KL}}(p\|u)\}$, where $u \in \Delta_{M+1}$ is uniform over the local candidates. Thus, $\beta > 0$ induces an upper-tail-seeking reweighting that places increasingly more mass on higher-value particles.

Theorem 3 (Amplification of high-value local refinements)**.** Fix a state $s \in \mathcal{S}$, anchor $c \in \mathcal{C}$, and corresponding particle cloud $\mathcal{B}_c(s)$. Let $\mathcal{H} \subset \{0, \dots, M\}$ be a nonempty proper subset and define $\delta_{\mathcal{H}} := \min_{k\in\mathcal{H}} Q^\star(s, z_{c,k}(s)) - \max_{\ell\in\{0,\dots,M\}\setminus\mathcal{H}} Q^\star(s, z_{c,\ell}(s))$. If the critic satisfies $\max_{0\le k\le M} |Q_\psi(s, z_{c,k}(s)) - Q^\star(s, z_{c,k}(s))| \le \epsilon_Q$, then

$$w_{c,\beta}(\mathcal{H} \mid s)/(1 - w_{c,\beta}(\mathcal{H} \mid s)) \ge \exp(\beta(\delta_{\mathcal{H}} - 2\epsilon_Q))|\mathcal{H}|/(M + 1 - |\mathcal{H}|),$$

where $w_{c,\beta}(\mathcal{H} \mid s) := \sum_{k\in\mathcal{H}} w_{c,\beta}(k \mid s)$. Hence, whenever $\delta_{\mathcal{H}} > 2\epsilon_Q$, value tilting increases the odds assigned to the truly high-value subset relative to uniform weighting, with amplification growing exponentially in both β and the value margin.

Theorem 3 shows that high-value local deviations exert greater influence on each refined candidate when their value advantage is large relative to critic error; the formal statement is Theorem 7, Appendix H. Under local smoothness, this guarantee extends to the refined latent: its loss relative to the best local particle is controlled by critic error, finite-temperature weighting, and weighted particle variance (Proposition 13). As aggressive tilting may amplify critic errors, we estimate local epistemic uncertainty from disagreement among the critic ensemble and set $\beta(s)$ inversely proportional to this disagreement. The weighting therefore remains diffuse where the critic is uncertain and becomes sharper as critic confidence increases (Theorem 9).

3 Experiment Results

We evaluate ANCHORQ in simulation on the LIBERO (Liu et al., 2023a) and RoboTwin (Chen et al., 2025) benchmarks, and in the real world on manipulation tasks using a Franka DROID setup (Khazatsky et al., 2024). We focus the analysis on three questions:

- Do flow-reversed anchors improve online RL over a continuous noise action space?
- How does particle tilting enable finer motions beyond those produced by the noise anchors?
- Does the same recipe work for steering a real robot?

3.1 Tasks & Setup

Benchmarks. We use the two evaluation splits of Tang et al. (2026), both drawn from LIBERO-90 (Liu et al., 2023a): a 15-task split and a 10-task hard split, none of them used to train the base policy. We also evaluate on 6 RoboTwin (Chen et al., 2025) benchmark tasks that requires bimanual manipulation. To evaluate the effectiveness of our approach in the real world, we additionally evaluate on three common household tasks using a Franka arm with Robotiq gripper in a DROID setup (Khazatsky et al., 2024). We show a visual overview of the tasks in Figure 6.

Baselines. Every method learns steering on the same frozen flow-matching VLA. In simulation we use $\pi_{0.5}$ (Black et al., 2025a), and on the real robot we use DROID-trained π_0 (Black et al., 2025b) for their out-of-the box robustness. **DSRL** (Wagenmaker et al., 2025) learns a noise-space policy over the continuous space that the flow policy samples from. On LIBERO 10-task split, we compare against a stronger RL variant of Tang et al. (2026), which pre-fills DSRL buffer with flow-reversed demonstrations. **Residual RL** follows Xiao et al. (2025) but only trains the residual actor on top of the frozen VLA, since we do not consider any offline learning. **VLM-FRS** queries a VLM to propose coarse actions, which are executed via flow reversal steering Tang et al. (2026).

Evaluation. Each RL run trains for a fixed budget of environment steps and is evaluated periodically on held-out configurations of its task. We report the mean over four seeds and over the tasks of a split at $200K$ and $500K$ environment interactions, with one standard error across seeds in our tables and show the full training progress in our figures.

3.2 Simulation Performance and Ablation Studies

We compare performance and sample efficiency on our simulation benchmarks and examine how particle tilting parameters, anchor composition, and the choice of Q function affect learning.

Table 1: **Performance on manipulation tasks in simulation.** Success rate (%) on the two held-out LIBERO-90 splits from Tang et al. (2026) and on RoboTwin (Chen et al., 2025). We report mean ± standard error over four seeds. Results marked with † use DSRL+FRS (Tang et al., 2026). ANCHORQ achieves the highest success rate across all three suites.

Method	**LIBERO-15T**		**LIBERO Hard-10**		**RoboTwin-6T**
$\pi_{0.5}$ zero-shot	8.5 ± 0.4		1.6 ± 0.4		47.7 ± 1.6
VLM+FRS zero-shot	21.9 ± 0.6		2.0 ± 0.3		54.2 ± 0.2
Online RL env. steps	250k	500k	250k	500k	500k
DSRL	37.2 ± 1.6	48.3 ± 3.8	$7.8 \pm 4.6^{\dagger}$	$25.2 \pm 3.4^{\dagger}$	47.2 ± 4.5
Residual RL	22.8 ± 0.3	26.7 ± 0.5	9.6 ± 2.1	14.6 ± 0.8	0.0 ± 0.0
ANCHORQ (Ours)	$\mathbf{86.0 \pm 0.8}$	$\mathbf{95.5 \pm 0.6}$	$\mathbf{52.1 \pm 2.7}$	$\mathbf{62.1 \pm 4.0}$	$\mathbf{70.5 \pm 3.0}$

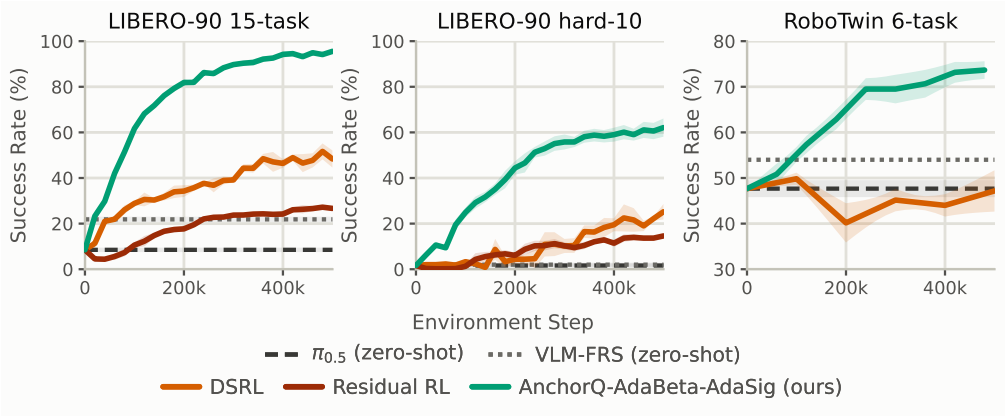

Figure 2: **Success rate on manipulation tasks in simulation**; mean ±1 standard error over four seeds. *(Left)* 15-task held-out LIBERO-90 split. *(Center)* 10-task hard split. *(Right)* 6-task bimanual suite in RoboTwin. ANCHORQ consistently outperforms the baselines across all three task suites.

ANCHORQ Outperforms Steering Baselines in Simulation. As shown in Table 1 and Figure 2, ANCHORQ outperforms other steering baselines by a wide margin. Especially on the LIBERO-90 15-task suite, it outperforms the closest baseline (DSRL) by at least 2× the success rate at *any* matched trajectory budget (over $250-2200$ trajectories). We attribute this huge improvement to ANCHORQ's efficient exploration thanks to its effective anchors and the particles that elicit meaningful exploratory behaviors. Similarly, on the hard-10 suite, ANCHORQ also excels at exploring and learning the right task-solving behavior, even if the base policy performance is close to 0. Finally, on the RoboTwin bimanual suite, we demonstrate that ANCHORQ can effectively learn dexterous behaviors that require the coordination of both arms, unaffected by the complex action space induced by the additional embodiment, whereas other baselines seem to struggle to beat the base policy's performance on these tasks. ANCHORQ achieves an average throughput of 294 completions/hr across all simulation tasks, compared to DSRL's 149, corresponding to 97% higher throughput.

Ablation studies. We consider several ablations. Tilting versus no tilting; tilting with an adaptive β and σ versus fixed values; the composition of the anchor set; and learning Q in the continuous noise space versus over the discrete set of anchor options using RAINBOW (Hessel et al., 2018) (Rainbow-FRS). Figure 3 shows the training progress of each ablation together with the corresponding success rates. Tilting has greatest effect when β is adaptive. An adaptive β is worth $+8.6$ points whereas tilting at a fixed $\beta = 0.1$ recovers only 4.0 points over not tilting at all. Adapting σ on top adds a further 1.3 points. Using discrete anchors achieve similar performance as no-tilt but suffers from the beginning of training and does not have the benefit of tilting. Noticeably, all of our ablations still outperforms DSRL, showing the powerful effect of leveraging flow-reversed anchors. We provide additional results and visualizations on per-task performance in Appendix B.

Method	Success rate	
	250k	500k
AdaBeta-AdaSig (ANCHORQ)	$\mathbf{86.0 \pm 1.0}$	$\mathbf{95.5 \pm 0.6}$
Particle tilting		
AdaBeta, $\sigma = 0.3$	80.8 ± 2.0	94.2 ± 0.7
$\beta = 0.1$, $\sigma = 0.3$	68.4 ± 1.4	85.6 ± 0.9
No tilting, anchors only	69.8 ± 1.4	81.6 ± 2.4
Discrete Q (Rainbow-FRS)		
w. ANCHORQ anchors	64.2 ± 2.1	79.4 ± 1.8
+ half-scale	68.3 ± 2.4	79.5 ± 1.7
+ rotations	63.4 ± 1.7	75.7 ± 2.9

Table 2: Success rate after 250k and 500k environment steps, reported as mean $\pm$ standard error over four seeds for ablations on particle tilting, adaptive β and σ, anchor design, and continuous versus discrete Q-learning on the 15-task LIBERO-90 split.

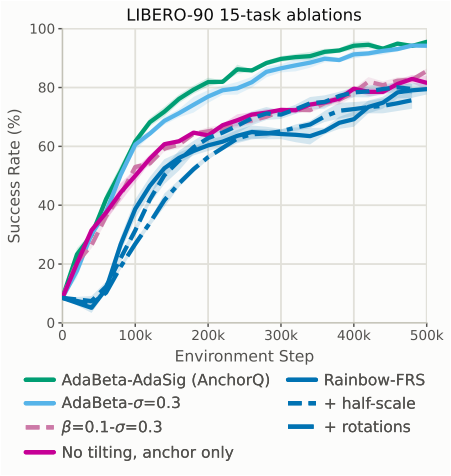

Figure 3: Ablation training progress: mean ± 1 standard error over four seeds.

3.3 ANALYSIS ON ANCHORS AND PARTICLE TILTING.

To illustrate how particle tilting helps with dexterous manipulation, we focus on task 81, which requires precise placement of a book into the front compartment of a caddy and overcoming the base policy's tendency to drop it into the wrong compartment. On this task, ANCHORQ achieves an 84% success rate, the anchor-only ablation gets 22%, and DSRL gets 0%. Figure 4 (a) measures behavior in robot space: the realized end-effector motion already deviates from the commanded anchor axis for both anchor-based arms, and tilting deviates markedly more, giving it the freedom to express fine, off-axis motions that the six discrete anchors cannot. Figure 4 (b) shows why this steering is reachable at all. The noise anchors sit at an intermediate magnitude, between the near-zero noise DSRL collapses to and $\pi_{0.5}$ prior, and ANCHORQ spreads its executed noise across this anchor band, tilting to refine directions that the discrete anchors alone cannot reach. DSRL is a continuous noise-space actor regularized toward the $\pi_{0.5}$ prior; however, it collapses to a near-zero noise vector and reproduces the base policy's failure. ANCHORQ refines within the anchor-supported region and explores beyond the discrete anchors, achieving the highest success rate.

3.4 REAL WORLD STEERING EXPERIMENTS

Setup and training. We evaluate ANCHORQ on three household manipulation tasks: stacking cups, toasting bread, and hanging a towel. We use a Franka arm with a DROID (Khazatsky et al., 2024) setup, with observations from a third-person camera and a wrist camera. All methods steer the same frozen DROID pretrained $\pi_{0.5}$ policy. We compare ANCHORQ against the base policy without steering and DSRL. We give both RL methods the same interaction budget of 100 online episodes per task. A human operator observes each rollout, provides a reward at the end of the episode, and resets the scene to one of a predefined set of initial configurations. These training configurations are separate from the configurations used for evaluation. RL training runs asynchronously with data collection, updating the steering parameters while keeping the base policy frozen.

ANCHORQ Outperforms Steering Baselines on a Real Robot. We evaluate each method on 20 held out configurations per task and report individual and average task success rates, with 95% Wilson score confidence intervals (Wilson, 1927) for the average. We show statistical significance using two sided t-tests comparing ANCHORQ with each baseline, testing the null hypothesis of equal mean success. As shown in Figure 5, ANCHORQ outperforms both $\pi_{0.5}$ and DSRL on all three tasks under the same online interaction budget. On the towel hanging task, it is the only method that improves over the base policy. These results demonstrate that the our approach can significantly improve manipulation performance on a real robot under limited environment interaction.

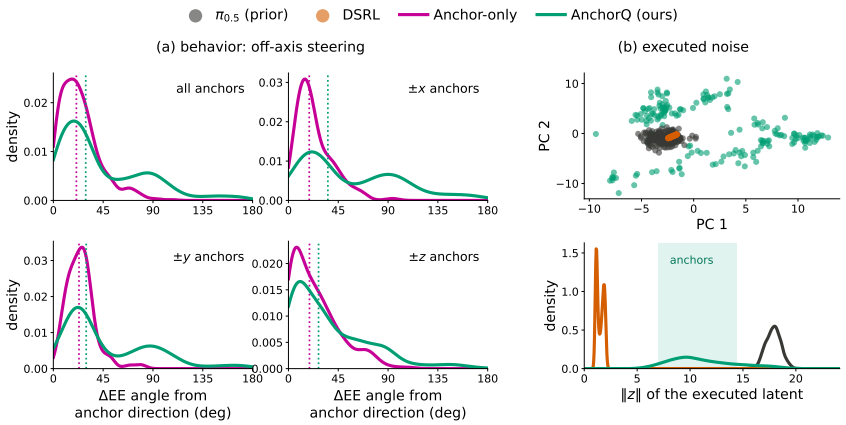

Figure 4: **Anchor & particle-tilting analysis on LIBERO task 81. (a)** Angle between realized end-effector displacement and anchor direction, grouped by anchor: ANCHORQ shows a heavier right tail on every axis, indicating finer steering beyond the discrete anchor. **(b)** Executed noise projected onto the top two pooled principal components (top) and noise magnitude $\|z\|$ (bottom): reverse-flow anchors span the shaded band, ANCHORQ explores across it, DSRL concentrates near zero, and the $\pi_{0.5}$ prior lies above.

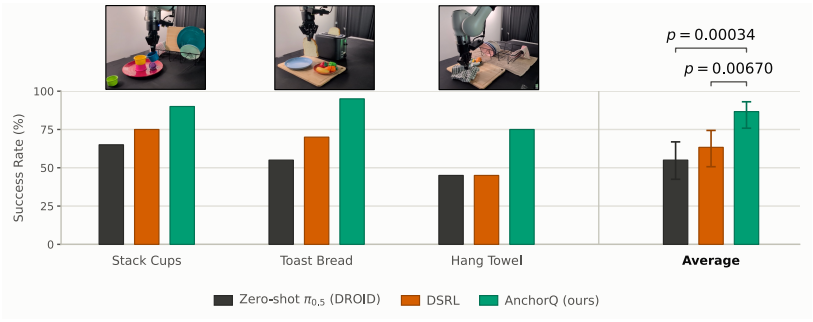

Figure 5: **Real world steering results.** We evaluate on three manipulation tasks using a Franka robot with a DROID setup. Each RL method receives 100 online training episodes per task and is evaluated on 20 held out configurations. Error bars on the aggregate success rate show 95% Wilson score confidence intervals. ANCHORQ achieves higher success rates than the frozen $\pi_{0.5}$ policy and DSRL on all three tasks.

4 CONCLUSION

We introduce ANCHORQ, a coarse-to-fine RL method that adapts frozen flow-matching policies through a structured and refinable action representation. Flow reversal maps reference behaviors into state-conditioned latent anchors, while critic-guided selection and local particle tilting enable fine-grained refinement. Our analysis shows how anchoring reduces identification complexity, how refinement trades support bias against action-class complexity, and when value tilting remains reliable under critic error. Across simulated and real-world manipulation tasks, ANCHORQ improves sample efficiency over post-training baselines in challenging manipulation tasks across simulation and real world benchmarks. These findings highlight action representation as a key design choice for efficient adaptation of frozen policies.

Future work. ANCHORQ requires additional memory from scoring particles at every decision. Distilling the adapted behavior into a policy could reduce the need for our sampling-based inference. Extending the current reference anchors to more complex embodiments and moving from task-specific to multitask post-training are also exciting future directions.

AI USE STATEMENT

In this work, we used generative AI tools for design or provide feedback on research methodology or experiments, implement methods. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, support qualitative and thematic data analysis, interpret results, and generate synthetic data sets. Assist with translation, clean and reformat dataset, are not applicable to this work. Additionally, we used generative AI tools for create or edit software code, suggest a structure for a research paper, and fix typos and grammar in writing. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by multiple authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or dataset releases. Although our method improves pretrained VLA policies, higher task success does not always guarantee safe behavior during exploration or deployment. We use human supervision and intervention to ensure safety during our real-world robot experiments. Future work should also incorporate appropriate human oversight and physical safeguards.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we include detailed descriptions of the implementation of our algorithm in Appendix C. We will release open-sourced code upon publication.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations (ICLR)*, 2018.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *Proceedings of The 9th Conference on Robot Learning (CoRL)*, 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems (RSS)*, 2025b.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science & Systems (RSS)*, 2023.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021a.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double Q-learning: Learning fast without a model. In *International Conference on Learning Representations (ICLR)*, 2021b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11), 2024.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.

Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2018.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2416–2425. IEEE, 2022.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2025.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, United Kingdom, 2020. ISBN 9781108486828. doi: 10.1017/9781108571401.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International conference on machine learning*, pp. 6131–6141. PMLR, 2021.

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. In *International Conference on Machine Learning (ICML)*, 2024.

Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Junjie Lu, Xinyao Qin, Yuhua Jiang, Kaixin Wang, Chuheng Zhang, Bin Liang, Jun Yang, Min Xu, and Li Zhao. UniSteer: Unified noise steering for efficient human-guided VLA adaptation. *arXiv preprint arXiv:2605.10821*, 2026.

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. *International Conference on Learning Representation (ICLR)*, 2026.

Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. In *Conference on Robot Learning (CoRL)*, 2024.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in neural information processing systems (NeurIPS)*, 2018.

Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-learning. In *International Conference on Machine Learning (ICML)*, 2025.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via Q-score matching. In *International Conference on Machine Learning (ICML)*, 2024.

Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

Yifan Ruan, Chenyang Cao, Andreas Burger, Ali Pesaranghader, Kaveh Kamali, Jaehong Kim, Nandita Vijaykumar, Alan Aspuru-Guzik, Igor Gilitschenski, and Nicholas Rhinehart. QPILOTS: Efficient test-time Q-steering for flow policies. *arXiv preprint arXiv:2606.14801*, 2026.

Younggyo Seo and Pieter Abbeel. Coarse-to-fine Q-network with action sequence for data-efficient reinforcement learning. *arXiv preprint arXiv:2411.12155*, 2024.

Younggyo Seo, Jafar Uruç, and Stephen James. Continuous control with coarse-to-fine reinforcement learning. *arXiv preprint arXiv:2407.07787*, 2024.

Nur Muhammad Mahi Shafiullah, Zichen Jeff Cui, Ariuntuya Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *arXiv preprint arXiv:2206.11251*, 2022.

Hikmet Simsir and Ozgur S. Oguz. Lagrangian perturbation diffusion steering: Latent reinforcement learning for generative policies. In *arXiv preprint arXiv:2606.01151*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Entong Su, Tyler Westenbroek, Anusha Nagabandi, and Abhishek Gupta. RFS: Reinforcement learning with residual flow steering for dexterous manipulation. In *International Conference on Learning Representations (ICLR)*, 2026.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999.

Derek Ming Siang Tan, Shailesh Shailesh, Srikrishna Iyer, William Wei Jie Teo, Yuanliang Ju, Qiao Gu, and Guillaume Sartoretti. RL^2-VLA: Adaptive RL latent compositional steering with test-time scaling for vision-language-action models. *arXiv preprint arXiv:2607.26991*, 2026.

Andy Tang, William Chen, Andrew Wagenmaker, Chelsea Finn, and Sergey Levine. Improving robotic generalist policies via flow reversal steering. *arXiv preprint arXiv:2606.13675*, 2026.

Yunhao Tang and Shipra Agrawal. Discretizing continuous action space for on-policy optimization. In *AAAI Conference on Artificial Intelligence*, 2020.

Arash Tavakoli, Fabio Pardo, and Petar Kormushev. Action branching architectures for deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pp. 4131–4138, 2018.

TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, Naveen Kuppuswamy, Kuan-Hui Lee, Katherine Liu, Dale McConachie, Ian McMahon, Haruki Nishimura, Calder Phillips-Grafflin, Charles Richter, Paarth Shah, Krishnan Srinivasan, Blake Wulfe, Chen Xu, Mengchao Zhang, Alex Alspach, Maya Angeles, Kushal Arora, Vitor Campagnolo Guizilini, Alejandro Castro, Dian Chen, Ting-Sheng Chu, Sam Creasey, Sean Curtis, Richard Denitto, Emma Dixon, Eric Dusel, Matthew Ferreira, Aimee Goncalves, Grant Gould, Damrong Guoy, Swati Gupta, Xuchen Han, Kyle Hatch, Brendan Hathaway, Allison Henry, Hillel Hochsztein, Phoebe Horgan, Shun Iwase, Donovon Jackson, Siddharth Karamcheti, Sedrick Keh, Joseph Masterjohn, Jean Mercat, Patrick Miller, Paul Mitiguy, Tony Nguyen, Jeremy Nimmer, Yuki Noguchi, Reko Ong, Aykut Onol, Owen Pfannenstiehl, Richard Poyner, Leticia Priebe Mendes Rocha, Gordon Richardson, Christopher Rodriguez, Derick Seale, Michael Sherman, Mariah Smith-Jones, David Tago, Pavel Tokmakov, Matthew Tran, Basile Van Hoorick, Igor Vasiljevic, Sergey Zakharov, Mark Zolotas, Rares Ambrus, Kerri Fetzer-Borelli, Benjamin Burchfiel, Hadas Kress-Gazit, Siyuan Feng, Stacie Ford, and Russ Tedrake. A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation. *arXiv preprint arxiv:2507.05331*, 2025.

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv preprint arXiv:2506.15799*, 2025.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

Shuhang Wang, Ziming Li, and Hui Cheng. RLMM-Flow: A flow-based mobile manipulation framework with latent-space reinforcement learning. *arXiv preprint arXiv:2607.26460*, 2026.

Yanwei Wang, Lirui Wang, Yilun Du, Balakumar Sundaralingam, Xuning Yang, Yu-Wei Chao, Claudia Perez-D'Arpino, Dieter Fox, and Julie Shah. Inference-time policy steering through human interactions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.2307/2276774.

Wenli Xiao, Haotian Lin, Andy Peng, Haoru Xue, Tairan He, Yuqi Xie, Fengyuan Hu, Jimmy Wu, Zhengyi Luo, Linxi "Jim" Fan, Guanya Shi, and Yuke Zhu. Self-improving vision-language-action models with data generation via residual RL. *arXiv preprint arXiv:2511.00091*, 2025.

Xiu Yuan, Tongzhou Mu, Stone Tao, Yunhao Fang, Zhang Zhang, and Hao Su. Policy decorator: Model-agnostic online refinement for large policy model. In *International Conference on Learning Representations*, volume 2025, pp. 28129–28164, 2025.

Tonghe Zhang, Chao Yu, Sichang Su, and Yu Wang. ReinFlow: Fine-tuning flow matching policy with online reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jianchao Zhao, Huoren Yang, Yusong Hu, Yuyang Gao, Qiguan Ou, Cong Wan, SongLin Dong, Zhiheng Ma, and Yihong Gong. Retrieve-then-steer: Online success memory for test-time adaptation of generative VLAs. *arXiv preprint arXiv:2605.10094*, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. *Proceedings of The 7th Conference on Robot Learning (CoRL)*, 2023.

APPENDIX CONTENTS

A Extended Related Work **17**

B Additional Experiment Results **18**

B.1 Per-Task Results 19

B.2 Additional LIBERO-90 Results 20

C Algorithm Details **21**

C.1 ANCHORQ Pseudocode 22

C.2 Implementation Details and Hyperparameters 24

D Overview of Supplemental Theoretical Results **25**

E Online Flow-Reversal Abstraction and Latent Search Cost **28**

E.1 Technical lemmas 29

E.2 Main result on online flow-reversal abstraction sample complexity 31

F The general-MDP price of the persistent FRS abstraction **34**

F.1 General latent MDP and the persistent flow-reversal action class 34

F.2 Main Result: Support gap and selection error 36

F.3 From critic accuracy to selection error 39

F.4 Sharpness and regime consequences 40

F.5 Interpretation and novelty 42

G ANCHORQ traces a bias–complexity frontier **43**

G.1 Continuous local action abstractions 43

G.2 Selection error within the reach-ρ abstraction 47

G.3 A geometric rate for the support gap 48

G.4 Behaviorally distinguishable local candidates 49

G.5 Finite behavioral discretization 52

G.6 A finite-sample frontier under a generative model 53

G.7 Relation to a finite Gaussian particle cloud 57

G.8 Interpretation, novelty, and connection to the preceding results 59

H Entropic weighting for local particle refinement **60**

H.1 The local particle law and its entropic tilt 61

H.2 The exact exponential-tilting theorem 62

H.3 Risk-sensitive local weighting and the correct random variable 66

H.4 Sharpness of the critic-separation condition 70

H.5 True-score improvement under a learned local weighting 71

H.6 Finite realized local particle clouds 72
H.7 Sub-optimality bound for hierarchical ANCHORQ 75
H.8 Connection to Theorems 1–3 79
H.9 Interpretation and novelty 80

I Uncertainty Adaptation: Epistemically adaptive local tilting **81**
I.1 Calibration of the epistemic scale 82
I.2 Finite-cloud implication 84
I.3 Automatic annealing with learning 84

A Extended Related Work

Generalist and generative robot policies. Policies trained on heterogeneous robot data generalize across scenes and instructions, from action-token transformers (Brohan et al., 2023; Zitkovich et al., 2023) to open checkpoints (Ghosh et al., 2024; Kim et al., 2025). The current generation couples a vision-language backbone to a flow-matching action expert (Black et al., 2025b;a; NVIDIA et al., 2025; Gemini Robotics Team et al., 2025), inheriting a generative toolkit that includes action chunking (Zhao et al., 2023), diffusion policies (Chi et al., 2024), and flow matching with rectified flow (Lipman et al., 2023; Liu et al., 2023b). We keep such a decoder frozen throughout and improve its behavior only through the latent that initializes action generation.

RL post-training of generative policies. A complementary line of work updates the generative policy itself from reward. Q-score matching trains a diffusion policy from a learned value function (Psenka et al., 2024), while DPPO applies policy optimization to diffusion-policy parameters (Ren et al., 2024). Flow-based counterparts include Flow Q-Learning (Park et al., 2025) and ReinFlow, which fine-tunes a flow-matching policy with online RL (Zhang et al., 2025). Flow Policy Optimization brings flow matching into an on-policy, PPO-style policy-gradient framework without requiring exact action likelihoods (McAllister et al., 2026). These methods retain a rich continuous action distribution and can alter the behaviors represented by the pretrained generator, but require updating or distilling the generative policy. Our setting instead treats the pretrained flow model as a fixed behavioral prior: interaction data trains only the critic, and improvement occurs by choosing and refining the decoder's input latent. This distinction is important for large generalist policies, where preserving pretrained behavior and avoiding backpropagation through the policy are primary design constraints.

Between direct policy fine-tuning and frozen-policy steering, PLD (Xiao et al., 2025) separates RL-based adaptation from updating the generalist. It first trains lightweight residual actors with a frozen VLA, then collects recovery trajectories aligned with the base policy's state distribution and distills them into the VLA through supervised fine-tuning. This shares our use of a frozen behavioral prior during online RL, but ultimately transfers the improvements into the generator's parameters. We instead retain the frozen generator and express adaptation through its input noise, connecting our approach to the steering methods discussed next.

Steering a frozen policy in its noise space. Closest are methods that fix the policy and search its latent input, reranking chunks by value (Nakamoto et al., 2024) or steering toward human intent (Wang et al., 2025). DSRL (Wagenmaker et al., 2025) makes the noise the RL action and trains a continuous-control agent (Haarnoja et al., 2018) in it; recent variants add trust-region perturbations (Simsir & Oguz, 2026), critic-guided denoising (Ruan et al., 2026), compositional steering (Tan et al., 2026), and retrieval priors (Zhao et al., 2026). FRS (Tang et al., 2026) contributes our inversion primitive, recovering the noise that elicits a reference motion and using reversed human demonstrations as noise-space training data. Its RL variant prefills DSRL's replay buffer and adds a behavior-cloning term, after which the actor still searches the continuous noise space.

UniSteer (Lu et al., 2026) independently combines a frozen flow decoder, noise-space RL, and action-to-noise inversion. Its reference actions are online human corrections, whereas ours are a fixed library requiring no human in the loop; moreover, it trains a continuous noise actor, while

ANCHORQ learns only a critic over state-conditioned anchors and their local neighborhoods. Thus our use of inversion is persistent and representational: it defines the structured action class used throughout online learning rather than only providing data or guidance for a continuous actor.

More generally, latent-steering methods differ in whether the learned object is an actor, a denoising-time guidance signal, or a selector over proposals. LP-DS learns constrained perturbations in the latent space (Simsir & Oguz, 2026), while Q-Pilots injects critic guidance into the generative process (Ruan et al., 2026). Residual flow steering (RFS) jointly optimizes latent-noise modulation for global adaptation and residual actions for local corrections (Su et al., 2026). Compositional and retrieval-based approaches instead construct guidance from reusable policy components or prior experience (Tan et al., 2026; Zhao et al., 2026). ANCHORQ occupies a different point in this design space: inversion first supplies a small set of state-conditioned semantic coordinates, and online value learning operates over that structured set before exposing continuous local refinements. Consequently, the critic need not discover useful modes and their latent locations simultaneously.

Reduced and coarse-to-fine action spaces for RL. Shrinking the effective action space can aid exploration through discretization (Tang & Agrawal, 2020), per-dimension branching (Tavakoli et al., 2018), clustered or vector-quantized primitives (Shafiullah et al., 2022; Lee et al., 2024), and temporal abstraction over options (Sutton et al., 1999) or chunks (Li et al., 2025). CQN and its action-sequence extension recursively refine intervals of normalized per-dimension action coordinates (Seo et al., 2024; Seo & Abbeel, 2024). Our hierarchy instead operates in a learned generative latent reached by an inversion operator, related to inversion methods used for generative editing (Song et al., 2021; Rout et al., 2024).

RLMM-Flow (Wang et al., 2026) also uses the phrase "coarse-to-fine latent steering," but its hierarchy varies latent dimensionality: the coarse level pools a continuous latent along the horizon, and the fine level relaxes it to a full-dimensional residual. It is offline and does not use inversion. In contrast, our coarse level is a finite set of inverted reference motions, and our fine level is an online, Q-weighted local particle refinement around the selected motion.

Uncertainty-aware and regularized value selection. Entropy- and KL-regularized RL trade value improvement against deviation from a reference distribution. This principle underlies relative-entropy policy improvement methods such as MPO (Abdolmaleki et al., 2018), maximum-entropy control (Haarnoja et al., 2018), and the general theory of regularized MDPs (Geist et al., 2019). SAC can also tune a global entropy temperature through a constrained objective (Haarnoja et al., 2018). Our temperature has a different role: it is state-dependent and determined by epistemic uncertainty rather than by a target policy entropy.

Randomized value functions and critic ensembles provide practical surrogates for epistemic uncertainty. Bootstrapped DQN uses randomized value functions for deep exploration (Osband et al., 2016); randomized prior functions preserve diversity away from observed data (Osband et al., 2018); and SUNRISE uses ensemble uncertainty in both Bellman backups and upper-confidence action selection (Lee et al., 2021). Our value-tilting rule combines this uncertainty machinery with KL regularization at the level of a realized local particle cloud: exponential weights are the exact optimizer of a KL-regularized value objective relative to the untilted cloud. Rather than adding ensemble uncertainty as an exploration bonus, we use it to set the state-dependent inverse temperature $\beta(s) = 1/u_{\mathrm{epi}}(s)$. The selector therefore remains close to the reference weighting where candidate values are poorly known and becomes sharper as epistemic uncertainty contracts. This provides an uncertainty-aware mechanism for deciding how strongly to exploit local refinements while preserving the support supplied by the anchors and particles.

B ADDITIONAL EXPERIMENT RESULTS

We supplement the experimental results in the main paper with per-task analyses of the simulation benchmarks and, in Figure 6, a visual overview of the tasks from LIBERO, RoboTwin, and the real-world robot experiments. Across these settings, ANCHORQ consistently outperforms the steering baselines and improves the frozen base policy, including on bimanual and real-world manipulation tasks. The following sections report the simulation results by task and examine the LIBERO tasks where particle tilting most clearly separates from anchor-only learning and baseline RL algorithms.

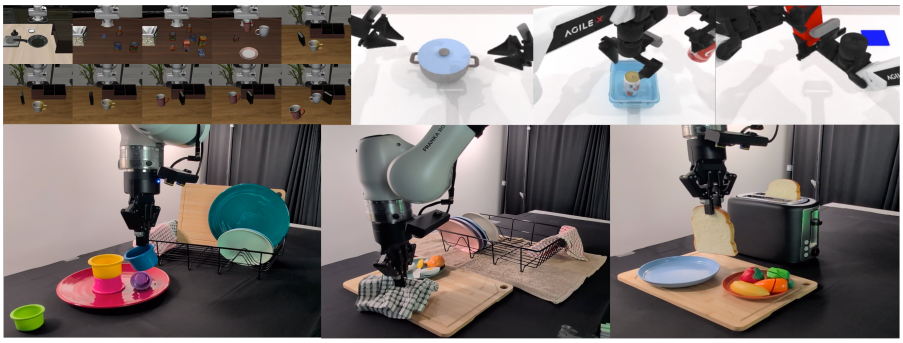

Figure 6: **Overview of the experimental tasks.** Representative tasks from the LIBERO simulation benchmark, the bimanual RoboTwin benchmark, and our real-world Franka setup.

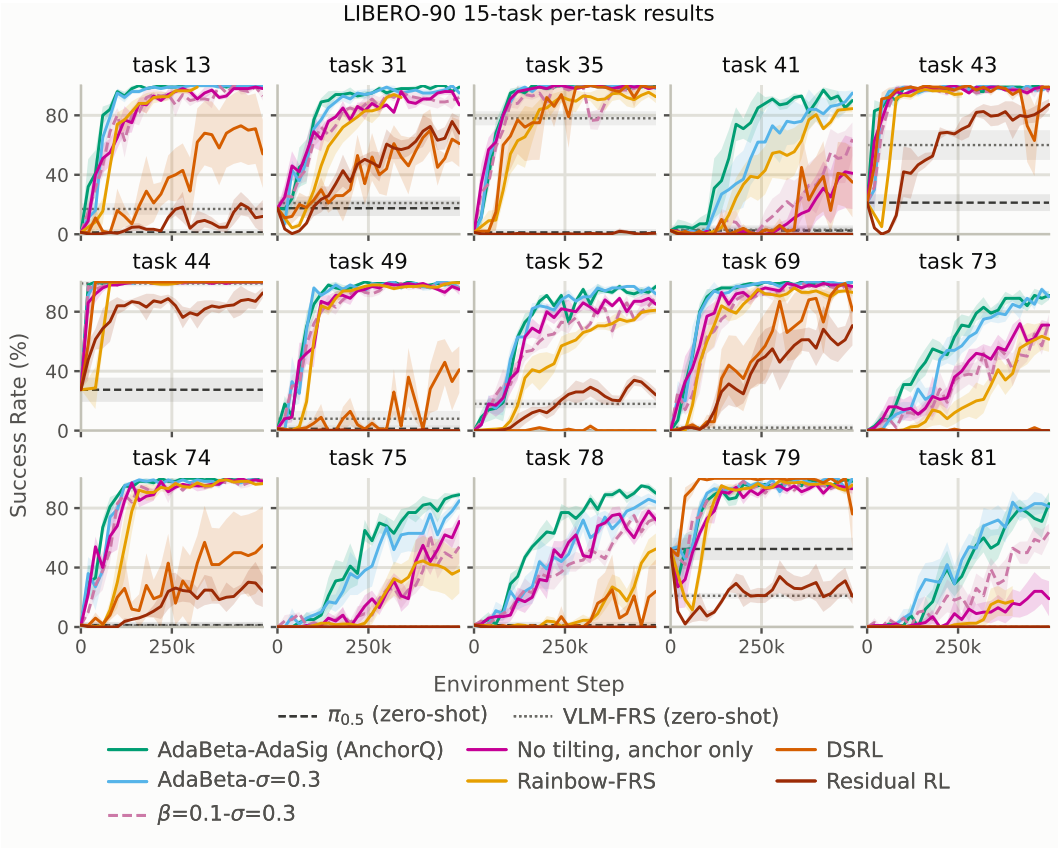

Figure 7: **Per-task performance on the LIBERO-90 15-task suite.** Success rate over training for the particle-tilting ablations and steering baselines. Curves show the mean and shaded regions show ± 1 standard error across seeds.

B.1 Per-Task Results

Figures 7–9 show the training progress for each task in the three simulation suites. On the LIBERO-90 15-task suite, the adaptive variants learn reliably across tasks, while fixed and no-tilting variants are less consistent. On the hard-10 suite, adaptive tilting succeeds on most tasks, although the hardest cases remain challenging for every method. On RoboTwin, at least one particle-tilting variant improves over the base policy on every task, with particularly clear gains on handover and object-placement tasks.

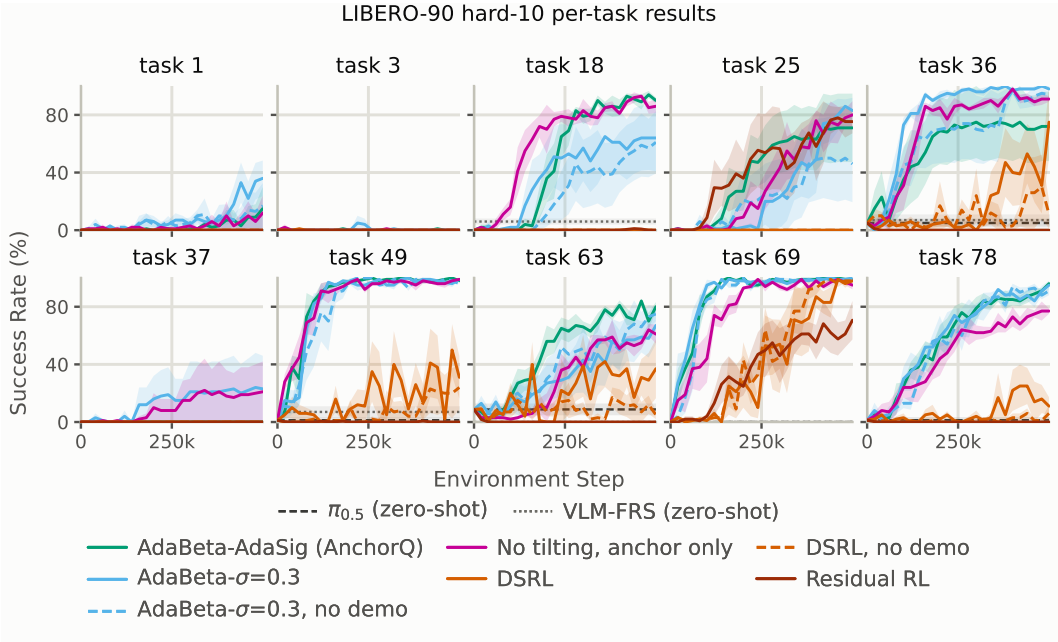

Figure 8: **Per-task performance on the LIBERO-90 hard-10 suite.** Success rate over training for the adaptive-tilting, anchor-only, DSRL, and Residual RL variants. Curves show the mean and shaded regions show ± 1 standard error across seeds.

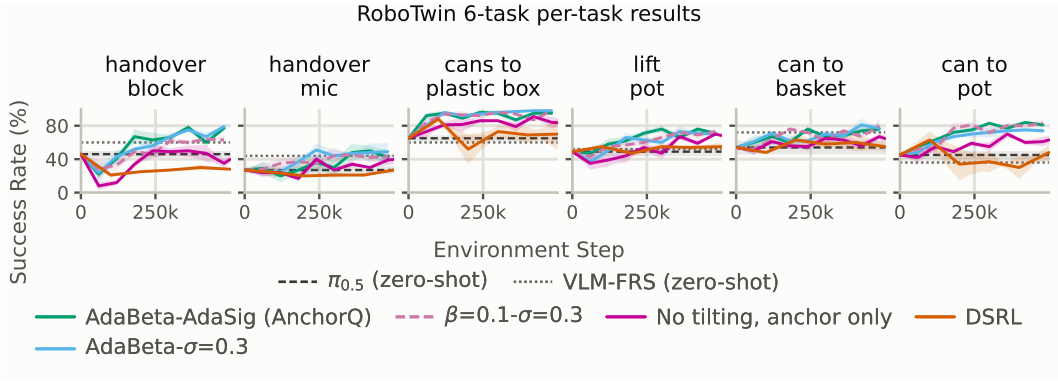

Figure 9: **Per-task performance on the RoboTwin 6-task suite.** Success rate over training for the particle-tilting ablations and steering baselines. Curves show the mean and shaded regions show ± 1 standard error across seeds.

B.2 Additional LIBERO-90 Results

Figure 10 and Table 3 report the four tasks with the clearest differences among particle tilting, anchor-only learning, and the steering baselines. Task 41 requires placing a frying pan on top of a cabinet, tasks 73 and 81 require placing a book into the front compartment of a caddy in different study scenes, and task 75 requires placing the book into the right compartment. The base policy performs poorly on all four tasks, which require precise end-effector motions to align the object with the target placement. These fine corrections are difficult to express using only the discrete flow-reversed anchors, giving particle tilting a larger advantage over the anchor-only ablation. Consequently, the adaptive tilting variants perform consistently across the four tasks, while anchor-only learning is task-dependent and the baselines (DSRL and Residual RL) make little progress.

Figure 11 illustrates this difference on task 81. The anchor-only rollout repeatedly approaches the caddy but ends with the book outside it, whereas ANCHORQ with adaptive tilting adjusts the approach and successfully inserts the book into the front compartment.

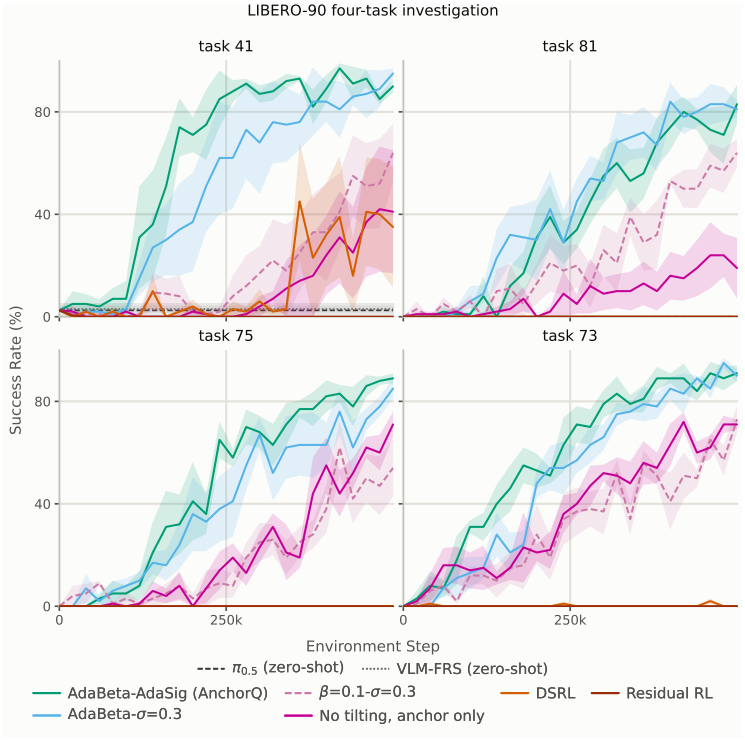

Figure 10: **Per-task ablations on four LIBERO-90 tasks.** ANCHORQ with adaptive tilting learns consistently across all four tasks. Fixed tilting, anchor-only learning, and the steering baselines show substantially greater variation across tasks. Curves show the mean and shaded regions show ± 1 standard error across seeds.

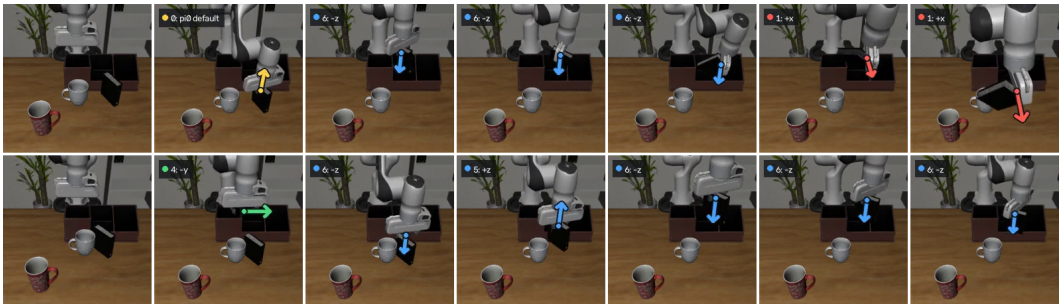

Figure 11: **Anchor-only learning versus adaptive particle tilting on LIBERO-90 task** 81**.** The task is to place the book into the front compartment of the caddy. **Top:** the no-tilting, anchor-only ablation fails to place the book. **Bottom:** ANCHORQ completes the insertion. Colored overlays identify the nearest anchor and its reference direction; yellow denotes a base-policy action.

C ALGORITHM DETAILS

We provide the ANCHORQ training procedure and the implementation choices used in our experiments, including the critic architecture and benchmark-specific hyperparameters.

Table 3: **Per-task success rate after 500k environment steps.** Success rate (%), reported as mean $\pm$ standard error across seeds. The best result on each task is bold.

Method	Task 41	Task 81	Task 75	Task 73
$\pi_{0.5}$ zero-shot	2.5 ± 2.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
VLM-FRS zero-shot	3.0 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
DSRL	35.0 ± 23.6	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Residual RL	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
ANCHORQ-AdaBeta-AdaSig	90.0 ± 2.6	$\mathbf{83.0 \pm 7.5}$	$\mathbf{89.0 \pm 1.9}$	$\mathbf{91.0 \pm 3.0}$
AdaBeta, σ=0.3	$\mathbf{95.0 \pm 1.9}$	81.0 ± 3.4	85.0 ± 4.1	90.0 ± 2.6
β=0.1, σ=0.3	64.0 ± 11.2	64.0 ± 5.4	54.0 ± 11.1	73.0 ± 5.3
No tilting, anchors only	41.0 ± 24.0	19.0 ± 11.8	71.0 ± 5.0	71.0 ± 1.9

C.1 ANCHORQ PSEUDOCODE

Algorithm 2 details the experience collection and critic updates summarized in Algorithm 1. Algorithms 3 and 4 specify local value-tilted refinement and adaptive particle reach, respectively. The accompanying text defines replay conventions and the uncertainty estimates used for adaptation.

Execution and replay. An outer decision executes at most n_{exec} low-level actions before replanning, with $1 \le n_{\text{exec}} \le h_a$; γ is the discount per outer decision. After the critic warmup W_Q, each collected transition triggers U minibatch gradient updates. Replay stores (s, z, r, s', d), where z is the executed latent, r is the reward for the decision, s' is the next observation, and d indicates true termination. Each observation carries its cached candidate bank, including the defer latent when enabled. The frozen VLA allows these banks to be reused across critic updates, avoiding repeated flow reversal. At a time-limit truncation, the final observation and its bank are retained for bootstrapping before the environment is reset.

Local refinement and action selection. For each newly collected observation, flow reversal constructs the noise anchors from the fixed reference commands, and the defer candidate is sampled from the VLA's base noise prior. This bank is cached with the observation. Each refinement call draws a fresh local cloud containing the candidate itself and M Gaussian perturbations; particle clouds and their weights are recomputed on demand. Exponential weights combine the particles into a refined latent. With probability ϵ_t, a candidate is selected uniformly; otherwise, the online ensemble mean selects the highest-valued refined latent. The selected latent is decoded by the frozen VLA and stored as the action in replay.

Critic ensemble and uncertainty signal. Write $Q_{\psi,i}$ for critic head i and $\bar{Q}_\psi(s,z) = N^{-1}\sum_{i=1}^{N} Q_{\psi,i}(s,z)$ for the online ensemble mean used for particle weighting and selection among refined latents. Uncertainty at a candidate latent is measured by the standard deviation across heads. Its median over a particle cloud provides the local uncertainty estimate used for adaptive tilting.

Adaptive particle reach. Particles are disabled before W_P. With adaptive scale disabled, $\sigma_p^{(t)}$ follows the configured schedule—e.g., $\sigma_p^{(t)} \equiv \sigma_p$ or a ramp to σ_p over W_σ decisions after warmup (an immediate switch if $W_\sigma = 0$). With adaptive scale enabled, the post-warmup scale is instead chosen from $[\sigma_{\min}, \sigma_{\max}]$ using a separate critic-only probe cloud around the current anchors at scale $\sigma_{\max}$:

$$u_{\text{probe}}(s) := \text{median}_{c,k}\, \text{Std}_{i=1,\dots,K}\, Q_{\psi,i}(s, z_{c,k}^{\text{probe}}),$$
$$\sigma(s) := \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\, \text{clip}(u_{\text{probe}}(s), 0, 1).$$

The clipping is fixed, not a running normalization: larger probe disagreement widens the search, while smaller disagreement contracts it. The probe requires no VLA decoding and serves only to set the scale; its particles do not compete for execution. A fresh refinement cloud is then drawn around each anchor at $\sigma(s)$. This optional reach heuristic is separate from the fixed-cloud adaptive-tilting guarantee.

Adaptive tilting intensity. For each anchor's realized refinement cloud, set

$$\widehat{u}_{\text{epi}}(s) := \text{median}_{0 \le k \le M}\, \text{Std}_{i=1,\dots,K}\, Q_{\psi,i}(s, z_k), \qquad \beta_{\text{loc}}(s) := \frac{1}{\max\{\widehat{u}_{\text{epi}}(s), u_{\min}\}}.$$

Algorithm 2 ANCHORQ: online anchors, local value tilting, and U critic updates per decision

Input: frozen π_{flow}, REVFLOW, reference map φ; K critics $\{Q_{\psi,i}\}$; prioritized replay $\mathcal{D}$
Input: execution length n_{exec}, updates U, critic warmup W_Q; $\gamma, \eta, \tau, N_{\text{targ}}$; exploration schedule ϵ_t
Input: M particles, particle warmup W_P, scale schedule $\sigma_p^{(t)}$; optional adaptive scale with $0 \le \sigma_{\min} \le \sigma_{\max}$
Input: fixed β or adaptive tilt with uncertainty floor $u_{\min} > 0$
Initialize $\bar{\psi} \leftarrow \psi$ and reset the environment to s_0
for each outer decision $t = 0, 1, \ldots$ **do**
    Retrieve cached $Z(s_t)$; if absent, construct and cache it by flow reversal of $\varphi(s_t, c)$ for each reference command
    When constructing a bank, include a base-prior candidate if defer is enabled
    if $M = 0$ or $t < W_P$ **then**
        $\sigma_t \leftarrow 0$
    else if adaptive scale is enabled **then**
        $\sigma_t \leftarrow \text{ADAPTIVESCALE}(s_t, Z(s_t))$
    else
        $\sigma_t \leftarrow \sigma_p^{(t)}$
    end if
    for each candidate $c \in Z(s_t)$ **do**
        $\tilde{z}_{t,c} \leftarrow \text{LOCALREFINE}(s_t, z_c(s_t), t, \sigma_t)$
    end for
    Choose c_t uniformly w.p. ϵ_t; otherwise $c_t \leftarrow \arg\max_{c \in Z(s_t)} \bar{Q}_\psi(s_t, \tilde{z}_{t,c})$
    $\tilde{z}_t \leftarrow \tilde{z}_{t,c_t}$
    Decode $a_t \leftarrow \pi_{\text{flow}}(s_t, \tilde{z}_t)$; execute its first n_{exec} actions, stopping early if the episode ends
    Observe outer reward r_t, final observation s_{t+1}, and terminal flag d_t
    Construct and cache $Z(s_{t+1})$ with the same anchor/defer construction
    Store $(s_t, \tilde{z}_t, r_t, s_{t+1}, d_t)$ in $\mathcal{D}$ ▷ observations include their cached banks
    if $t + 1 \ge W_Q$ and $\mathcal{D}$ can supply a minibatch **then**
        for $u = 1, \ldots, U$ **do**
            Sample $\mathcal{B} = \{(s_b, z_b, r_b, s'_b, d_b)\}$ with IS weights ω_b
            Draw target-head subset $\mathcal{I}$, $|\mathcal{I}| = N_{\text{targ}}$
            for each replay row b (batched) **do**
                $y_b \leftarrow r_b$
                if $d_b = 0$ **then**
                    Retrieve cached $Z(s'_b)$ from replay
                    if $M = 0$ or $t < W_P$ **then**
                        $\sigma'_b \leftarrow 0$
                    else if adaptive scale is enabled **then**
                        $\sigma'_b \leftarrow \text{ADAPTIVESCALE}(s'_b, Z(s'_b))$
                    else
                        $\sigma'_b \leftarrow \sigma_p^{(t)}$
                    end if
                    for each candidate $c \in Z(s'_b)$ **do**
                        $\tilde{z}'_{b,c} \leftarrow \text{LOCALREFINE}(s'_b, z_c(s'_b), t, \sigma'_b)$
                    end for
                    $c'_b \leftarrow \arg\max_{c \in Z(s'_b)} \bar{Q}_\psi(s'_b, \tilde{z}'_{b,c})$
                    $\tilde{z}'_b \leftarrow \tilde{z}'_{b,c'_b}$
                    $y_b \leftarrow r_b + \gamma \min_{i \in \mathcal{I}} Q_{\bar{\psi},i}(s'_b, \tilde{z}'_b)$
                end if
            end for
            $\mathcal{L}_Q \leftarrow (K|\mathcal{B}|)^{-1} \sum_{b \in \mathcal{B}} \sum_{i=1}^{K} \omega_b \big(Q_{\psi,i}(s_b, z_b) - \text{stopgrad}(y_b)\big)^2$
            Take an optimizer step on ψ with learning rate η
            Update replay priorities from absolute TD errors
            $\bar{\psi} \leftarrow (1 - \tau)\bar{\psi} + \tau\psi$ ▷ frozen VLA and any random priors remain fixed
        end for
    end if
    If the episode ended, reset before the next decision
end for

This instantiates (135); when adaptation is disabled, the routine instead uses the fixed intensity $\beta = 1/T$. The adaptive weights compare value gaps in units of the floored uncertainty scale: $w_j/w_k = \exp((q_j - q_k)/u_{\text{epi}}(s))$, where $q_k = \bar{Q}_\psi(s, z_k)$ and $u_{\text{epi}}(s) = \max\{\widehat{u}_{\text{epi}}(s), u_{\min}\}$. Large uncertainty therefore softens the weighting, while small uncertainty permits stronger preference for

Algorithm 3 Local refinement subroutine for Algorithm 2

function LOCALREFINE(s, z_c, t, σ)
  if $M = 0$ or $t < W_P$ or $\sigma = 0$ **then**
    return z_c ▷ exact-anchor mode
  end if
  Set $z_0 \leftarrow z_c$; draw $\xi_k \sim \mathcal{N}(0, I)$ and set $z_k \leftarrow z_c + \sigma\xi_k$ for $k = 1, \dots, M$ ▷ fresh local cloud
  Evaluate the critic ensemble on the cloud and set $q_k \leftarrow \bar{Q}_\psi(s, z_k)$ for $k = 0, \dots, M$
  if adaptive tilt is enabled **then**
    $\widehat{u}_{\text{epi}} \leftarrow \text{median}_{0 \le k \le M} \, \text{Std}_{i=1,\dots,K} \, Q_{\psi,i}(s, z_k)$
    $\beta_{\text{loc}} \leftarrow 1/\max\{\widehat{u}_{\text{epi}}, u_{\min}\}$
  else
    $\beta_{\text{loc}} \leftarrow \beta$
  end if
  $w_k \leftarrow \exp\big(\beta_{\text{loc}}(q_k - \max_j q_j)\big) \Big/ \sum_{\ell=0}^{M} \exp\big(\beta_{\text{loc}}(q_\ell - \max_j q_j)\big)$ for $k = 0, \dots, M$
  return $\sum_{k=0}^{M} w_k z_k$ ▷ barycentric refinement
end function

Algorithm 4 Adaptive particle scale

function ADAPTIVESCALE($s, Z(s)$)
  for each candidate $c \in Z(s)$ **do**
    $z_{c,0}^{\text{probe}} \leftarrow z_c(s)$
    Draw $\xi_{c,k}^{\text{probe}} \sim \mathcal{N}(0, I)$ and set $z_{c,k}^{\text{probe}} \leftarrow z_c(s) + \sigma_{\max}\xi_{c,k}^{\text{probe}}$, $k = 1, \dots, M$
  end for
  $u_{\text{probe}} \leftarrow \text{median}_{c,\, 0 \le k \le M} \, \text{Std}_{i=1,\dots,K} \, Q_{\psi,i}(s, z_{c,k}^{\text{probe}})$
  return $\sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \, \text{clip}(u_{\text{probe}}, 0, 1)$
end function

high-scoring particles. The floor limits β_{loc} to $1/u_{\min}$, preventing a near-zero disagreement estimate from producing an arbitrarily sharp tilt.

Theorem 9 interprets these weights as an uncertainty-weighted KL-regularized improvement rule. Its reliability requires the calibration condition in Assumption 8: local critic error must be controlled by the uncertainty scale. Neither a small ensemble standard deviation nor the use of random priors establishes that condition automatically. Corollary 13 additionally accounts for executing the weighted latent, with a separate interpolation penalty under local smoothness. Adaptive reach and adaptive tilt can be enabled independently. Reach uses disagreement over the probe cloud, while tilt uses disagreement recomputed on each anchor's refinement cloud. Their realized values need not change monotonically during learning, since the critic and clouds also change.

Critic updates and next-state refinement. For each nonterminal replay row, we retrieve the cached next-state bank and refine all candidates with fresh particles, then select the highest-valued refined latent using the online ensemble mean. Adaptive reach, tilt, and particle weights are recomputed using the current online ensemble at the next state. At each update, we sample a subset $\mathcal{I}$ of N_{targ} target heads and form the detached target $y = r + \gamma \min_{i \in \mathcal{I}} Q_{\bar{\psi},i}(s', \tilde{z}')$, where $\tilde{z}'$ is the selected refined latent; true terminal rows use $y = r$. Prioritized replay supplies importance-sampling weights for the squared TD loss. Each optimizer step is followed by a replay-priority refresh and a Polyak target update; the VLA and random prior networks remain frozen.

C.2 Implementation Details and Hyperparameters

Following FRS (Tang et al., 2026), we use six cardinal Cartesian reference motions and a *defer* candidate sampled from the VLA's base noise prior. For bimanual RoboTwin tasks, the Cartesian product of the six directions for each arm yields 36 reference motions; adding defer gives 37 candidates. In simulation, ANCHORQ and DSRL use REDQ's ensemble update scheme (Chen et al., 2021a), with multiple critic updates per decision and a minimum over randomly sampled target heads. Real-world training uses a three-step TD target (multi-step Q-learning) in place of the one-step target in Algorithm 2, and averages the sampled target heads. Table 4 lists the settings for each benchmark.

Critic architecture. The critic evaluates $Q(s, z)$ from state features and the full latent noise trajectory. On LIBERO-90, two frozen 384-dimensional DINOv2-S embeddings from the third-person and wrist cameras form a 768-dimensional visual input, alongside an 8-dimensional proprioceptive vector. Separate linear layers with ReLU project the visual and proprioceptive inputs to 256 dimensions each. Their concatenation passes through a $512 \to 256 \to 256$ state encoder. The 10×32 noise trajectory is flattened and encoded by a $320 \to 256 \to 256$ MLP. Input dimensions follow each benchmark's observation and latent spaces.

With randomized prior functions enabled, each of the $N = 10$ heads has independent trainable featurizers and encoders. An output MLP maps the concatenated state and noise features through widths $512 \to 256 \to 256 \to 1$. Hidden layers use ReLU, without layer normalization or dropout. Each head adds an independently initialized, frozen prior network: $Q_{\psi,i}(s, z) = f_{\psi_i}(s, z) + \kappa p_i(s, z)$, with prior scale κ given in Table 4. Disagreement across heads supplies the uncertainty signal for local adaptation.

For DROID, we stack four frames of proprioceptive and DINO features and use shared trainable encoders with independent scalar output heads, without random priors. The critic applies layer normalization to the flattened 15×32 noise input; execution uses the original latent.

Table 4: **Hyperparameters by benchmark.** Settings for LIBERO-90 (Liu et al., 2023a), RoboTwin (Chen et al., 2025), and the real-world experiments.

Symbol	Short Key	LIBERO-90	RoboTwin	Real world	Meaning
Environment and action space					
n_{exec}	`pi0_replan_steps`	10	30	10	substeps per outer decision
K	`pi0_flow_steps`	10	10	10	Euler steps in the flow ODE
C	`num_anchors`	7	37	7	reference candidates including defer
(H, D)	`noise_shape`	$(10, 32)$	$(50, 32)$	$(15, 32)$	noise trajectory shape
Critic, replay, and exploration					
γ	`gamma`	0.999^{10}	0.999^{30}	0.99	per-outer-decision discount
η	`learning_rate`	3×10^{-4}	3×10^{-4}	3×10^{-4}	Adam step size
B	`batch_size`	16	16	64	transitions per gradient step
N	`num_critics`	10	10	10	critic heads
N_{targ}	`n_critics_to_sample`	2	2	2	random target subset size
τ	`tau`	0.005	0.005	0.005	Polyak rate
U	`num_updates_per_train_step`	10	10	5	gradient steps / decision
W_Q	`learning_starts`	1000	1000	0	critic warmup (outer decisions)
ϵ	`eps_start` $\to$ `eps_end`	$1.0 \to 0.05$	$1.0 \to 0.05$	$0.6 \to 0.05$	exploration schedule
	`eps_decay_steps`	10,000	5000	2000	decisions for exploration decay
κ	`rpf_prior_scale`	2	20	2	random-prior scale
Particle tilting					
M	`particle_count`	32	10	8	sampled particles per anchor
σ_p	`particle_sigma`	0.5	0.3	0.0	particle scale
β	`tilt_beta`	1.0	0.1	1.0	fixed tilting strength
W_P	`particle_warmup`	10,000	2000	2000	particle warmup (outer decisions)
$(\sigma_{\min}, \sigma_{\max})$	`sigma_min/_max`	$(0.05, 0.3)$	$(0.05, 0.3)$	$(0.05, 0.3)$	adaptive particle-scale range

D Overview of Supplemental Theoretical Results

The theoretical results in this appendix develop a unified view of ANCHORQ as a *structured online action abstraction with local value-guided refinement*. The results answer a sequence of complementary questions: why retaining flow-reversed action anchors as the online action class can reduce the statistical burden of learning; what price is paid for that restriction in terms of representation bias; how local particles trade this representation bias against increased learning complexity; how exponential value-tilting exploits the resulting candidate support; and, finally, how the tilting intensity can be adapted to epistemic uncertainty in the learned critic. Taken together, the theory separates three distinct roles:

- *Flow-reversal action anchor abstraction*: determines the structured online action class;
- *Local particle refinement*: relax that class and enlarge the available latent support;
- *Value tilting*: reallocate posterior sampling probability toward high-value candidates.

The uncertainty-adaptive rule then determines how aggressively the learned critic should be trusted when performing the final reweighting.

Online flow-reversal can reduce identification complexity. We begin in Appendix E by isolating the statistical effect of retaining the structured latent action class induced by the flow-reversed action anchors throughout online learning, compared with learning over a broader latent class even when *a priori* information is given about the restricted set.

Theorem 4 considers a stylized one-decision latent-mode problem with N behaviorally distinct latent modes, of which only $C < N$ belong to the online, state-conditioned flow-reversed action anchor abstraction. When the flow-reversed action class is sufficiently accurate, a learner restricted to this class requires

$$\mathcal{O}\left(\frac{C}{\epsilon^2}\log\frac{C}{\alpha}\right)$$

online samples. On the other hand, a learner that retains a uniform guarantee over the larger class must, on an instance for which the flow-reversal action anchor abstraction has zero bias, spend additional online samples ruling out the $N - C$ non-flow-reversed modes. Importantly, this lower bound persists even if the learner is given the exact reward laws of all flow-reversed modes before online interaction.

Thus the statistical benefit is attributable to the *persistent online restriction* itself: information about flow-reversed behaviors, whether supplied through prefill or even their exact reward laws, cannot cannot certify that unexplored non-flow-reversed behaviors are suboptimal.

The statistical gain has an explicit support price. The smaller action class is beneficial only when it contains sufficiently good actions. In Appendix F, Theorem 5 therefore moves to a general discounted latent MDP and defines the one-step flow-reversal support gap

$$\Delta_{\mathrm{F}}(s) := V^\star(s) - \max_{c\in\mathcal{C}} Q^\star\big(s, z_c(s)\big), \quad \text{and} \quad \bar{\Delta}_{\mathrm{F}} := \sup_{s\in\mathcal{S}} \Delta_{\mathrm{F}}(s).$$

It shows that the value price of remaining inside the flow-reversed anchor bank at every decision is controlled by this local support deficiency,

$$\Delta_{\mathrm{F}}(s) \le V^\star(s) - V^\star_{\mathrm{F}}(s) \le \frac{\bar{\Delta}_{\mathrm{F}}}{1-\gamma},$$

and, more importantly, gives the exact decomposition

$$V^\star - V^\omega = \underbrace{V^\star - V^\star_{\mathrm{F}}}_{\text{persistent flow-reversal support price}} + \underbrace{(I-\gamma P^\omega)^{-1}\,\mathrm{err}_{\mathrm{sel}}(\omega;\cdot)}_{\text{selection error within the flow-reversed class}}.$$

The first term is structural: no amount of additional critic data can recover a useful behavior that is absent from the flow-reversed action support. The second is statistical and can decrease as the learned critic and selector improve. Together with Theorem 4, this identifies the fundamental trade-off behind ANCHORQ: the smaller online class can lower identification complexity, but only at the cost of an explicit and measurable support bias.

Particles trace a bias–complexity frontier. In Appendix G, we show that ANCHORQ provides a controlled way to relax the exact-anchor restriction without immediately reverting to unrestricted latent-space control. Theorem 6 considers a local action class of deterministic local reach ρ around the state-dependent flow-reversed anchors. Increasing ρ can only reduce the corresponding support gap $\bar{\Delta}_\rho$; under smooth local value geometry, the remaining gap also admits an explicit local quadratic bound. At the same time, increasing ρ enlarges the number of actions $A_\rho(\epsilon_{\mathrm{B}})$ that are distinguishable at resolution ϵ_{B} in the Bellman behavioral pseudometric. The finite-sample result makes both effects explicit:

$$\left\|V^\star - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}}\right\|_\infty \le \frac{\bar{\Delta}_\rho + \epsilon_{\mathrm{B}} + 2\epsilon_Q}{1-\gamma},$$

while the required generative-model sample count grows with $A_\rho(\epsilon_{\mathrm{B}})$. Hence local expansion is not free exploration: it trades reduced structural bias against a larger behavioral action class that must be represented and valued. The deterministic reach ρ used in this analysis is an analytical local envelope and is not identified with the Gaussian particle scale used by the implementation; Appendix G.7 separately connects a finite Gaussian particle cloud to coverage of such a local envelope.

Exponential tilting exploits the available support. Once the candidate class has been generated, the remaining problem is how to use value information to allocate execution probability within it.

Indeed, the preceding results complete the coarse-to-fine picture developed in Theorems 4–6. Those results characterize the action representation made available to online learning; Appendix H characterizes how value information is used to refine that representation after a coarse anchor has been selected.

Theorem 7 gives the exact population-level characterization of this operation. Conditional on a state s and selected anchor c, the exponentially tilted law is the unique optimizer of a KL-regularized value-improvement problem relative to the untilted local particle law. The theorem further shows that increasing the tilt intensity β monotonically increases both the expected local-particle score and the KL departure from the reference law, while preserving its support. Thus the particle construction determines which local refinements are available, whereas β determines how aggressively value information reallocates weight among those refinements. Corollary 8 makes the corresponding upper-tail preference explicit: for small β, the entropic objective equals the mean local score plus a positive variance correction to first order.

The same analysis quantifies when this value-directed weighting remains reliable under an imperfect learned critic. Theorem 7*.e.* shows that a truly high-value local region with population-score separation $\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}})$—i.e., worst score inside A minus the best score outside A, ignoring $\nu_{s,c}$-measure-zero exceptions—has its odds amplified by at least the factor

$$\exp\left(\beta\left(\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) - 2\epsilon_{\mathrm{crit}}\right)\right).$$

Consequently, strict amplification is guaranteed whenever the true local value separation exceeds $2\epsilon_{\mathrm{crit}}$.

Finite-cloud results connect these population identities to the implemented particle weighting rule. With critic error, Corollary 11 gives the corresponding population-score gap $\log(M+1)/\beta + 2\epsilon_{\mathrm{crit}}$. This still does not by itself characterize the action executed by ANCHORQ, because the implementation forms the barycentric latent rather than sampling one of the weighted particles. Proposition 13 closes this gap under local smoothness of the population value function: the loss from replacing the weighted particle values by the value of their weighted-average latent is at most $L_{\mathrm{sm}}\mathcal{V}_{c,\beta}(s)/2$, where $\mathcal{V}_{c,\beta}(s)$ is the weighted particle variance. Corollary 12 therefore separates the local refinement error into finite-particle representation error, finite-temperature weighting error, critic error, and barycentric interpolation error.

Finally, Theorem 8 propagates these local guarantees through the complete hierarchical ANCHORQ policy. The resulting bound

$$\|V^\star - V^{\pi_{\mathrm{A}}}\|_\infty \le \frac{\bar{\Delta}_\rho + \bar{\epsilon}_{\mathrm{sel}} + \bar{\epsilon}_{\mathrm{rep}} + \beta^{-1}\log(M+1) + 2\bar{\epsilon}_{\mathrm{crit}} + \frac{L_{\mathrm{sm}}}{2}\bar{\mathcal{V}}_\beta}{1-\gamma}$$

decomposes the sequential-control loss into six distinct terms: the structural support gap $\bar{\Delta}_\rho$, the coarse-anchor selection error $\bar{\epsilon}_{\mathrm{sel}}$, the finite-particle representation error $\bar{\epsilon}_{\mathrm{rep}}$, the finite-temperature term $\beta^{-1}\log(M+1)$, the critic-error term $2\bar{\epsilon}_{\mathrm{crit}}$, and the barycentric interpolation penalty $(L_{\mathrm{sm}}/2)\bar{\mathcal{V}}_\beta$. The decomposition makes explicit where ANCHORQ can lose value. The anchor and particle representation determines whether sufficiently good actions are available, through the support gap $\bar{\Delta}_\rho$ and representation error $\bar{\epsilon}_{\mathrm{rep}}$; learning determines whether those actions are identified reliably, through $\bar{\epsilon}_{\mathrm{sel}}$ and $\bar{\epsilon}_{\mathrm{crit}}$; and the value-tilted refinement mechanism introduces the finite-temperature and interpolation terms.

Thus, the final guarantee connects the method's coarse-to-fine representation directly to its online decision rule: *richer local search reduces structural bias, while reliable value information is needed to exploit that additional flexibility.*

Epistemic uncertainty determines how aggressively to tilt. The final result addresses this reliability–aggressiveness tradeoff directly. In Section I, the tilting intensity is chosen statewise according to

$$\beta(s) = 1/u_{\mathrm{epi}}(s),$$

where $u_{\mathrm{epi}}(s)$ is an ensemble-based surrogate for local epistemic uncertainty. Theorem 9 isolates the benefit of adapting the tilting intensity to epistemic uncertainty. The resulting particle distribution is

the exact optimizer of an uncertainty-weighted KL-regularized improvement problem,

$$\max_{\pi} \left\{ \mathbb{E}_{\pi}[\widehat{q}] - u_{\mathrm{epi}}(s) D_{\mathrm{KL}}(\pi \| \mu) \right\}.$$

Thus epistemic uncertainty directly determines the penalty for departing from the flow-reversed reference distribution.

Under the finite-bank calibration condition that critic error is proportional to $u_{\mathrm{epi}}(s)$, the same theorem gives a true-value selection bound whose error contracts linearly with the epistemic scale. It also shows that reliable preference between two candidates is governed by the standardized value gap $(q_j - q_k)/u_{\mathrm{epi}}(s)$, rather than by the unnormalized critic difference alone. Moreover, when critic perturbations occur on the same uncertainty scale, adaptive $\beta(s) = 1/u_{\mathrm{epi}}(s)$ yields a scale-free stability bound: sensitivity of the tilted weights depends on relative critic error rather than growing with the absolute uncertainty level. Corollary 14 gives the most direct finite-cloud interpretation of adaptive β. The gap between the best represented local particle and the population value achieved by the learned weighting is proportional to the epistemic uncertainty scale:

$$\text{adaptive weighting error} = \mathcal{O}\left(u_{\mathrm{epi}}(s)\right).$$

Thus, when the critic is poorly known, the method remains deliberately soft; as critic uncertainty contracts, the finite-temperature and critic-estimation terms contract together.

Overall picture. The results therefore form a single structured refinement hierarchy:

Mechanism	Effect
online flow-reversal	small structured online action class
local particles	controlled relaxation of flow-reversed anchor support
exponential tilting	value-directed exploitation of that support
adaptive β	epistemically calibrated exploitation

The first two results characterize the benefit and cost of the online flow-reversed action anchor abstraction; the particle result describes how that restriction can be relaxed; and the tilting results describe how the resulting capacity should be used once it is available. In this sense, flow-reversal of the coarse action anchors determines the structured hypothesis class, particles determine how much local expressiveness is exposed to learning, and epistemically adaptive value tilting determines how strongly the learned critic should exploit that expressiveness.

E Online Flow-Reversal Abstraction and Latent Search Cost

We first isolate the statistical consequence of using the state-conditioned bank of flow-reversed action anchors as a *persistent action abstraction*. The result is stated for a one-decision episodic latent Markov decision process (MDP). This is exactly a fixed-confidence Bernoulli best-arm identification problem; a one-step episodic MDP (see, e.g., Lattimore & Szepesvári (2020); Kaufmann et al. (2016)). The construction separates the cost of identifying a useful latent mode from state-distribution shift, temporal credit assignment, and function approximation.

Latent-mode MDPs. Fix integers C and N such that $1 \le C < N$. Let $[N] := \{1, \dots, N\}$ and $\mathcal{C} := [C]$. The set $\mathcal{C}$ indexes the C latent modes generated by the flow-reversed action anchors abstraction, while $[N] \setminus \mathcal{C}$ indexes the remaining latent modes available to a learner that retains the full latent action class. It suffices to analyze this finite latent-mode subclass: any learner that is uniformly correct over the full continuous latent space must also be correct on this N-point restriction, so any lower bound here carries to unrestricted continuous latent-space learning.

Choose distinct latent actions $z_1, \dots, z_N \in \mathcal{Z}$. For each mean vector

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_N) \in [0,1]^N,$$

define the one-decision episodic latent MDP—namely, the learner chooses one latent action, observes binary success or failure, and the episode ends—$\mathcal{M}_{\boldsymbol{\mu}}$ as follows. The state space consists of an initial state s_0 and a terminal state s_{term}. At s_0, selecting latent z_i produces a reward

$R \sim \mathrm{Bernoulli}(\mu_i)$ and transitions deterministically to s_{term}. Each episode therefore produces one online reward observation.

At the unique decision state s_0, define the flow-reversed anchor map F by

$$F(s_0, c) = z_c, \qquad c \in \mathcal{C},$$

so that $z_1, \ldots, z_C$ are precisely the latent actions available under the online flow-reversal abstraction.

For this MDP, set

$$Q^\star_{\boldsymbol{\mu}}(s_0, z_i) = \mu_i \qquad \text{and} \qquad V^\star_{\boldsymbol{\mu}}(s_0) = \mu^\star(\boldsymbol{\mu}), \quad \text{where } \mu^\star(\boldsymbol{\mu}) := \max_{i \in [N]} \mu_i.$$

The best mean available within the *online flow-reversal abstraction* is $\mu^\star_F(\boldsymbol{\mu}) := \max_{c \in \mathcal{C}} \mu_c$. The flow-reversal abstraction bias on $\mathcal{M}_{\boldsymbol{\mu}}$ is

$$\Delta_F(\boldsymbol{\mu}) := \mu^\star(\boldsymbol{\mu}) - \mu^\star_F(\boldsymbol{\mu}) = V^\star_{\boldsymbol{\mu}}(s_0) - \max_{c \in \mathcal{C}} Q^\star_{\boldsymbol{\mu}}(s_0, F(s_0, c)). \tag{1}$$

Online learners and the correctness criterion. An online learner is specified by a sequence of action-selection kernels and a terminal decision rule. Before online interaction, the learner may be given the exact flow-reversed mode Bernoulli laws, equivalently the vector $W_F(\boldsymbol{\mu}) := (\mu_c)_{c \in \mathcal{C}}$. This side information is at least as informative as any finite FRS prefill consisting only of samples from the flow-reversed modes. In Theorem 4 below, part *a.* uses no such side information; part *b.* grants it to the full-class learner in order to strengthen the lower bound.

At online round $t \in [n]$, the learner selects an index $I_t \in [N]$ as a possibly randomized function of $W_F(\boldsymbol{\mu})$ and the previous online observations, and then observes $R_t \,|\, I_t \sim \mathrm{Bernoulli}(\mu_{I_t})$. Conditionally on the selected indices, the rewards are independent across rounds. After n online rounds, the learner returns an index $\widehat{I}_n \in [N]$. For each $i \in [N]$, define the online sample count

$$N_i(n) := \sum_{t=1}^{n} \mathbf{1}\{I_t = i\}.$$

A learner is *flow-reversal-restricted* if $I_t \in \mathcal{C}$ for every t and $\widehat{I}_n \in \mathcal{C}$ almost surely. A learner is *full-class* if it may select and return any index in $[N]$.

Let $\mathfrak{M}$ be a class of latent-mode MDPs. For accuracy $\epsilon > 0$ and failure probability $\alpha \in (0, 1)$, a learner is (ϵ, α)-correct on $\mathfrak{M}$ if, for every $\mathcal{M}_{\boldsymbol{\mu}} \in \mathfrak{M}$, the guarantee holds:

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\mu^\star(\boldsymbol{\mu}) - \mu_{\widehat{I}_n} \le \epsilon\right) \ge 1 - \alpha.$$

E.1 Technical lemmas

We need some technical lemmas before we prove the first main result.

For $p, q \in [0, 1]$, write

$$\mathrm{kl}(p, q) := p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$$

for the binary relative entropy, with the standard conventions

$$0 \log \frac{0}{q} := 0, \qquad p \log \frac{p}{0} := +\infty \text{ for } p > 0,$$

and the analogous conventions for the second term.

The following is a slight extension of the standard adaptive KL divergence decomposition to include instance-independent side information and a terminal randomized output; see Lattimore & Szepesvári (2020, Lemma 15.1) and Kaufmann et al. (2016, Lemma 1). We state it here in an abstract form and specialize it to the Bernoulli latent-mode construction below.

Lemma 1 (Adaptive change of measure)**.** Let θ and θ' be two problem instances with a common finite action set $[N] := \{1, \ldots, N\}$. For each action $i \in [N]$, let $P_{\theta,i}$ and $P_{\theta',i}$ denote the corresponding observation distributions under the two instances. Consider an adaptive learner run for

exactly n rounds. Before the first round, the learner receives side information W having the same distribution under θ and θ'. At round t, the learner selects an action $A_t \in [N]$ according to a possibly randomized kernel depending on the history

$$H_{t-1} := \big(W, A_1, Y_1, \ldots, A_{t-1}, Y_{t-1}\big),$$

and then observes $Y_t \sim P_{\theta,A_t}$ under instance θ, and $Y_t \sim P_{\theta',A_t}$ under instance θ'. After round n, the learner may produce a possibly randomized terminal output $\widehat{U}$ based on the interaction history. Assume that the learner uses the same action-selection and terminal-output kernels under the two instances. Define the complete transcript

$$\mathcal{T}_n := \big(W, A_1, Y_1, \ldots, A_n, Y_n, \widehat{U}\big),$$

and let $\mathbb{P}_\theta^{(n)}$ and $\mathbb{P}_{\theta'}^{(n)}$ denote the distributions of $\mathcal{T}_n$ under θ and θ', respectively. Also define $N_i(n) := \sum_{t=1}^n \mathbf{1}\{A_t = i\}$. Then

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{(n)} \,\middle\|\, \mathbb{P}_{\theta'}^{(n)}\right) = \sum_{i=1}^N \mathbb{E}_\theta[N_i(n)] D_{\mathrm{KL}}\left(P_{\theta,i} \,\|\, P_{\theta',i}\right), \tag{2}$$

so that, for every event $\mathcal{E}$ measurable with respect to $\mathcal{T}_n$, the lower bound holds:

$$\sum_{i=1}^N \mathbb{E}_\theta[N_i(n)] D_{\mathrm{KL}}\left(P_{\theta,i} \,\|\, P_{\theta',i}\right) \geq \mathrm{kl}\left(\mathbb{P}_\theta^{(n)}(\mathcal{E}), \mathbb{P}_{\theta'}^{(n)}(\mathcal{E})\right). \tag{3}$$

Proof. Fix a round t. Conditional on a realized history $H_{t-1} = h$, the learner uses the same conditional distribution for A_t under instances θ and θ'. Hence the action-selection step contributes zero conditional relative entropy.

Conditional on $(H_{t-1}, A_t = i)$, the observation distribution is $P_{\theta,i}$ under instance θ and $P_{\theta',i}$ under instance θ'. The chain rule for relative entropy therefore gives

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{(n)} \,\middle\|\, \mathbb{P}_{\theta'}^{(n)}\right) = \sum_{t=1}^n \mathbb{E}_\theta\left[D_{\mathrm{KL}}\left(P_{\theta,A_t} \,\|\, P_{\theta',A_t}\right)\right]. \tag{4}$$

The initial side information contributes zero relative entropy because it has the same distribution under the two instances. Likewise, the terminal output contributes zero conditional relative entropy because, conditional on the preceding interaction history, the learner uses the same terminal-output kernel under θ and θ'.

Expanding the expectation in (4) over the selected action gives

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}_\theta\left[D_{\mathrm{KL}}\left(P_{\theta,A_t} \,\|\, P_{\theta',A_t}\right)\right] &= \sum_{t=1}^n \sum_{i=1}^N \mathbb{P}_\theta(A_t = i) D_{\mathrm{KL}}\left(P_{\theta,i} \,\|\, P_{\theta',i}\right) \\
&= \sum_{i=1}^N \mathbb{E}_\theta\left[\sum_{t=1}^n \mathbf{1}\{A_t = i\}\right] D_{\mathrm{KL}}\left(P_{\theta,i} \,\|\, P_{\theta',i}\right) \\
&= \sum_{i=1}^N \mathbb{E}_\theta[N_i(n)] D_{\mathrm{KL}}\left(P_{\theta,i} \,\|\, P_{\theta',i}\right),
\end{aligned}$$

which proves (2).

Now fix any event $\mathcal{E}$ measurable with respect to the complete transcript $\mathcal{T}_n$. Mapping the transcript to the binary indicator $\mathbf{1}\{\mathcal{E}\}$ is a measurable transformation. By the data-processing inequality,

$$D_{\mathrm{KL}}\left(\mathbb{P}_\theta^{(n)} \,\middle\|\, \mathbb{P}_{\theta'}^{(n)}\right) \geq \mathrm{kl}\left(\mathbb{P}_\theta^{(n)}(\mathcal{E}), \mathbb{P}_{\theta'}^{(n)}(\mathcal{E})\right),$$

because under the two instances the indicator $\mathbf{1}\{\mathcal{E}\}$ is Bernoulli with parameters $\mathbb{P}_\theta^{(n)}(\mathcal{E})$ and $\mathbb{P}_{\theta'}^{(n)}(\mathcal{E})$, respectively. Combining this inequality with (2) proves (3). □

The next technical lemma are elementary bounds on binary relative entropy, and are standard; see, e.g., Lattimore & Szepesvári (2020, Lemma 10.2) for its basic monotonicity properties and Kaufmann et al. (2016) for their use in fixed-confidence bandit lower bounds.

Lemma 2 (Binary relative-entropy bounds)**.** Let $\alpha \in (0, 1/2)$. If $p \geq 1 - \alpha$ and $q \leq \alpha$, then

$$\mathrm{kl}(p, q) \geq \mathrm{kl}(1 - \alpha, \alpha). \tag{5}$$

If, in addition, $\alpha \leq 1/4$, then

$$\mathrm{kl}(1 - \alpha, \alpha) \geq \log(3/(4\alpha))/2. \tag{6}$$

Finally, for every $\epsilon \in (0, 1/8]$, the estimate holds:

$$\mathrm{kl}\left(1/2 - 2\epsilon, 1/2 + 2\epsilon\right) \leq 128\epsilon^2/3. \tag{7}$$

Proof. First suppose $p, q \in (0, 1)$. Directly differentiating $\mathrm{kl}(p, q)$ yields

$$\frac{\partial}{\partial p}\,\mathrm{kl}(p, q) = \log\frac{p(1-q)}{q(1-p)} \quad \text{and} \quad \frac{\partial}{\partial q}\,\mathrm{kl}(p, q) = \frac{q-p}{q(1-q)}.$$

On the region $p > q$, the first derivative is positive and the second derivative is negative. Since $p \geq 1 - \alpha > \alpha \geq q$, the minimum over interior points satisfying $p \geq 1 - \alpha$ and $q \leq \alpha$ is attained at $(p, q) = (1 - \alpha, \alpha)$. If $p = 1$ or $q = 0$, the same inequality follows by taking limits from the interior; in the case $q = 0 < p$, the left-hand side is $+\infty$. This proves (5) for all $p, q \in [0, 1]$ covered by the statement.

A direct calculation yields $\mathrm{kl}(1 - \alpha, \alpha) = (1 - 2\alpha)\log\frac{1-\alpha}{\alpha}$. If $\alpha \leq 1/4$, then $1 - 2\alpha \geq \frac{1}{2}$ and $1 - \alpha \geq \frac{3}{4}$. Therefore

$$\mathrm{kl}(1 - \alpha, \alpha) \geq \frac{1}{2}\log\frac{3}{4\alpha},$$

proving (6).

Set $x := 4\epsilon$. Since $\epsilon \leq 1/8$, we have $x \in (0, 1/2]$. Using

$$1 - (1/2 - 2\epsilon) = 1/2 + 2\epsilon \quad \text{and} \quad 1 - (1/2 + 2\epsilon) = 1/2 - 2\epsilon,$$

we obtain

$$\mathrm{kl}\left(1/2 - 2\epsilon, 1/2 + 2\epsilon\right) = 4\epsilon\log\frac{1 + 4\epsilon}{1 - 4\epsilon} = x\log\frac{1 + x}{1 - x}.$$

For $x \in [0, 1)$, define

$$g(x) := 2x/(1 - x^2) - \log((1 + x)/(1 - x)).$$

Then $g(0) = 0$ and $g'(x) = 4x^2/(1 - x^2)^2 \geq 0$. Hence

$$\log((1 + x)/(1 - x)) \leq 2x/(1 - x^2).$$

It follows that

$$x\log((1 + x)/(1 - x)) \leq 2x^2/(1 - x^2) \leq 8x^2/3 = 128\epsilon^2/3,$$

where the second inequality uses $x^2 \leq 1/4$. This proves (7). □

E.2 Main result on online flow-reversal abstraction sample complexity

With the technical lemmas in hand, we now state the main sample complexity separation result.

Theorem 4 (Online flow-reversed action anchor abstraction sample-complexity)**.** Fix integers $1 \leq C < N$, accuracy $\epsilon \in (0, 1/8]$, and failure probability $\alpha \in (0, 1/4]$.

a. **Upper bound for an online flow-reversal abstraction.** Consider the class

$$\mathfrak{M}_F(\epsilon/2) := \{\mathcal{M}_{\boldsymbol{\mu}} : \Delta_F(\boldsymbol{\mu}) \leq \epsilon/2\}.$$

There exists an flow-reversal-restricted learner that uses no prior knowledge and is (ϵ, α)-correct on $\mathfrak{M}_F(\epsilon/2)$ after

$$n_{\mathrm{FR}} := C\left\lceil \frac{8}{\epsilon^2}\log\frac{2C}{\alpha}\right\rceil \quad \text{episodes.}$$

b. **Lower bound for retaining the full latent class.** Define a base instance $\mathcal{M}_{\boldsymbol{\mu}^{(0)}}$ by

$$\mu_1^{(0)} := 1/2, \qquad \mu_i^{(0)} := 1/2 - 2\epsilon, \quad i \in [N] \setminus \{1\}. \tag{8}$$

For each $j \in [N] \setminus \mathcal{C}$, define an alternative instance $\mathcal{M}_{\boldsymbol{\mu}^{(j)}}$ by

$$\mu_i^{(j)} := \begin{cases} 1/2 + 2\epsilon, & i = j, \\ \mu_i^{(0)}, & i \neq j. \end{cases} \tag{9}$$

Let $\mathfrak{M}_{\text{full}} := \left\{\mathcal{M}_{\boldsymbol{\mu}^{(0)}}\right\} \cup \left\{\mathcal{M}_{\boldsymbol{\mu}^{(j)}} : j \in [N] \setminus \mathcal{C}\right\}$. Consider any full-class learner that is given the exact flow-reversed-mode laws $W_F(\boldsymbol{\mu})$ before online interaction and is (ϵ, α)-correct on $\mathfrak{M}_{\text{full}}$ after n online episodes. Then, on the zero-bias base instance $\mathcal{M}_{\boldsymbol{\mu}^{(0)}}$, every non-flow-reversed mode satisfies

$$\mathbb{E}_{\boldsymbol{\mu}^{(0)}}[N_j(n)] \geq \mathrm{kl}(1-\alpha, \alpha) / \mathrm{kl}\left(1/2 - 2\epsilon, 1/2 + 2\epsilon\right), \qquad j \in [N] \setminus \mathcal{C}, \tag{10}$$

so that

$$n \geq \frac{3(N-C)}{256\epsilon^2} \log \frac{3}{4\alpha}. \tag{11}$$

c. **Zero-bias separation.** The base instance satisfies $\Delta_F(\boldsymbol{\mu}^{(0)}) = 0$. Thus the lower bound in part *b.* holds on an instance for which restricting the action class to the online flow-reversal abstraction incurs no control bias. The additional samples are required solely because a learner that is correct on $\mathfrak{M}_{\text{full}}$ must rule out every alternative in which a non-flow-reversed mode is superior.

Proof. We prove the three parts in order.

Proof of part *a*. Let $m := \left\lceil \frac{8}{\epsilon^2} \log \frac{2C}{\alpha} \right\rceil$. The learner samples each flow-reversed mode $c \in \mathcal{C}$ exactly m times. Let $\widehat{\mu}_c := \frac{1}{m} \sum_{\ell=1}^{m} R_{c,\ell}$ be the empirical mean of the m rewards obtained from flow-reversed mode c. The learner returns $\widehat{I} \in \arg\max_{c \in \mathcal{C}} \widehat{\mu}_c$, with any fixed deterministic tie-breaking rule.

Define the event

$$\mathcal{E} := \bigcap_{c \in \mathcal{C}} \left\{|\widehat{\mu}_c - \mu_c| \leq \epsilon/4\right\}.$$

By Höeffding's inequality, for each $c \in \mathcal{C}$, we have that

$$\mathbb{P}_{\boldsymbol{\mu}}\left(|\widehat{\mu}_c - \mu_c| > \epsilon/4\right) \leq 2\exp\left(-2m\epsilon^2/16\right) = 2\exp\left(-m\epsilon^2/8\right).$$

A union bound and the definition of m yield

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}^c) \leq 2C \exp\left(-m\epsilon^2/8\right) \leq \alpha.$$

Fix $c_F^\star \in \arg\max_{c \in \mathcal{C}} \mu_c$. On the event $\mathcal{E}$, we have that

$$\mu_{\widehat{I}} \geq \widehat{\mu}_{\widehat{I}} - \frac{\epsilon}{4} \geq \widehat{\mu}_{c_F^\star} - \frac{\epsilon}{4} \geq \mu_{c_F^\star} - \frac{\epsilon}{2} = \mu_F^\star(\boldsymbol{\mu}) - \frac{\epsilon}{2}.$$

Therefore, on $\mathcal{E}$, we have that

$$\mu^\star(\boldsymbol{\mu}) - \mu_{\widehat{I}} \leq \mu^\star(\boldsymbol{\mu}) - \mu_F^\star(\boldsymbol{\mu}) + \frac{\epsilon}{2} = \Delta_F(\boldsymbol{\mu}) + \frac{\epsilon}{2} \leq \epsilon,$$

where the final inequality uses $\Delta_F(\boldsymbol{\mu}) \leq \epsilon/2$. Since $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) \geq 1 - \alpha$, the learner is (ϵ, α)-correct on $\mathfrak{M}_F(\epsilon/2)$. It uses exactly $Cm = n_{\text{FR}}$ online episodes.

Proof of part *b*. Since $\epsilon \leq 1/8$, all means in (8)–(9) belong to $[0, 1]$. Moreover, all instances in $\mathfrak{M}_{\text{full}}$ agree on the flow-reversed modes—namely, we have that

$$W_F(\boldsymbol{\mu}^{(0)}) = W_F(\boldsymbol{\mu}^{(j)}) \qquad \text{for every } j \in [N] \setminus \mathcal{C}.$$

Thus the exact flow-reversal side information is identical under the base and every alternative.

On the base instance, mode 1 has mean $1/2$, while every other mode has mean $1/2 - 2\epsilon$. Hence mode 1 is the unique ϵ-optimal mode, since every other mode has suboptimality gap $2\epsilon > \epsilon$. Since the learner is (ϵ, α)-correct, we have that

$$\mathbb{P}_{\boldsymbol{\mu}^{(0)}}\left(\widehat{I}_n = 1\right) \geq 1 - \alpha. \tag{12}$$

Fix any $j \in [N] \setminus \mathcal{C}$. On the alternative instance $\mathcal{M}_{\boldsymbol{\mu}^{(j)}}$, mode j has mean $1/2 + 2\epsilon$, mode 1 has mean $1/2$, and every remaining mode has mean $1/2 - 2\epsilon$. Thus mode j is the unique ϵ-optimal mode. Correctness on $\mathcal{M}_{\boldsymbol{\mu}^{(j)}}$ therefore implies $\mathbb{P}_{\boldsymbol{\mu}^{(j)}}(\widehat{I}_n = j) \geq 1 - \alpha$, and hence

$$\mathbb{P}_{\boldsymbol{\mu}^{(j)}}\left(\widehat{I}_n = 1\right) \leq \alpha. \tag{13}$$

Now, we apply Lemma 1 with $\boldsymbol{\mu} = \boldsymbol{\mu}^{(0)}$, $\boldsymbol{\lambda} = \boldsymbol{\mu}^{(j)}$, and $\mathcal{E} = \{\widehat{I}_n = 1\}$. The two mean vectors differ only at coordinate j. Therefore

$$\mathbb{E}_{\boldsymbol{\mu}^{(0)}}[N_j(n)]\,\mathrm{kl}\left(\frac{1}{2} - 2\epsilon, \frac{1}{2} + 2\epsilon\right) \geq \mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}^{(0)}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\mu}^{(j)}}(\mathcal{E})\right).$$

Combining expressions (12) and (13) with Lemma 2 yields

$$\mathrm{kl}\left(\mathbb{P}_{\boldsymbol{\mu}^{(0)}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\mu}^{(j)}}(\mathcal{E})\right) \geq \mathrm{kl}(1 - \alpha, \alpha).$$

Dividing by the positive Bernoulli relative entropy proves (10).

Summing (10) over the $N - C$ indices in $[N] \setminus \mathcal{C}$ yields

$$\mathbb{E}_{\boldsymbol{\mu}^{(0)}}\left[\textstyle\sum_{j \in [N]\setminus\mathcal{C}} N_j(n)\right] \geq (N - C)\,\mathrm{kl}(1 - \alpha, \alpha)/\,\mathrm{kl}(1/2 - 2\epsilon, 1/2 + 2\epsilon)$$

Applying (7), we further deduce that

$$\mathbb{E}_{\boldsymbol{\mu}^{(0)}}\left[\textstyle\sum_{j \in [N]\setminus\mathcal{C}} N_j(n)\right] \geq \frac{3(N - C)}{128\epsilon^2}\,\mathrm{kl}(1 - \alpha, \alpha) \geq \frac{3(N - C)}{256\epsilon^2}\log\frac{3}{4\alpha} \tag{14}$$

where the last inequality follows by applying (6).

Finally, on every realized transcript, we have that $\sum_{j \in [N]\setminus\mathcal{C}} N_j(n) \leq \sum_{i=1}^{N} N_i(n) = n$. Taking expectations under $\boldsymbol{\mu}^{(0)}$ and combining with (14) proves (11).

Proof of part *c*. On the base instance, mode 1 belongs to $\mathcal{C}$ and has mean $1/2$, while every other mode has mean $1/2 - 2\epsilon$. Consequently, we have that $\mu^\star(\boldsymbol{\mu}^{(0)}) = \mu_F^\star(\boldsymbol{\mu}^{(0)}) = \frac{1}{2}$. By (1), we have that $\Delta_F(\boldsymbol{\mu}^{(0)}) = 0$. The lower bound in part *b.* therefore holds on a zero-bias instance. Each alternative $\boldsymbol{\mu}^{(j)}$ agrees with the base instance on every flow-reversed mode and differs only on the non-flow-reversed mode j. Hence neither the exact flow-reversed-mode laws nor any prefill consisting only of flow-reversed-mode samples can distinguish the base instance from alternative j. A learner that is correct on the full family must obtain the information required to rule out that alternative through online samples of mode j. This is precisely the per-mode requirement in (10). □

Corollary 1 (Full-class learning with bank-restricted prefill)**.** Consider any learner whose online action class contains all N latent modes and whose prefill, if present, is generated using only interactions with modes in $\mathcal{C}$. If the learner is (ϵ, α)-correct on $\mathfrak{M}_{\mathrm{full}}$, then the lower bounds (10)–(11) apply.

This result covers both residual RL without prefill and learning initialized with bank-restricted data; pure DSRL and FRS-prefill followed by DSRL are respective examples (Wagenmaker et al., 2025; Tang et al., 2026).

Proof. Suppose, for contradiction, that such a learner violates one of the lower bounds. Construct a learner that is given the exact flow-reversed-mode laws $W_F(\boldsymbol{\mu})$. Using independent internal randomness and these exact Bernoulli laws, it first simulates the original learner's entire prefill-generation procedure, including any adaptive choices among modes in $\mathcal{C}$, and then runs the original learner's online procedure on the resulting simulated prefill. The simulated prefill has exactly the same distribution as the original prefill under every instance in $\mathfrak{M}_{\mathrm{full}}$. The constructed learner therefore has the same correctness probability and the same distribution of online sample counts as the original learner. This contradicts Theorem 4*b.*, which applies to every learner given the exact flow-reversed-mode laws. Hence the stated lower bounds also apply to the original learner. □

This is a worst-case result for latent-mode problems with no assumed smoothness or generalization between modes. It does not imply that every practical DSRL run must explicitly test $N - C$ additional modes. Rather, it shows that FRS-only prefill cannot, without further structural assumptions, guarantee removal of the exploration cost associated with retaining the full latent action class.

Remark 1 (What the separation establishes)**.** Theorem 4 is a nested-model or adaptation result. The flow-reversal-restricted learner commits to the hypothesis that the online flow-reversal action abstraction has bias at most $\epsilon/2$. The full-class learner retains a correctness guarantee on instances in which a non-flow-reversed latent mode is better. The lower bound shows that retaining this broader guarantee has a statistical cost even on the base instance, where the flow-reversal abstraction is exactly sufficient. A full-class learner can avoid that cost only by declining to rule out the non-flow-reversed alternatives, in which case it no longer has the stated correctness guarantee on $\mathfrak{M}_{\text{full}}$.

Remark 2 (Why buffer prefill with flow-reversed action anchors cannot remove the lower bound)**.** The lower bound grants the learner the exact Bernoulli reward law of every flow-reversed mode before any online interaction. This is more information than any finite flow-reversal prefill generated solely by those modes. Since the base and alternative instances have identical reward laws on all flow-reversed modes, no amount of such prefill changes the likelihood ratio between them. The lower bound therefore applies, *a fortiori*, to FRS-prefill followed by DSRL Tang et al. (2026).

Remark 3 (Relation to continuous latent spaces)**.** The finite latent set $\{z_1, \dots, z_N\}$ may be embedded in any continuous latent space $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ containing N distinct points. More generally, each index may represent a measurable latent region whose elements induce the same reward law. The theorem therefore models a latent space with N behaviorally distinguishable modes and makes no smoothness assumption linking the rewards of different modes. Since the construction is a one-state, one-decision episodic MDP, it is also a standard Bernoulli multi-armed bandit and can be inserted as a decision state in a longer-horizon MDP without changing the identification lower bound at that state.

F THE GENERAL-MDP PRICE OF THE PERSISTENT FRS ABSTRACTION

Theorem 4 isolates the statistical benefit of solving a smaller online latent-action problem. We now quantify the control price of that restriction in a general discounted latent MDP. The result separates two errors: an irreducible support error caused by the flow-reversal action anchor abstraction and a reducible selection error caused by finite-data learning within that abstraction.

F.1 GENERAL LATENT MDP AND THE PERSISTENT FLOW-REVERSAL ACTION CLASS

Let $\mathcal{M} = (\mathcal{S}, \mathcal{Z}, P, r, \gamma)$ be a discounted latent-action MDP. Assume that $(\mathcal{S}, \Sigma_{\mathcal{S}})$ and $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$ are standard Borel spaces, that $\gamma \in [0, 1)$, that $r : \mathcal{S} \times \mathcal{Z} \to \mathbb{R}$ is bounded and measurable, and that $P(\cdot \mid s, z)$ is a stochastic kernel on $\mathcal{S}$ given $(s, z) \in \mathcal{S} \times \mathcal{Z}$.

For a bounded measurable function $V : \mathcal{S} \to \mathbb{R}$, define

$$Q_V(s, z) := r(s, z) + \gamma \int_{\mathcal{S}} V(s')P(\mathrm{d}s' \mid s, z),$$

and define the full latent-action Bellman optimality operator

$$(\mathcal{T}V)(s) := \sup_{z \in \mathcal{Z}} Q_V(s, z).$$

Assume that $\mathcal{T}$ maps bounded measurable functions to bounded measurable functions and that the full optimal value $V^\star$ is its unique bounded fixed point—namely, $V^\star = \mathcal{T}V^\star$. Define the corresponding full optimal action-value function by

$$Q^\star(s, z) := Q_{V^\star}(s, z) = r(s, z) + \gamma \int_{\mathcal{S}} V^\star(s')P(\mathrm{d}s' \mid s, z).$$

Recall that $\mathcal{C} = \{1, \dots, C\}$ is the finite set of coarse commands, and let $F : \mathcal{S} \times \mathcal{C} \to \mathcal{Z}$ be the measurable online flow-reversal map. As in the method definition, write

$$z_c(s) := F(s, c) = H_\theta^K\big(s, \varphi(s, c)\big).$$

Recall the state-dependent flow-reversed action bank is $\mathcal{Z}(s) := \{z_c(s) : c \in \mathcal{C}\}$.

For a bounded measurable V, define the flow-reversal-restricted Bellman optimality operator

$$(\mathcal{T}_F V)(s) := \max_{c \in \mathcal{C}} Q_V\big(s, z_c(s)\big). \tag{15}$$

Since $\mathcal{C}$ is finite, the maximum in (15) is attained at every state.

A stationary flow-reversal selector is a measurable Markov kernel $\omega(\cdot \mid s)$ on $\mathcal{C}$. The selector draws a command $c \sim \omega(\cdot \mid s)$ and executes the actual latent $z_c(s)$. Define its reward by

$$r^\omega(s) := \sum_{c \in \mathcal{C}} \omega(c \mid s) r\big(s, z_c(s)\big),$$

and transition kernel by

$$P^\omega(B \mid s) := \sum_{c \in \mathcal{C}} \omega(c \mid s) P\big(B \mid s, z_c(s)\big), \quad \text{where} \quad B \in \Sigma_{\mathcal{S}}.$$

For a bounded measurable V, write

$$(P^\omega V)(s) := \int_{\mathcal{S}} V(s') P^\omega(\mathrm{d}s' \mid s).$$

For an initial state s, let $\mathbb{E}_s^\omega[\cdot]$ denote expectation under the Markov chain $(S_t)_{t \geq 0}$ satisfying $S_{t+1} \sim P^\omega(\cdot \mid S_t)$ with $S_0 = s$. Define the policy Bellman operator

$$(\mathcal{T}^\omega V)(s) := r^\omega(s) + \gamma (P^\omega V)(s) = \sum_{c \in \mathcal{C}} \omega(c \mid s) Q_V\big(s, z_c(s)\big).$$

The value V^ω is the unique bounded fixed point of $\mathcal{T}^\omega$.

Lemma 3 (Restricted Bellman fixed point and optimal stationary selector)**.** The operator $\mathcal{T}_F$ and, for every stationary flow-reversal selector ω, the operator $\mathcal{T}^\omega$ are monotone γ-contractions in the supremum norm. Hence $\mathcal{T}_F$ has a unique bounded fixed point $V_F^\star$, and $\mathcal{T}^\omega$ has a unique bounded fixed point V^ω—namely,

$$V_F^\star = \mathcal{T}_F V_F^\star, \qquad \text{and} \qquad V^\omega = \mathcal{T}^\omega V^\omega.$$

There exists a measurable deterministic stationary flow-reversal selector $\omega_F^\star$ such that $V^{\omega_F^\star} = V_F^\star$. Moreover, for every stationary flow-reversal selector ω, the bounds hold:

$$V^\omega(s) \leq V_F^\star(s) \leq V^\star(s) \qquad \text{for every } s \in \mathcal{S}.$$

Proof. Let V, W be bounded measurable functions. For each $s \in \mathcal{S}$, we have that

$$\begin{aligned} |(\mathcal{T}_F V)(s) - (\mathcal{T}_F W)(s)| &\leq \max_{c \in \mathcal{C}} \big| Q_V\big(s, z_c(s)\big) - Q_W\big(s, z_c(s)\big) \big| \\ &= \gamma \max_{c \in \mathcal{C}} \left| \int_{\mathcal{S}} \big(V(s') - W(s')\big) P(\mathrm{d}s' \mid s, z_c(s)) \right| \\ &\leq \gamma \|V - W\|_\infty. \end{aligned}$$

Taking the supremum over s proves

$$\|\mathcal{T}_F V - \mathcal{T}_F W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Thus $\mathcal{T}_F$ is a contraction. It is monotone because $V \leq W$ pointwise implies $Q_V(s, z) \leq Q_W(s, z)$ for every (s, z), and taking a maximum over c preserves this inequality. The contraction mapping theorem gives a unique bounded fixed point $V_F^\star$.

Fix a stationary flow-reversal selector ω. For bounded measurable V and W, and every state s,

$$|(\mathcal{T}^\omega V)(s) - (\mathcal{T}^\omega W)(s)| \leq \sum_{c \in \mathcal{C}} \omega(c \mid s) \big| Q_V\big(s, z_c(s)\big) - Q_W\big(s, z_c(s)\big) \big| \leq \gamma \|V - W\|_\infty.$$

Taking the supremum over s shows that $\mathcal{T}^\omega$ is a γ-contraction. The same pointwise-order argument used for $\mathcal{T}_F$ shows that $\mathcal{T}^\omega$ is monotone. The contraction mapping theorem therefore gives a unique bounded fixed point V^ω.

For every bounded V and state s, we have that

$$(\mathcal{T}^{\omega}V)(s) = \sum_{c\in\mathcal{C}} \omega(c \mid s) Q_V\big(s, z_c(s)\big) \le \max_{c\in\mathcal{C}} Q_V\big(s, z_c(s)\big) = (\mathcal{T}_F V)(s).$$

In particular, we have that $V^{\omega} = \mathcal{T}^{\omega}V^{\omega} \le \mathcal{T}_F V^{\omega}$. Iterating the monotone operator $\mathcal{T}_F$ yields

$$V^{\omega} \le \mathcal{T}_F V^{\omega} \le \mathcal{T}_F^2 V^{\omega} \le \cdots.$$

Since $\mathcal{T}_F$ is a contraction, the sequence $\mathcal{T}_F^k V^{\omega}$ converges in supremum norm to its unique fixed point $V_F^{\star}$. Passing to the limit yields $V^{\omega} \le V_F^{\star}$.

Define

$$c_F^{\star}(s) := \min \arg\max_{c\in\mathcal{C}} Q_{V_F^{\star}}\big(s, z_c(s)\big).$$

For each $c \in \mathcal{C}$, set $q_c^F(s) := Q_{V_F^{\star}}\big(s, z_c(s)\big)$. Every function q_c^F is measurable. For a fixed $c \in \mathcal{C}$,

$$\{s : c_F^{\star}(s) = c\} = \Big(\bigcap_{j\in\mathcal{C}} \{s : q_c^F(s) \ge q_j^F(s)\}\Big) \cap \Big(\bigcap_{\ell<c} \{s : q_\ell^F(s) < q_c^F(s)\}\Big).$$

This is a finite intersection of measurable sets. Hence $c_F^{\star}$ is measurable and defines a deterministic stationary selector $\omega_F^{\star}$. By construction, we have that

$$(\mathcal{T}^{\omega_F^{\star}} V_F^{\star})(s) = \max_{c\in\mathcal{C}} Q_{V_F^{\star}}\big(s, z_c(s)\big) = (\mathcal{T}_F V_F^{\star})(s) = V_F^{\star}(s).$$

Therefore $V_F^{\star}$ is a fixed point of $\mathcal{T}^{\omega_F^{\star}}$. By uniqueness of the policy fixed point, we have that $V^{\omega_F^{\star}} = V_F^{\star}$.

Finally, for every bounded V, the bound $\mathcal{T}_F V \le \mathcal{T} V$ holds pointwise since the flow-reversed latent action bank is a subset of the full latent action space. Hence we deduce that

$$\mathcal{T}_F V^{\star} \le \mathcal{T} V^{\star} = V^{\star}.$$

Applying $\mathcal{T}_F$ repeatedly and using monotonicity gives

$$\mathcal{T}_F^k V^{\star} \le V^{\star} \qquad \text{for every } k \ge 1.$$

The left-hand side converges in supremum norm to $V_F^{\star}$, so $V_F^{\star} \le V^{\star}$. □

Define the optimal restricted action-value function on the actual flow-reversed latents by

$$Q_F^{\star}\big(s, z_c(s)\big) := r\big(s, z_c(s)\big) + \gamma \int_{\mathcal{S}} V_F^{\star}(s') P(\mathrm{d}s' \mid s, z_c(s)).$$

Then the optimal value is given by

$$V_F^{\star}(s) = \max_{c\in\mathcal{C}} Q_F^{\star}\big(s, z_c(s)\big).$$

F.2 Main Result: Support Gap and Selection Error

We now define some key quantities that help decompose the error.

Definition 1 (One-step flow-reversal support gap)**.** For each state $s \in \mathcal{S}$, define $\Delta_F(s) := V^{\star}(s) - \max_{c\in\mathcal{C}} Q^{\star}\big(s, z_c(s)\big)$, and define the worst-case support gap by $\bar{\Delta}_F := \sup_{s\in\mathcal{S}} \Delta_F(s)$.

Since $V^{\star}(s) = \sup_{z\in\mathcal{Z}} Q^{\star}(s, z)$ and every $z_c(s)$ belongs to $\mathcal{Z}$, we have that $\Delta_F(s) \ge 0$. The quantity $\Delta_F(s)$ forces only the current latent to lie in the latent bank generated via flow-reversal; the continuation value inside $Q^{\star}$ still assumes globally optimal future control over the full latent space.

Definition 2 (Restricted selection error)**.** For a stationary flow-reversal selector ω, define its pointwise selection error by

$$\mathrm{err}_{\mathrm{sel}}(\omega; s) := V_F^{\star}(s) - (\mathcal{T}^{\omega} V_F^{\star})(s) = V_F^{\star}(s) - \sum_{c\in\mathcal{C}} \omega(c \mid s) Q_F^{\star}\big(s, z_c(s)\big). \tag{16}$$

Its uniform selection error is

$$\mathrm{err}_{\mathrm{sel}}(\omega) := \sup_{s\in\mathcal{S}} \mathrm{err}_{\mathrm{sel}}(\omega; s). \tag{17}$$

Since $V_F^\star(s)$ is the maximum of $Q_F^\star(s, z_c(s))$ over $c \in \mathcal{C}$ and $\omega(\cdot \mid s)$ is a probability distribution, both quantities in (16)–(17) are nonnegative.

The following theorem characterizes the price of abstraction and finite-sample selection error.

Theorem 5 (General-MDP abstraction price and finite-sample selection error)**.** The following statements hold.

a. **Oracle price of the online flow-reversal abstraction.** For every state $s \in \mathcal{S}$, the bounds hold:

$$\Delta_F(s) \le V^\star(s) - V_F^\star(s) \le \bar{\Delta}_F/(1-\gamma). \tag{18}$$

b. **Exact decomposition for an arbitrary learned flow-reversal selector.** Fix any stationary flow-reversal selector ω. Then, as an identity between bounded functions, we have the decomposition

$$V^\star - V^\omega = \big(V^\star - V_{\mathrm{F}}^\star\big) + (I - \gamma P^\omega)^{-1}\,\mathrm{err}_{\mathrm{sel}}(\omega;\cdot). \tag{19}$$

Consequently, for every $s \in \mathcal{S}$, the bounds hold:

$$\Delta_F(s) \le V^\star(s) - V^\omega(s) \le \frac{\bar{\Delta}_F + \mathrm{err}_{\mathrm{sel}}(\omega)}{1-\gamma}. \tag{20}$$

c. **Finite-sample consequence.** Let $\widehat{\omega}_n$ be any random stationary flow-reversal selector returned after n online environment interactions. Suppose that, for some $\alpha \in (0,1)$ and deterministic number $\epsilon_{\mathrm{sel}}(n,\alpha) \ge 0$, the bound holds:

$$\mathbb{P}\left(\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n) \le \epsilon_{\mathrm{sel}}(n,\alpha)\right) \ge 1-\alpha. \tag{21}$$

Then, with probability at least $1-\alpha$, simultaneously for every $s \in \mathcal{S}$,

$$\Delta_F(s) \le V^\star(s) - V^{\widehat{\omega}_n}(s) \le (\bar{\Delta}_F + \epsilon_{\mathrm{sel}}(n,\alpha))/(1-\gamma). \tag{22}$$

Proof. We prove the three parts in order.

Proof of part *a.*: lower bound. By Lemma 3, we have that $V_F^\star \le V^\star$. The operator $\mathcal{T}_F$ is monotone, so that $\mathcal{T}_F V_F^\star \le \mathcal{T}_F V^\star$. Using the fixed-point equation for $V_F^\star$ and the definition of $\Delta_F(s)$, we deduce that

$$V_F^\star(s) = (\mathcal{T}_F V_F^\star)(s) \le (\mathcal{T}_F V^\star)(s) = \max_{c\in\mathcal{C}} Q^\star\big(s, z_c(s)\big) = V^\star(s) - \Delta_F(s).$$

Rearranging yields $\Delta_F(s) \le V^\star(s) - V_F^\star(s)$.

Proof of part *a.*: resolvent identity and upper bound. For each $c \in \mathcal{C}$, define the measurable function $q_c^\dagger(s) := Q^\star\big(s, z_c(s)\big)$. Moreover, let

$$c^\dagger(s) := \min\left\{ c \in \mathcal{C} \,:\, Q^\star\big(s, z_c(s)\big) = \max_{j\in\mathcal{C}} Q^\star(s, z_j(s)) \right\}.$$

The smallest-maximizer selector is measurable since, for every $c \in \mathcal{C}$, we have that

$$\{s : c^\dagger(s) = c\} = \left(\bigcap_{j\in\mathcal{C}} \{s : q_c^\dagger(s) \ge q_j^\dagger(s)\} \right) \cap \left(\bigcap_{\ell<c} \{s : q_\ell^\dagger(s) < q_c^\dagger(s)\} \right),$$

which is a finite intersection of measurable sets. Let $\omega^\dagger$ be the associated deterministic FRS selector that assigns probability one to $c^\dagger(s)$. By construction, we have that

$$(\mathcal{T}^{\omega^\dagger} V^\star)(s) = Q^\star\big(s, z_{c^\dagger(s)}(s)\big) = \max_{c\in\mathcal{C}} Q^\star\big(s, z_c(s)\big) = V^\star(s) - \Delta_F(s). \tag{23}$$

Define $D^\dagger := V^\star - V^{\omega^\dagger}$. Using the Bellman fixed-point equation $V^{\omega^\dagger} = \mathcal{T}^{\omega^\dagger} V^{\omega^\dagger}$, adding and subtracting $\mathcal{T}^{\omega^\dagger} V^\star$, and applying (23), we obtain, for every state s,

$$\begin{aligned} D^\dagger(s) &= V^\star(s) - (\mathcal{T}^{\omega^\dagger} V^{\omega^\dagger})(s) \\ &= V^\star(s) - (\mathcal{T}^{\omega^\dagger} V^\star)(s) + (\mathcal{T}^{\omega^\dagger} V^\star)(s) - (\mathcal{T}^{\omega^\dagger} V^{\omega^\dagger})(s) \\ &= \Delta_F(s) + \gamma\big(P^{\omega^\dagger} D^\dagger\big)(s). \end{aligned}$$

Hence, as an identity between bounded functions, we have that

$$(I - \gamma P^{\omega^\dagger}) D^\dagger = \Delta_F(\cdot). \tag{24}$$

Recall also that a Markov operator is nonexpansive in the supremum norm—i.e., $\|P^{\omega^\dagger} V\|_\infty \le \|V\|_\infty$. Therefore the Neumann series $\sum_{t=0}^{\infty} \gamma^t (P^{\omega^\dagger})^t$ converges in operator norm and is the inverse of $I - \gamma P^{\omega^\dagger}$. Applying this inverse to (24) yields

$$D^\dagger = (I - \gamma P^{\omega^\dagger})^{-1} \Delta_F(\cdot) = \sum_{t=0}^{\infty} \gamma^t (P^{\omega^\dagger})^t \Delta_F(\cdot).$$

For every initial state s, the t-th term satisfies $\big((P^{\omega^\dagger})^t \Delta_F(\cdot)\big)(s) = \mathbb{E}_s^{\omega^\dagger}[\Delta_F(S_t)]$. The support-gap function is bounded and nonnegative. Therefore the monotone convergence theorem permits interchange of the expectation and the increasing sequence of nonnegative partial sums. Since $\Delta_F(s) \le \bar{\Delta}_F$ for every state, we have that

$$D^\dagger(s) = \mathbb{E}_s^{\omega^\dagger}\left[\sum_{t=0}^{\infty} \gamma^t \Delta_F(S_t)\right] \le \sum_{t=0}^{\infty} \gamma^t \bar{\Delta}_F = \frac{\bar{\Delta}_F}{1-\gamma}.$$

The selector $\omega^\dagger$ is an admissible flow-reversal selector. By Lemma 3, the bound $V^{\omega^\dagger} \le V_F^\star$ holds, so that

$$V^\star(s) - V_F^\star(s) \le V^\star(s) - V^{\omega^\dagger}(s) = D^\dagger(s).$$

Combining this inequality with the preceding occupancy and uniform bounds proves (33) and the upper bound in (18).

Proof of part *b*. Fix an arbitrary stationary flow-reversal selector ω, and define $D_F^\omega := V_F^\star - V^\omega$. Using the Bellman fixed-point equation $V^\omega = \mathcal{T}^\omega V^\omega$, adding and subtracting $\mathcal{T}^\omega V_F^\star$, and using Definition 2, we obtain, for every state s, we deduce that

$$\begin{aligned} D_F^\omega(s) &= V_F^\star(s) - (\mathcal{T}^\omega V^\omega)(s) \\ &= V_F^\star(s) - (\mathcal{T}^\omega V_F^\star)(s) + (\mathcal{T}^\omega V_F^\star)(s) - (\mathcal{T}^\omega V^\omega)(s) \\ &= \mathrm{err}_{\mathrm{sel}}(\omega; s) + \gamma\big(P^\omega D_F^\omega\big)(s). \end{aligned}$$

Hence $(I - \gamma P^\omega) D_F^\omega = \mathrm{err}_{\mathrm{sel}}(\omega; \cdot)$. As above, the Neumann series yields

$$D_F^\omega = (I - \gamma P^\omega)^{-1} \mathrm{err}_{\mathrm{sel}}(\omega; \cdot) = \sum_{t=0}^{\infty} \gamma^t (P^\omega)^t \mathrm{err}_{\mathrm{sel}}(\omega; \cdot). \tag{25}$$

The selection-error function is bounded and nonnegative. Thus the monotone convergence theorem again permits interchange of expectation and the nonnegative discounted series. Therefore, for every initial state s, we have that

$$V_F^\star(s) - V^\omega(s) = \mathbb{E}_s^\omega\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{err}_{\mathrm{sel}}(\omega; S_t)\right] \le \mathrm{err}_{\mathrm{sel}}(\omega)/(1-\gamma). \tag{26}$$

The algebraic identity $V^\star - V^\omega = \big(V^\star - V_F^\star\big) + \big(V_F^\star - V^\omega\big)$ combined with (25) proves (19). Its occupancy form—namely, for every $s \in \mathcal{S}$, the form $V^\star(s) - V^\omega(s) = V^\star(s) - V_F^\star(s) + \mathbb{E}_s^\omega\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{err}_{\mathrm{sel}}(\omega; S_t)\right]$—follows from the first equality in (26).

Since $V^\omega \le V_F^\star$, part *a.* implies that $V^\star(s) - V^\omega(s) \ge V^\star(s) - V_F^\star(s) \ge \Delta_F(s)$. For the upper bound, part *a.* and (26) together yield

$$V^\star(s) - V^\omega(s) = V^\star(s) - V_F^\star(s) + V_F^\star(s) - V^\omega(s) \le \frac{\bar{\Delta}_F + \mathrm{err}_{\mathrm{sel}}(\omega)}{1-\gamma}.$$

This proves (20).

Proof of part *c*. Define the event $\mathcal{E}_n := \{\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n) \le \epsilon_{\mathrm{sel}}(n, \alpha)\}$. By assumption, we have $\mathbb{P}(\mathcal{E}_n) \ge 1 - \alpha$. On every outcome in $\mathcal{E}_n$, the realized selector $\widehat{\omega}_n$ is a stationary flow-reversal selector satisfying $\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n) \le \epsilon_{\mathrm{sel}}(n, \alpha)$. Applying the deterministic bound (20) to that realized selector yields, simultaneously for every state s, that

$$\Delta_F(s) \le V^\star(s) - V^{\widehat{\omega}_n}(s) \le (\bar{\Delta}_F + \epsilon_{\mathrm{sel}}(n, \alpha))/(1-\gamma).$$

Since this statement holds on an event of probability at least $1 - \alpha$, it proves (22). □

F.3 From critic accuracy to selection error

Theorem 5 does not assume a particular critic objective or candidate-selection rule. The following corollary shows how a conventional finite-sample critic bound can be inserted into the abstract selection term.

In particular, Theorem 5 has an abstract term $\mathrm{err}_{\mathrm{sel}}$; the next corollary shows what that abstract error means for an ordinary critic-based implementation. Indeed, it shows

$$\text{selection error} \le 2 \cdot \text{critic error} + \text{failure to maximize the critic}.$$

The last sentence is simply identifying the common special case where the final term vanishes.

Corollary 2 (Approximate critic and approximate candidate selection)**.** Let $\widehat{Q}_n$ be any random score defined on the state-dependent latent bank generated via flow-reversal, and let $\widehat{\omega}_n$ be any random stationary flow-reversal selector. Suppose that, with probability at least $1-\alpha$, the following two inequalities hold simultaneously:

$$\sup_{s\in\mathcal{S}} \max_{c\in\mathcal{C}} \left|\widehat{Q}_n\big(s, z_c(s)\big) - Q_F^\star\big(s, z_c(s)\big)\right| \le \epsilon_Q(n,\alpha), \tag{27}$$

and, for every $s \in \mathcal{S}$,

$$\sum_{c\in\mathcal{C}} \widehat{\omega}_n(c \mid s)\widehat{Q}_n\big(s, z_c(s)\big) \ge \max_{c\in\mathcal{C}} \widehat{Q}_n\big(s, z_c(s)\big) - \epsilon_{\mathrm{opt}}(n,\alpha). \tag{28}$$

Then, on the same event,

$$\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n) \le 2\epsilon_Q(n,\alpha) + \epsilon_{\mathrm{opt}}(n,\alpha). \tag{29}$$

Consequently, with probability at least $1-\alpha$, simultaneously for every state s,

$$\Delta_F(s) \le V^\star(s) - V^{\widehat{\omega}_n}(s) \le \frac{\bar{\Delta}_F + 2\epsilon_Q(n,\alpha) + \epsilon_{\mathrm{opt}}(n,\alpha)}{1-\gamma}. \tag{30}$$

In particular, exact greedy selection with respect to $\widehat{Q}_n$ corresponds to $\epsilon_{\mathrm{opt}}(n,\alpha) = 0$.

Proof. Work on an outcome on which (27) and (28) both hold. Fix an arbitrary state $s \in \mathcal{S}$. For brevity, write

$$q_c := Q_F^\star\big(s, z_c(s)\big) \qquad \text{and} \qquad \widehat{q}_c := \widehat{Q}_n\big(s, z_c(s)\big).$$

By (27), we have that

$$q_c \le \widehat{q}_c + \epsilon_Q(n,\alpha) \qquad \text{for every } c \in \mathcal{C}.$$

Taking the maximum over c gives us that

$$\max_{c\in\mathcal{C}} q_c \le \max_{c\in\mathcal{C}} \widehat{q}_c + \epsilon_Q(n,\alpha). \tag{31}$$

By (28), we have that

$$\max_{c\in\mathcal{C}} \widehat{q}_c \le \sum_{c\in\mathcal{C}} \widehat{\omega}_n(c \mid s)\widehat{q}_c + \epsilon_{\mathrm{opt}}(n,\alpha).$$

The other direction of the uniform critic bound yields $\widehat{q}_c \le q_c + \epsilon_Q(n,\alpha)$ for every $c \in \mathcal{C}$. Averaging with respect to $\widehat{\omega}_n(\cdot \mid s)$ then gives us that

$$\sum_{c\in\mathcal{C}} \widehat{\omega}_n(c \mid s)\widehat{q}_c \le \sum_{c\in\mathcal{C}} \widehat{\omega}_n(c \mid s)q_c + \epsilon_Q(n,\alpha). \tag{32}$$

Combining (31)–(32) yields

$$\max_{c\in\mathcal{C}} q_c \le \sum_{c\in\mathcal{C}} \widehat{\omega}_n(c \mid s)q_c + 2\epsilon_Q(n,\alpha) + \epsilon_{\mathrm{opt}}(n,\alpha).$$

Since $\max_{c\in\mathcal{C}} q_c = V_F^\star(s)$, rearranging the above inequality proves

$$\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n; s) \le 2\epsilon_Q(n,\alpha) + \epsilon_{\mathrm{opt}}(n,\alpha).$$

The state s is arbitrary, so taking the supremum over s proves (29). Applying Theorem 5.*c.* with $\epsilon_{\mathrm{sel}}(n,\alpha) = 2\epsilon_Q(n,\alpha) + \epsilon_{\mathrm{opt}}(n,\alpha)$ proves (30). □

The corollary is basically a bridge from *my critic is accurate* to *my learned* ANCHORQ *policy has good control performance.*

It says the total error comes from:

$$\bar{\Delta}_{\mathrm{F}} + 2\epsilon_Q(n, \alpha) + \epsilon_{\text{opt}}(n, \alpha),$$

where the first term captures that flow-reversing action anchors does not contain the right behavior, the second term captures that critic has not learned the available behaviors accurately, and the last term captures that we do not fully exploit the learned critic. Those are three very different failure modes. The first is a structural error. More data does not fix this. If the right behavior is missing from the latent bank generated by flow-reversing anchors, even a perfect critic and perfect selector cannot recover it. The second is a statistical learning error. More/better data can reduce this. The right candidate is available, but the critic cannot yet tell which candidate is good. And, the last term is the candidate-selection error. This depends on how aggressively you optimize the learned scores. Exact greedy selection removes it completely: $\epsilon_{\text{opt}} = 0$.

F.4 SHARPNESS AND REGIME CONSEQUENCES

Theorem 5 bounds the loss of the flow-reversal-restricted policy using a one-step support gap and a one-step selection error. A natural question is whether the resulting factors are merely artifacts of the Bellman argument. The following proposition shows that they are not: the one-step support gap can equal the entire control loss in a one-decision problem, and both support error and selection error can accumulate over the full effective discounted horizon $1/(1-\gamma)$.

Proposition 1 (Worst-case tightness of the abstraction and selection bounds)**.** Each of the three bounds appearing in Theorem 5 is attained by some discounted latent MDP:

a. **Tightness of the one-step lower bound.** For every $d > 0$, there exists a one-decision latent MDP and an flow-reversal action set for which

$$\Delta_F(s_0) = V^\star(s_0) - V_F^\star(s_0) = d.$$

b. **Tightness of support-error accumulation.** For every $\gamma \in [0, 1)$ and $d > 0$, there exists a one-state discounted latent MDP and an flow-reversal action set for which

$$\Delta_F(s) = d, \qquad V^\star(s) - V_F^\star(s) = \frac{d}{1-\gamma}.$$

c. **Tightness of selection-error accumulation.** For every $\gamma \in [0, 1)$ and $\xi > 0$, there exists a one-state flow-reversal-restricted latent MDP and an flow-reversal selector ω for which

$$\mathrm{err}_{\mathrm{sel}}(\omega; s) = \xi, \qquad V_F^\star(s) - V^\omega(s) = \frac{\xi}{1-\gamma}.$$

Proof. We prove each part separately.

Proof of part *a*. Consider an MDP with one decision state s_0 and one terminal state. At s_0, the full latent action set contains two latents, $z^\star$ and z_F. Executing either latent transitions immediately to the terminal state. Set

$$r(s_0, z^\star) = d \qquad \text{and} \qquad r(s_0, z_F) = 0,$$

and let the latent bank generated via flow-reversal contain only z_F. Then

$$V^\star(s_0) = d, \qquad V_F^\star(s_0) = 0, \qquad \text{and } \; Q^\star(s_0, z_F) = 0.$$

Therefore

$$\Delta_F(s_0) = d - 0 = d = V^\star(s_0) - V_F^\star(s_0).$$

Proof of part *b*. Consider an MDP with a single nonterminal state s. Both available latents return deterministically to s. Let the full action set contain $z^\star$ and z_F, set

$$r(s, z^\star) = d \qquad \text{and} \qquad r(s, z_F) = 0,$$

and let the flow-reversed latent bank contain only z_F. The full optimal policy repeatedly executes $z^\star$, so

$$V^\star(s) = d + \gamma V^\star(s) = \frac{d}{1-\gamma}.$$

The restricted policy repeatedly executes z_F, so

$$V_F^\star(s) = 0 + \gamma V_F^\star(s) = 0.$$

When the current latent is forced to be z_F but the continuation is valued by the full optimum,

$$Q^\star(s, z_F) = 0 + \gamma V^\star(s) = \frac{\gamma d}{1-\gamma}.$$

Hence

$$\Delta_F(s) = V^\star(s) - Q^\star(s, z_F) = \frac{d}{1-\gamma} - \frac{\gamma d}{1-\gamma} = d,$$

while

$$V^\star(s) - V_F^\star(s) = \frac{d}{1-\gamma}.$$

Proof of part *c*. Consider a one-state restricted MDP whose flow-reversed bank contains two latents z_1 and z_2. Both return deterministically to the same state s. Set

$$r(s, z_1) = \xi \qquad \text{and} \qquad r(s, z_2) = 0.$$

The restricted optimal policy repeatedly executes z_1, and hence

$$V_F^\star(s) = \xi + \gamma V_F^\star(s) = \frac{\xi}{1-\gamma}.$$

Let ω select z_2 with probability one. Then $V^\omega(s) = 0$. Moreover, we have that

$$Q_F^\star(s, z_2) = 0 + \gamma V_F^\star(s) = \frac{\gamma\xi}{1-\gamma},$$

so that

$$\mathrm{err}_{\mathrm{sel}}(\omega; s) = V_F^\star(s) - Q_F^\star(s, z_2) = \frac{\xi}{1-\gamma} - \frac{\gamma\xi}{1-\gamma} = \xi,$$

while $V_F^\star(s) - V^\omega(s) = \xi/(1-\gamma)$. □

Next we investigate the small and large bias regimes.

Corollary 3 (Bias regimes)**.** Fix a target control accuracy $\epsilon_{\mathrm{ctrl}} > 0$ and initial state s_0.

a. **Large bias.** If $\Delta_F(s_0) > \epsilon_{\mathrm{ctrl}}$, then no policy that always executes a latent from the flow-reversed bank can be ϵ_{ctrl}-optimal at s_0.

b. **Small bias.** Suppose that, for some $\alpha \in (0,1)$ and deterministic $\epsilon_{\mathrm{sel}}(n,\alpha) \geq 0$, the bound $\mathbb{P}\left(\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n) \leq \epsilon_{\mathrm{sel}}(n,\alpha)\right) \geq 1-\alpha$ holds. If, in addition, $\bar{\Delta}_F + \epsilon_{\mathrm{sel}}(n,\alpha) \leq (1-\gamma)\epsilon_{\mathrm{ctrl}}$, then, with probability at least $1-\alpha$,

$$V^\star(s) - V^{\widehat{\omega}_n}(s) \leq \epsilon_{\mathrm{ctrl}} \qquad \text{for every } s \in \mathcal{S}.$$

Proof. For part *a.*, let π be any possibly randomized or history-dependent policy that always selects a latent from the FRS bank. Let Z_0 denote its latent action at the initial state s_0. Since $V^\star$ upper-bounds the value of any continuation policy, we have

$$\begin{aligned} V^\pi(s_0) &\leq \mathbb{E}^\pi\left[r(s_0, Z_0) + \gamma V^\star(S_1) \,|\, S_0 = s_0\right] \\ &= \mathbb{E}^\pi\left[Q^\star(s_0, Z_0) \,|\, S_0 = s_0\right] \\ &\leq \max_{c \in \mathcal{C}} Q^\star\big(s_0, z_c(s_0)\big) = V^\star(s_0) - \Delta_F(s_0). \end{aligned}$$

Therefore we deduce that

$$V^\star(s_0) - V^\pi(s_0) \geq \Delta_F(s_0) > \epsilon_{\mathrm{ctrl}}.$$

Thus no policy restricted to the flow-reversed bank can be ϵ_{ctrl}-optimal at s_0.

For part *b.*, apply (22) and the assumed inequality to get that

$$V^\star(s) - V^{\widehat{\omega}_n}(s) \leq \frac{\bar{\Delta}_F + \epsilon_{\mathrm{sel}}(n,\alpha)}{1-\gamma} \leq \epsilon_{\mathrm{ctrl}}$$

simultaneously for all states on the event in (21), which has probability at least $1-\alpha$. □

F.5 INTERPRETATION AND NOVELTY

The support gap is an oracle diagnostic of the FRS abstraction. The quantity

$$\Delta_F(s) = V^\star(s) - \max_{c \in \mathcal{C}} Q^\star\big(s, z_c(s)\big)$$

asks a deliberately local question: if future decisions could still use the globally optimal full latent policy, how much is lost by requiring only the current latent to come from the online flow-reversed bank? It therefore measures the quality of the state-dependent flow-reversed action support independently of critic estimation, replay coverage, hard versus soft backups, or actor parameterization.

One-step support error and repeated restriction are not the same quantity. The actual oracle ANCHORQ comparator must remain inside the flow-reversed bank at every future state. Its value loss is therefore generally larger than the one-step gap. Part (a) shows exactly how the local support errors may accumulate:

$$\Delta_F(s) \le V^\star(s) - V_F^\star(s) \le \bar{\Delta}_F/(1-\gamma).$$

The state-dependent occupancy expression in

$$V^\star(s) - V_F^\star(s) \le \mathbb{E}_s^{\omega^\dagger}\left[\sum_{t=0}^{\infty} \gamma^t \Delta_F(S_t)\right] \le \bar{\Delta}_F/(1-\gamma). \tag{33}$$

is more informative than the uniform bound: large support gaps matter only when they occur at states visited by the full-$Q^\star$-greedy flow-reversal selector. Proposition 1 shows that neither the lower coefficient nor the factor $1/(1-\gamma)$ can be improved uniformly over discounted MDPs.

The theorem separates structural bias from finite-sample learning error. The following expression (also given in (19)) is the central decomposition:

$$\underbrace{V^\star - V^\omega}_{\text{total full-MDP loss}} = \underbrace{V^\star - V_F^\star}_{\text{online flow-reversal support price}} + \underbrace{(I - \gamma P^\omega)^{-1}\, \mathrm{err}_{\mathrm{sel}}(\omega;\cdot)}_{\text{discounted selection error within the bank}} .$$

The first term is irreducible for anchor-only ANCHORQ: more data cannot remove a behavior that is not present in the bank. The second term is reducible: it vanishes as the learned selector approaches the optimal selector for the restricted MDP. Corollary 2 shows that critic estimation and candidate optimization enter only through this second term.

The result is backend agnostic. Theorem 5 makes no assumption about hard versus soft Bellman backups, critic ensembles, subsampling scheme (e.g., REDQ Chen et al. (2021b)), entropy regularization, or how the selector is parameterized. Any learning analysis that establishes

$$\mathrm{err}_{\mathrm{sel}}(\widehat{\omega}_n) \le \epsilon_{\mathrm{sel}}(n, \alpha)$$

can be substituted directly into (22). These implementation choices affect how quickly the selection error decreases; they do not change the structural support term $\bar{\Delta}_F$.

Connection to the sample-complexity separation. Theorem 4 shows that committing to a C-mode persistent FRS abstraction can reduce online identification complexity. Theorem 5 bounds the value loss caused by the FRS restriction and selection error. Corollary 3 translates this bound into conditions for achieving a target value-suboptimality tolerance ϵ_{ctrl} relative to the optimal full-latent policy. In the small-bias regime,

$$\bar{\Delta}_F + \epsilon_{\mathrm{sel}}(n, \alpha) \le (1-\gamma)\epsilon_{\mathrm{ctrl}},$$

the restricted learner can meet the desired full-MDP control tolerance while retaining the smaller action problem. In the large-bias regime, if $\Delta_F(s_0) > \epsilon_{\mathrm{ctrl}}$, no amount of ANCHORQ data can reach the target accuracy at s_0.

Persistent action abstraction versus continuous noise-space RL. The distinction concerns the online action class, rather than a particular initialization scheme. In the anchor-restricted setting analyzed here, ANCHORQ receives the statistical benefit of a smaller action class and pays the persistent support price quantified above. Continuous noise-space RL instead learns an actor or perturbation that selects latents beyond a fixed anchor bank. Examples include DSRL (Wagenmaker et al., 2025), FRS-prefill followed by DSRL (Tang et al., 2026), human-guided noise-actor learning

in UniSteer (Lu et al., 2026), and the trust-region noise perturbations of LP-DS (Simsir & Oguz, 2026). These methods use different sources of guidance and constraints, but do not impose the persistent finite FRS action restriction studied here.

When the online class retains all latent modes, it need not incur the fixed FRS-bank support gap. Theorem 4 shows that guaranteeing correctness over this broader class can nevertheless require additional online exploration, even when the FRS bank is sufficient on the instance encountered. FRS-only prefill is a concrete example: it provides information about bank modes without ruling out better modes outside the bank. This is a worst-case statement under the theorem's information and model assumptions, not a method-specific lower bound for every noise-space RL algorithm; additional guidance, structural assumptions, or restrictions on reachable latents require separate analysis. Together, the two results identify the tradeoff between lower identification complexity with explicit support bias and broader expressiveness with a potentially higher cost of verifying that no better latent mode exists.

Technical Novelty. The contraction, Bellman-residual, and resolvent arguments used in the proof are standard tools. The contribution is their organization around the state-dependent online flow-reversal abstraction. In particular, the theorem introduces a full-$Q^\star$ one-step support gap for the latent bank generated via flow-reversal, proves a sharp bridge from that local geometric quantity to the infinite-horizon value price of persistent flow-reversal restriction, and separates it exactly from the finite-sample error of selecting among the current-state flow-reversed latents. This decomposition is the general-MDP counterpart to the adaptation separation in Theorem 4 and provides the bias coordinate for the particle bias–complexity frontier developed next.

G ANCHORQ TRACES A BIAS–COMPLEXITY FRONTIER

Theorem 4 shows that a online flow-reversal action anchor abstraction can reduce online identification complexity when a small structured action class is sufficient. Theorem 5 shows that the same restriction incurs an explicit support price. Exact-anchor ANCHORQ therefore represents one endpoint of a statistical tradeoff: it has the smallest persistent action class, but it can exclude useful latents that lie near an flow-reversed anchor rather than exactly on the anchor bank.

Particle ANCHORQ relaxes the abstraction locally. The purpose of this section is to formalize the resulting frontier. We first define a deterministic local envelope (whose size is parameterized by ρ) around the state-dependent flow-reversed anchors. We then show that increasing ρ monotonically reduces the one-step support gap while monotonically increasing, at every fixed behavioral resolution, the maximal number of pairwise behaviorally distinguishable local candidates. Finally, under a finite-state generative model, we give an explicit finite-sample bound in which both quantities appear.

We use $\rho \geq 0$ for the deterministic *local-reach radius* in the analysis. The symbol $\epsilon_{\mathrm{B}} > 0$ denotes resolution in the Bellman behavioral pseudometric, whereas $\epsilon_z > 0$ denotes Euclidean resolution in latent space. The deterministic local reach ρ around flow-reversed action anchors is also not identical to the Gaussian particle scale σ_p used by the implementation. A Gaussian scale determines a distribution of finitely many random candidates and has unbounded support. A separate result at the end of this section gives a formal connection between a finite Gaussian cloud and a bounded local envelope of reach ρ.

G.1 CONTINUOUS LOCAL ACTION ABSTRACTIONS

Throughout this section, retain the discounted latent-action MDP $\mathcal{M} = (\mathcal{S}, \mathcal{Z}, P, r, \gamma)$ from Section F, with the following additional assumptions.

Assumption 1 (Finite-state continuous latent model)**.** The state space $\mathcal{S} = \{1, \ldots, S\}$ is finite with $S < \infty$, the latent action space is $\mathcal{Z} = \mathbb{R}^{d_z}$, and for some finite $r_{\max} > 0$, rewards satisfy $r(s, z) \in [0, r_{\max}]$ for every $(s, z) \in \mathcal{S} \times \mathcal{Z}$. For every state s and next state s', the maps

$$z \mapsto r(s, z) \qquad \text{and} \qquad z \mapsto P(s' \mid s, z)$$

are continuous. The full optimal value $V^\star$ is the unique bounded fixed point of the full Bellman operator $\mathcal{T}$, and $V^\star(s) = \sup_{z \in \mathcal{Z}} Q^\star(s, z)$ for every state s.

Recall that $\mathcal{C} = \{1, \dots, C\}$ is the fixed coarse-command set and let $z_c(s) = F(s, c)$ be the corresponding online flow-reversed anchor, as in Theorem 5.

Definition 3 (Local-reach-ρ flow-reversed action set)**.** For every reach $\rho \geq 0$ and state $s \in \mathcal{S}$, define

$$\mathcal{Z}_\rho(s) := \bigcup_{c \in \mathcal{C}} B_2\big(z_c(s), \rho\big),$$

where $B_2(z, \rho) := \big\{\tilde{z} \in \mathbb{R}^{d_z} : \|\tilde{z} - z\|_2 \leq \rho\big\}$. For a bounded function $V : \mathcal{S} \to \mathbb{R}$, define the reach-$\rho$ Bellman optimality operator

$$(\mathcal{T}_\rho V)(s) := \max_{z \in \mathcal{Z}_\rho(s)} \left\{ r(s, z) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, z) V(s') \right\}. \tag{34}$$

Since $\mathcal{Z}_\rho(s)$ is a finite union of compact Euclidean balls and the maximand in (34) is continuous in z, the maximum is attained.

A stationary reach-ρ policy is a Markov kernel $\pi(\mathrm{d}z \mid s)$ supported on $\mathcal{Z}_\rho(s)$ at every state. Let V^π denote its value. Write $V_\rho^\star$ for the unique fixed point of $\mathcal{T}_\rho$.

Lemma 4 (Well-posedness of the reach-ρ control problem)**.** Suppose Assumption 1 holds. For every $\rho \geq 0$, the operator $\mathcal{T}_\rho$ is a monotone γ-contraction in the supremum norm. Hence it has a unique bounded fixed point $V_\rho^\star$. There exists a deterministic stationary reach-ρ policy $\pi_\rho^\star$ satisfying $V^{\pi_\rho^\star} = V_\rho^\star$. Moreover, the bounds $V^\pi(s) \leq V_\rho^\star(s) \leq V^\star(s)$ hold for every stationary reach-ρ policy π and every state $s \in \mathcal{S}$.

Proof. Fix bounded functions $V, W : \mathcal{S} \to \mathbb{R}$. For any state $s \in \mathcal{S}$,

$$|(\mathcal{T}_\rho V)(s) - (\mathcal{T}_\rho W)(s)| \leq \max_{z \in \mathcal{Z}_\rho(s)} \gamma \left| \sum_{s' \in \mathcal{S}} P(s' \mid s, z)\big(V(s') - W(s')\big) \right| \leq \gamma \|V - W\|_\infty.$$

Taking the maximum over s yeilds

$$\|\mathcal{T}_\rho V - \mathcal{T}_\rho W\|_\infty \leq \gamma \|V - W\|_\infty.$$

Thus $\mathcal{T}_\rho$ is a contraction. If $V \leq W$ pointwise, then the Bellman maximand for V is no larger than the Bellman maximand for W for every (s, z), so $\mathcal{T}_\rho V \leq \mathcal{T}_\rho W$. Hence $\mathcal{T}_\rho$ is monotone. The contraction mapping theorem gives a unique bounded fixed point $V_\rho^\star$.

For each state s, continuity and compactness imply that the set

$$\arg\max_{z \in \mathcal{Z}_\rho(s)} \left\{ r(s, z) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, z) V_\rho^\star(s') \right\}$$

is nonempty. Since $\mathcal{S}$ is finite, select one maximizer $z_\rho^\star(s)$ at every state and define the deterministic stationary policy

$$\pi_\rho^\star(\cdot \mid s) := \delta_{z_\rho^\star(s)}(\cdot).$$

Its Bellman operator $\mathcal{T}^{\pi_\rho^\star}$ satisfies $\mathcal{T}^{\pi_\rho^\star} V_\rho^\star = \mathcal{T}_\rho V_\rho^\star = V_\rho^\star$. The policy Bellman operator is also a γ-contraction, so its bounded fixed point is unique. Therefore $V^{\pi_\rho^\star} = V_\rho^\star$.

For an arbitrary stationary reach-ρ policy π and any bounded V,

$$\mathcal{T}^\pi V \leq \mathcal{T}_\rho V$$

pointwise, since $\mathcal{T}^\pi V$ averages Bellman values over actions in $\mathcal{Z}_\rho(s)$ whereas $\mathcal{T}_\rho V$ takes their maximum. Starting from $V_\rho^\star$ and iterating $\mathcal{T}^\pi$ gives

$$(\mathcal{T}^\pi)^k V_\rho^\star \leq V_\rho^\star \qquad \text{for every } k \geq 1.$$

The left-hand side converges in supremum norm to V^π, so $V^\pi \leq V_\rho^\star$.

Finally, for every bounded V, we have that $\mathcal{T}_\rho V \leq \mathcal{T} V$ as $\mathcal{Z}_\rho(s) \subseteq \mathcal{Z}$. In particular, we have

$$\mathcal{T}_\rho V^\star \leq \mathcal{T} V^\star = V^\star.$$

Monotonicity of $\mathcal{T}_\rho$ gives $\mathcal{T}_\rho^k V^\star \leq V^\star$ for every $k \geq 1$. The left-hand side converges to $V_\rho^\star$, proving $V_\rho^\star \leq V^\star$. □

Definition 4 (Reach-ρ one-step support gap)**.** For every $\rho \geq 0$ and state $s \in \mathcal{S}$, define

$$\Delta_\rho(s) := V^\star(s) - \max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s, z),$$

and define the corresponding uniform support gap by $\bar{\Delta}_\rho := \max_{s \in \mathcal{S}} \Delta_\rho(s)$.

Since $\mathcal{Z}_\rho(s) \subseteq \mathcal{Z}$, every $\Delta_\rho(s)$ is nonnegative. At zero reach, we have $\mathcal{Z}_0(s) = \{z_c(s) : c \in \mathcal{C}\}$, so that

$$\Delta_0(s) = \Delta_F(s), \qquad \bar{\Delta}_0 = \bar{\Delta}_F, \qquad \text{and} \qquad V_0^\star = V_F^\star. \tag{35}$$

Proposition 2 (Monotone expressiveness under local reach)**.** Let $0 \leq \rho_1 \leq \rho_2$. Then, for every state s, the following hold:

$$\mathcal{Z}_{\rho_1}(s) \subseteq \mathcal{Z}_{\rho_2}(s), \quad \Delta_{\rho_2}(s) \leq \Delta_{\rho_1}(s), \quad \text{and} \quad V_{\rho_1}^\star(s) \leq V_{\rho_2}^\star(s) \leq V^\star(s),$$

so that $\bar{\Delta}_{\rho_2} \leq \bar{\Delta}_{\rho_1}$. In addition, for every state s, the limits are such that

$$\lim_{\rho \to \infty} \Delta_\rho(s) = 0, \quad \text{and} \quad \lim_{\rho \to \infty} \bar{\Delta}_\rho = 0, \tag{36}$$

since $\mathcal{S}$ is finite.

Proof. For each anchor $z_c(s)$, we have that

$$B_2\big(z_c(s), \rho_1\big) \subseteq B_2\big(z_c(s), \rho_2\big).$$

Taking the union over $c \in \mathcal{C}$ proves $\mathcal{Z}_{\rho_1}(s) \subseteq \mathcal{Z}_{\rho_2}(s)$. Maximizing the same function $Q^\star(s, \cdot)$ over a larger set cannot decrease its maximum. Therefore

$$\max_{z \in \mathcal{Z}_{\rho_2}(s)} Q^\star(s, z) \geq \max_{z \in \mathcal{Z}_{\rho_1}(s)} Q^\star(s, z),$$

which proves $\Delta_{\rho_2}(s) \leq \Delta_{\rho_1}(s)$ after subtracting both sides from $V^\star(s)$. Taking the maximum over the finite state set proves the uniform inequality.

The admissible policy class at radius ρ_1 is contained in the admissible policy class at radius ρ_2. Therefore optimizing over the larger class gives $V_{\rho_1}^\star(s) \leq V_{\rho_2}^\star(s)$. The upper bound $V_{\rho_2}^\star \leq V^\star$ follows from Lemma 4.

To prove (36), fix a state s and an arbitrary command $c_0 \in \mathcal{C}$. Since $C \geq 1$, such a command exists. For every latent $z \in \mathbb{R}^{d_z}$, we have $z \in B_2\left(z_{c_0}(s), \|z - z_{c_0}(s)\|_2\right)$. Hence

$$\bigcup_{\rho \geq 0} \mathcal{Z}_\rho(s) = \mathbb{R}^{d_z}.$$

Define $q_\rho(s) := \max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s, z)$. The quantity $q_\rho(s)$ is nondecreasing in ρ and bounded above by $V^\star(s)$. Fix $\eta > 0$. By $V^\star(s) = \sup_{z \in \mathbb{R}^{d_z}} Q^\star(s, z)$, there exists $z_\eta \in \mathbb{R}^{d_z}$ such that $Q^\star(s, z_\eta) \geq V^\star(s) - \eta$. Whenever $\rho \geq \|z_\eta - z_{c_0}(s)\|_2$, the point z_η belongs to $\mathcal{Z}_\rho(s)$, so

$$q_\rho(s) \geq V^\star(s) - \eta.$$

Since also $q_\rho(s) \leq V^\star(s)$, it follows that $\lim_{\rho \to \infty} q_\rho(s) = V^\star(s)$. Therefore

$$\lim_{\rho \to \infty} \Delta_\rho(s) = \lim_{\rho \to \infty} \big(V^\star(s) - q_\rho(s)\big) = 0,$$

proving (36).

Finally, fix $\eta > 0$. For every $s \in \mathcal{S}$, pointwise convergence gives a finite $\rho_s(\eta)$ such that

$$\Delta_\rho(s) \leq \eta \qquad \text{whenever } \rho \geq \rho_s(\eta).$$

Since $\mathcal{S}$ is finite, define

$$\rho_\eta := \max_{s \in \mathcal{S}} \rho_s(\eta) < \infty.$$

Thus, for every $\rho \geq \rho_\eta$, we have that

$$\bar{\Delta}_\rho = \max_{s \in \mathcal{S}} \Delta_\rho(s) \leq \eta.$$

Since $\eta > 0$ was arbitrary, $\lim_{\rho \to \infty} \bar{\Delta}_\rho = 0$ follows. □

Proposition 3 (Reach-ρ abstraction price)**.** For every $\rho \geq 0$ and state s,

$$\Delta_\rho(s) \leq V^\star(s) - V_\rho^\star(s) \leq \bar{\Delta}_\rho/(1-\gamma). \tag{37}$$

More precisely, let $z_\rho^\dagger(s)$ be any maximizer of $Q^\star(s,\cdot)$ over $\mathcal{Z}_\rho(s)$ and let $\pi_\rho^\dagger$ be the deterministic policy selecting $z_\rho^\dagger(s)$. Then

$$V^\star - V^{\pi_\rho^\dagger} = (I - \gamma P^{\pi_\rho^\dagger})^{-1}\Delta_\rho(\cdot) = \sum_{t=0}^{\infty} \gamma^t (P^{\pi_\rho^\dagger})^t \Delta_\rho(\cdot),$$

and hence

$$V^\star(s) - V_\rho^\star(s) \leq \mathbb{E}_s^{\pi_\rho^\dagger}\left[\sum_{t=0}^{\infty} \gamma^t \Delta_\rho(S_t)\right]. \tag{38}$$

Proof. By Lemma 4, we have that $V_\rho^\star \leq V^\star$. Monotonicity of $\mathcal{T}_\rho$ therefore yields $\mathcal{T}_\rho V_\rho^\star \leq \mathcal{T}_\rho V^\star$. Using the fixed-point identity for $V_\rho^\star$ and the definition of Δ_ρ,

$$V_\rho^\star(s) = (\mathcal{T}_\rho V_\rho^\star)(s) \leq (\mathcal{T}_\rho V^\star)(s) = \max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s,z) = V^\star(s) - \Delta_\rho(s).$$

Rearranging proves the lower bound in (37).

For every state s, the maximum defining $z_\rho^\dagger(s)$ is attained by compactness and continuity. Since $\mathcal{S}$ is finite, the deterministic selector is well defined. By construction,

$$(\mathcal{T}^{\pi_\rho^\dagger} V^\star)(s) = V^\star(s) - \Delta_\rho(s). \tag{39}$$

Define $D_\rho^\dagger := V^\star - V^{\pi_\rho^\dagger}$. Using the Bellman fixed-point identity $V^{\pi_\rho^\dagger} = \mathcal{T}^{\pi_\rho^\dagger} V^{\pi_\rho^\dagger}$, adding and subtracting $\mathcal{T}^{\pi_\rho^\dagger} V^\star$, and applying (39), we obtain

$$\begin{aligned} D_\rho^\dagger(s) &= V^\star(s) - (\mathcal{T}^{\pi_\rho^\dagger} V^{\pi_\rho^\dagger})(s) \\ &= V^\star(s) - (\mathcal{T}^{\pi_\rho^\dagger} V^\star)(s) + (\mathcal{T}^{\pi_\rho^\dagger} V^\star)(s) - (\mathcal{T}^{\pi_\rho^\dagger} V^{\pi_\rho^\dagger})(s) \\ &= \Delta_\rho(s) + \gamma (P^{\pi_\rho^\dagger} D_\rho^\dagger)(s). \end{aligned}$$

Hence $(I - \gamma P^{\pi_\rho^\dagger}) D_\rho^\dagger = \Delta_\rho(\cdot)$. A Markov operator has supremum-norm operator norm at most one. Since $\gamma < 1$, the Neumann series $\sum_{t=0}^{\infty} \gamma^t (P^{\pi_\rho^\dagger})^t$ converges in operator norm and equals $(I - \gamma P^{\pi_\rho^\dagger})^{-1}$. This proves $V^\star - V^{\pi_\rho^\dagger} = (I - \gamma P^{\pi_\rho^\dagger})^{-1}\Delta_\rho(\cdot)$.

Evaluating the Neumann series at state s gives

$$V^\star(s) - V^{\pi_\rho^\dagger}(s) = \mathbb{E}_s^{\pi_\rho^\dagger}\left[\sum_{t=0}^{\infty} \gamma^t \Delta_\rho(S_t)\right].$$

Since $\pi_\rho^\dagger$ is an admissible reach-ρ policy, $V_\rho^\star(s) \geq V^{\pi_\rho^\dagger}(s)$. Therefore

$$V^\star(s) - V_\rho^\star(s) \leq V^\star(s) - V^{\pi_\rho^\dagger}(s),$$

proving (38). Finally, $\Delta_\rho(S_t) \leq \bar{\Delta}_\rho$ for every t, so that

$$V^\star(s) - V_\rho^\star(s) \leq \sum_{t=0}^{\infty} \gamma^t \bar{\Delta}_\rho = \bar{\Delta}_\rho/(1-\gamma).$$

□

G.2 SELECTION ERROR WITHIN THE REACH-ρ ABSTRACTION

For $z \in \mathcal{Z}_\rho(s)$, define the optimal action-value function of the reach-ρ restricted MDP by

$$Q^\star_\rho(s,z) := r(s,z) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,z) V^\star_\rho(s'). \tag{40}$$

Then $V^\star_\rho(s) = \max_{z \in \mathcal{Z}_\rho(s)} Q^\star_\rho(s,z)$. For a stationary reach-ρ policy π, define

$$r^\pi(s) := \int_{\mathcal{Z}_\rho(s)} r(s,z)\pi(\mathrm{d}z \mid s),$$

$$P^\pi(s' \mid s) := \int_{\mathcal{Z}_\rho(s)} P(s' \mid s,z)\pi(\mathrm{d}z \mid s),$$

and define its pointwise selection error inside the reach-ρ abstraction by

$$e_{\mathrm{sel},\rho}(\pi; s) := V^\star_\rho(s) - \int_{\mathcal{Z}_\rho(s)} Q^\star_\rho(s,z)\pi(\mathrm{d}z \mid s). \tag{41}$$

Let $\overline{e}_{\mathrm{sel},\rho}(\pi) := \max_{s \in \mathcal{S}} e_{\mathrm{sel},\rho}(\pi; s)$. Since the integral in (41) is an average of values no larger than their maximum, $e_{\mathrm{sel},\rho}(\pi; s) \geq 0$.

Proposition 4 (Exact support–selection decomposition at reach ρ)**.** For every stationary reach-ρ policy π,

$$V^\star - V^\pi = \big(V^\star - V^\star_\rho\big) + (I - \gamma P^\pi)^{-1} e_{\mathrm{sel},\rho}(\pi; \cdot). \tag{42}$$

Equivalently, for every state s, the decomposition

$$V^\star(s) - V^\pi(s) = V^\star(s) - V^\star_\rho(s) + \mathbb{E}^\pi_s \left[\sum_{t=0}^{\infty} \gamma^t e_{\mathrm{sel},\rho}(\pi; S_t) \right], \tag{43}$$

holds so that

$$V^\star(s) - V^\pi(s) \in [\Delta_\rho(s), (\bar{\Delta}_\rho + \overline{e}_{\mathrm{sel},\rho}(\pi))/(1-\gamma)] \tag{44}$$

for every state s.

Proof. Let $\mathcal{T}^\pi$ denote the Bellman operator of π:

$$(\mathcal{T}^\pi V)(s) = r^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} P^\pi(s' \mid s) V(s').$$

By (40) and (41), we have that

$$e_{\mathrm{sel},\rho}(\pi; s) = V^\star_\rho(s) - (\mathcal{T}^\pi V^\star_\rho)(s).$$

Since $V^\pi = \mathcal{T}^\pi V^\pi$, adding and subtracting $\mathcal{T}^\pi V^\star_\rho$ gives

$$\begin{aligned} V^\star_\rho(s) - V^\pi(s) &= V^\star_\rho(s) - (\mathcal{T}^\pi V^\star_\rho)(s) + (\mathcal{T}^\pi V^\star_\rho)(s) - (\mathcal{T}^\pi V^\pi)(s) \\ &= e_{\mathrm{sel},\rho}(\pi; s) + \gamma \sum_{s' \in \mathcal{S}} P^\pi(s' \mid s)\big(V^\star_\rho(s') - V^\pi(s')\big). \end{aligned}$$

Hence, in vector notation, $(I - \gamma P^\pi)\big(V^\star_\rho - V^\pi\big) = e_{\mathrm{sel},\rho}(\pi; \cdot)$. Since P^π is a Markov operator and $\gamma < 1$, we have that $(I - \gamma P^\pi)^{-1} = \sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$ in the supremum-norm operator topology. Therefore

$$V^\star_\rho - V^\pi = (I - \gamma P^\pi)^{-1} e_{\mathrm{sel},\rho}(\pi; \cdot). \tag{45}$$

Adding $V^\star - V^\star_\rho$ to both sides of (45) proves (42). Evaluating the Neumann series at state s proves (43).

The lower bound in (44) follows since $V^\pi \leq V^\star_\rho$ by Lemma 4 and

$$V^\star(s) - V^\star_\rho(s) \geq \Delta_\rho(s)$$

by Proposition 3. For the upper bound, the same proposition yields

$$V^\star(s) - V^\star_\rho(s) \leq \bar{\Delta}_\rho/(1-\gamma),$$

while the Neumann series and nonnegativity give

$$\big((I - \gamma P^\pi)^{-1} e_{\mathrm{sel},\rho}(\pi; \cdot)\big)(s) \leq \sum_{t=0}^{\infty} \gamma^t \overline{e}_{\mathrm{sel},\rho}(\pi) = \overline{e}_{\mathrm{sel},\rho}(\pi)/(1-\gamma).$$

Combining the two inequalities proves the result. □

Remark 4 (Recovery of Theorem 5)**.** At zero reach, $\rho = 0$, the local action set is the exact FRS anchor bank by (35). Proposition 4 therefore reduces to the support–selection decomposition of Theorem 5.

G.3 A GEOMETRIC RATE FOR THE SUPPORT GAP

Monotonicity alone does not quantify how quickly the support gap decreases. The following assumption gives a local quadratic rate when a full-MDP optimal latent lies near the flow-reversed anchor bank.

Assumption 2 (Smooth local optimal-value geometry)**.** For every state s, there exists a latent

$$z^\star(s) \in \arg\max_{z \in \mathcal{Z}} Q^\star(s, z).$$

The map $z \mapsto Q^\star(s, z)$ is differentiable on every line segment joining $z^\star(s)$ to an flow-reversed anchor $z_c(s)$ and has an L-Lipschitz gradient on those segments. Moreover,

$$\nabla_z Q^\star\big(s, z^\star(s)\big) = 0.$$

Here $z^\star(s)$ is an optimizer over the full latent space $\mathcal{Z}$, not over the reach-ρ action class. The latter is the constrained approximation whose loss is measured by $\Delta_\rho(s)$. Since in our setting $\mathcal{Z} = \mathbb{R}^{d_z}$, differentiability at an attained full-space optimum implies $\nabla_z Q^\star(s, z^\star(s)) = 0$.

Define the distance from a full-MDP optimal latent to the nearest flow-reversed anchor by

$$\mathrm{dist}_F(s) := \min_{c \in \mathcal{C}} \|z^\star(s) - z_c(s)\|_2 . \tag{46}$$

Proposition 5 (Quadratic decay of local support bias)**.** Under Assumption 2, for every $\rho \geq 0$ and state s, the upper bound

$$\Delta_\rho(s) \leq L\big(\mathrm{dist}_F(s) - \rho\big)_+^2/2, \tag{47}$$

holds where $(x)_+ := \max\{x, 0\}$, so that

$$\bar{\Delta}_\rho \leq L \max_{s \in \mathcal{S}} \big(\mathrm{dist}_F(s) - \rho\big)_+^2/2. \tag{48}$$

Proof. Fix a state $s \in \mathcal{S}$. Let $c^\star(s)$ be an index attaining the minimum in (46), and abbreviate

$$z^\star := z^\star(s), \qquad z_c := z_{c^\star(s)}(s), \qquad d := \mathrm{dist}_F(s) = \|z^\star - z_c\|_2.$$

If $d = 0$, then $z^\star = z_c \in \mathcal{Z}_\rho(s)$ for every $\rho \geq 0$. Hence

$$\max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s, z) = V^\star(s)$$

and $\Delta_\rho(s) = 0$. The claimed bound follows.

Suppose $d > 0$. If $\rho \geq d$, then $z^\star \in B_2(z_c, \rho)$, so again $\Delta_\rho(s) = 0$, which equals the right-hand side of (47).

It remains to consider $0 \leq \rho < d$. Define $z_\rho := z_c + \frac{\rho}{d}\big(z^\star - z_c\big)$. Then $\|z_\rho - z_c\|_2 = \rho$, so $z_\rho \in \mathcal{Z}_\rho(s)$. Also, we have $z_\rho - z^\star = -\big(1 - \frac{\rho}{d}\big)\big(z^\star - z_c\big)$, and therefore $\|z_\rho - z^\star\|_2 = d - \rho$.

The standard smoothness inequality, applied to the function $z \mapsto Q^\star(s, z)$ at $z^\star$ and z_ρ, gives

$$Q^\star(s, z_\rho) \geq Q^\star(s, z^\star) + \langle \nabla_z Q^\star(s, z^\star), z_\rho - z^\star \rangle - \frac{L}{2}\|z_\rho - z^\star\|_2^2 = V^\star(s) - \frac{L}{2}(d - \rho)^2,$$

where the equality uses $\nabla_z Q^\star(s, z^\star) = 0$. Since $z_\rho \in \mathcal{Z}_\rho(s)$, the lower bound holds:

$$\max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s, z) \geq Q^\star(s, z_\rho).$$

Subtracting from $V^\star(s)$ yields $\Delta_\rho(s) \leq L(d - \rho)^2/2$. This is exactly (47) in the case $\rho < d$. Taking the maximum over the finite state set proves (48). □

G.4 BEHAVIORALLY DISTINGUISHABLE LOCAL CANDIDATES

A Euclidean perturbation need not correspond to a distinct closed-loop behavior. We therefore measure local action complexity using a Bellman behavioral pseudometric rather than by counting raw latent coordinates.

Let $V_{\max} := r_{\max}/(1-\gamma)$ and define $\mathcal{V}_{\max} := \{V : \mathcal{S} \to [0, V_{\max}]\}$.

Definition 5 (Bellman behavioral pseudometric)**.** For every state s and latents $z, \tilde{z} \in \mathcal{Z}$, define

$$d_s^{\mathrm{B}}(z, \tilde{z}) := \sup_{V \in \mathcal{V}_{\max}} |Q_V(s, z) - Q_V(s, \tilde{z})| \,. \tag{49}$$

Thus two latents are close in d_s^{B} when no bounded continuation value can distinguish their one-step reward and transition consequences by more than the prescribed resolution.

Lemma 5 (Basic properties of the behavioral pseudometric)**.** For each state s, the function d_s^{B} is a pseudometric on $\mathcal{Z}$. For every $V \in \mathcal{V}_{\max}$, the upper bound holds:

$$|Q_V(s, z) - Q_V(s, \tilde{z})| \le d_s^{\mathrm{B}}(z, \tilde{z}). \tag{50}$$

In particular,

$$|Q^\star(s, z) - Q^\star(s, \tilde{z})| \le d_s^{\mathrm{B}}(z, \tilde{z}). \tag{51}$$

Moreover,

$$d_s^{\mathrm{B}}(z, \tilde{z}) \le |r(s, z) - r(s, \tilde{z})| + \gamma V_{\max} \sum_{s' \in \mathcal{S}} |P(s' \mid s, z) - P(s' \mid s, \tilde{z})|. \tag{52}$$

Consequently, $d_s^{\mathrm{B}}(z_n, z) \to 0$ whenever $\|z_n - z\|_2 \to 0$. Every compact Euclidean subset of $\mathcal{Z}$ therefore has a finite behavioral ϵ_{B}-net for every $\epsilon_{\mathrm{B}} > 0$.

Proof. Nonnegativity and symmetry follow immediately from the absolute value in (49). Also, $d_s^{\mathrm{B}}(z, z) = 0$. Fix $z_1, z_2, z_3 \in \mathcal{Z}$. For every $V \in \mathcal{V}_{\max}$, we have that

$$\begin{aligned} |Q_V(s, z_1) - Q_V(s, z_3)| &\le |Q_V(s, z_1) - Q_V(s, z_2)| + |Q_V(s, z_2) - Q_V(s, z_3)| \\ &\le d_s^{\mathrm{B}}(z_1, z_2) + d_s^{\mathrm{B}}(z_2, z_3). \end{aligned}$$

Taking the supremum over $V \in \mathcal{V}_{\max}$ proves the triangle inequality. Distinct latents may have identical reward and transition laws, in which case their distance is zero; hence the object is a pseudometric rather than necessarily a metric.

Inequality (50) follows directly because the left-hand side is one of the quantities over which the supremum is taken. Since rewards are in $[0, r_{\max}]$, the full optimal value satisfies $V^\star(s) \in [0, V_{\max}]$ for every state, so $V^\star \in \mathcal{V}_{\max}$. Substituting $V = V^\star$ into (50) proves (51).

For every $V \in \mathcal{V}_{\max}$, we have that

$$\begin{aligned} |Q_V(s, z) - Q_V(s, \tilde{z})| &\le |r(s, z) - r(s, \tilde{z})| + \gamma \sum_{s' \in \mathcal{S}} V(s') |P(s' \mid s, z) - P(s' \mid s, \tilde{z})| \\ &\le |r(s, z) - r(s, \tilde{z})| + \gamma V_{\max} \sum_{s' \in \mathcal{S}} |P(s' \mid s, z) - P(s' \mid s, \tilde{z})|. \end{aligned}$$

Taking the supremum over $V \in \mathcal{V}_{\max}$ proves (52). By Assumption 1, the right-hand side converges to zero whenever $\tilde{z} \to z$ in Euclidean norm. Thus the identity map from any Euclidean compact subset of $\mathcal{Z}$ to the same set equipped with d_s^{B} is continuous. A continuous image of a compact set is compact and therefore totally bounded in the pseudometric quotient. Hence the compact set admits a finite behavioral ϵ_{B}-net for every $\epsilon_{\mathrm{B}} > 0$. □

For a pseudometric space (A, d) and resolution $\epsilon_{\mathrm{B}} > 0$, define its ϵ_{B}-packing number by

$$\mathcal{M}(A, d, \epsilon_{\mathrm{B}}) := \sup \{|\mathcal{P}| : \mathcal{P} \subseteq A \text{ is finite, and } d(x, y) > \epsilon_{\mathrm{B}} \text{ for all distinct } x, y \in \mathcal{P}\} \,. \tag{53}$$

The value may be infinite in a general pseudometric space. A set counted by (53) consists of candidates that are pairwise behaviorally distinguishable at resolution ϵ_{B}.

Define the maximal behavioral packing complexity of the reach-ρ action abstraction by

$$A_\rho(\epsilon_{\mathrm{B}}) := \max_{s \in \mathcal{S}} \mathcal{M}\left(\mathcal{Z}_\rho(s), d_s^{\mathrm{B}}, \epsilon_{\mathrm{B}}\right).$$

Proposition 6 (Monotone behavioral complexity)**.** For every fixed resolution $\epsilon_{\mathrm{B}} > 0$, the map $\rho \mapsto A_\rho(\epsilon_{\mathrm{B}})$ is nondecreasing. In particular, if $0 \le \rho_1 \le \rho_2$, then $A_{\rho_1}(\epsilon_{\mathrm{B}}) \le A_{\rho_2}(\epsilon_{\mathrm{B}})$.

Proof. Fix a state $s \in \mathcal{S}$. By Proposition 2, the inclusion holds:

$$\mathcal{Z}_{\rho_1}(s) \subseteq \mathcal{Z}_{\rho_2}(s).$$

Every finite ϵ_{B}-separated subset of $\mathcal{Z}_{\rho_1}(s)$ is also an ϵ_{B}-separated subset of $\mathcal{Z}_{\rho_2}(s)$. Taking the supremum over all such finite subsets gives

$$\mathcal{M}\left(\mathcal{Z}_{\rho_1}(s), d_s^{\mathrm{B}}, \epsilon_{\mathrm{B}}\right) \le \mathcal{M}\left(\mathcal{Z}_{\rho_2}(s), d_s^{\mathrm{B}}, \epsilon_{\mathrm{B}}\right).$$

Taking the maximum over the finite state set proves the result. □

The monotonicity is weak. If newly admitted latents are behaviorally redundant at resolution ϵ_{B}, then $A_\rho(\epsilon_{\mathrm{B}})$ may remain constant.

Assumption 3 (Behavioral Lipschitzness)**.** There exists $L_{\mathrm{B}} > 0$ such that, for every state s and all $z, \tilde{z} \in \mathbb{R}^{d_z}$, the bound holds:

$$d_s^{\mathrm{B}}(z, \tilde{z}) \le L_{\mathrm{B}} \|z - \tilde{z}\|_2.$$

Lemma 6 (A sufficient condition for behavioral Lipschitzness)**.** Suppose there are constants $L_r, L_P \ge 0$ such that, for all states s and latents $z, \tilde{z}$,

$$|r(s,z) - r(s,\tilde{z})| \le L_r \|z - \tilde{z}\|_2, \quad \text{and} \quad \sum_{s' \in \mathcal{S}} |P(s' \mid s, z) - P(s' \mid s, \tilde{z})| \le L_P \|z - \tilde{z}\|_2.$$

Then Assumption 3 holds with $L_{\mathrm{B}} := L_r + \gamma V_{\max} L_P$.

Proof. Fix $V \in \mathcal{V}_{\max}$. Then

$$\begin{aligned} |Q_V(s,z) - Q_V(s,\tilde{z})| &\le |r(s,z) - r(s,\tilde{z})| + \gamma \left| \sum_{s' \in \mathcal{S}} V(s') \big(P(s' \mid s, z) - P(s' \mid s, \tilde{z})\big) \right| \\ &\le L_r \|z - \tilde{z}\|_2 + \gamma V_{\max} \sum_{s' \in \mathcal{S}} |P(s' \mid s, z) - P(s' \mid s, \tilde{z})| \\ &\le \big(L_r + \gamma V_{\max} L_P\big) \|z - \tilde{z}\|_2. \end{aligned}$$

Taking the supremum over $V \in \mathcal{V}_{\max}$ proves the claim. □

Assumption 4 (Local behavioral richness)**.** There exist a state s_0, a command c_0, a radius $\rho_{\mathrm{rich}} > 0$, an integer $d_{\mathrm{eff}} \in \{1, \dots, d_z\}$, a d_{eff}-dimensional linear subspace $U \subseteq \mathbb{R}^{d_z}$, and a constant $\ell_{\mathrm{B}} > 0$ such that

$$d_{s_0}^{\mathrm{B}}\big(z_{c_0}(s_0) + u, z_{c_0}(s_0) + v\big) \ge \ell_{\mathrm{B}} \|u - v\|_2$$

for all $u, v \in U$ satisfying $\|u\|_2 \le \rho_{\mathrm{rich}}$ and $\|v\|_2 \le \rho_{\mathrm{rich}}$.

Proposition 7 (Polynomial upper and lower bounds on local behavioral complexity)**.** Under Assumption 3, for every $\rho \ge 0$ and $\epsilon_{\mathrm{B}} > 0$,

$$A_\rho(\epsilon_{\mathrm{B}}) \le C \left(1 + \frac{2 L_{\mathrm{B}} \rho}{\epsilon_{\mathrm{B}}}\right)^{d_z}.$$

If Assumption 4 also holds, then, for every $\rho \in (0, \rho_{\mathrm{rich}}]$ and $\epsilon_{\mathrm{B}} > 0$,

$$A_\rho(\epsilon_{\mathrm{B}}) \ge \left(\frac{\ell_{\mathrm{B}} \rho}{\epsilon_{\mathrm{B}}}\right)^{d_{\mathrm{eff}}}. \tag{54}$$

The lower bound is nontrivial when $\ell_{\mathrm{B}} \rho > \epsilon_{\mathrm{B}}$.

Proof. We first prove the upper bound. Fix a state s and let $\mathcal{P} \subseteq \mathcal{Z}_\rho(s)$ be any finite set whose distinct points have pairwise behavioral distance strictly larger than ϵ_{B}. For each $z \in \mathcal{P}$, choose one command $c(z) \in \mathcal{C}$ such that

$$z \in B_2\big(z_{c(z)}(s), \rho\big).$$

This is possible by the definition of $\mathcal{Z}_\rho(s)$. For $c \in \mathcal{C}$, let $\mathcal{P}_c := \{z \in \mathcal{P} : c(z) = c\}$ so that the sets $\mathcal{P}_c$ partition $\mathcal{P}$.

If $z, \tilde{z} \in \mathcal{P}_c$ are distinct, then

$$\epsilon_{\mathrm{B}} < d_s^{\mathrm{B}}(z, \tilde{z}) \le L_{\mathrm{B}} \|z - \tilde{z}\|_2,$$

so $\|z - \tilde{z}\|_2 > \epsilon_{\mathrm{B}}/L_{\mathrm{B}}$. Hence the Euclidean balls centered at points of $\mathcal{P}_c$ with radius $\epsilon_{\mathrm{B}}/(2L_{\mathrm{B}})$ are pairwise disjoint. They are all contained in

$$B_2\left(z_c(s), \rho + \epsilon_{\mathrm{B}}/(2L_{\mathrm{B}})\right).$$

Comparing d_z-dimensional Euclidean volumes gives

$$|\mathcal{P}_c| \operatorname{vol}\left(B_2\left(0, \frac{\epsilon_{\mathrm{B}}}{2L_{\mathrm{B}}}\right)\right) \le \operatorname{vol}\left(B_2\left(0, \rho + \frac{\epsilon_{\mathrm{B}}}{2L_{\mathrm{B}}}\right)\right).$$

Therefore $|\mathcal{P}_c| \le \left(1 + \frac{2L_{\mathrm{B}}\rho}{\epsilon_{\mathrm{B}}}\right)^{d_z}$. Summing over $c \in \mathcal{C}$ yields $|\mathcal{P}| \le C\left(1 + \frac{2L_{\mathrm{B}}\rho}{\epsilon_{\mathrm{B}}}\right)^{d_z}$. Since $\mathcal{P}$ was arbitrary, the same upper bound holds for the packing number at state s, and then for $A_\rho(\epsilon_{\mathrm{B}})$ after maximizing over s.

We now prove the lower bound. Fix $\rho \in (0, \rho_{\mathrm{rich}}]$ and work at the state, anchor, and subspace in Assumption 4. Let

$$D_\rho := \{u \in U : \|u\|_2 \le \rho\}.$$

Choose an inclusion-maximal finite subset $\{u_1, \dots, u_N\} \subseteq D_\rho$ whose distinct points have pairwise Euclidean distance strictly larger than $\epsilon_{\mathrm{B}}/\ell_{\mathrm{B}}$. Such a finite maximal set exists because D_ρ is compact and has finite Euclidean packing number at every positive resolution. Maximality implies that the closed Euclidean balls in U of radius $\epsilon_{\mathrm{B}}/\ell_{\mathrm{B}}$ centered at the u_i cover D_ρ: otherwise an uncovered point could be added while preserving the strict separation.

Let $v_{d_{\mathrm{eff}}}$ denote the volume of the unit Euclidean ball in U. By subadditivity of volume, we have

$$v_{d_{\mathrm{eff}}} \rho^{d_{\mathrm{eff}}} \le N v_{d_{\mathrm{eff}}} \left(\epsilon_{\mathrm{B}}/\ell_{\mathrm{B}}\right)^{d_{\mathrm{eff}}}.$$

Canceling the unit-ball volume gives $N \ge \left(\ell_{\mathrm{B}}\rho/\epsilon_{\mathrm{B}}\right)^{d_{\mathrm{eff}}}$. For distinct i, j, Assumption 4 yields

$$d_{s_0}^{\mathrm{B}}\left(z_{c_0}(s_0) + u_i, z_{c_0}(s_0) + u_j\right) \ge \ell_{\mathrm{B}} \|u_i - u_j\|_2 > \epsilon_{\mathrm{B}}.$$

The translated points lie in $\mathcal{Z}_\rho(s_0)$ and form a behavioral ϵ_{B}-packing. Hence $A_\rho(\epsilon_{\mathrm{B}}) \ge N$, proving (54). □

For a pseudometric space (A, d), let $\mathcal{N}_{\mathrm{int}}(A, d, \epsilon_{\mathrm{B}})$ denote the smallest cardinality of an *internal* ϵ_{B}-net: a set $\mathcal{A} \subseteq A$ such that every $z \in A$ has some $a \in \mathcal{A}$ with $d(z, a) \le \epsilon_{\mathrm{B}}$.

Lemma 7 (Packing–net sandwich)**.** Suppose $\mathcal{M}(A, d, \epsilon_{\mathrm{B}}) < \infty$. Then

$$\mathcal{M}(A, d, 2\epsilon_{\mathrm{B}}) \le \mathcal{N}_{\mathrm{int}}(A, d, \epsilon_{\mathrm{B}}) \le \mathcal{M}(A, d, \epsilon_{\mathrm{B}}). \tag{55}$$

In particular, under Assumption 3, for every state s there exists a behavioral ϵ_{B}-net $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s) \subseteq \mathcal{Z}_\rho(s)$ satisfying $|\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)| \le A_\rho(\epsilon_{\mathrm{B}})$. If Assumption 4 also holds and $\rho \in (0, \rho_{\mathrm{rich}}]$, then every behavioral ϵ_{B}-net of $\mathcal{Z}_\rho(s_0)$ has cardinality at least

$$\left(\ell_{\mathrm{B}}\rho/(2\epsilon_{\mathrm{B}})\right)^{d_{\mathrm{eff}}}. \tag{56}$$

Proof. Let $\mathcal{P} \subseteq A$ be any finite set whose distinct points have pairwise distance strictly larger than $2\epsilon_{\mathrm{B}}$, and let $\mathcal{A} \subseteq A$ be an internal ϵ_{B}-net. No single point $a \in \mathcal{A}$ can be within distance ϵ_{B} of two distinct points of $\mathcal{P}$; otherwise the triangle inequality would make those two points have distance at most $2\epsilon_{\mathrm{B}}$. Consequently $|\mathcal{A}| \ge |\mathcal{P}|$. Taking the supremum over all such $\mathcal{P}$ and then the infimum over all internal ϵ_{B}-nets proves the left inequality in (55).

For the right inequality, construct an inclusion-maximal ϵ_{B}-separated subset $\mathcal{P}_{\max} \subseteq A$. Because the ϵ_{B}-packing number is finite, this construction terminates after finitely many points. Maximality implies that $\mathcal{P}_{\max}$ is an internal ϵ_{B}-net: if some $x \in A$ were farther than ϵ_{B} from every point of $\mathcal{P}_{\max}$, then x could be added while preserving ϵ_{B}-separation. Moreover,

$$|\mathcal{P}_{\max}| \le \mathcal{M}(A, d, \epsilon_{\mathrm{B}}).$$

This proves the right inequality.

Under Assumption 3, Proposition 7 makes the relevant packing numbers finite, so the existence claim follows after maximizing over states. Finally, (56) follows from the left inequality in (55) and Proposition 7 applied at resolution $2\epsilon_{\mathrm{B}}$. □

G.5 FINITE BEHAVIORAL DISCRETIZATION

Fix $\rho \geq 0$ and $\epsilon_{\mathrm{B}} > 0$. For every state s, choose a behavioral ϵ_{B}-net $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s) \subseteq \mathcal{Z}_\rho(s)$ with $|\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)| \leq A_\rho(\epsilon_{\mathrm{B}})$. Define the finite-net Bellman operator

$$(\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}} V)(s) := \max_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} Q_V(s,a),$$

and let $V^\star_{\rho,\epsilon_{\mathrm{B}}}$ denote its unique fixed point.

Define the one-step support gap of the finite net by

$$\Delta_{\rho,\epsilon_{\mathrm{B}}}(s) := V^\star(s) - \max_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} Q^\star(s,a).$$

Lemma 8 (Discretization adds at most ϵ_{B} one-step error)**.** For every state s, we have that

$$\Delta_{\rho,\epsilon_{\mathrm{B}}}(s) \leq \Delta_\rho(s) + \epsilon_{\mathrm{B}}, \tag{57}$$

so that

$$V^\star(s) - V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) \in [0, (\bar{\Delta}_\rho + \epsilon_{\mathrm{B}})/(1-\gamma)] \tag{58}$$

for every state $s \in \mathcal{S}$.

Proof. Fix a state $s \in \mathcal{S}$. Since $\mathcal{Z}_\rho(s)$ is compact and $Q^\star(s,\cdot)$ is continuous, choose

$$z^\star_\rho(s) \in \arg\max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s,z).$$

Since $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$ is an ϵ_{B}-net under d^{B}_s, there exists $a_{\rho,\epsilon_{\mathrm{B}}}(s) \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$ satisfying

$$d^{\mathrm{B}}_s\big(z^\star_\rho(s), a_{\rho,\epsilon_{\mathrm{B}}}(s)\big) \leq \epsilon_{\mathrm{B}}.$$

Lemma 5 gives us that

$$Q^\star\big(s, a_{\rho,\epsilon_{\mathrm{B}}}(s)\big) \geq Q^\star\big(s, z^\star_\rho(s)\big) - \epsilon_{\mathrm{B}}.$$

Therefore

$$\max_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} Q^\star(s,a) \geq \max_{z \in \mathcal{Z}_\rho(s)} Q^\star(s,z) - \epsilon_{\mathrm{B}}.$$

Subtracting both sides from $V^\star(s)$ proves (57).

The finite sets $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$ define a persistent state-dependent action abstraction of the form analyzed in Theorem 5. Applying the oracle upper bound from that theorem with support gap $\Delta_{\rho,\epsilon_{\mathrm{B}}}$ gives

$$V^\star(s) - V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) \leq \max_{x \in \mathcal{S}} \Delta_{\rho,\epsilon_{\mathrm{B}}}(x)/(1-\gamma).$$

By (57), we have that

$$\max_{x \in \mathcal{S}} \Delta_{\rho,\epsilon_{\mathrm{B}}}(x) \leq \bar{\Delta}_\rho + \epsilon_{\mathrm{B}}.$$

This proves the upper bound in (58). The lower bound follows from $V^\star_{\rho,\epsilon_{\mathrm{B}}} \leq V^\star$. □

For $a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$, define

$$Q^\star_{\rho,\epsilon_{\mathrm{B}}}(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V^\star_{\rho,\epsilon_{\mathrm{B}}}(s'). \tag{59}$$

Let $\omega(\cdot \mid s)$ be any stationary Markov kernel supported on $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$, and define its pointwise finite-net selection error by

$$e_{\rho,\epsilon_{\mathrm{B}}}(\omega; s) := V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) - \sum_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} \omega(a \mid s) Q^\star_{\rho,\epsilon_{\mathrm{B}}}(s,a). \tag{60}$$

Proposition 8 (Finite-net support–selection decomposition)**.** Let

$$\bar{e}_{\rho,\epsilon_{\mathrm{B}}}(\omega) := \max_{s\in\mathcal{S}} e_{\rho,\epsilon_{\mathrm{B}}}(\omega; s).$$

Then $e_{\rho,\epsilon_{\mathrm{B}}}(\omega; s) \geq 0$ for every state and

$$V^\star - V^\omega = \big(V^\star - V^\star_{\rho,\epsilon_{\mathrm{B}}}\big) + (I - \gamma P^\omega)^{-1} e_{\rho,\epsilon_{\mathrm{B}}}(\omega; \cdot), \tag{61}$$

so that

$$0 \leq V^\star(s) - V^\omega(s) \leq (\bar{\Delta}_\rho + \epsilon_{\mathrm{B}} + \bar{e}_{\rho,\epsilon_{\mathrm{B}}}(\omega))/(1-\gamma) \tag{62}$$

for every state s. In particular, if a learned selector $\widehat{\omega}_n$ satisfies

$$\mathbb{P}\left(\bar{e}_{\rho,\epsilon_{\mathrm{B}}}(\widehat{\omega}_n) \leq \epsilon_{\mathrm{sel}}(n,\alpha)\right) \geq 1-\alpha, \tag{63}$$

then, with probability at least $1-\alpha$, simultaneously for every state,

$$V^\star(s) - V^{\widehat{\omega}_n}(s) \leq (\bar{\Delta}_\rho + \epsilon_{\mathrm{B}} + \epsilon_{\mathrm{sel}}(n,\alpha))/(1-\gamma). \tag{64}$$

Proof. Since

$$V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) = \max_{a\in\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} Q^\star_{\rho,\epsilon_{\mathrm{B}}}(s,a),$$

the average in (60) cannot exceed $V^\star_{\rho,\epsilon_{\mathrm{B}}}(s)$. Hence $e_{\rho,\epsilon_{\mathrm{B}}}(\omega; s) \geq 0$.

Let $\mathcal{T}^\omega$ be the policy Bellman operator. By (60), we have that

$$e_{\rho,\epsilon_{\mathrm{B}}}(\omega; \cdot) = V^\star_{\rho,\epsilon_{\mathrm{B}}} - \mathcal{T}^\omega V^\star_{\rho,\epsilon_{\mathrm{B}}}.$$

Since $V^\omega = \mathcal{T}^\omega V^\omega$, we have

$$V^\star_{\rho,\epsilon_{\mathrm{B}}} - V^\omega = V^\star_{\rho,\epsilon_{\mathrm{B}}} - \mathcal{T}^\omega V^\star_{\rho,\epsilon_{\mathrm{B}}} + \mathcal{T}^\omega V^\star_{\rho,\epsilon_{\mathrm{B}}} - \mathcal{T}^\omega V^\omega = e_{\rho,\epsilon_{\mathrm{B}}}(\omega; \cdot) + \gamma P^\omega\big(V^\star_{\rho,\epsilon_{\mathrm{B}}} - V^\omega\big).$$

The Markov operator P^ω has supremum-norm operator norm at most one, so $I - \gamma P^\omega$ is invertible with Neumann-series inverse $(I-\gamma P^\omega)^{-1} = \sum_{t=0}^\infty \gamma^t (P^\omega)^t$. Therefore

$$V^\star_{\rho,\epsilon_{\mathrm{B}}} - V^\omega = (I - \gamma P^\omega)^{-1} e_{\rho,\epsilon_{\mathrm{B}}}(\omega; \cdot). \tag{65}$$

Adding $V^\star - V^\star_{\rho,\epsilon_{\mathrm{B}}}$ to both sides proves (61).

Lemma 8 yields $0 \leq V^\star(s) - V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) \leq (\bar{\Delta}_\rho + \epsilon_{\mathrm{B}})/(1-\gamma)$. Moreover, evaluating the Neumann series in (65) at state s yields

$$V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) - V^\omega(s) = \mathbb{E}^\omega_s\left[\sum_{t=0}^\infty \gamma^t e_{\rho,\epsilon_{\mathrm{B}}}(\omega; S_t)\right] \leq \frac{\bar{e}_{\rho,\epsilon_{\mathrm{B}}}(\omega)}{1-\gamma}.$$

Combining the two bounds proves (62). Finally, on the event in (63), substitute

$$\bar{e}_{\rho,\epsilon_{\mathrm{B}}}(\widehat{\omega}_n) \leq \epsilon_{\mathrm{sel}}(n,\alpha)$$

into (62). The event has probability at least $1-\alpha$, proving (64). □

G.6 A FINITE-SAMPLE FRONTIER UNDER A GENERATIVE MODEL

We now instantiate the finite-sample selection term using a synchronous empirical Q-iteration scheme. This specialization is used only to make the dependence on the number of behaviorally distinguishable local candidates explicit.

Assumption 5 (Generative model)**.** For every state s and latent z, a call to the generative model returns a random pair $(\mathsf{R}, \mathsf{S}')$ satisfying

$$0 \leq \mathsf{R} \leq r_{\max}, \qquad \mathbb{E}[\mathsf{R}] = r(s,z), \qquad \mathbb{P}(\mathsf{S}' = s') = P(s' \mid s, z).$$

All generative-model calls are independent conditional on their queried state–latent pairs.

For the finite action sets $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$, define the optimal Q-Bellman operator

$$(\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}Q)(s,a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s')} Q(s',a').$$

Its unique fixed point is the function $Q^{\star}_{\rho,\epsilon_{\mathrm{B}}}$ defined in (59).

Initialize $\widehat{Q}_0(s,a) := 0$ for every finite-net state–action pair. For an integer $J \geq 1$, at each iteration $k \in \{0,\dots,J-1\}$ and for every (s,a), draw m fresh generative samples

$$\big(\mathsf{R}_{k,i}(s,a), \mathsf{S}'_{k,i}(s,a)\big), \qquad i = 1,\dots,m,$$

and set

$$\widehat{Q}_{k+1}(s,a) := \frac{1}{m}\sum_{i=1}^{m}\left[\mathsf{R}_{k,i}(s,a) + \gamma \max_{a' \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(\mathsf{S}'_{k,i}(s,a))} \widehat{Q}_k\big(\mathsf{S}'_{k,i}(s,a), a'\big)\right]. \tag{66}$$

After J iterations, let $\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}$ be a deterministic policy satisfying

$$\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}(s) \in \operatorname*{arg\,max}_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} \widehat{Q}_J(s,a).$$

Theorem 6 (QFRS bias–complexity frontier)**.** Suppose Assumptions 1, 3, and 5 hold. Fix the following constants:

$$\rho \geq 0, \quad \epsilon_{\mathrm{B}} > 0, \quad \epsilon_Q > 0, \quad \text{and} \quad \alpha \in (0,1).$$

For every state, choose a behavioral ϵ_{B}-net $\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)$ with cardinality at most $A_\rho(\epsilon_{\mathrm{B}})$. Let $J \geq 1$ satisfy $\gamma^J V_{\max} \leq \epsilon_Q/2$, and set

$$m := \left\lceil \frac{2V_{\max}^2}{(1-\gamma)^2\epsilon_Q^2} \log\left(\frac{2SA_\rho(\epsilon_{\mathrm{B}})J}{\alpha}\right)\right\rceil. \tag{67}$$

Then, with probability at least $1-\alpha$, we have that

$$\|V^{\star} - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}}\|_\infty \leq \frac{\bar{\Delta}_\rho + \epsilon_{\mathrm{B}} + 2\epsilon_Q}{1-\gamma}. \tag{68}$$

The number of generative-model samples used by (66) is

$$N_{\mathrm{gen}} = Jm\sum_{s\in\mathcal{S}}|\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)| \leq SJA_\rho(\epsilon_{\mathrm{B}})\left\lceil \frac{2V_{\max}^2}{(1-\gamma)^2\epsilon_Q^2} \log\left(\frac{2SA_\rho(\epsilon_{\mathrm{B}})J}{\alpha}\right)\right\rceil. \tag{69}$$

Furthermore, we have the following consequences:

a. For every fixed behavioral resolution ϵ_{B}, estimation accuracy ϵ_Q, confidence level α, and admissible iteration count J, increasing the reach ρ moves the two structural quantities in opposite directions: $\bar{\Delta}_\rho$ is nonincreasing in ρ, whereas $A_\rho(\epsilon_{\mathrm{B}})$ is nondecreasing in ρ. Consequently, for fixed ϵ_{B}, ϵ_Q, α, and J, the sample-complexity upper bound on the right-hand side of (69) is nondecreasing in ρ.

b. If Assumption 2 also holds, then

$$\left\|V^{\star} - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}}\right\|_\infty \leq \frac{1}{1-\gamma}\left[\frac{L}{2}\max_{s\in\mathcal{S}}\big(\mathrm{dist}_F(s)-\rho\big)_+^2 + \epsilon_{\mathrm{B}} + 2\epsilon_Q\right] \tag{70}$$

with probability at least $1-\alpha$. Under the standing behavioral Lipschitzness assumption, the bound holds:

$$A_\rho(\epsilon_{\mathrm{B}}) \leq C\left(1 + \frac{2L_{\mathrm{B}}\rho}{\epsilon_{\mathrm{B}}}\right)^{d_z}. \tag{71}$$

c. If Assumption 4 also holds and $\rho \in (0, \rho_{\mathrm{rich}}]$, then every behavioral ϵ_{B}-net at the distinguished state s_0 contains at least $(\ell_{\mathrm{B}}\rho/(2\epsilon_{\mathrm{B}}))^{d_{\mathrm{eff}}}$ candidates.

Proof. We divide the proof into five steps.

Step 1: boundedness of the empirical iterates and targets. We prove by induction that

$$0 \le \widehat{Q}_k(s,a) \le V_{\max} \tag{72}$$

for every $k \in \{0,\dots,J\}$ and every finite-net state–action pair. The claim holds for $k=0$ because $\widehat{Q}_0 = 0$. Suppose it holds at iteration k. Every summand in (66) is nonnegative. It is also bounded above by

$$r_{\max} + \gamma V_{\max} = r_{\max} + \gamma \frac{r_{\max}}{1-\gamma} = \frac{r_{\max}}{1-\gamma} = V_{\max}.$$

Therefore their average lies in $[0, V_{\max}]$, proving (72) at iteration $k+1$.

Step 2: a uniform one-step empirical Bellman bound. Let $\mathcal{F}_k$ be the sigma-algebra generated by all samples drawn before iteration k. Then $\widehat{Q}_k$ is $\mathcal{F}_k$-measurable. By the fresh-sample assumption, conditional on $\mathcal{F}_k$, the m summands in (66) for a fixed pair (s,a) are independent, lie in $[0, V_{\max}]$, and have common conditional expectation $(\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widehat{Q}_k)(s,a)$. Höeffding's inequality therefore gives, for every $\xi > 0$, the bound

$$\mathbb{P}\left(\left|\widehat{Q}_{k+1}(s,a) - (\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widehat{Q}_k)(s,a)\right| > \xi \,\middle|\, \mathcal{F}_k\right) \le 2\exp\left(-2m\xi^2/V_{\max}^2\right). \tag{73}$$

Taking expectations of both sides gives the same unconditional bound. Across all iterations, there are at most $SJA_\rho(\epsilon_{\mathrm{B}})$ state–action–iteration triples. A union bound applied to (73) yields

$$\mathbb{P}\left(\max_{0\le k<J}\left\|\widehat{Q}_{k+1} - \mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widehat{Q}_k\right\|_\infty > \xi\right) \le 2SJA_\rho(\epsilon_{\mathrm{B}})\exp\left(-2m\xi^2/V_{\max}^2\right). \tag{74}$$

Set $\xi := (1-\gamma)\epsilon_Q/2$. By (67), we have that

$$\frac{2m\xi^2}{V_{\max}^2} = \frac{m(1-\gamma)^2\epsilon_Q^2}{2V_{\max}^2} \ge \log\left(\frac{2SA_\rho(\epsilon_{\mathrm{B}})J}{\alpha}\right).$$

Substituting into (74) shows that the event

$$\mathcal{E}_{\mathrm{Bell}} := \left\{\max_{0\le k<J}\left\|\widehat{Q}_{k+1} - \mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widehat{Q}_k\right\|_\infty \le \frac{(1-\gamma)\epsilon_Q}{2}\right\}$$

has probability at least $1-\alpha$.

Step 3: error of empirical Q-iteration on the finite net. The operator $\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}$ is a γ-contraction in supremum norm. Indeed, for any two finite-net Q-functions Q and $\widetilde{Q}$,

$$\begin{aligned}\left|(\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}Q)(s,a) - (\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widetilde{Q})(s,a)\right| &\le \gamma\sum_{s'\in\mathcal{S}} P(s' \mid s,a)\left|\max_{a'\in\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s')} Q(s',a') - \max_{a'\in\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s')}\widetilde{Q}(s',a')\right| \\ &\le \gamma\|Q-\widetilde{Q}\|_\infty.\end{aligned}$$

Taking the maximum over (s,a) proves the contraction property. Its fixed point is $Q^\star_{\rho,\epsilon_{\mathrm{B}}}$.

Work on the event $\mathcal{E}_{\mathrm{Bell}}$. Define $e_k := \|\widehat{Q}_k - Q^\star_{\rho,\epsilon_{\mathrm{B}}}\|_\infty$. For $k < J$, we have that

$$e_{k+1} \le \|\widehat{Q}_{k+1} - \mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widehat{Q}_k\|_\infty + \|\mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}\widehat{Q}_k - \mathcal{T}_{\rho,\epsilon_{\mathrm{B}}}Q^\star_{\rho,\epsilon_{\mathrm{B}}}\|_\infty \le \frac{(1-\gamma)\epsilon_Q}{2} + \gamma e_k.$$

Iterating this inequality yields

$$e_J \le \gamma^J e_0 + \frac{(1-\gamma)\epsilon_Q}{2}\sum_{j=0}^{J-1}\gamma^j \le \gamma^J e_0 + \frac{\epsilon_Q}{2}.$$

Since $\widehat{Q}_0 = 0$ and $0 \le Q^\star_{\rho,\epsilon_{\mathrm{B}}} \le V_{\max}$, we have that $e_0 \le V_{\max}$. Since $\gamma^J V_{\max} \le \frac{\epsilon_Q}{2}$, we have $\gamma^J e_0 \le \frac{\epsilon_Q}{2}$. Therefore

$$\|\widehat{Q}_J - Q^\star_{\rho,\epsilon_{\mathrm{B}}}\|_\infty \le \epsilon_Q. \tag{75}$$

Step 4: performance of the greedy finite-net policy. Fix a state s and let

$$a^\star(s) \in \operatorname*{arg\,max}_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} Q^\star_{\rho,\epsilon_{\mathrm{B}}}(s,a).$$

By the definition of $\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}$ and (75), we have that

$$\begin{aligned} Q^\star_{\rho,\epsilon_{\mathrm{B}}}\big(s, \widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}(s)\big) &\geq \widehat{Q}_J\big(s, \widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}(s)\big) - \epsilon_Q \\ &= \max_{a \in \mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)} \widehat{Q}_J(s,a) - \epsilon_Q \\ &\geq \widehat{Q}_J\big(s, a^\star(s)\big) - \epsilon_Q \\ &\geq Q^\star_{\rho,\epsilon_{\mathrm{B}}}\big(s, a^\star(s)\big) - 2\epsilon_Q \\ &= V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) - 2\epsilon_Q. \end{aligned}$$

Define the finite-net greedy residual $e_{\mathrm{greedy}}(s) := V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) - Q^\star_{\rho,\epsilon_{\mathrm{B}}}\big(s, \widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}(s)\big)$. The preceding inequality yields $0 \leq e_{\mathrm{greedy}}(s) \leq 2\epsilon_Q$.

Let $D_{\mathrm{greedy}} := V^\star_{\rho,\epsilon_{\mathrm{B}}} - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}}$. Adding and subtracting the finite-net policy Bellman operator applied to $V^\star_{\rho,\epsilon_{\mathrm{B}}}$ gives

$$\begin{aligned} D_{\mathrm{greedy}}(s) &= V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) - (\mathcal{T}^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}})(s) \\ &= V^\star_{\rho,\epsilon_{\mathrm{B}}}(s) - (\mathcal{T}^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} V^\star_{\rho,\epsilon_{\mathrm{B}}})(s) + \gamma(P^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} D_{\mathrm{greedy}})(s) \\ &= e_{\mathrm{greedy}}(s) + \gamma(P^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} D_{\mathrm{greedy}})(s). \end{aligned}$$

Taking the supremum norm yields

$$\|D_{\mathrm{greedy}}\|_\infty \leq 2\epsilon_Q + \gamma \|D_{\mathrm{greedy}}\|_\infty,$$

so that rearranging gives us that

$$\left\| V^\star_{\rho,\epsilon_{\mathrm{B}}} - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} \right\|_\infty \leq 2\epsilon_Q/(1-\gamma). \tag{76}$$

Step 5: combine approximation, estimation, and monotonicity. By Lemma 8, we have that

$$\left\| V^\star - V^\star_{\rho,\epsilon_{\mathrm{B}}} \right\|_\infty \leq (\bar{\Delta}_\rho + \epsilon_{\mathrm{B}})/(1-\gamma).$$

Combining this with (76) and the triangle inequality gives

$$\left\| V^\star - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} \right\|_\infty \leq \left\| V^\star - V^\star_{\rho,\epsilon_{\mathrm{B}}} \right\|_\infty + \left\| V^\star_{\rho,\epsilon_{\mathrm{B}}} - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}} \right\|_\infty \leq (\bar{\Delta}_\rho + \epsilon_{\mathrm{B}} + 2\epsilon_Q)/(1-\gamma).$$

This proves (68) on the event $\mathcal{E}_{\mathrm{Bell}}$, which has probability at least $1-\alpha$.

The sample-count identity follows because iteration k draws m samples for every finite-net state–action pair. Therefore

$$N_{\mathrm{gen}} = Jm \sum_{s \in \mathcal{S}} |\mathcal{A}_{\rho,\epsilon_{\mathrm{B}}}(s)| \leq SJmA_\rho(\epsilon_{\mathrm{B}}).$$

Proposition 2 proves that $\bar{\Delta}_\rho$ is nonincreasing in ρ, while Proposition 6 proves that $A_\rho(\epsilon_{\mathrm{B}})$ is nondecreasing. This proves that $\bar{\Delta}_\rho$ and $A_\rho(\epsilon_{\mathrm{B}})$ are monotonic. The right-hand side of (69) is a nondecreasing function of $A_\rho(\epsilon_{\mathrm{B}})$, so its monotonicity in ρ follows as well.

Finally, substituting (48) into (68) proves (70), and (71) follows from Proposition 7. The lower bound $(\ell_{\mathrm{B}}\rho/(2\epsilon_{\mathrm{B}}))^{d_{\mathrm{eff}}}$ is (56) from Lemma 7. □

Corollary 4 (A target-accuracy reach choice)**.** Fix a desired full-MDP control tolerance $\epsilon_{\mathrm{ctrl}} > 0$ and a confidence level $\alpha \in (0,1)$. Suppose there exist $\rho \geq 0$, $\epsilon_{\mathrm{B}} > 0$ and $\epsilon_Q > 0$ such that

$$\bar{\Delta}_\rho \leq (1-\gamma)\epsilon_{\mathrm{ctrl}}/3, \qquad \epsilon_{\mathrm{B}} \leq (1-\gamma)\epsilon_{\mathrm{ctrl}}/3, \qquad 2\epsilon_Q \leq (1-\gamma)\epsilon_{\mathrm{ctrl}}/3. \tag{77}$$

If J and m satisfy $\gamma^J V_{\max} \leq \epsilon_Q/2$ and (67), then with probability at least $1-\alpha$,

$$\|V^\star - V^{\widehat{\pi}_{\rho,\epsilon_{\mathrm{B}},J}}\|_\infty \leq \epsilon_{\mathrm{ctrl}}.$$

Among two reaches $\rho_1 \le \rho_2$ that both satisfy the support-bias requirement in (77), the smaller reach has no larger behavioral packing complexity:

$$A_{\rho_1}(\epsilon_{\mathrm{B}}) \le A_{\rho_2}(\epsilon_{\mathrm{B}}).$$

Thus, if a smallest radius meeting a prescribed support-bias budget exists, it minimizes the behavioral packing complexity among all larger feasible radii.

Proof. The performance conclusion follows by substituting the three inequalities in (77) into (68). The monotonicity conclusion follows from Proposition 6. □

Corollary 5 (Nested-radius adaptation cost in the latent-mode specialization)**.** Consider the one-decision Bernoulli latent-mode construction of Theorem 4. Suppose a resolution-ϵ_{B} representation of a smaller reach ρ_1 contains A_1 behaviorally distinct modes and a representation of a larger reach $\rho_2 \ge \rho_1$ contains a nested set of $A_2 > A_1$ modes. Fix $\epsilon \in (0, 1/8]$ and $\alpha \in (0, 1/4]$. There exists a zero-bias instance for the smaller reach on which any learner that remains (ϵ, α)-correct over the larger A_2-mode class must collect, in expectation, at least

$$\frac{3(A_2 - A_1)}{256\epsilon^2} \log \frac{3}{4\alpha} \tag{78}$$

online samples from modes admitted by the larger reach but excluded by the smaller reach.

Proof. Identify the A_1 smaller-reach modes with the persistent abstraction in Theorem 4, and identify the remaining $A_2 - A_1$ larger-reach modes with the non-abstraction modes in that theorem. The base instance used in Theorem 4 has its optimal mode inside the smaller set, so the smaller reach has zero abstraction bias. For each of the $A_2 - A_1$ additional modes, the corresponding alternative instance changes only that mode and makes it uniquely ϵ-optimal. Part (b) of Theorem 4, with $C = A_1$, and $N = A_2$, gives exactly (78). □

G.7 Relation to a finite Gaussian particle cloud

The implemented particle construction uses

$$z_{c,k}(s) = z_c(s) + \sigma_p \xi_{c,k}, \qquad \xi_{c,k} \sim \mathcal{N}(0, I_{d_z}), \tag{79}$$

with an exact center particle $z_{c,0}(s) = z_c(s)$. Since a Gaussian distribution has full support for every $\sigma_p > 0$, the scale σ_p does not define a hard support radius. The following proposition, which leverages standard covering arguments (Wainwright, 2019; Vershynin, 2019), instead quantifies how many particles are sufficient to form a Euclidean net of a prescribed reach-ρ envelope.

Proposition 9 (Finite Gaussian coverage of a bounded local envelope)**.** Fix constants

$$\rho > 0, \qquad \epsilon_z > 0, \qquad \sigma_p > 0, \qquad \alpha \in (0, 1).$$

Fix a state $s \in \mathcal{S}$. For every command $c \in \mathcal{C}$, draw M independent particles according to (79). Define

$$p_{\min}(\rho, \epsilon_z, \sigma_p) := \inf_{\|u\|_2 \le \rho} \mathbb{P}\left(\|\sigma_p \xi - u\|_2 \le \frac{\epsilon_z}{2} \right), \qquad \xi \sim \mathcal{N}(0, I_{d_z}).$$

Then $p_{\min}(\rho, \epsilon_z, \sigma_p) > 0$. Let $N_{\rho,\epsilon_z} := (1 + 4\rho/\epsilon_z)^{d_z}$. If

$$M \ge \frac{1}{p_{\min}(\rho, \epsilon_z, \sigma_p)} \log\left(C N_{\rho,\epsilon_z} / \alpha \right), \tag{80}$$

then, with probability at least $1 - \alpha$, simultaneously for every command c, the sampled particles around $z_c(s)$ form a Euclidean ϵ_z-net of $B_2(z_c(s), \rho)$. Moreover, the lower bound holds:

$$p_{\min}(\rho, \epsilon_z, \sigma_p) \ge \mathrm{vol}\left(B_2\left(0, \frac{\epsilon_z}{2}\right) \right) (2\pi\sigma_p^2)^{-d_z/2} \exp\left(-\frac{(\rho + \epsilon_z/2)^2}{2\sigma_p^2} \right). \tag{81}$$

Proof. Let $\{u_1, \dots, u_N\}$ be a Euclidean $\epsilon_z/2$-net of $B_2(0, \rho)$. A standard Euclidean volumetric argument, with resolution $\epsilon_z/2$, permits the choice

$$N \le \left(1 + \frac{4\rho}{\epsilon_z} \right)^{d_z} = N_{\rho,\epsilon_z}.$$

Fix a command c and net point u_j. Each perturbation $\sigma_p \xi_{c,k}$ falls in $B_2(u_j, \epsilon_z/2)$ with probability at least $p_{\min}(\rho, \epsilon_z, \sigma_p)$. Therefore the probability that none of the M particles falls in this ball is at most

$$\big(1 - p_{\min}\big)^M \le \exp(-M p_{\min}),$$

where the final inequality uses $1 - x \le e^{-x}$ for $x \in [0, 1]$. A union bound over all CN command–net-point pairs gives failure probability at most

$$CN \exp(-M p_{\min}) \le C N_{\rho,\epsilon_z} \exp(-M p_{\min}).$$

Under (80), this quantity is at most α.

On the complementary event, fix a command c and arbitrary $u \in B_2(0, \rho)$. Choose a net point u_j satisfying $\|u - u_j\|_2 \le \frac{\epsilon_z}{2}$. By the event just established, there exists a particle perturbation $\sigma_p \xi_{c,k}$ satisfying $\|\sigma_p \xi_{c,k} - u_j\|_2 \le \frac{\epsilon_z}{2}$. The triangle inequality yields $\|\sigma_p \xi_{c,k} - u\|_2 \le \epsilon_z$. Translating by $z_c(s)$ proves that the candidate particles form an ϵ_z-net of $B_2(z_c(s), \rho)$.

It remains to prove (81). Fix any u with $\|u\|_2 \le \rho$. The density of $\sigma_p \xi$ at $x \in \mathbb{R}^{d_z}$ is

$$f_{\sigma_p}(x) = (2\pi\sigma_p^2)^{-d_z/2} \exp\left(-\frac{\|x\|_2^2}{2\sigma_p^2}\right).$$

If $x \in B_2(u, \epsilon_z/2)$, then

$$\|x\|_2 \le \|u\|_2 + \|x - u\|_2 \le \rho + \frac{\epsilon_z}{2}.$$

Hence throughout $B_2(u, \epsilon_z/2)$,

$$f_{\sigma_p}(x) \ge (2\pi\sigma_p^2)^{-d_z/2} \exp\left(-\frac{(\rho + \epsilon_z/2)^2}{2\sigma_p^2}\right).$$

Integrating this lower bound over $B_2(u, \epsilon_z/2)$ yields

$$\mathbb{P}\left(\|\sigma_p \xi - u\|_2 \le \frac{\epsilon_z}{2}\right) \ge \mathrm{vol}\left(B_2\left(0, \frac{\epsilon_z}{2}\right)\right) \cdot (2\pi\sigma_p^2)^{-d_z/2} \exp\left(-\frac{(\rho + \epsilon_z/2)^2}{2\sigma_p^2}\right).$$

The right-hand side is strictly positive and independent of u. Taking the infimum over $\|u\|_2 \le \rho$ proves both positivity and (81). □

Corollary 6 (From Euclidean particle coverage to behavioral coverage)**.** Suppose Assumption 3 holds, so that

$$d_s^{\mathrm{B}}(z, \tilde{z}) \le L_{\mathrm{B}} \|z - \tilde{z}\|_2$$

for every state and pair of latents. Fix a behavioral resolution $\epsilon_{\mathrm{B}} > 0$ and choose a Euclidean resolution $\epsilon_z > 0$ satisfying $L_{\mathrm{B}} \epsilon_z \le \epsilon_{\mathrm{B}}$. If the particle count satisfies (80) with this ϵ_z, then, with probability at least $1 - \alpha$, the sampled particles form a behavioral ϵ_{B}-net of $\mathcal{Z}_\rho(s)$.

Proof. On the event in Proposition 9, for every $z \in \mathcal{Z}_\rho(s)$ there exists a sampled particle $\tilde{z}$ satisfying

$$\|z - \tilde{z}\|_2 \le \epsilon_z.$$

Assumption 3 therefore gives

$$d_s^{\mathrm{B}}(z, \tilde{z}) \le L_{\mathrm{B}} \epsilon_z \le \epsilon_{\mathrm{B}}.$$

This is exactly the definition of a behavioral ϵ_{B}-net. □

From behavioral coverage to representation error. The preceding corollary shows that the finite Gaussian cloud approximates the idealized reach-ρ class geometrically. To connect this coverage statement to the control bounds above, we next translate behavioral approximation into value approximation. Because the Gaussian sampler has unbounded support, a realized particle may lie outside $\mathcal{Z}_\rho(s)$ and may even have value larger than the best action in the reach-ρ class. We therefore measure only the one-sided loss incurred when the finite bank fails to represent the best action available in $\mathcal{Z}_\rho(s)$.

For a realized particle bank $\mathcal{B}_P(s)$, define its one-step representation error relative to the reach-ρ class by

$$\epsilon_{\mathrm{rep}}\big(s, \mathcal{B}_P(s); \rho\big) := \left[\sup_{z \in \mathcal{Z}_\rho(s)} Q^\star(s, z) - \max_{\tilde{z} \in \mathcal{B}_P(s)} Q^\star(s, \tilde{z})\right]_+,$$

where $[x]_+ := \max\{x, 0\}$.

Corollary 7 (Gaussian particles imply bounded representation error)**.** Suppose the conditions of Corollary 6 hold, so that with probability at least $1-\alpha$ the sampled particle bank is a behavioral ϵ_{B}-net of $\mathcal{Z}_\rho(s)$. Then, with probability at least $1-\alpha$, the bound holds: $\epsilon_{\mathrm{rep}}\big(s, \mathcal{B}_P(s); \rho\big) \le \epsilon_{\mathrm{B}}$, which is equivalent to

$$\max_{\tilde{z}\in\mathcal{B}_P(s)} Q^\star(s,\tilde{z}) \ge \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s,z) - \epsilon_{\mathrm{B}}.$$

Proof. Let $\mathcal{E}$ denote the event that $\mathcal{B}_P(s)$ is a behavioral ϵ_{B}-net of $\mathcal{Z}_\rho(s)$. By Corollary 6, we have that $\mathbb{P}(\mathcal{E}) \ge 1-\alpha$. We prove the claimed value bound on $\mathcal{E}$.

Fix an arbitrary $\eta > 0$. By the definition of the supremum, there exists $z_\eta \in \mathcal{Z}_\rho(s)$ such that

$$Q^\star(s,z_\eta) \ge \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s,z) - \eta. \tag{82}$$

Since $\mathcal{E}$ holds and the particle bank is a behavioral ϵ_{B}-net of $\mathcal{Z}_\rho(s)$, there exists $\tilde{z}_\eta \in \mathcal{B}_P(s)$ satisfying $d_s^{\mathrm{B}}(z_\eta, \tilde{z}_\eta) \le \epsilon_{\mathrm{B}}$. By Lemma 5, we have that

$$Q^\star(s,\tilde{z}_\eta) \ge Q^\star(s,z_\eta) - d_s^{\mathrm{B}}(z_\eta,\tilde{z}_\eta) \ge Q^\star(s,z_\eta) - \epsilon_{\mathrm{B}}.$$

Combining this inequality with (82) yields

$$Q^\star(s,\tilde{z}_\eta) \ge \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s,z) - \eta - \epsilon_{\mathrm{B}}.$$

Because $\tilde{z}_\eta$ is a member of the realized particle bank,

$$\max_{\tilde{z}\in\mathcal{B}_P(s)} Q^\star(s,\tilde{z}) \ge \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s,z) - \eta - \epsilon_{\mathrm{B}}.$$

Since $\eta > 0$ was arbitrary, letting $\eta \downarrow 0$ gives

$$\max_{\tilde{z}\in\mathcal{B}_P(s)} Q^\star(s,\tilde{z}) \ge \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s,z) - \epsilon_{\mathrm{B}}.$$

Rearranging and taking the positive part proves $\epsilon_{\mathrm{rep}}\big(s, \mathcal{B}_P(s); \rho\big) \le \epsilon_{\mathrm{B}}$. The result therefore holds on an event of probability at least $1-\alpha$. □

Remark 5 (What the Gaussian coverage result does and does not imply)**.** Proposition 9, Corollary 6, and Corollary 7 provide the bridge from the implemented Gaussian particle mechanism to the deterministic reach-ρ analysis above. In particular, under the stated regularity conditions, the representation-error term may be taken to satisfy $\epsilon_{\mathrm{rep}} \le \epsilon_{\mathrm{B}}$ with probability at least $1-\alpha$. These coverage and representation-error guarantees are stated for a single fixed state s. A simultaneous guarantee over all states requires an additional union bound, for example by replacing α with α/S in the preceding particle-count condition. These results do not identify ρ with σ_p or ϵ_z with ϵ_{B}. The former pair controls envelope size versus random sampling scale; the latter pair is connected only through a behavioral regularity condition such as $d_s^{\mathrm{B}}(z,\tilde{z}) \le L_{\mathrm{B}}\|z-\tilde{z}\|_2$.

The result also shows why finite-particle coverage can be difficult in a high-dimensional isotropic latent space: the lower bound on $p_{\min}$, and hence the required particle count, can deteriorate rapidly with dimension. The parameter d_{eff} in Proposition 7 measures local behavioral richness through a packing lower bound. It does not by itself establish a low-dimensional upper bound on behavioral complexity or improve the Gaussian coverage guarantee.

G.8 Interpretation, novelty, and connection to the preceding results

Theorem 6 describes a reach-parameterized bias–complexity tradeoff connecting the two regimes identified by Theorems 4 and 5.

At zero reach, $\rho = 0$, the local action set is exactly the online flow-reversed anchor bank, so the structural bias is $\bar{\Delta}_F$ and the action class has at most C candidates before any further discretization. Increasing ρ admits local residual corrections. The oracle support gap $\bar{\Delta}_\rho$ can only decrease, and under smooth local geometry it is bounded above by a quadratic function of the residual distance from the nearest anchor to an optimal latent. At the same time, the maximal number of pairwise

behaviorally distinguishable actions at resolution ϵ_{B} can only increase. Under nondegenerate local geometry, every behavioral ϵ_{B}-net must contain at least order $(\rho/\epsilon_{\mathrm{B}})^{d_{\mathrm{eff}}}$ candidates.

The resulting finite-sample sub-optimality guarantee (68) separates three error contributions, each scaled by $1/(1-\gamma)$:

$$\bar{\Delta}_\rho + \epsilon_{\mathrm{B}} + 2\epsilon_Q,$$

namely, the remaining local support bias, behavior discretization error, and finite-sample value-selection error, respectively. The same reach that reduces the first term enlarges the behavioral action class appearing in the sample count. This is the precise sense in which ANCHORQ traces a bias–complexity frontier rather than simply providing more exploration.

The connection to Theorem 4 is statistical. In a one-decision specialization, enlarging the local reach adds latent modes that must be ruled out if the learner remains correct over the larger class. Corollary 5 shows that the additional verification cost is proportional to the number of newly admitted modes, even on a zero-bias instance for the smaller reach.

The connection to Theorem 5 is structural. That theorem is the zero-reach endpoint $\rho = 0$. Proposition 3 applies the same one-step $V^\star - Q^\star$ support-gap logic to every local reach. Particle expansion therefore changes the support term itself; it does not merely improve selection among the original anchors.

The contraction, packing, Höeffding, and approximate-value-iteration arguments are standard. The Q-FRS-specific contribution is their organization around the state-dependent local FRS action family. In particular, the theorem

i. identifies a monotone path from exact-anchor Q-FRS toward broader latent control;

ii. measures complexity in terms of behaviorally distinguishable candidates rather than raw particle count;

iii. places support bias and finite-data selection error in the same full-MDP performance bound;

iv. distinguishes deterministic local reach from the stochastic Gaussian scale used to instantiate a finite particle cloud.

The finite-sample guarantee above applies to synchronous empirical Q-iteration on fixed behavioral nets under a generative model. The Gaussian coverage results connect finite particle clouds to the representation-error term, but do not by themselves provide a sample-complexity guarantee for the implemented hierarchical policy. Theorem 8 separately bounds the suboptimality of that policy, accounting for coarse-anchor selection, particle weighting, critic error, and barycentric refinement.

The reach parameter and the value-tilting parameter have different roles. Reach determines which local behaviors are made available; value tilting reallocates probability within the behaviors that are available. The next section studies the latter operation and its connection to entropic risk.

H ENTROPIC WEIGHTING FOR LOCAL PARTICLE REFINEMENT

Theorem 6 separates the coarse behavioral representation supplied by the flow-reversed anchors from the local degrees of freedom supplied by particles. In the implemented method, a coarse anchor is selected first. Exponential tilting is then applied only to the local particle cloud around that anchor, and the resulting weights are used to construct a weighted latent refinement. This section studies this local weighting operation and its effect on the executed barycentric refinement.

Throughout this section, $\beta \geq 0$ has one meaning only: it is the exponential-tilting intensity, equivalently an inverse temperature. It is distinct from the confidence parameter α, the MDP discount γ, the deterministic local reach ρ, the Gaussian particle scale σ_p, and the flow-integration step size h. We analyze only the positive, reward-seeking tilt $\beta > 0$; no negative or lower-tail tilt is considered.

Notation introduced in this section. To avoid collisions with the preceding results, the probability law before value reweighting is ν_s, its exponential tilt is $\nu^q_{s,\beta}$, $\mathrm{CE}_\beta(q; s)$ is the entropic candidate evaluation, and ϵ_{crit} is a uniform score-estimation error. A realized finite bank is indexed by a finite set $\mathcal{J}_s$; its cardinality is written explicitly as $|\mathcal{J}_s|$ rather than introducing another symbol. None of this notation is used for the deterministic reach R, Gaussian particle scale σ_p, behavioral resolution

ϵ_{B}, Euclidean latent resolution ϵ_z, flow step size h, discount factor γ, confidence level α, or sample count n defined earlier.

Throughout, we will use the standard notation for moment generating functions. For instance, $\mathcal{M}_X(a) = \mathbb{E}[e^{aX}]$, where $\mathcal{M}_X$ is the moment-generating function of the random variable X. For a conditional moment-generating function, we use the notation $\mathcal{M}_{X|Y}(a|y) := \mathbb{E}[e^{aX}|Y = y]$ for random variables X and Y.

H.1 The local particle law and its entropic tilt

Fix a state $s \in \mathcal{S}$ and a coarse anchor $c \in \mathcal{C}$ selected by the anchor-level ANCHORQ rule. We study only the subsequent local refinement around $z_c(s)$. As in the particle-refinement analysis, let $\mathcal{Z} = \mathbb{R}^{d_z}$ equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{Z})$.

Let $\nu_{s,c}$ be a probability measure on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ describing the local particle law around the selected anchor before value-based reweighting. For example, the population particle construction

$$Z_{s,c} = z_c(s) + \sigma_p \xi_a, \qquad \xi_a \sim \mathcal{N}(0, I_{d_z}),$$

induces a Gaussian local law centered at $z_c(s)$. For the implemented finite particle cloud, after conditioning on a realized bank

$$b = \big(z_{c,0}(s), \dots, z_{c,M}(s)\big),$$

the corresponding reference law is the uniform law over its indexed particles,

$$\nu_{s,c,b} := \frac{1}{M+1} \sum_{k=0}^{M} \delta_{z_{c,k}(s)}.$$

The abstract results below are stated for $\nu_{s,c}$; the finite realized-bank specialization is developed in Section H.6.

Assumption 6 (Bounded measurable local score)**.** The score $q : \mathcal{S} \times \mathcal{Z} \to \mathbb{R}$ is measurable. There exist finite constants $q_{\min} < q_{\max}$ such that, for every state and selected anchor under consideration,

$$q_{\min} \le q(s,z) \le q_{\max} \qquad \text{for } \nu_{s,c}\text{-almost every } z.$$

Boundedness is natural when q is an action value in the bounded-reward discounted MDP of Theorems 5 and 6. It guarantees that the exponential moments and derivatives used below are finite. The same results extend under weaker exponential-moment assumptions, but boundedness avoids unnecessary domain qualifications.

Definition 6 (Entropic certainty equivalent[2] of local-particle value)**.** For $\beta > 0$, define

$$\mathrm{CE}_\beta(q; s, c) := \frac{1}{\beta} \log \int_{\mathcal{Z}} \exp\big(\beta q(s,z)\big)\, \nu_{s,c}(\mathrm{d}z).$$

At zero tilt, define $\mathrm{CE}_0(q; s, c) := \int_{\mathcal{Z}} q(s,z)\, \nu_{s,c}(\mathrm{d}z)$.

Corollary 8 below shows that the zero-tilt definition is the continuous extension as $\beta \downarrow 0$. The quantity $\mathrm{CE}_\beta(q; s, c)$ is the reward-oriented entropic certainty equivalent of the values of local refinements around the selected anchor. Positive β places increasing emphasis on the upper tail of those local scores.

Definition 7 (Exponentially tilted local weighting law)**.** For $\beta > 0$, define the probability measure $\nu^q_{s,c,\beta}$ by

$$\frac{\mathrm{d}\nu^q_{s,c,\beta}}{\mathrm{d}\nu_{s,c}}(z) := \exp\left(\beta\left(q(s,z) - \mathrm{CE}_\beta(q; s, c)\right)\right). \tag{83}$$

At zero tilt, set $\nu^q_{s,c,0} := \nu_{s,c}$.

[2]The notation CE denotes a certainty equivalent. If $Y^q_{s,c} := q(s, Z_{s,c})$ with $Z_{s,c} \sim \nu_{s,c}$, then $\mathrm{CE}_\beta(q; s, c)$ is the deterministic scalar satisfying $\exp\big(\beta\, \mathrm{CE}_\beta(q; s, c)\big) = \mathbb{E}\left[\exp\big(\beta Y^q_{s,c}\big)\right]$. Solving for this scalar gives exactly $\mathrm{CE}_\beta(q; s, c) = \beta^{-1} \log \mathbb{E}[\exp(\beta Y^q_{s,c})]$.

The normalization in (83) is exact:

$$\int_{\mathcal{Z}} \exp\left(\beta\left[q(s,z) - \mathrm{CE}_\beta(q;s,c)\right]\right)\nu_{s,c}(\mathrm{d}z)$$
$$= \exp\left(-\beta \mathrm{CE}_\beta(q;s,c)\right)\int_{\mathcal{Z}} e^{\beta q(s,z)}\nu_{s,c}(\mathrm{d}z) = 1.$$

Let $Z_{s,c} \sim \nu_{s,c}$ and define the scalar local-score random variable $Y^q_{s,c} := q(s, Z_{s,c})$. Then

$$\mathrm{CE}_\beta(q;s,c) = \frac{1}{\beta}\log \mathbb{E}\left[e^{\beta Y^q_{s,c}}\right]. \tag{84}$$

Thus the randomness in the entropic evaluation is the randomness of the local particle around the already-selected anchor. Unless q is itself defined through a conditional entropic transform of realized return, $Y^q_{s,c}$ is a random local candidate score, not the environment return.

Risk-sensitive interpretation. Equation (84) is the reward-oriented entropic certainty equivalent of the local particle score $Y^q_{s,c}$. Under the reward convention used here, $\beta > 0$ is upper-tail seeking: increasing β places greater emphasis on high-value local refinements.

Importantly, $\nu^q_{s,c,\beta}$ is an auxiliary *weighting law*, not the distribution of the action executed by AnchorQ. In the implemented finite-cloud method, its probability masses are the exponential weights

$$w_{c,\beta}(k \mid s) \propto \exp\left(\beta Q_\psi(s, z_{c,k}(s))\right),$$

and these weights are used to construct the executed barycentric refinement

$$\tilde{z}_{c,\beta}(s) = \sum_{k=0}^{M} w_{c,\beta}(k \mid s) z_{c,k}(s).$$

Thus the entropic transform governs how strongly each local particle influences the executed refinement; it does not define a policy that samples one of those particles for execution.

Theorem 7 makes the variational interpretation exact. The entropic certainty equivalent and the exponentially tilted weighting law are two views of the same optimization:

$$\mathrm{CE}_\beta(q;s,c) = \sup_{\mu \ll \nu_{s,c}} \left\{\int_{\mathcal{Z}} q(s,z)\,\mu(\mathrm{d}z) - \frac{1}{\beta} D_{\mathrm{KL}}(\mu\|\nu_{s,c})\right\},$$

whose unique optimizer is $\nu^q_{s,c,\beta}$. Hence the exponential weights are the optimal change of measure in the variational representation of the entropic risk functional: they favor high-value local deviations while the KL penalty limits departure from the untilted local particle law.

H.2 The exact exponential-tilting theorem

For probability measures μ and ν on $\mathcal{Z}$, define the extended relative entropy

$$D_{\mathrm{KL}}(\mu\|\nu) := \begin{cases} \displaystyle\int_{\mathcal{Z}} f(z)\log f(z)\,\nu(\mathrm{d}z), & \text{if } \mu \ll \nu \text{ with } f = \mathrm{d}\mu/\mathrm{d}\nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

Here $0 \log 0 := 0$. For $\beta > 0$, define the local KL-regularized score functional

$$\mathcal{F}_{\beta,s,c}(\mu;q) := \int_{\mathcal{Z}} q(s,z)\,\mu(\mathrm{d}z) - \frac{1}{\beta} D_{\mathrm{KL}}(\mu\|\nu_{s,c}),$$

with the convention $\mathcal{F}_{\beta,s,c}(\mu;q) = -\infty$ whenever $D_{\mathrm{KL}}(\mu\|\nu_{s,c}) = +\infty$.

For a Borel set $A \subseteq \mathcal{Z}$ and probability measure μ satisfying $0 < \mu(A) < 1$, define

$$\mathrm{Odds}_\mu(A) := \frac{\mu(A)}{1-\mu(A)}.$$

For $0 < \nu_{s,c}(A) < 1$, define the essential score separation

$$\Xi_{\nu_{s,c}}(A;q) := \operatorname{ess\,inf}^{\nu_{s,c}}_{z\in A} q(s,z) - \operatorname{ess\,sup}^{\nu_{s,c}}_{z\in A^c} q(s,z).$$

Theorem 7 (Entropic local weighting: duality, support, monotonicity, and robust high-value amplification). Suppose Assumption 6 holds. Fix a state $s \in \mathcal{S}$ and an anchor $c \in \mathcal{C}$.

a. **Exact KL-regularized optimizer.** For every $\beta > 0$ and every probability measure μ on $\mathcal{Z}$ satisfying $D_{\mathrm{KL}}(\mu \| \nu_{s,c}) < \infty$, the local KL-regularized score functional satisfies

$$\mathcal{F}_{\beta,s,c}(\mu; q) = \mathrm{CE}_\beta(q; s, c) - \frac{1}{\beta} D_{\mathrm{KL}}\left(\mu \,\middle\|\, \nu^q_{s,c,\beta}\right), \tag{85}$$

so that

$$\mathrm{CE}_\beta(q; s, c) = \sup_\mu \left\{ \int_{\mathcal{Z}} q(s, z)\, \mu(\mathrm{d}z) - \frac{1}{\beta} D_{\mathrm{KL}}(\mu \| \nu_{s,c}) \right\}, \tag{86}$$

and the unique optimizer is $\nu^q_{s,c,\beta}$. In particular,

$$\mathrm{CE}_\beta(q; s, c) = \int_{\mathcal{Z}} q(s, z)\, \nu^q_{s,c,\beta}(\mathrm{d}z) - \frac{1}{\beta} D_{\mathrm{KL}}\left(\nu^q_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right). \tag{87}$$

b. **Support preservation of the weighting law.** For every $\beta > 0$, the measures $\nu^q_{s,c,\beta}$ and $\nu_{s,c}$ are mutually absolutely continuous (i.e., $\nu^q_{s,c,\beta} \sim \nu_{s,c}$) and hence

$$\mathrm{supp}(\nu^q_{s,c,\beta}) = \mathrm{supp}(\nu_{s,c}). \tag{88}$$

Thus exponential tilting neither removes nor adds particles to the local weighting support. This statement concerns the weighting law: the barycentric latent executed by ANCHORQ may lie in the convex hull of this support rather than in the support itself.

c. **Monotone score and KL path.** Define

$$\mathcal{S}_{q,s,c}(\beta) := \int_{\mathcal{Z}} q(s, z)\, \nu^q_{s,c,\beta}(\mathrm{d}z), \quad \text{and} \quad \mathsf{KL}_{q,s,c}(\beta) := D_{\mathrm{KL}}\left(\nu^q_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right).$$

Then the following bounds hold:

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \mathcal{S}_{q,s,c}(\beta) = \mathrm{Var}_{\nu^q_{s,c,\beta}}\left(q(s, Z)\right) \geq 0, \tag{89}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \mathsf{KL}_{q,s,c}(\beta) = \beta \mathrm{Var}_{\nu^q_{s,c,\beta}}\left(q(s, Z)\right) \geq 0, \tag{90}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \mathrm{CE}_\beta(q; s, c) = \frac{\mathsf{KL}_{q,s,c}(\beta)}{\beta^2} \geq 0. \tag{91}$$

Moreover, we have the mean score improvement bound:

$$\mathcal{S}_{q,s,c}(\beta) - \mathcal{S}_{q,s,c}(0) \geq \frac{1}{\beta} \mathsf{KL}_{q,s,c}(\beta) \geq 0. \tag{92}$$

If $q(s, \cdot)$ is not constant $\nu_{s,c}$-almost everywhere, then all three maps are strictly increasing on $(0, \infty)$.

d. **Exponential amplification of high-score weight.** Let $A \in \mathcal{B}(\mathcal{Z})$ satisfy $0 < \nu_{s,c}(A) < 1$. Then

$$\frac{\mathrm{Odds}_{\nu^q_{s,c,\beta}}(A)}{\mathrm{Odds}_{\nu_{s,c}}(A)} = \frac{\mathbb{E}_{\nu_{s,c}(\cdot|A)}\left[e^{\beta q(s,Z)}\right]}{\mathbb{E}_{\nu_{s,c}(\cdot|A^c)}\left[e^{\beta q(s,Z)}\right]}, \tag{93}$$

and therefore

$$\mathrm{Odds}_{\nu^q_{s,c,\beta}}(A) \geq e^{\beta \Xi_{\nu_{s,c}}(A;q)}\, \mathrm{Odds}_{\nu_{s,c}}(A), \tag{94}$$

so that

$$\nu^q_{s,c,\beta}(A) \geq \frac{\nu_{s,c}(A) e^{\beta \Xi_{\nu_{s,c}}(A;q)}}{1 - \nu_{s,c}(A) + \nu_{s,c}(A) e^{\beta \Xi_{\nu_{s,c}}(A;q)}}. \tag{95}$$

Hence a positive score margin $\Xi_{\nu_{s,c}}(A; q) > 0$ yields strict exponential amplification of the weight assigned to A relative to its complement.

e. **Reliability under learned-score error.** Let q_{pop} be a bounded population score and let $\hat{q}$ be a bounded learned score satisfying

$$\|\hat{q}(s,\cdot) - q_{\mathrm{pop}}(s,\cdot)\|_{L^\infty(\nu_{s,c})} \le \epsilon_{\mathrm{crit}}. \tag{96}$$

Then

$$|\mathrm{CE}_\beta(\hat{q}; s, c) - \mathrm{CE}_\beta(q_{\mathrm{pop}}; s, c)| \le \epsilon_{\mathrm{crit}}. \tag{97}$$

Moreover, the bounds hold:

$$e^{-2\beta\epsilon_{\mathrm{crit}}} \le \frac{\mathrm{d}\nu^{\hat{q}}_{s,c,\beta}}{\mathrm{d}\nu^{q_{\mathrm{pop}}}_{s,c,\beta}}(z) \le e^{2\beta\epsilon_{\mathrm{crit}}} \tag{98}$$

for $\nu_{s,c}$-almost every z, and consequently

$$D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \,\middle\|\, \nu^{q_{\mathrm{pop}}}_{s,c,\beta}\right) \le 2\beta\epsilon_{\mathrm{crit}},$$
$$D_{\mathrm{KL}}\left(\nu^{q_{\mathrm{pop}}}_{s,c,\beta} \,\middle\|\, \nu^{\hat{q}}_{s,c,\beta}\right) \le 2\beta\epsilon_{\mathrm{crit}}.$$

The learned weighting law is also $2\epsilon_{\mathrm{crit}}$-optimal for the population KL-regularized objective:

$$\mathcal{F}_{\beta,s,c}\left(\nu^{q_{\mathrm{pop}}}_{s,c,\beta}; q_{\mathrm{pop}}\right) - \mathcal{F}_{\beta,s,c}\left(\nu^{\hat{q}}_{s,c,\beta}; q_{\mathrm{pop}}\right) \le 2\epsilon_{\mathrm{crit}}. \tag{99}$$

Finally, let A satisfy $0 < \nu_{s,c}(A) < 1$ and define its population-score separation[3]

$$\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) := \operatorname{ess\,inf}^{\nu_{s,c}}_{z\in A} q_{\mathrm{pop}}(s,z) - \operatorname{ess\,sup}^{\nu_{s,c}}_{z\in A^c} q_{\mathrm{pop}}(s,z).$$

Then

$$\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A) \ge \exp\left(\beta\left[\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) - 2\epsilon_{\mathrm{crit}}\right]\right)\mathrm{Odds}_{\nu_{s,c}}(A). \tag{100}$$

Therefore strict high-value weight amplification is guaranteed whenever

$$\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) > 2\epsilon_{\mathrm{crit}}. \tag{101}$$

Proof. We prove each part separately.

Part a: exact KL-regularized optimizer. Fix $\beta > 0$. Boundedness gives

$$q_{\min} \le \mathrm{CE}_\beta(q; s, c) \le q_{\max}.$$

Hence

$$0 < e^{-\beta(q_{\max}-q_{\min})} \le \frac{\mathrm{d}\nu^{q}_{s,c,\beta}}{\mathrm{d}\nu_{s,c}}(z) \le e^{\beta(q_{\max}-q_{\min})} < \infty$$

for $\nu_{s,c}$-almost every z. Thus $\nu^{q}_{s,c,\beta} \sim \nu_{s,c}$.

Let μ satisfy $D_{\mathrm{KL}}(\mu\|\nu_{s,c}) < \infty$. Then $\mu \ll \nu^{q}_{s,c,\beta}$ and

$$\log\frac{\mathrm{d}\mu}{\mathrm{d}\nu^{q}_{s,c,\beta}}(z) = \log\frac{\mathrm{d}\mu}{\mathrm{d}\nu_{s,c}}(z) - \beta q(s,z) + \beta\mathrm{CE}_\beta(q; s, c).$$

[3]The quantity $\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}})$ measures the population-score margin separating A from its complement under the local proposal law $\nu_{s,c}$: it is the lowest population score attained on A minus the highest population score attained on A^c, ignoring sets of $\nu_{s,c}$-measure zero. More precisely, the essential infimum is the largest value a such that $q_{\mathrm{pop}}(s,z) \ge a$ for $\nu_{s,c}$-almost every $z \in A$, while the essential supremum is the smallest value b such that $q_{\mathrm{pop}}(s,z) \le b$ for $\nu_{s,c}$-almost every $z \in A^c$. Thus exceptional subsets of A or A^c having zero probability under the local proposal law do not determine the separation margin. Consequently, $\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) > 0$ means that, $\nu_{s,c}$-almost surely, every candidate in A has larger population score than every candidate outside A. For a finite particle bank with positive mass on every particle, the essential infimum over A is the minimum population score among bank particles lying in A, and the essential supremum over A^c is the maximum population score among bank particles lying in A^c. Thus $\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}})$ compares only the represented particles; latent points outside the bank do not affect this separation.

Integrating with respect to μ gives

$$D_{\mathrm{KL}}\left(\mu \,\middle\|\, \nu^q_{s,c,\beta}\right) = D_{\mathrm{KL}}(\mu\|\nu_{s,c}) - \beta \int q(s,z)\,\mu(\mathrm{d}z) + \beta \mathrm{CE}_\beta(q;s,c).$$

Rearranging proves (85). Nonnegativity of KL gives the variational principle, and equality occurs if and only if $\mu = \nu^q_{s,c,\beta}$. This proves uniqueness and (87).

Part b: support preservation of the weighting law. The density bound above is strictly positive and finite $\nu_{s,c}$-almost everywhere, so $\nu^q_{s,c,\beta} \sim \nu_{s,c}$. Mutually absolutely continuous Borel measures have the same null open sets, and therefore the same topological support, proving (88).

Part c: monotone score and KL path. Let $Y^q_{s,c} := q(s, Z_{s,c})$, with $Z_{s,c} \sim \nu_{s,c}$, and define its cumulant-generating function

$$\mathcal{K}_{Y^q_{s,c}}(\beta) := \log \mathcal{M}_{Y^q_{s,c}}(\beta) = \log \int_{\mathcal{Z}} e^{\beta q(s,z)} \nu_{s,c}(\mathrm{d}z). \tag{102}$$

Boundedness justifies differentiation under the integral. Hence

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\mathcal{K}_{Y^q_{s,c}}(\beta) = \mathcal{S}_{q,s,c}(\beta),$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\mathcal{K}_{Y^q_{s,c}}(\beta) = \mathrm{Var}_{\nu^q_{s,c,\beta}}\left(q(s,Z)\right).$$

Differentiating the first identity proves (89).

Furthermore, basic calculus yields

$$\log \frac{\mathrm{d}\nu^q_{s,c,\beta}}{\mathrm{d}\nu_{s,c}}(z) = \beta q(s,z) - \mathcal{K}_{Y^q_{s,c}}(\beta),$$

so that

$$\mathsf{KL}_{q,s,c}(\beta) = \beta \mathcal{S}_{q,s,c}(\beta) - \mathcal{K}_{Y^q_{s,c}}(\beta). \tag{103}$$

Differentiating this expression yields

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\mathsf{KL}_{q,s,c}(\beta) = \beta \mathrm{Var}_{\nu^q_{s,c,\beta}}\left(q(s,Z)\right),$$

proving (90).

Since $\mathrm{CE}_\beta(q;s,c) = \mathcal{K}_{Y^q_{s,c}}(\beta)/\beta$, the quotient rule together with (103) yields

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\mathrm{CE}_\beta(q;s,c) = \frac{\mathsf{KL}_{q,s,c}(\beta)}{\beta^2}.$$

This proves (91).

Finally, applying the variational principle with $\mu = \nu_{s,c}$ yields $\mathrm{CE}_\beta(q;s,c) \geq \mathcal{S}_{q,s,c}(0)$. Combining this with (87) proves (92). Strictness follows whenever $q(s,\cdot)$ is nonconstant $\nu_{s,c}$-almost everywhere.

Part d: exponential amplification of high-score weight. Let A satisfy $0 < \nu_{s,c}(A) < 1$. By definition, we immediately have that

$$\mathrm{Odds}_{\nu^q_{s,c,\beta}}(A) = \frac{\int_A e^{\beta q(s,z)} \nu_{s,c}(\mathrm{d}z)}{\int_{A^c} e^{\beta q(s,z)} \nu_{s,c}(\mathrm{d}z)}.$$

Factoring out the base masses yields

$$\mathrm{Odds}_{\nu^q_{s,c,\beta}}(A) = \mathrm{Odds}_{\nu_{s,c}}(A) \mathbb{E}_{\nu_{s,c}(\cdot|A)}[e^{\beta q(s,Z)}] / \mathbb{E}_{\nu_{s,c}(\cdot|A^c)}[e^{\beta q(s,Z)}],$$

which proves (93).

On A, we have the lower bound

$$q(s,z) \geq \operatorname{ess\,inf}_{u\in A}^{\nu_{s,c}} q(s,u)$$

almost everywhere, whereas on A^c, we have the upper bound

$$q(s,z) \leq \operatorname{ess\,sup}_{u\in A^c}^{\nu_{s,c}} q(s,u)$$

almost everywhere. Substitution gives

$$\operatorname{Odds}_{\nu^q_{s,c,\beta}}(A) \geq e^{\beta \Xi_{\nu_{s,c}}(A;q)} \operatorname{Odds}_{\nu_{s,c}}(A),$$

proving (94). Solving this odds inequality for $\nu^q_{s,c,\beta}(A)$ yields (95).

Part e: reliability under learned-score error. Condition (96) gives

$$q_{\mathrm{pop}}(s,z) - \epsilon_{\mathrm{crit}} \leq \hat{q}(s,z) \leq q_{\mathrm{pop}}(s,z) + \epsilon_{\mathrm{crit}}$$

for $\nu_{s,c}$-almost every z. Therefore

$$e^{-\beta\epsilon_{\mathrm{crit}}} \int e^{\beta q_{\mathrm{pop}}(s,z)} \nu_{s,c}(\mathrm{d}z) \leq \int e^{\beta \hat{q}(s,z)} \nu_{s,c}(\mathrm{d}z) \leq e^{\beta\epsilon_{\mathrm{crit}}} \int e^{\beta q_{\mathrm{pop}}(s,z)} \nu_{s,c}(\mathrm{d}z).$$

Taking logarithms and dividing by β proves (97).

Next, basic calculus yields

$$\log \frac{\mathrm{d}\nu^{\hat{q}}_{s,c,\beta}}{\mathrm{d}\nu^{q_{\mathrm{pop}}}_{s,c,\beta}}(z) = \beta\left(\hat{q}(s,z) - q_{\mathrm{pop}}(s,z)\right) - \beta\left(\mathrm{CE}_\beta(\hat{q};s,c) - \mathrm{CE}_\beta(q_{\mathrm{pop}};s,c)\right).$$

Both differences have absolute value at most ϵ_{crit}, so that

$$-2\beta\epsilon_{\mathrm{crit}} \leq \log \frac{\mathrm{d}\nu^{\hat{q}}_{s,c,\beta}}{\mathrm{d}\nu^{q_{\mathrm{pop}}}_{s,c,\beta}}(z) \leq 2\beta\epsilon_{\mathrm{crit}}.$$

Exponentiating proves (98). Integrating the log-density ratio under either measure gives the two KL bounds.

For any μ of finite KL relative to $\nu_{s,c}$, we have that

$$|\mathcal{F}_{\beta,s,c}(\mu;\hat{q}) - \mathcal{F}_{\beta,s,c}(\mu;q_{\mathrm{pop}})| \leq \epsilon_{\mathrm{crit}}.$$

Applying this first to the population optimizer and then to the learned optimizer yields

$$\mathcal{F}_{\beta,s,c}\left(\nu^{q_{\mathrm{pop}}}_{s,c,\beta};q_{\mathrm{pop}}\right) \leq \mathcal{F}_{\beta,s,c}\left(\nu^{\hat{q}}_{s,c,\beta};q_{\mathrm{pop}}\right) + 2\epsilon_{\mathrm{crit}},$$

which proves (99).

Finally, we have that $\operatorname{ess\,inf}_A^{\nu_{s,c}} \hat{q} \geq \operatorname{ess\,inf}_A^{\nu_{s,c}} q_{\mathrm{pop}} - \epsilon_{\mathrm{crit}}$, and $\operatorname{ess\,sup}_{A^c}^{\nu_{s,c}} \hat{q} \leq \operatorname{ess\,sup}_{A^c}^{\nu_{s,c}} q_{\mathrm{pop}} + \epsilon_{\mathrm{crit}}$. Hence we deduce that

$$\Xi_{\nu_{s,c}}(A;\hat{q}) \geq \Xi_{\nu_{s,c}}(A;q_{\mathrm{pop}}) - 2\epsilon_{\mathrm{crit}}.$$

Applying part *d* to $\hat{q}$ proves (100). If $\Xi_{\nu_{s,c}}(A;q_{\mathrm{pop}}) > 2\epsilon_{\mathrm{crit}}$, the multiplicative factor is strictly larger than one. □

H.3 Risk-sensitive local weighting and the correct random variable

Theorem 7 shows that, conditional on an anchor c, exponential weighting is exactly the optimizer associated with an entropic risk-sensitive objective over the random local-particle score

$$Y^q_{s,c} := q(s, Z_{s,c}), \qquad Z_{s,c} \sim \nu_{s,c}.$$

Under our reward convention, $\beta > 0$ is upper-tail seeking: increasing β places greater emphasis on local particles having large score. Importantly, this statement concerns the auxiliary weighting law $\nu^q_{s,c,\beta}$. ANCHORQ does not sample its executed latent from this law; in the finite-cloud implementation, the corresponding weights are used to construct the barycentric refinement

$$\tilde{z}_{c,\beta}(s) = \sum_{k=0}^{M} w_{c,\beta}(k \mid s)\, z_{c,k}(s).$$

The small-β expansion below makes the upper-tail preference of the weighting objective explicit.

Corollary 8 (Mean–variance expansion with an explicit remainder)**.** Under Assumption 6, let $\mathrm{osc}(q) := q_{\max} - q_{\min}$. Then, for every $\beta > 0$,

$$\left|\mathrm{CE}_\beta(q; s, c) - \mathbb{E}_{\nu_{s,c}}[q(s, Z)] - \frac{\beta}{2}\mathrm{Var}_{\nu_{s,c}}(q(s, Z))\right| \le \frac{\beta^2 \,\mathrm{osc}(q)^3}{6}. \tag{104}$$

Consequently, as $\beta \downarrow 0$, the expansion holds:

$$\mathrm{CE}_\beta(q; s, c) = \mathbb{E}_{\nu_{s,c}}[q(s, Z)] + \frac{\beta}{2}\mathrm{Var}_{\nu_{s,c}}(q(s, Z)) + \mathcal{O}(\beta^2). \tag{105}$$

Proof. Retain the cumulant-generating function $\mathcal{K}_{Y^q_{s,c}}$ introduced in (102). Its first two derivatives at zero are

$$\mathcal{K}'_{Y^q_{s,c}}(0) = \mathbb{E}_{\nu_{s,c}}[q(s, Z)], \qquad \mathcal{K}''_{Y^q_{s,c}}(0) = \mathrm{Var}_{\nu_{s,c}}(q(s, Z)).$$

For any $\theta \ge 0$, differentiating once more, justified by boundedness, gives

$$\mathcal{K}'''_{Y^q_{s,c}}(\theta) = \mathbb{E}_{\nu^q_{s,c,\theta}}\left[\left(q(s, Z) - \mathcal{S}_{q,s,c}(\theta)\right)^3\right].$$

Both $q(s, Z)$ and its expectation under every tilted weighting law lie in $[q_{\min}, q_{\max}]$. Hence $|q(s, Z) - \mathcal{S}_{q,s,c}(\theta)| \le \mathrm{osc}(q)$, and therefore $|\mathcal{K}'''_{Y^q_{s,c}}(\theta)| \le \mathrm{osc}(q)^3$.

Applying Taylor's theorem, for some $\theta_\beta \in (0, \beta)$, yields

$$\mathcal{K}_{Y^q_{s,c}}(\beta) = \mathcal{K}_{Y^q_{s,c}}(0) + \beta\mathcal{K}'_{Y^q_{s,c}}(0) + \frac{\beta^2}{2}\mathcal{K}''_{Y^q_{s,c}}(0) + \frac{\beta^3}{6}\mathcal{K}'''_{Y^q_{s,c}}(\theta_\beta).$$

Since $\mathcal{K}_{Y^q_{s,c}}(0) = 0$ and $\mathrm{CE}_\beta(q; s, c) = \mathcal{K}_{Y^q_{s,c}}(\beta)/\beta$, dividing by β and applying the third-derivative bound proves (104). The asymptotic expansion (105) follows immediately. □

Risk preference as a local refinement mechanism. Conditional on the selected anchor c, the law $\nu_{s,c}$ describes the local variation supplied by the particle construction. The random variable

$$Y^q_{s,c} = q(s, Z_{s,c}), \qquad Z_{s,c} \sim \nu_{s,c},$$

therefore describes variation in score across possible local deviations around that coarse behavior. Equation (105) shows that positive β does more than optimize their mean score: to first order in β, it additionally rewards variance in local scores, and hence the presence of upper-tail value opportunities.

The resulting exponential change of measure converts an undirected local particle cloud into a value-directed weighting of that cloud. In the implemented method, larger weight does not correspond to larger probability of executing an individual particle. Rather, it gives that particle greater influence on the barycentric refinement $\tilde{z}_{c,\beta}(s)$. Thus the three stages of ANCHORQ have distinct roles:

- *FRS anchor generation*: reverse coarse behavior modes to latent anchors.
- *Local particles*: expose local degrees of freedom.
- *Risk-sensitive weighting*: direct the local refinement, and select coarse behavioral mode.

As $\beta \downarrow 0$, the weighting approaches the untilted local law. As β increases, the weighting concentrates progressively on the highest-scoring particles; in a finite cloud with a unique score maximizer, it converges to unit weight on that particle as $\beta \to \infty$.

Score risk and critic uncertainty are distinct. The upper-tail preference across local particle scores should be distinguished from uncertainty in the value estimate of a fixed particle. The random variable in the entropic objective is $q(s, Z_{s,c})$ with $Z_{s,c} \sim \nu_{s,c}$, so its variability is across local candidate behaviors.

A common set of heuristics in practice involve leveraging a critic ensemble. If the score supplied to the weighting rule is itself constructed conservatively from a critic ensemble, for example

$$\hat{q}_{\mathrm{pes}}(s, z) := \min_{i \in \mathcal{I}} Q_i(s, z),$$

then the two operations act along different axes:

- pessimism across value estimates at fixed z: guard against overestimation;
- upper-tail weighting across local particles: favor promising refinements.

The reliability condition from Theorem 7—namely, $\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) > 2\epsilon_{\mathrm{crit}}$—formalizes the corresponding tradeoff: aggressive local weighting is certifiably aligned with the population score only when the value separation of the high-score region dominates critic error.

Beyond entropic weighting. The entropic functional is one member of a broader family of local risk-sensitive objectives. Let $\mathfrak{R}_{\nu_{s,c}}(q)$ be a reward-oriented functional of the random local score $q(s, Z_{s,c})$. Whenever it admits a variational representation

$$\mathfrak{R}_{\nu_{s,c}}(q) = \sup_{\mu \ll \nu_{s,c}} \left\{ \int_{\mathcal{Z}} q(s,z)\, \mu(\mathrm{d}z) - \mathcal{P}_{\mathfrak{R}}(\mu, \nu_{s,c}) \right\},$$

an optimizer $\mu^\star$ defines a corresponding local weighting rule. The entropic case is recovered by

$$\mathcal{P}_{\mathfrak{R}}(\mu, \nu_{s,c}) = \frac{1}{\beta} D_{\mathrm{KL}}(\mu \| \nu_{s,c}),$$

for which $\mu^\star = \nu^q_{s,c,\beta}$ has the closed-form exponential density derived in Theorem 7.

In a population version of the local-refinement construction, such a weighting law naturally induces the barycentric refinement

$$\tilde{z}_{\mathfrak{R}}(s,c) := \int_{\mathcal{Z}} z\, \mu^\star(\mathrm{d}z),$$

whenever the integral exists. Thus alternative risk functionals can change how the same local particle law is combined without changing how FRS constructs the coarse anchor or how the local particle cloud is generated.

Return-level risk. As noted above, the variance in (105) is the variance of local candidate scores under $Z_{s,c} \sim \nu_{s,c}$. It is not automatically return variance, and it is not uncertainty across critic heads. The following proposition identifies the exact score transformation under which the local entropic functional becomes an entropic functional of realized return.

Proposition 10 (Exact return-level entropic interpretation under a conditional transform)**.** Fix a state s, a selected anchor c, and $\beta > 0$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ carry a pair $(Z_{s,c}, G_{s,c})$, where $Z_{s,c} \sim \nu_{s,c}$ and $G_{s,c}$ is a real-valued discounted return satisfying $\mathbb{E}\left[e^{\beta G_{s,c}}\right] < \infty$. Let

$$\mathcal{M}_{G_{s,c}|Z_{s,c}}(\beta \mid z) := \mathbb{E}\left[e^{\beta G_{s,c}} \,\middle|\, Z_{s,c} = z\right]$$

denote the conditional moment-generating function, and define

$$q^c_{\mathrm{ret},\beta}(s,z) := \frac{1}{\beta} \log \mathcal{M}_{G_{s,c}|Z_{s,c}}(\beta \mid z) \tag{106}$$

for $\nu_{s,c}$-almost every z. Then

$$\mathrm{CE}_\beta\left(q^c_{\mathrm{ret},\beta}; s, c\right) = \frac{1}{\beta} \log \mathbb{E}\left[e^{\beta G_{s,c}}\right]. \tag{107}$$

Moreover, with the joint exponentially tilted measure $\mathbb{P}^G_{s,c,\beta}$ defined by

$$\mathrm{d}\mathbb{P}^G_{s,c,\beta} / \mathrm{d}\mathbb{P} := e^{\beta G_{s,c}} / \mathbb{E}[e^{\beta G_{s,c}}],$$

the marginal law of $Z_{s,c}$ under this measure is $\mathcal{L}_{\mathbb{P}^G_{s,c,\beta}}(Z_{s,c}) = \nu^{q^c_{\mathrm{ret},\beta}}_{s,c,\beta}$.

Proof. Since $e^{\beta G_{s,c}} > 0$ almost surely, we have that

$$\mathcal{M}_{G_{s,c}|Z_{s,c}}(\beta \mid Z_{s,c}) = \mathbb{E}\left[e^{\beta G_{s,c}} \,\middle|\, Z_{s,c}\right]$$

is strictly positive almost surely. Furthermore,

$$\mathbb{E}\left[\mathcal{M}_{G_{s,c}|Z_{s,c}}(\beta \mid Z_{s,c})\right] = \mathbb{E}\left[e^{\beta G_{s,c}}\right] < \infty,$$

so the conditional moment is finite almost surely.

By (106), we have

$$e^{\beta q^c_{\mathrm{ret},\beta}(s,Z_{s,c})} = \mathbb{E}\left[e^{\beta G_{s,c}} \,\middle|\, Z_{s,c}\right]$$

almost surely. Therefore, by the tower property,

$$\int_{\mathcal{Z}} e^{\beta q^c_{\mathrm{ret},\beta}(s,z)} \nu_{s,c}(\mathrm{d}z) = \mathbb{E}\left[e^{\beta q^c_{\mathrm{ret},\beta}(s,Z_{s,c})}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\beta G_{s,c}} \,\middle|\, Z_{s,c}\right]\right] = \mathbb{E}\left[e^{\beta G_{s,c}}\right].$$

Taking logarithms and dividing by β proves (107).

For any Borel set $A \subseteq \mathcal{Z}$, elementry calculations yield

$$\begin{aligned}
\mathbb{P}^G_{s,c,\beta}(Z_{s,c} \in A) &= \frac{\mathbb{E}\left[\mathbf{1}\{Z_{s,c} \in A\} e^{\beta G_{s,c}}\right]}{\mathbb{E}[e^{\beta G_{s,c}}]} \\
&= \frac{\mathbb{E}\left[\mathbf{1}\{Z_{s,c} \in A\}\, \mathbb{E}[e^{\beta G_{s,c}} \mid Z_{s,c}]\right]}{\mathbb{E}[e^{\beta G_{s,c}}]} \\
&= \frac{\displaystyle\int_A e^{\beta q^c_{\mathrm{ret},\beta}(s,z)} \nu_{s,c}(\mathrm{d}z)}{\displaystyle\int_{\mathcal{Z}} e^{\beta q^c_{\mathrm{ret},\beta}(s,z)} \nu_{s,c}(\mathrm{d}z)} \\
&= \nu^{q^c_{\mathrm{ret},\beta}}_{s,c,\beta}(A).
\end{aligned}$$

Since this holds for every Borel set A, the marginal laws agree. □

Proposition 10 identifies the precise distinction between local candidate-value risk and realized-return risk. If the score is the ordinary conditional mean return, the outer entropic functional is applied to variation in conditional mean values across local particles. If the score is $q^c_{\mathrm{ret},\beta}$, the same outer transform is exactly the entropic certainty equivalent of the realized return.

This identity concerns the risk functional and its associated exponentially tilted weighting law. In the bounded-return setting, the transformed score is bounded, and Theorem 7 also establishes KL-regularized optimality of this law. It does *not* imply that the actual barycentric latent executed by AnchorQ has the return distribution $\mathbb{P}^G_{s,c,\beta}$, because AnchorQ executes $\tilde{z}_{c,\beta}$ rather than sampling $Z_{s,c}$ from the tilted law. Relating the value of the executed barycenter to the weighted candidate values requires an additional local regularity condition; this connection is developed in Section H.6.

Corollary 9 (Mean-return candidate risk is generally not return risk)**.** In the setting of Proposition 10, assume additionally that $\mathbb{E}[|G_{s,c}|] < \infty$, and define

$$q^c_{\mathrm{mean}}(s,z) := \mathbb{E}\left[G_{s,c} \,\middle|\, Z_{s,c} = z\right].$$

Then

$$\mathrm{CE}_\beta\left(q^c_{\mathrm{mean}}; s, c\right) \le \frac{1}{\beta} \log \mathbb{E}\left[e^{\beta G_{s,c}}\right]. \tag{108}$$

Equality holds if and only if, for $\nu_{s,c}$-almost every z, the conditional return $G_{s,c} \mid Z_{s,c} = z$ is almost surely constant.

Proof. Conditional Jensen's inequality gives

$$e^{\beta q^c_{\mathrm{mean}}(s,Z_{s,c})} = e^{\beta\, \mathbb{E}[G_{s,c} \mid Z_{s,c}]} \le \mathbb{E}\left[e^{\beta G_{s,c}} \,\middle|\, Z_{s,c}\right]$$

almost surely. Taking expectations,

$$\int_{\mathcal{Z}} e^{\beta q^c_{\mathrm{mean}}(s,z)} \nu_{s,c}(\mathrm{d}z) \le \mathbb{E}\left[e^{\beta G_{s,c}}\right].$$

Taking logarithms and dividing by $\beta > 0$ proves (108).

Because $x \mapsto e^{\beta x}$ is strictly convex, equality in conditional Jensen holds almost surely if and only if $G_{s,c}$ is conditionally almost surely constant given $Z_{s,c}$, which gives the stated condition. □

H.4 SHARPNESS OF THE CRITIC-SEPARATION CONDITION

The factor $2\epsilon_{\mathrm{crit}}$ in (101) is not an artifact of the proof. A uniform critic error of size ϵ_{crit} can underestimate the score of a truly high-value local particle by ϵ_{crit} while simultaneously overestimating a lower-value particle by ϵ_{crit}, reducing their apparent separation by $2\epsilon_{\mathrm{crit}}$. The following two-point construction shows that this loss is worst-case sharp.

Proposition 11 (Worst-case sharpness of the local value-separation threshold)**.** Fix a state s, a selected anchor c, a tilt intensity $\beta > 0$, a critic-error level $\epsilon_{\mathrm{crit}} > 0$, and a population-score separation $\vartheta_{\mathrm{score}} \geq 0$. There exists a two-point local particle space

$$\mathcal{Z}_{\mathrm{loc}} = \{z_{\mathrm{hi}}, z_{\mathrm{lo}}\},$$

a local reference law $\nu_{s,c}$ assigning strictly positive mass to both points, a population score q_{pop}, and a learned score $\hat{q}$ such that

$$\|\hat{q}(s,\cdot) - q_{\mathrm{pop}}(s,\cdot)\|_{L^\infty(\nu_{s,c})} = \epsilon_{\mathrm{crit}}, \tag{109}$$

the point z_{hi} has population-score separation exactly $\vartheta_{\mathrm{score}}$ from z_{lo}, and, for $A := \{z_{\mathrm{hi}}\}$,

$$\frac{\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A)}{\mathrm{Odds}_{\nu_{s,c}}(A)} = \exp\left(\beta\left[\vartheta_{\mathrm{score}} - 2\epsilon_{\mathrm{crit}}\right]\right). \tag{110}$$

Consequently, if $\vartheta_{\mathrm{score}} < 2\epsilon_{\mathrm{crit}}$, an admissible learned critic can reverse the ordering of the two local particles and cause exponential weighting to *decrease* the relative weight assigned to the truly high-value particle. At $\vartheta_{\mathrm{score}} = 2\epsilon_{\mathrm{crit}}$, an admissible critic can eliminate any strict preference between them.

Proof. Fix the two-point local particle space $\mathcal{Z}_{\mathrm{loc}} = \{z_{\mathrm{hi}}, z_{\mathrm{lo}}\}$, and let $\nu_{s,c}$ assign strictly positive probability to both points. Set $A := \{z_{\mathrm{hi}}\}$. Define the population score by

$$q_{\mathrm{pop}}(s, z_{\mathrm{hi}}) := \vartheta_{\mathrm{score}}, \qquad q_{\mathrm{pop}}(s, z_{\mathrm{lo}}) := 0.$$

Hence the population-score separation between the designated high-value particle and its complement is exactly

$$\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) = \vartheta_{\mathrm{score}}.$$

Now define the learned score by

$$\hat{q}(s, z_{\mathrm{hi}}) := \vartheta_{\mathrm{score}} - \epsilon_{\mathrm{crit}}, \qquad \hat{q}(s, z_{\mathrm{lo}}) := \epsilon_{\mathrm{crit}}.$$

The absolute score error at each point is exactly ϵ_{crit}, so (109) holds.

The learned score difference is

$$\hat{q}(s, z_{\mathrm{hi}}) - \hat{q}(s, z_{\mathrm{lo}}) = \vartheta_{\mathrm{score}} - 2\epsilon_{\mathrm{crit}}.$$

On a two-point space, the exact odds identity (93) gives

$$\frac{\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A)}{\mathrm{Odds}_{\nu_{s,c}}(A)} = \frac{e^{\beta\hat{q}(s,z_{\mathrm{hi}})}}{e^{\beta\hat{q}(s,z_{\mathrm{lo}})}} = \exp\left(\beta\left[\hat{q}(s, z_{\mathrm{hi}}) - \hat{q}(s, z_{\mathrm{lo}})\right]\right) = \exp\left(\beta\left[\vartheta_{\mathrm{score}} - 2\epsilon_{\mathrm{crit}}\right]\right),$$

proving (110).

If $\vartheta_{\mathrm{score}} < 2\epsilon_{\mathrm{crit}}$, the exponent is negative, so the tilted weighting law assigns smaller odds to the truly high-value point than the untilted local law. If $\vartheta_{\mathrm{score}} = 2\epsilon_{\mathrm{crit}}$, the exponent is zero, so the odds are unchanged. This proves the final claims. □

Thus the threshold $\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) > 2\epsilon_{\mathrm{crit}}$ is necessary for a uniform worst-case guarantee that exponential weighting strictly amplifies the relative weight of a truly high-value local region under only a supremum-norm critic-error assumption. The result concerns the *weighting law*; translating this weighting guarantee into a guarantee for the value of the barycentric latent executed by ANCHORQ additionally requires the local regularity conditions developed in Section H.6. The threshold is not claimed to be necessary on every particular instance.

H.5 TRUE-SCORE IMPROVEMENT UNDER A LEARNED LOCAL WEIGHTING

The robust amplification result in Theorem 7 is local to a separated high-value subset of the particle law. The next result gives a complementary global statement: despite critic error, the learned exponential weighting can improve the population score averaged over the entire local particle distribution.

Corollary 10 (Population-score lower bound for learned local weighting)**.** Under the assumptions of part (e) of Theorem 7, fix a state $s \in \mathcal{S}$ and an anchor $c \in \mathcal{C}$. Then

$$\int_{\mathcal{Z}} q_{\mathrm{pop}}(s,z)\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z) \geq \int_{\mathcal{Z}} q_{\mathrm{pop}}(s,z)\,\nu_{s,c}(\mathrm{d}z) + \frac{1}{\beta} D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right) - 2\epsilon_{\mathrm{crit}}. \tag{111}$$

Consequently, the learned weighting has strictly larger average population score than the untilted local law whenever

$$D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right) > 2\beta\epsilon_{\mathrm{crit}}. \tag{112}$$

Proof. By Theorem 7, $\nu^{\hat{q}}_{s,c,\beta}$ is the unique optimizer of the KL-regularized objective associated with the learned score $\hat{q}$. Since the untilted local law $\nu_{s,c}$ is feasible,

$$\int_{\mathcal{Z}} \hat{q}(s,z)\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z) - \frac{1}{\beta} D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right) \geq \int_{\mathcal{Z}} \hat{q}(s,z)\,\nu_{s,c}(\mathrm{d}z). \tag{113}$$

The uniform critic-error assumption $\|\hat{q}(s,\cdot) - q_{\mathrm{pop}}(s,\cdot)\|_{L^\infty(\nu_{s,c})} \leq \epsilon_{\mathrm{crit}}$ implies

$$\int_{\mathcal{Z}} q_{\mathrm{pop}}(s,z)\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z) \geq \int_{\mathcal{Z}} \hat{q}(s,z)\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z) - \epsilon_{\mathrm{crit}}. \tag{114}$$

Indeed, the integrand satisfies $q_{\mathrm{pop}}(s,z) \geq \hat{q}(s,z) - \epsilon_{\mathrm{crit}}$ $\nu_{s,c}$-almost everywhere, and $\nu^{\hat{q}}_{s,c,\beta} \ll \nu_{s,c}$ by Theorem 7.

Analogously, we have that

$$\int_{\mathcal{Z}} \hat{q}(s,z)\,\nu_{s,c}(\mathrm{d}z) \geq \int_{\mathcal{Z}} q_{\mathrm{pop}}(s,z)\,\nu_{s,c}(\mathrm{d}z) - \epsilon_{\mathrm{crit}}. \tag{115}$$

Rearranging (113) yields

$$\int_{\mathcal{Z}} \hat{q}(s,z)\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z) \geq \int_{\mathcal{Z}} \hat{q}(s,z)\,\nu_{s,c}(\mathrm{d}z) + \frac{1}{\beta} D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right).$$

Combining this inequality with (114) and (115) yields

$$\int_{\mathcal{Z}} q_{\mathrm{pop}}(s,z)\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z) \geq \int_{\mathcal{Z}} q_{\mathrm{pop}}(s,z)\,\nu_{s,c}(\mathrm{d}z) + \frac{1}{\beta} D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \,\middle\|\, \nu_{s,c}\right) - 2\epsilon_{\mathrm{crit}},$$

proving (111).

Under the condition of (112), we have that $\frac{1}{\beta} D_{\mathrm{KL}}(\nu^{\hat{q}}_{s,c,\beta} \| \nu_{s,c}) - 2\epsilon_{\mathrm{crit}} > 0$, so the weighted population score is strictly larger than its value under the untilted local law. □

Corollary 10 is a statement about the *weighted average* of the population scores of local particles. In the implemented ANCHORQ policy, however, the tilted law is used to construct the barycentric latent

$$\tilde{z}_{c,\beta}(s) = \int_{\mathcal{Z}} z\,\nu^{\hat{q}}_{s,c,\beta}(\mathrm{d}z)$$

in the population idealization, provided that $\int_{\mathcal{Z}} \|z\|_2\,\nu_{s,c}(\mathrm{d}z) < \infty$. Under the bounded-score assumptions, the tilted density relative to $\nu_{s,c}$ is bounded, so this condition also ensures that the tilted barycenter exists. It holds for the Gaussian particle law. For a realized finite cloud, the corresponding weighted sum is always well-defined. Therefore (111) does not by itself imply an improvement in $q_{\mathrm{pop}}(s, \tilde{z}_{c,\beta}(s))$. The finite-cloud results below show how local smoothness of the population action value converts the weighted-score guarantee into a guarantee for the executed barycentric refinement.

H.6 FINITE REALIZED LOCAL PARTICLE CLOUDS

The implemented local-refinement rule operates on a finite particle cloud generated around a selected FRS anchor. Since the Gaussian perturbations are resampled at each refinement step, this cloud is random. In this subsection we condition on a realized cloud and analyze the weighting and barycentric refinement induced by that realization. Conditional on the cloud, both the exponential weights and the executed barycentric latent are deterministic. High-probability guarantees over the random Gaussian construction itself are developed separately in Section G.7.

Fix a state $s \in \mathcal{S}$ and an anchor $c \in \mathcal{C}$, and condition on a realized local cloud

$$\mathcal{B}_c(s) := (z_{c,0}(s), z_{c,1}(s), \dots, z_{c,M}(s)).$$

Different particle indices are treated as distinct even if two realized latent vectors coincide.

Set $q_{c,k} := q\big(s, z_{c,k}(s)\big)$ for each $k \in \{0, \dots, M\}$. and let the vector u defined such that $u(k) := 1/(M+1)$ for $k \in \{0, \dots, M\}$ denote uniform weighting of the indexed cloud. Let

$$\Delta_{M+1} := \left\{ w \in [0,1]^{M+1} : \sum_{k=0}^{M} w_k = 1 \right\}.$$

Proposition 12 (Exact entropic weights on a realized local cloud)**.** For every $\beta > 0$, define

$$w_{c,k}^{q,\beta}(s) := \exp\left(\beta q_{c,k}\right) / \textstyle\sum_{\ell=0}^{M} \exp\left(\beta q_{c,\ell}\right), \quad \forall\, k \in \{0, 1, \dots, M\}. \tag{116}$$

Then the finite-cloud entropic evaluation

$$\mathrm{CE}_{\beta}^{\mathrm{cloud}}(q; s, c) := \frac{1}{\beta} \log \left[\frac{1}{M+1} \sum_{k=0}^{M} \exp(\beta q_{c,k}) \right] \tag{117}$$

satisfies

$$\mathrm{CE}_{\beta}^{\mathrm{cloud}}(q; s, c) = \max_{w \in \Delta_{M+1}} \left\{ \sum_{k=0}^{M} w_k q_{c,k} - \frac{1}{\beta} D_{\mathrm{KL}}(w \| u) \right\}, \tag{118}$$

where $D_{\mathrm{KL}}(w\|u) := \sum_{k=0}^{M} w_k \log \frac{w_k}{u(k)}$. The unique optimizer is $w_c^{q,\beta}(s) := \big(w_{c,k}^{q,\beta}(s)\big)_{k=0}^{M}$. In particular, the weights are non-negative:

$$w_{c,k}^{q,\beta}(s) > 0 \qquad \text{for every } k \in \{0, 1, \dots, M\}.$$

Moreover, the bounds hold:

$$\max_{0 \le k \le M} q_{c,k} - \frac{\log(M+1)}{\beta} \le \mathrm{CE}_{\beta}^{\mathrm{cloud}}(q; s, c) \le \max_{0 \le k \le M} q_{c,k}, \tag{119}$$

and

$$0 \le \max_{0 \le k \le M} q_{c,k} - \sum_{k=0}^{M} w_{c,k}^{q,\beta}(s) q_{c,k} \le \frac{\log(M+1)}{\beta}. \tag{120}$$

Proof. Apply Theorem 7 to the finite index set $[M]$, equipped with the uniform probability mass function $u(k) = 1/(M+1)$, and score $k \mapsto q(s, z_{c,k}(s))$. Its exponentially tilted law is

$$w_{c,k}^{q,\beta}(s) = e^{\beta q_{c,k}} / \textstyle\sum_{\ell=0}^{M} e^{\beta q_{c,\ell}},$$

proving (116). The variational identity (118) is the finite-space specialization of (86), and uniqueness follows from Theorem 7. Positivity follows immediately from the exponential form.

Let $\bar{q}_{\max} := \max_{0 \le k \le M} q_{c,k}$. Since every summand in (117) is at most $e^{\beta \bar{q}_{\max}}$,

$$\frac{1}{M+1} \sum_{k=0}^{M} e^{\beta q_{c,k}} \le e^{\beta \bar{q}_{\max}},$$

which gives the upper bound in (119). Since at least one summand equals $e^{\beta \bar{q}_{\max}}$,

$$\frac{1}{M+1}\sum_{k=0}^{M} e^{\beta q_{c,k}} \geq \frac{1}{M+1} e^{\beta \bar{q}_{\max}},$$

giving the lower bound.

Finally, the optimizer identity gives

$$\mathrm{CE}_{\beta}^{\mathrm{cloud}}(q; s, c) = \sum_{k=0}^{M} w_{c,k}^{q,\beta}(s) q_{c,k} - \frac{1}{\beta} D_{\mathrm{KL}}\left(w_{c}^{q,\beta}(s) \,\middle\|\, u\right) \leq \sum_{k=0}^{M} w_{c,k}^{q,\beta}(s) q_{c,k}.$$

Combining this with the lower bound in (119) yields

$$\sum_{k=0}^{M} w_{c,k}^{q,\beta}(s) q_{c,k} \geq \bar{q}_{\max} - \frac{\log(M+1)}{\beta}.$$

The reverse inequality follows because a convex combination cannot exceed the maximum component, proving (120). □

The log-mean-exp in (117) is the optimal value of the *auxiliary entropic weighting problem*. It is not, in general, the value of the barycentric latent executed by ANCHORQ, nor is it used as the Bellman target in the implemented method. The next results connect the entropic weighting guarantees to the actual local refinement.

Corollary 11 (Population-score error of learned local weights)**.** Let $q_{c,k}^{\mathrm{pop}} := q_{\mathrm{pop}}\big(s, z_{c,k}(s)\big)$ be population scores and let $\hat{q}_{c,k} := \hat{q}\big(s, z_{c,k}(s)\big)$ satisfy

$$\max_{0 \leq k \leq M} |\hat{q}_{c,k} - q_{c,k}^{\mathrm{pop}}| \leq \epsilon_{\mathrm{crit}}. \tag{121}$$

Let $\widehat{w}_{c,k}^{\beta}(s) := e^{\beta \hat{q}_{c,k}} / \sum_{\ell=0}^{M} e^{\beta \hat{q}_{c,\ell}}$. Then

$$\max_{0 \leq k \leq M} q_{c,k}^{\mathrm{pop}} - \sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta}(s) q_{c,k}^{\mathrm{pop}} \leq \frac{\log(M+1)}{\beta} + 2\epsilon_{\mathrm{crit}}. \tag{122}$$

Proof. The uniform score-error bound gives

$$\max_{k} q_{c,k}^{\mathrm{pop}} \leq \max_{k} \hat{q}_{c,k} + \epsilon_{\mathrm{crit}}. \tag{123}$$

Proposition 12, applied to the learned scores, gives

$$\max_{k} \hat{q}_{c,k} - \sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta}(s) \hat{q}_{c,k} \leq \frac{\log(M+1)}{\beta}.$$

Finally, we have that

$$\sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta}(s) \hat{q}_{c,k} \leq \sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta}(s) q_{c,k}^{\mathrm{pop}} + \epsilon_{\mathrm{crit}}, \tag{124}$$

because the weights are nonnegative and sum to one. Combining (123)–(124) proves (122). □

From weighted particle values to the executed barycenter. The preceding corollary controls the population score averaged under the learned local weights. ANCHORQ, however, executes the weighted latent

$$\tilde{z}_{c,\beta}(s) := \sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta}(s) z_{c,k}(s).$$

The next proposition gives the regularity condition required to transfer the weighted-score guarantee to this executed latent.

Define the weighted particle variance

$$\mathcal{V}_{c,\beta}(s) := \sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta}(s) \left\| z_{c,k}(s) - \tilde{z}_{c,\beta}(s) \right\|_2^2.$$

Proposition 13 (Value of the executed barycentric refinement)**.** Suppose $q_{\mathrm{pop}}(s,\cdot)$ is differentiable on a neighborhood of $\operatorname{conv}\mathcal{B}_c(s)$ and has L_{sm}-Lipschitz gradient there. Consequently, for every $z,\tilde{z}\in\operatorname{conv}\mathcal{B}_c(s)$, the bound holds:

$$q_{\mathrm{pop}}(s,\tilde{z})\le q_{\mathrm{pop}}(s,z)+\nabla_z q_{\mathrm{pop}}(s,z)^\top(\tilde{z}-z)+\frac{L_{\mathrm{sm}}}{2}\|\tilde{z}-z\|_2^2. \tag{125}$$

Then

$$q_{\mathrm{pop}}\left(s,\tilde{z}_{c,\beta}(s)\right)\ge\sum_{k=0}^{M}\widehat{w}_{c,k}^{\beta}(s)q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)-\frac{L_{\mathrm{sm}}}{2}\mathcal{V}_{c,\beta}(s). \tag{126}$$

Moreover, under the uniform critic-error condition (121),

$$\max_{0\le k\le M}q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)-q_{\mathrm{pop}}\left(s,\tilde{z}_{c,\beta}(s)\right)\le\frac{\log(M+1)}{\beta}+2\epsilon_{\mathrm{crit}}+\frac{L_{\mathrm{sm}}}{2}\mathcal{V}_{c,\beta}(s). \tag{127}$$

Proof. For each k, apply (125) with $z=\tilde{z}_{c,\beta}(s)$ and $\tilde{z}=z_{c,k}(s)$. This yields

$$\begin{aligned}q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)\le&q_{\mathrm{pop}}\left(s,\tilde{z}_{c,\beta}(s)\right)+\nabla_z q_{\mathrm{pop}}\left(s,\tilde{z}_{c,\beta}(s)\right)^\top\left(z_{c,k}(s)-\tilde{z}_{c,\beta}(s)\right)\\&+\frac{L_{\mathrm{sm}}}{2}\left\|z_{c,k}(s)-\tilde{z}_{c,\beta}(s)\right\|_2^2.\end{aligned}$$

Multiply by $\widehat{w}_{c,k}^{\beta}(s)$ and sum over $k\in\{0,\dots,M\}$ so that, by the definition of the barycenter, we deduce that

$$\sum_{k=0}^{M}\widehat{w}_{c,k}^{\beta}(s)\left(z_{c,k}(s)-\tilde{z}_{c,\beta}(s)\right)=0.$$

Hence the first-order term vanishes, yielding

$$\sum_{k=0}^{M}\widehat{w}_{c,k}^{\beta}(s)q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)\le q_{\mathrm{pop}}\left(s,\tilde{z}_{c,\beta}(s)\right)+\frac{L_{\mathrm{sm}}}{2}\mathcal{V}_{c,\beta}(s).$$

Rearranging proves (126).

Corollary 11 gives

$$\max_k q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)\le\sum_{k=0}^{M}\widehat{w}_{c,k}^{\beta}(s)q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)+\frac{\log(M+1)}{\beta}+2\epsilon_{\mathrm{crit}}.$$

Substituting (126) proves (127). □

Connection to the ideal reach-ρ class. The preceding proposition compares the executed barycentric refinement with the best particle in the realized local cloud. The particle theory of Section G, however, is stated relative to the ideal reach-ρ neighborhood around the selected anchor. We therefore separate the remaining gap into two pieces: the error from representing this continuous local neighborhood by finitely many sampled particles, and the error from combining those particles through learned entropic weights. The following corollary makes this decomposition explicit.

Corollary 12 (From ideal local reach to the executed refinement)**.** Let $\mathcal{Z}_{\rho,c}(s):=B_2\left(z_c(s),\rho\right)$ denote the ideal reach-ρ neighborhood of the selected anchor. Define the realized local representation error

$$\epsilon_{\mathrm{rep}}\left(s,c,\mathcal{B}_c(s);\rho\right):=\left[\sup_{z\in\mathcal{Z}_{\rho,c}(s)}q_{\mathrm{pop}}(s,z)-\max_{0\le k\le M}q_{\mathrm{pop}}\left(s,z_{c,k}(s)\right)\right]_+. \tag{128}$$

Under the smoothness assumption of Proposition 13 and the uniform critic-error condition 121, for $\beta>0$, we have that

$$\sup_{z\in\mathcal{Z}_{\rho,c}(s)}q_{\mathrm{pop}}(s,z)-q_{\mathrm{pop}}\left(s,\tilde{z}_{c,\beta}(s)\right)\le\epsilon_{\mathrm{rep}}\left(s,c,\mathcal{B}_c(s);\rho\right)+\frac{\log(M+1)}{\beta}+2\epsilon_{\mathrm{crit}}+\frac{L_{\mathrm{sm}}}{2}\mathcal{V}_{c,\beta}(s).$$

Proof. By definition of (128), we have that

$$\sup_{z\in\mathcal{Z}_{\rho,c}(s)} q_{\mathrm{pop}}(s,z) \le \max_k q_{\mathrm{pop}}\left(s, z_{c,k}(s)\right) + \epsilon_{\mathrm{rep}}\left(s, c, \mathcal{B}_c(s); \rho\right).$$

Subtract $q_{\mathrm{pop}}(s, \tilde{z}_{c,\beta}(s))$ from both sides and apply (127) to conclude the proof. □

Thus, conditional on the selected anchor, the gap between the best behavior available in the ideal reach-ρ neighborhood and the latent actually executed by ANCHORQ decomposes into finite-particle representation error, critic error, finite-temperature weighting error, and barycentric interpolation error.

H.7 SUB-OPTIMALITY BOUND FOR HIERARCHICAL ANCHORQ

The preceding results characterize local refinement within any fixed anchor neighborhood. We now combine this local analysis with the support-gap decomposition from Theorems 5 and 6 to obtain a sub-optimality bound for the complete hierarchical ANCHORQ policy. At each state, the policy samples a local particle cloud around every FRS anchor, forms one value-tilted barycentric candidate per cloud, and then uses the critic to select among these refined candidates. Only the selected candidate is executed through the frozen VLA.

For each state $s \in \mathcal{S}$ and anchor $c \in \mathcal{C}$, define the ideal reach-ρ neighborhood $\mathcal{Z}_{\rho,c}(s) := B_2(z_c(s), \rho)$ and its best unrestricted action value

$$Q^\star_{\rho,c}(s) := \sup_{z\in\mathcal{Z}_{\rho,c}(s)} Q^\star(s,z).$$

The union of these local neighborhoods is $\mathcal{Z}_\rho(s) := \bigcup_{c\in\mathcal{C}} \mathcal{Z}_{\rho,c}(s)$, so

$$Q^\star_\rho(s) := \sup_{z\in\mathcal{Z}_\rho(s)} Q^\star(s,z) = \max_{c\in\mathcal{C}} Q^\star_{\rho,c}(s).$$

Recall the corresponding support gap

$$\Delta_\rho(s) := V^\star(s) - Q^\star_\rho(s).$$

Let

$$\mathbf{B}_s := (\mathbf{B}_{s,c})_{c\in\mathcal{C}}, \qquad \mathbf{B}_{s,c} := (Z_{s,c,0}, \dots, Z_{s,c,M}),$$

denote the joint collection of random local particle clouds, where $Z_{s,c,0} = z_c(s)$. For a realization $b = (b_c)_{c\in\mathcal{C}}$, with $b_c = (z_{c,0}, \dots, z_{c,M})$, define

$$\widehat{w}_{c,k}(s,b) := \frac{\exp(\beta Q_\psi(s, z_{c,k}))}{\sum_{\ell=0}^M \exp(\beta Q_\psi(s, z_{c,\ell}))}, \qquad k = 0, \dots, M,$$

where $\beta > 0$. The corresponding local barycentric candidate is

$$\mathcal{T}_{\beta,c}(s,b) := \sum_{k=0}^M \widehat{w}_{c,k}(s,b) z_{c,k}.$$

Weights are normalized separately within each cloud; particles associated with different anchors are not pooled into a common tilted distribution.

After constructing all candidates, the greedy selector chooses

$$\widehat{c}(s,b) \in \arg\max_{c\in\mathcal{C}} Q_\psi(s, \mathcal{T}_{\beta,c}(s,b)), \tag{129}$$

with a fixed measurable tie-breaking rule. The executed latent is

$$\mathcal{T}_\beta(s,b) := \mathcal{T}_{\beta,\widehat{c}(s,b)}(s,b).$$

Unlike a selector applied to the unrefined anchors, $\widehat{c}(s,b)$ depends on the realized particle clouds through their refined candidates.

Define the resulting *refined-candidate selection error* by

$$\epsilon_{\mathrm{sel}}(s,b) := \max_{c\in\mathcal{C}} Q^\star(s, \mathcal{T}_{\beta,c}(s,b)) - Q^\star(s, \mathcal{T}_\beta(s,b)) \ge 0.$$

This measures the loss from choosing among the realized refined candidates using the learned critic rather than their true values. Also define the critic error at these candidates by

$$\epsilon_{\mathrm{crit,sel}}(s,b) := \max_{c\in\mathcal{C}} \left|Q_\psi(s,\mathcal{T}_{\beta,c}(s,b)) - Q^\star(s,\mathcal{T}_{\beta,c}(s,b))\right|.$$

Greedy selection implies

$$\epsilon_{\mathrm{sel}}(s,b) \le 2\epsilon_{\mathrm{crit,sel}}(s,b). \tag{130}$$

Accuracy at the sampled particles alone does not imply accuracy at their barycenters, so this error is distinguished from the particle-level error below.

For the joint realization b, define the representation error relative to the union of ideal local neighborhoods by

$$\epsilon_{\mathrm{rep}}(s,b) := \left[Q^\star_\rho(s) - \max_{\substack{c\in\mathcal{C}\\ 0\le k\le M}} Q^\star(s,z_{c,k}) \right]_+,$$

and the maximum critic error on the sampled particles by

$$\epsilon_{\mathrm{crit}}(s,b) := \max_{\substack{c\in\mathcal{C}\\ 0\le k\le M}} \left|Q_\psi(s,z_{c,k}) - Q^\star(s,z_{c,k})\right|.$$

The maximum over all sampled particles in the representation error is an analytical benchmark, not an additional action-selection step. The positive part keeps the comparison valid even when particles fall outside the ideal reach-ρ neighborhoods.

For each cloud, define its tilted particle variance by

$$\mathcal{V}_{\beta,c}(s,b) := \sum_{k=0}^{M} \widehat{w}_{c,k}(s,b)\,\|z_{c,k} - \mathcal{T}_{\beta,c}(s,b)\|_2^2,$$

and use the uniform within-cloud bound

$$\mathcal{V}_\beta(s,b) := \max_{c\in\mathcal{C}} \mathcal{V}_{\beta,c}(s,b).$$

This is the largest local tilted variance, not the variance of a distribution that mixes particles across anchors.

Assumption 7. Consider the discounted latent-action MDP induced by the frozen VLA, with bounded rewards and discount factor $\gamma \in [0,1)$. All sampled particles and refined candidates are admissible latent actions. For every state s and almost every realized joint collection b, the function $Q^\star(s,\cdot)$ has L_{sm}-Lipschitz gradient on $\mathrm{conv}(b_c)$ for every $c \in \mathcal{C}$.

Theorem 8 (Sub-optimality decomposition for hierarchical ANCHORQ)**.** Suppose Assumption 7 holds. Fix the learned critic, the state-conditioned particle-sampling rule, and $\beta > 0$. Define the stationary randomized policy π_{A} to sample $\mathbf{B}_s$, form all local barycentric candidates, and execute $\mathcal{T}_\beta(s,\mathbf{B}_s)$ using the greedy selector in (129). Then, for every state s,

$$\begin{aligned} 0 &\le V^\star(s) - \mathbb{E}_{\mathbf{B}_s}\left[Q^\star(s,\mathcal{T}_\beta(s,\mathbf{B}_s))\right] \\ &\le \Delta_\rho(s) + \mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{sel}}(s,\mathbf{B}_s)] + \mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{rep}}(s,\mathbf{B}_s)] \\ &\quad + \frac{\log(M+1)}{\beta} + 2\mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{crit}}(s,\mathbf{B}_s)] + \frac{L_{\mathrm{sm}}}{2}\mathbb{E}_{\mathbf{B}_s}[\mathcal{V}_\beta(s,\mathbf{B}_s)]. \end{aligned} \tag{131}$$

Define

$$\begin{aligned} \bar{\Delta}_\rho &:= \sup_s \Delta_\rho(s), & \bar{\epsilon}_{\mathrm{sel}} &:= \sup_s \mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{sel}}(s,\mathbf{B}_s)], \\ \bar{\epsilon}_{\mathrm{rep}} &:= \sup_s \mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{rep}}(s,\mathbf{B}_s)], & \bar{\epsilon}_{\mathrm{crit}} &:= \sup_s \mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{crit}}(s,\mathbf{B}_s)], \\ \bar{\mathcal{V}}_\beta &:= \sup_s \mathbb{E}_{\mathbf{B}_s}[\mathcal{V}_\beta(s,\mathbf{B}_s)], & \bar{\epsilon}_{\mathrm{crit,sel}} &:= \sup_s \mathbb{E}_{\mathbf{B}_s}[\epsilon_{\mathrm{crit,sel}}(s,\mathbf{B}_s)]. \end{aligned}$$

Whenever these quantities are finite, the policy satisfies

$$\|V^\star - V^{\pi_{\mathrm{A}}}\|_\infty \le \frac{\bar{\Delta}_\rho + \bar{\epsilon}_{\mathrm{sel}} + \bar{\epsilon}_{\mathrm{rep}} + \beta^{-1}\log(M+1) + 2\bar{\epsilon}_{\mathrm{crit}} + \frac{L_{\mathrm{sm}}}{2}\bar{\mathcal{V}}_\beta}{1-\gamma}. \tag{132}$$

Moreover, $\bar{\epsilon}_{\mathrm{sel}} \le 2\bar{\epsilon}_{\mathrm{crit,sel}}$, yielding a bound entirely in terms of the support gap, representation error, temperature, particle-level critic error, candidate-level critic error, and within-cloud variance.

Proof. Fix a state s and a joint cloud realization b for which Assumption 7 holds. For each anchor c, optimality of the Gibbs weights in the KL-regularized objective, compared with a point mass on a highest-scoring particle, gives

$$\sum_{k=0}^{M} \widehat{w}_{c,k}(s,b) Q_\psi(s, z_{c,k}) \geq \max_{0\leq k\leq M} Q_\psi(s, z_{c,k}) - \frac{\log(M+1)}{\beta}.$$

Using the particle-level critic error twice therefore yields

$$\sum_{k=0}^{M} \widehat{w}_{c,k}(s,b) Q^\star(s, z_{c,k}) \geq \max_{0\leq k\leq M} Q^\star(s, z_{c,k}) - \frac{\log(M+1)}{\beta} - 2\epsilon_{\mathrm{crit}}(s,b).$$

Smoothness on $\mathrm{conv}(b_c)$ and the barycenter identity $\sum_k \widehat{w}_{c,k}(z_{c,k} - \mathcal{T}_{\beta,c}) = 0$ imply

$$Q^\star(s, \mathcal{T}_{\beta,c}(s,b)) \geq \sum_{k=0}^{M} \widehat{w}_{c,k}(s,b) Q^\star(s, z_{c,k}) - \frac{L_{\mathrm{sm}}}{2} \mathcal{V}_{\beta,c}(s,b).$$

Consequently, for every c,

$$\begin{aligned} Q^\star(s, \mathcal{T}_{\beta,c}(s,b)) \geq & \max_{0\leq k\leq M} Q^\star(s, z_{c,k}) - \frac{\log(M+1)}{\beta} \\ & - 2\epsilon_{\mathrm{crit}}(s,b) - \frac{L_{\mathrm{sm}}}{2} \mathcal{V}_{\beta,c}(s,b). \end{aligned}$$

Apply this inequality to an anchor whose cloud contains a highest-valued particle among all clouds. Since $\mathcal{V}_{\beta,c}(s,b) \leq \mathcal{V}_\beta(s,b)$, it follows that

$$\begin{aligned} \max_{c\in\mathcal{C}} Q^\star(s, \mathcal{T}_{\beta,c}(s,b)) \geq & \max_{\substack{c\in\mathcal{C} \\ 0\leq k\leq M}} Q^\star(s, z_{c,k}) - \frac{\log(M+1)}{\beta} \\ & - 2\epsilon_{\mathrm{crit}}(s,b) - \frac{L_{\mathrm{sm}}}{2} \mathcal{V}_\beta(s,b). \end{aligned} \tag{133}$$

For the complete policy, add and subtract the best ideal-neighborhood value, the best sampled-particle value, and the best refined-candidate value:

$$\begin{aligned} & V^\star(s) - Q^\star(s, \mathcal{T}_\beta(s,b)) \\ & = \Delta_\rho(s) + \left[Q^\star_\rho(s) - \max_{\substack{c\in\mathcal{C} \\ 0\leq k\leq M}} Q^\star(s, z_{c,k}) \right] \\ & \quad + \left[\max_{\substack{c\in\mathcal{C} \\ 0\leq k\leq M}} Q^\star(s, z_{c,k}) - \max_{c\in\mathcal{C}} Q^\star(s, \mathcal{T}_{\beta,c}(s,b)) \right] + \epsilon_{\mathrm{sel}}(s,b). \end{aligned}$$

Bound the first bracket by its positive part $\epsilon_{\mathrm{rep}}(s,b)$ and the second using (133). This gives

$$\begin{aligned} V^\star(s) - Q^\star(s, \mathcal{T}_\beta(s,b)) \leq & \Delta_\rho(s) + \epsilon_{\mathrm{sel}}(s,b) + \epsilon_{\mathrm{rep}}(s,b) \\ & + \frac{\log(M+1)}{\beta} + 2\epsilon_{\mathrm{crit}}(s,b) + \frac{L_{\mathrm{sm}}}{2} \mathcal{V}_\beta(s,b). \end{aligned}$$

Taking expectation over the joint cloud collection proves the upper bound in (131). Its lower bound follows from $Q^\star(s,z) \leq V^\star(s)$ for every admissible latent z.

To verify (130), let c^+ maximize the true value among the realized refined candidates. Then

$$\begin{aligned} \epsilon_{\mathrm{sel}}(s,b) &= Q^\star\big(s, \mathcal{T}_{\beta,c^+}(s,b)\big) - Q^\star\big(s, \mathcal{T}_{\beta,\widehat{c}(s,b)}(s,b)\big) \\ &\leq Q_\psi\big(s, \mathcal{T}_{\beta,c^+}(s,b)\big) - Q_\psi\big(s, \mathcal{T}_{\beta,\widehat{c}(s,b)}(s,b)\big) + 2\epsilon_{\mathrm{crit,sel}}(s,b) \\ &\leq 2\epsilon_{\mathrm{crit,sel}}(s,b), \end{aligned}$$

where the last inequality uses greedy selection.

Finally, let $\mathcal{T}^{\pi_A}$ and P^{π_A} denote the Bellman and transition operators of the stationary policy. Since

$$Q^\star(s,z) = r(s,z) + \gamma \mathbb{E}_{S' \sim P(\cdot|s,z)}[V^\star(S')],$$

we have

$$\mathcal{T}^{\pi_A} V^\star(s) = \mathbb{E}_{\mathbf{B}_s}\left[Q^\star(s, \mathcal{T}_\beta(s, \mathbf{B}_s))\right].$$

Define $\xi_a := V^\star - \mathcal{T}^{\pi_A} V^\star$. The policy Bellman equation gives

$$V^\star - V^{\pi_A} = \xi_a + \gamma P^{\pi_A}(V^\star - V^{\pi_A}),$$

and hence

$$V^\star - V^{\pi_A} = (I - \gamma P^{\pi_A})^{-1} \xi_a.$$

Because P^{π_A} is a Markov operator,

$$\left\|(I - \gamma P^{\pi_A})^{-1}\right\|_{\infty \to \infty} \le \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}.$$

Taking the supremum of the one-step bound over states establishes (132). Taking expectations and suprema in (130) also gives $\bar{\epsilon}_{\mathrm{sel}} \le 2\bar{\epsilon}_{\mathrm{crit,sel}}$. □

Theorem 8 separates the performance loss into errors arising at each stage of the refinement-first construction. The term $\bar{\Delta}_\rho$ is structural: it measures the largest one-step gap between the unrestricted optimum and the best latent in the union of reach-ρ neighborhoods around the FRS anchors. The representation term $\bar{\epsilon}_{\mathrm{rep}}$ measures how well the joint collection of finite particle clouds represents the best value in this union. There is no additional term for committing to one anchor before refinement, because every anchor is refined before selection.

The remaining terms distinguish local refinement from final selection. The finite-temperature term $\log(M+1)/\beta$ is the price of soft entropic weighting within each cloud. It depends on the number of particles per cloud, not the total number across anchors, because tilting is local. The term $2\bar{\epsilon}_{\mathrm{crit}}$ captures the effect of particle-level critic errors on the tilted weights. The variance term $(L_{\mathrm{sm}}/2)\bar{\mathcal{V}}_\beta$ accounts for the difference between the weighted average of particle values and the value of their weighted-average latent. Smoothness is needed only within each cloud's convex hull, not between different anchor neighborhoods.

The term $\bar{\epsilon}_{\mathrm{sel}}$ is the price of choosing among the resulting barycentric candidates using the learned critic. It vanishes when the selector chooses a truly highest-valued refined candidate and is at most $2\bar{\epsilon}_{\mathrm{crit,sel}}$. Particle-level and candidate-level critic errors are distinct because a barycenter need not be one of the sampled particles. Both are measured relative to the unrestricted optimal action value $Q^\star$. They can include estimation error and bias from the candidate-based bootstrap, and need not vanish merely because the critic accurately fits its Bellman target. The theorem is therefore a performance decomposition conditional on accuracy relative to $Q^\star$, not a guarantee that training drives these terms to zero.

The bound exposes complementary tradeoffs. Increasing the ideal reach ρ can reduce the structural support gap but enlarges the neighborhood union that the particles must represent. For fixed anchors and realized clouds, representation error is nondecreasing in ρ; adding particles while retaining the existing clouds cannot increase that representation error. These statements concern the representation term, not monotonic improvement of the entire bound or the executed policy. Increasing β reduces the explicit soft-weighting term, but reliable refinement still requires sufficiently accurate local value rankings, as quantified by Theorem 7. Thus good performance requires both useful local candidates and accurate selection among their refinements.

The same pointwise argument accommodates a common adaptive tilt $\beta(s,b) > 0$ shared across anchors. In that case, all weights, barycenters, and variances use the realized tilt, and the temperature term in the uniform value bound becomes

$$\sup_s \mathbb{E}_{\mathbf{B}_s}\left[\frac{\log(M+1)}{\beta(s,\mathbf{B}_s)}\right],$$

provided it is finite. This allows uncertainty-dependent tilting without changing the decomposition or dispensing with either critic-error term.

The theorem evaluates greedy selection after refinement. For an ϵ-greedy execution rule that selects an anchor uniformly with probability ϵ, define

$$D_{\mathrm{unif}}(s,b) := \max_{c\in\mathcal{C}} Q^\star(s,\mathcal{T}_{\beta,c}(s,b)) - \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} Q^\star(s,\mathcal{T}_{\beta,c}(s,b))\,.$$

The same decomposition applies to the resulting randomized policy after replacing the conditional selection error by

$$\epsilon^{\epsilon}_{\mathrm{sel}}(s,b) := (1-\epsilon)\epsilon_{\mathrm{sel}}(s,b) + \epsilon D_{\mathrm{unif}}(s,b),$$

and averaging over the selection randomness as well as the particle clouds. This makes the additional exploration loss explicit rather than attributing it to critic error.

H.8 Connection to Theorems 1–3

The preceding results complete the coarse-to-fine picture developed in Theorems 4–6. Those results characterize the action representation made available to online learning; the present section characterizes how value information is used to refine that representation after a coarse anchor has been selected.

Connection to Theorem 4. Theorem 4 isolates the statistical benefit of restricting the coarse online decision to a small set of state-conditioned FRS anchors. The local weighting mechanism does not alter this first-stage identification problem: AnchorQ still selects among the C coarse anchors before constructing a local refinement. Particles then restore additional expressivity only around the selected anchor.

Conditional on this local cloud, Theorem 7 shows that value information directs the refinement toward reliably better local particles. In particular, for a high-value local region A,

$$\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A) \geq \exp\left(\beta\left[\Xi_{\nu_{s,c}}(A;q_{\mathrm{pop}}) - 2\epsilon_{\mathrm{crit}}\right]\right)\mathrm{Odds}_{\nu_{s,c}}(A). \tag{134}$$

Thus the FRS construction reduces the coarse search problem, while particles and value weighting introduce a controlled local refinement problem. The latter is useful only once the critic separates genuinely useful refinements from inferior ones with sufficient accuracy.

Connection to Theorem 5. Theorem 5 shows that restricting the available latent actions incurs a structural support price. Value weighting cannot by itself remove this price. It changes the relative influence of particles within each realized local cloud and can thereby affect which refined anchor candidate is ultimately selected, but it does not enlarge the union of local regions accessible around the FRS anchors.

For each anchor c, the barycentric refinement

$$\tilde{z}_{c,\beta}(s) = \sum_{k=0}^{M} w_{c,\beta}(k\,|\,s), z_{c,k}(s)$$

need not coincide with any sampled particle; rather, it lies in their convex hull. Consequently, if the realized local cloud around anchor c lies in the convex reach-ρ neighborhood $B_2(z_c(s),\rho)$, then

$$\tilde{z}_{c,\beta}(s) \in B_2(z_c(s),\rho).$$

Thus, if this condition holds for every anchor, the candidate ultimately selected for execution lies in

$$\mathcal{Z}_\rho(s) = \bigcup_{c\in\mathcal{C}} B_2(z_c(s),\rho).$$

Local value weighting can therefore improve the refinement within each accessible region and influence selection among the resulting candidates, but it cannot overcome the structural gap between the union of these regions and globally better latent behavior outside it. This distinction is captured by the support term Δ_ρ in Theorem 8.

Connection to Theorem 6. Theorem 6 characterizes the tradeoff created by enlarging the local action representation. The ideal reach ρ controls how much latent behavior is accessible in the union of local neighborhoods around the FRS anchors, while the implemented parameters (σ_p, M)

determine the finite random clouds used to probe these neighborhoods. The tilt intensity β plays a different role: it does not enlarge the accessible region, but controls how strongly particle values influence each local refinement.

Theorem 8 combines these effects for the actual hierarchical policy:

$$\|V^\star - V^{\pi_{\mathrm{A}}}\|_\infty \le \frac{\bar{\Delta}_\rho + \bar{\epsilon}_{\mathrm{sel}} + \bar{\epsilon}_{\mathrm{rep}} + \beta^{-1}\log(M+1) + 2\bar{\epsilon}_{\mathrm{crit}} + L_{\mathrm{sm}}\bar{\mathcal{V}}_\beta/2}{1-\gamma}.$$

The terms have distinct origins: $\bar{\Delta}_\rho$ is the structural price of restricting behavior to the union of the local action regions; $\bar{\epsilon}_{\mathrm{rep}}$ is the finite-particle representation error across those regions; $\log(M+1)/\beta$ is the finite-temperature weighting error; $2\bar{\epsilon}_{\mathrm{crit}}$ captures the effect of particle-level critic errors on the local tilting; $(L_{\mathrm{sm}}/2)\bar{\mathcal{V}}_\beta$ controls the discrepancy between high weighted particle value and the value of the corresponding barycentric latent; and $\bar{\epsilon}_{\mathrm{sel}}$ is the error from selecting among the resulting refined candidates using the learned critic. In particular, $\bar{\epsilon}_{\mathrm{sel}} \le 2\bar{\epsilon}_{\mathrm{crit,sel}}$, where $\bar{\epsilon}_{\mathrm{crit,sel}}$ measures critic error at the refined barycentric candidates.

The bound therefore exposes three coupled tuning axes. Increasing the ideal reach ρ can reduce structural support bias, but enlarges the region that the particle clouds must represent. For fixed anchors and realized clouds, adding particles while retaining the existing particles cannot increase the representation error, although it may affect the other terms in the bound. Increasing β reduces the explicit finite-temperature term, but more aggressive weighting is trustworthy only when local value differences are large relative to particle-level critic error. Indeed, (98) shows that sensitivity of the weighting law to critic perturbations grows with $\beta\epsilon_{\mathrm{crit}}$, while (134) guarantees correct high-value amplification when the true local margin exceeds $2\epsilon_{\mathrm{crit}}$. Finally, because selection occurs only after each anchor has been locally refined, reliable execution also requires accurate value comparisons across the resulting barycentric candidates, as captured by $\bar{\epsilon}_{\mathrm{sel}}$.

H.9 Interpretation and novelty

The Gibbs variational identity and the associated exponential-family derivatives are standard. The contribution here is their use within the specific coarse-to-fine structure of AnchorQ, together with results that connect the resulting risk-sensitive weights to the barycentric latent actually executed by the robot.

First, the analysis identifies the precise random variable to which the entropic functional is applied. Conditional on state s and selected anchor c,

$$Y^q_{s,c} = q(s, Z_{s,c}), \qquad Z_{s,c} \sim \nu_{s,c},$$

is the score of a random *local refinement* of that anchor. Positive β is therefore upper-tail seeking over local particle values, not over coarse FRS commands, critic-head uncertainty, or realized environment returns. It becomes an exact entropic functional of realized return only under the conditional transformation of Proposition 10.

Second, the exponential weights have an exact optimization interpretation. The tilted local law $\nu^q_{s,c,\beta}$ is the unique optimizing change of measure in the KL-regularized variational representation of the entropic functional. Thus the weights are not an ad hoc softmax heuristic: they are the optimal risk-sensitive reweighting of the local particle law under a KL penalty from its untilted reference distribution. This interpretation also separates the role of β from the particle parameters: (σ_p, M) control which local refinements are represented, whereas β controls how aggressively value concentrates their influence.

Third, the high-value amplification result provides a reliability guarantee for this local weighting. If a nonempty proper subset of the particle indices has true value margin $\delta_{\mathcal{H}}$ over its complement and the critic error is bounded by ϵ_{crit}, then its odds relative to that complement are at least a factor

$$\exp\left(\beta(\delta_{\mathcal{H}} - 2\epsilon_{\mathrm{crit}})\right)$$

times the corresponding odds under uniform weighting. The threshold $\delta_{\mathcal{H}} > 2\epsilon_{\mathrm{crit}}$ is worst-case sharp under a uniform critic-error model by Proposition 11. Thus aggressive tilting is theoretically justified only when the critic resolves the relevant local value differences reliably.

Fourth, the implemented policy does not sample from this tilted law. It leverages the barycentric latent

$$\tilde{z}_{c,\beta}(s) = \sum_{k=0}^{M} w_{c,\beta}(k \mid s)\, z_{c,k}(s).$$

This creates a new theoretical question absent from ordinary exponential candidate selection: high weighted particle value need not imply high value at their weighted-average latent. Proposition 13 shows that local smoothness closes this gap, with interpolation penalty $L_{\mathrm{sm}} \cdot \mathcal{V}_{c,\beta}(s)/2$. Hence the theory explicitly distinguishes the quality of the weighting rule from the geometry of latent interpolation used to realize that weighting.

Finally, Theorem 8 combines the full construction in a single sequential-control bound. Its decomposition separates the structural support bias, coarse-anchor error, finite-particle representation error, critic error, finite-temperature weighting error, and barycentric interpolation error.

The resulting picture is that the anchor and particle construction determines the available local behavior, while risk-sensitive weighting determines how that local capacity is used.

I Uncertainty Adaptation: Epistemically adaptive local tilting

The preceding section treated the exponential-tilting intensity β as fixed. There, positive β was shown to control the strength of upper-tail weighting over local particle values: increasing β places progressively more influence on high-value local refinements, while the limit $\beta \downarrow 0$ recovers the un-tilted local particle law. The same analysis also exposes the principal limitation of aggressive tilting with a learned critic. If the local critic error is bounded by ϵ_{crit}, Theorem 7 yields $e^{-2\beta\epsilon_{\mathrm{crit}}} \leq \mathrm{d}\nu^{\hat{q}}_{s,c,\beta}/\mathrm{d}\nu^{q_{\mathrm{pop}}}_{s,c,\beta} \leq e^{2\beta\epsilon_{\mathrm{crit}}}$. Thus the effect of critic error on the tilted weighting law is governed by the product $\beta\epsilon_{\mathrm{crit}}$.

A large fixed β can therefore amplify spurious score differences precisely when the critic is least reliable, as may occur early in online learning or in weakly covered regions of the state–latent space.

This motivates adapting the tilting intensity to *epistemic uncertainty* in the learned critic. For a fixed state s, selected anchor c, and realized local particle cloud, let $\widehat{u}_{\mathrm{epi}}(s) \in [0, \infty)$ denote a finite scalar surrogate obtained from disagreement across the critic ensemble. We suppress its dependence on the selected anchor and realized cloud in the notation. We define

$$u_{\mathrm{epi}}(s) := \max\{\widehat{u}_{\mathrm{epi}}(s), u_{\min}\}, \qquad \beta(s) := 1/u_{\mathrm{epi}}(s). \tag{135}$$

The floor $u_{\min} > 0$ bounds the maximum tilting intensity when the estimated uncertainty is spuriously small.

To distinguish the resulting adaptive weights from the fixed-β weights defined in Proposition 12, we use the shorthand

$$\widehat{w}^{\mathrm{ad}}_{c,k}(s) := \widehat{w}^{\beta(s)}_{c,k}(s) = \frac{\exp\left(\beta(s)\hat{q}_{c,k}(s)\right)}{\sum_{\ell=0}^{M} \exp\left(\beta(s)\hat{q}_{c,\ell}(s)\right)}, \qquad k = 0, \ldots, M,$$

where $\hat{q}_{c,k}(s)$ denotes the learned score of the kth particle in the realized local cloud, as in Section H.6. Thus $\widehat{w}^{\mathrm{ad}}_{c,k}(s)$ is not a new weighting rule; it is the same entropic weighting analyzed above, with its tilting intensity now chosen state-wise from epistemic uncertainty. We similarly write

$$\tilde{z}^{\mathrm{ad}}_{c}(s) := \sum_{k=0}^{M} \widehat{w}^{\mathrm{ad}}_{c,k}(s) z_{c,k}(s)$$

for the corresponding adaptive barycentric refinement. We similarly define the adaptive weighted particle variance by

$$\mathcal{V}^{\mathrm{ad}}_{c}(s) := \sum_{k=0}^{M} \widehat{w}^{\mathrm{ad}}_{c,k}(s) \left\| z_{c,k}(s) - \tilde{z}^{\mathrm{ad}}_{c}(s) \right\|_2^2. \tag{136}$$

The interpretation is direct. Exponential weighting remains the upper-tail-seeking local refinement mechanism developed in Section H; adaptation changes only *how strongly that mechanism is applied*. Large epistemic uncertainty yields weak tilting and keeps the weighting law close to the FRS

local particle law. As epistemic uncertainty contracts, the same mechanism becomes progressively more aggressive in amplifying high-value local refinements.

Throughout this section, u_{epi} refers to uncertainty in the learned score of a fixed particle. It is distinct from the variation of $q(s, Z)$ across local particles analyzed by the entropic functional in the preceding section, and from aleatoric randomness in the realized return conditional on a fixed latent.

I.1 Calibration of the epistemic scale

Randomized-prior critic ensembles provide a practical surrogate for epistemic uncertainty, but ensemble disagreement is not automatically a calibrated confidence radius for deep Q-learning. We therefore state explicitly the condition required by the adaptive analysis.

Assumption 8 (Local epistemic calibration)**.** Let q_{pop} denote the population local score and $\hat{q}$ the learned score. There exists a finite constant $c_{\mathrm{cal}} > 0$, independent of the state s and selected anchor c, such that, for every state-anchor pair under consideration,

$$\|\hat{q}(s, \cdot) - q_{\mathrm{pop}}(s, \cdot)\|_{L^\infty(\nu_{s,c})} \le c_{\mathrm{cal}} u_{\mathrm{epi}}(s).$$

When interpreted probabilistically, this inequality is assumed to hold simultaneously over the state-anchor pairs under consideration on a common high-probability event.

Assumption 8 says that the ensemble-derived scale need not equal the critic error, but it must track that error up to a state-independent multiplicative constant. Under this condition, choosing β inversely proportional to u_{epi} normalizes the aggressiveness of the local tilt by the reliability of the critic.

Theorem 9 (Epistemically adaptive local tilting)**.** Suppose Assumption 6 and 8 hold. Fix a state s and selected anchor c, and set $u := u_{\mathrm{epi}}(s)$ and $\beta := \beta(s) = 1/u$. Let $\nu^{\hat{q}}_{s,c,\beta}$ and $\nu^{q_{\mathrm{pop}}}_{s,c,\beta}$ denote the adaptive weighting laws constructed from the learned and population scores, respectively, using the same reference law $\nu_{s,c}$ and the same realized intensity $\beta = 1/u_{\mathrm{epi}}(s)$.

a. **Uncertainty-weighted variational objective.** The learned adaptive weighting law is the unique optimizer of

$$\nu^{\hat{q}}_{s,c,\beta} = \operatorname*{arg\,max}_{\mu \ll \nu_{s,c}} \left\{ \int_{\mathcal{Z}} \hat{q}(s, z)\mu(\mathrm{d}z) - u_{\mathrm{epi}}(s) D_{\mathrm{KL}}(\mu \| \nu_{s,c}) \right\}.$$

Thus epistemic uncertainty is exactly the coefficient penalizing departure from the untilted FRS local-particle law.

b. **Scale-free sensitivity.** The adaptive learned and population weighting laws satisfy

$$e^{-2c_{\mathrm{cal}}} \le \frac{\mathrm{d}\nu^{\hat{q}}_{s,c,\beta}}{\mathrm{d}\nu^{q_{\mathrm{pop}}}_{s,c,\beta}}(z) \le e^{2c_{\mathrm{cal}}} \tag{137}$$

for $\nu_{s,c}$-almost every[4] z, so that

$$D_{\mathrm{KL}}\left(\nu^{\hat{q}}_{s,c,\beta} \| \nu^{q_{\mathrm{pop}}}_{s,c,\beta}\right) \le 2c_{\mathrm{cal}}, \quad \text{and} \quad D_{\mathrm{KL}}\left(\nu^{q_{\mathrm{pop}}}_{s,c,\beta} \| \nu^{\hat{q}}_{s,c,\beta}\right) \le 2c_{\mathrm{cal}}. \tag{138}$$

These bounds are independent of the absolute magnitude of $u := u_{\mathrm{epi}}(s)$.

[4]For two probability measures μ and ν, we write $\mu \sim \nu$ to denote *equivalence of measures*, i.e., $\mu \ll \nu$ and $\nu \ll \mu$, where $\mu \ll \nu$ denotes absolute continuity. Equivalently, μ and ν have the same measure-zero sets. In our setting, the tilted law satisfies

$$\frac{\mathrm{d}\nu^{q}_{s,c,\beta}}{\mathrm{d}\nu_{s,c}}(z) = \exp\left(\beta\left(q(s, z) - \mathrm{CE}_\beta(q; s, c)\right)\right).$$

Under the bounded-score assumption and finite β, this density is strictly positive and finite $\nu_{s,c}$-almost everywhere. Hence $\nu^{q}_{s,c,\beta} \ll \nu_{s,c}$ and also $\nu_{s,c} \ll \nu^{q}_{s,c,\beta}$, so $\nu^{q}_{s,c,\beta} \sim \nu_{s,c}$. Consequently, tilted laws constructed from different bounded scores, such as $\hat{q}$ and q_{pop}, are also mutually equivalent and have the same null sets as $\nu_{s,c}$.

c. **Standardized high-value amplification.** For any Borel set $A \subseteq \mathcal{Z}$ with $0 < \nu_{s,c}(A) < 1$,

$$\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A) \geq \exp\left(\frac{\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}})}{u_{\mathrm{epi}}(s)} - 2c_{\mathrm{cal}}\right) \mathrm{Odds}_{\nu_{s,c}}(A). \tag{139}$$

In particular, if for some $m \geq 0$ we have

$$\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) \geq \big(2c_{\mathrm{cal}} + m\big) u_{\mathrm{epi}}(s), \tag{140}$$

then

$$\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A) \geq e^m \, \mathrm{Odds}_{\nu_{s,c}}(A). \tag{141}$$

Proof. Part *a* is Theorem 7*(a)* with $1/\beta = u_{\mathrm{epi}}(s)$.

For part *b*, Theorem 7*(e)* gives us that

$$e^{-2\beta\epsilon_{\mathrm{crit}}} \leq \frac{\mathrm{d}\nu^{\hat{q}}_{s,c,\beta}}{\mathrm{d}\nu^{q_{\mathrm{pop}}}_{s,c,\beta}} \leq e^{2\beta\epsilon_{\mathrm{crit}}}.$$

Under Assumption 8, we have that $\epsilon_{\mathrm{crit}} \leq c_{\mathrm{cal}} u_{\mathrm{epi}}(s)$, and with $\beta = \frac{1}{u_{\mathrm{epi}}(s)}$, we obtain $\beta\epsilon_{\mathrm{crit}} \leq c_{\mathrm{cal}}$. Substitution gives (137), and integrating the log-density ratio under either measure yields (138).

For part *c*, Theorem 7*(e)* gives us that

$$\mathrm{Odds}_{\nu^{\hat{q}}_{s,c,\beta}}(A) \geq \exp\left(\beta\left(\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) - 2\epsilon_{\mathrm{crit}}\right)\right) \mathrm{Odds}_{\nu_{s,c}}(A).$$

Using $\beta = 1/u_{\mathrm{epi}}(s)$ and $\epsilon_{\mathrm{crit}} \leq c_{\mathrm{cal}} u_{\mathrm{epi}}(s)$ yields

$$\beta\left(\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}}) - 2\epsilon_{\mathrm{crit}}\right) \geq \frac{\Xi_{\nu_{s,c}}(A; q_{\mathrm{pop}})}{u_{\mathrm{epi}}(s)} - 2c_{\mathrm{cal}},$$

which proves (139). Under (140), the exponent is at least m, proving (141). □

The following corollary applies uncertainty adaptation to the barycentric refinement used in AnchorQ. Recall that for a fixed state s and coarse anchor c, we have the adaptive weights by $\widehat{w}^{\mathrm{ad}}_{c,k}(s) := \widehat{w}^{\beta(s)}_{c,k}(s)$ for each $k \in \{0, \ldots, M\}$. The adaptive barycentric refinement is $\tilde{z}^{\mathrm{ad}}_c(s) := \sum_{k=0}^{M} \widehat{w}^{\mathrm{ad}}_{c,k}(s) z_{c,k}(s)$, with weighted particle variance defined in (136), where $\beta(s) := 1/u_{\mathrm{epi}}(s)$.

Corollary 13 (Value of the adaptive barycentric refinement)**.** Fix a state s, a selected coarse anchor c, and its realized local particle cloud $\mathcal{B}_c(s) = \{z_{c,0}(s), \ldots, z_{c,M}(s)\}$. Set $\beta(s) := 1/u_{\mathrm{epi}}(s)$. Suppose the local smoothness condition of Proposition 13 holds on $\mathrm{conv}\,\mathcal{B}_c(s)$, and suppose the critic error on the realized cloud is calibrated to the epistemic scale—i.e.,

$$\max_{0 \leq k \leq M} \left|\hat{q}\left(s, z_{c,k}(s)\right) - q_{\mathrm{pop}}\left(s, z_{c,k}(s)\right)\right| \leq c_{\mathrm{cal}} u_{\mathrm{epi}}(s). \tag{142}$$

Then

$$\max_{0 \leq k \leq M} q_{\mathrm{pop}}\left(s, z_{c,k}(s)\right) - q_{\mathrm{pop}}\left(s, \tilde{z}^{\mathrm{ad}}_c(s)\right) \leq u_{\mathrm{epi}}(s)\left(\log(M+1) + 2c_{\mathrm{cal}}\right) + \frac{L_{\mathrm{sm}}}{2} \mathcal{V}^{\mathrm{ad}}_c(s).$$

Proof. Apply Proposition 13 with $\beta = \beta(s) = 1/u_{\mathrm{epi}}(s)$. The calibration condition (142) implies that the uniform critic-error condition used in (127) holds with $\epsilon_{\mathrm{crit}} = c_{\mathrm{cal}} u_{\mathrm{epi}}(s)$. Therefore

$$\max_{0 \leq k \leq M} q_{\mathrm{pop}}\left(s, z_{c,k}(s)\right) - q_{\mathrm{pop}}\left(s, \tilde{z}^{\mathrm{ad}}_c(s)\right) \leq \frac{\log(M+1)}{\beta(s)} + 2c_{\mathrm{cal}} u_{\mathrm{epi}}(s) + \frac{L_{\mathrm{sm}}}{2} \mathcal{V}^{\mathrm{ad}}_c(s).$$

Since $1/\beta(s) = u_{\mathrm{epi}}(s)$, the claimed bound follows. □

Interpretation. Corollary 13 extends the adaptive tilting guarantee from the particle weights to the latent actually executed by the barycenter implementation. The bound separates two effects:

$$\underbrace{u_{\mathrm{epi}}(s)\left(\log(M+1)+2c_{\mathrm{cal}}\right)}_{\text{epistemically tempered selection error}}+\underbrace{L_{\mathrm{sm}}\mathcal{V}_c^{\mathrm{ad}}(s)/2}_{\text{barycentric interpolation error}}.$$

Thus the value-selection component contracts with calibrated epistemic uncertainty, while the additional cost of executing a weighted latent rather than sampling a particle is controlled by the local curvature of the population score and the geometric spread of the tilted cloud.

Theorem 9 isolates the benefit of adapting the tilting intensity to epistemic uncertainty. With fixed β, the sensitivity of the weighting law to critic error scales as $\beta\epsilon_{\mathrm{crit}}$. Yet, under the adaptive choice $\beta(s) = 1/u_{\mathrm{epi}}(s)$, and a calibrated uncertainty scale $\epsilon_{\mathrm{crit}} \lesssim u_{\mathrm{epi}}$, this dependence becomes scale free. The same normalization yields a standardized criterion for aggressive local weighting: a high-value region is amplified according to its population value separation measured in units of epistemic uncertainty.

Consequently, adaptation does not remove the upper-tail preference established in the preceding section. Rather, it determines when that preference can be trusted. Large uncertainty keeps the local weighting conservative; as uncertainty contracts, progressively smaller true value differences become sufficient to justify strong exponential amplification.

I.2 Finite-cloud implication

The same substitution gives an immediate implementation-level consequence for the finite particle cloud used by AnchorQ. In the fixed-temperature analysis, Corollary 11 gives us that

$$\max_{0\le k\le M} q_{c,k}^{\mathrm{pop}} - \sum_{k=0}^{M} \widehat{w}_{c,k}^{\beta} q_{c,k}^{\mathrm{pop}} \le \frac{\log(M+1)}{\beta} + 2\epsilon_{\mathrm{crit}}.$$

Under adaptive tilting, both terms are controlled by the same epistemic scale.

Corollary 14 (Finite-cloud error under epistemically adaptive tilting)**.** Fix a realized local particle cloud and assume (142) holds on that cloud. Let $\beta(s) = 1/u_{\mathrm{epi}}(s)$. Then

$$\max_{0\le k\le M} q_{c,k}^{\mathrm{pop}} - \sum_{k=0}^{M} \widehat{w}_{c,k}^{\mathrm{ad}}(s) q_{c,k}^{\mathrm{pop}} \le u_{\mathrm{epi}}(s)\left(\log(M+1)+2c_{\mathrm{cal}}\right).$$

If, in addition, the local smoothness assumptions of Proposition 13 hold, then the executed barycentric refinement satisfies

$$\max_{0\le k\le M} q_{\mathrm{pop}}\left(s, z_{c,k}(s)\right) - q_{\mathrm{pop}}\left(s, \tilde{z}_c^{\mathrm{ad}}(s)\right) \le u_{\mathrm{epi}}(s)\left(\log(M+1)+2c_{\mathrm{cal}}\right) + \frac{L_{\mathrm{sm}}}{2}\mathcal{V}_c^{\mathrm{ad}}(s).$$

Proof. Substitute $\beta(s) = 1/u_{\mathrm{epi}}(s)$ and $\epsilon_{\mathrm{crit}} \le c_{\mathrm{cal}} u_{\mathrm{epi}}(s)$ into Corollary 11 and Proposition 13. □

Corollary 14 gives the most direct finite-cloud interpretation of adaptive β. Ignoring the separate barycentric interpolation term, the gap between the best represented local particle and the population value achieved by the learned weighting is bounded above by a quantity linear in $u_{\mathrm{epi}}(s)$, for fixed M and c_{cal}:

$$\text{adaptive weighting error} = \mathcal{O}\left(u_{\mathrm{epi}}(s)\right).$$

As calibrated epistemic uncertainty contracts, the finite-temperature term and the upper bound on the critic-estimation contribution contract together, until the uncertainty floor becomes active.

I.3 Automatic annealing with learning

The adaptive rule also induces an automatic transition from conservative to aggressive local refinement. Suppose, for a fixed state, local particle law, and score landscape, the epistemic uncertainty decreases over training—i.e., $u_{\mathrm{epi},t}(s) \downarrow$. Then $\beta^t(s) = \frac{1}{u_{\mathrm{epi},t}(s)} \uparrow$.

By the monotonicity results of Theorem 7, the average score under the tilted weighting and the KL departure from the local reference law are nondecreasing as uncertainty contracts. Both increase strictly when the score is nonconstant $\nu_{s,c}$-almost surely and the tilting intensity increases strictly. Hence the same fixed exponential-tilting mechanism automatically moves from a weakly tilted, reference-like regime toward increasingly concentrated upper-tail refinement.

This statement isolates the effect of uncertainty adaptation: during actual online learning the critic scores and particle clouds also change, so the observed weighting statistics need not be monotone in training time. Conditional on a fixed local score landscape, however, contraction of epistemic uncertainty necessarily strengthens the risk-seeking local tilt.

Role of the uncertainty floor. The floor in (135) imposes $\beta(s) \leq 1/u_{\min}$. It therefore caps the risk-seeking intensity, preventing a vanishing ensemble-disagreement estimate from making $\beta(s)$ diverge. Once the floor becomes active, uncertainty-driven annealing saturates at this maximum intensity.