BREAKING THE CURSE OF REPULSION IN OFF-POLICY REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Off-policy policy optimization reuses historical behavior, including negative-advantage samples that provide useful corrective feedback. We identify learner-relative remoteness as a key state variable of reuse: repeated updates from a fixed negative sample can make it increasingly remote without weakening its score response. This creates a self-reinforcing remoteness–reuse loop. Aggregate analysis shows how this feedback competes with positive attraction, ranging from stable displacement beyond the positive-only solution to persistent drift and loss of finite stable equilibria. The framework covers both continuous Gaussian and discrete categorical policies, with distinct far-field manifestations: Gaussian score response grows with standardized distance, whereas categorical score response remains bounded while suppression persists. Dynamic Remoteness-Aware Policy Optimization (DRPO) exponentially attenuates remote negative updates while preserving near-field feedback, with Gaussian ultimate-boundedness guarantees. Controlled experiments separately isolate policy geometry as the source of the far-field effect and establish its causal transmission to instability. External diagnostics in D4RL locomotion and structured generation recover the predicted near-to-far pattern, while task-level evaluations show that DRPO achieves higher aggregate scores than reported matched controls.

1 INTRODUCTION

Policy-based methods learn from signed feedback: positive updates pull the policy toward actions estimated to be preferable, whereas negative updates suppress those estimated to be undesirable. Negative feedback can encode useful boundaries and exclusion information absent from positive-only learning (Zhu et al., 2025; Song, 2025). This motivates a different question under off-policy reuse. As the policy changes, the same historical negative sample can be revisited under a different current-policy geometry, so its effect need not remain the same as when it was collected. We therefore ask how repeated off-policy reuse changes the behavior of otherwise useful negative feedback; we study the zero-refresh limit for controlled isolation.

Learner-relative remoteness is the state variable linking successive reuse steps. For a fixed negative coefficient, our isolated-reuse analysis shows that negative updates increase an action's remoteness while its score norm does not decrease. Repetition therefore sustains or amplifies repulsion from the same historical action—the curse of repulsion.

The system-level outcome of this reuse loop depends on the balance between positive and negative feedback masses, each combining sample frequency and coefficient magnitude. When positive mass dominates, repulsion can displace the policy beyond the positive-only target while preserving a finite stable equilibrium. At balance, the restoring force disappears and persistent drift emerges; under negative dominance, no finite stable equilibrium remains. These regimes distinguish stable extrapolation from persistent drift and instability.

This remoteness–reuse framework applies to both continuous Gaussian and discrete categorical policies, while policy geometry determines its far-field manifestation. For fixed-covariance Gaussian policies, score response grows with standardized distance. For categorical policies, surprisal can diverge while the logit-score response remains bounded, yielding persistent suppression without Gaussian-style unbounded amplification.

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.

We therefore introduce Dynamic Remoteness-Aware Policy Optimization (DRPO) to break this remote-reuse feedback loop. DRPO recomputes current-policy remoteness and exponentially attenuates negative updates beyond a remoteness threshold. The taper dominates every finite-order far-field response, restores eventual inward Gaussian drift for any fixed finite negative-to-positive mass ratio, and makes categorical suppression self-limiting. Remoteness controls geometric exposure, not directional utility.

Learner-relative remoteness is typically confounded with advantage magnitude and data quality in observational data. We therefore match those factors, vary remoteness independently, and intervene separately on near- and far-field negative updates.

We contribute: **(1)** a characterization of policy-relative remoteness under repeated negative reuse and aggregate regimes for stable extrapolation, persistent drift, and loss of finite stable equilibria; **(2)** DRPO, a selective remoteness-aware exponential-taper method with Gaussian ultimate-boundedness and self-limiting categorical suppression guarantees; and **(3)** controlled experiments that separate coefficient magnitude from policy geometry, identify a far-field causal pathway, and test the stability–utility trade-off, together with matched external diagnostics and task-performance evaluations in D4RL locomotion and structured generation.

2 Related Work

Off-policy historical data across policy domains. Off-policy policy optimization often relies on historical data, from replay and offline-RL datasets in continuous control to fixed preference and verifier-labeled corpora in language-model post-training. Such data are commonly handled through importance weighting, conservative or behavior-regularized objectives, and reference-regularized optimization (Precup et al., 2000; Fujimoto et al., 2019; Kumar et al., 2020; Fujimoto & Gu, 2021; Kostrikov et al., 2021; Tang et al., 2026). Truncated importance ratios also underlie corrected off-policy return estimators such as Retrace and V-trace (Munos et al., 2016; Espeholt et al., 2018), primarily for distribution correction rather than learner-relative negative-tail control. Our focus is narrower: repeated historical signed feedback under an evolving current policy, especially its negative branch.

Controlling negative updates. Negative-update controls span advantage-weighted policy fitting, which downweights low-advantage samples (Peng et al., 2019; Nair et al., 2020; Kostrikov et al., 2021); fixed global down-weighting of negative advantages, as in HPO's hysteretic weight (Sana et al., 2026); baseline-based positive–negative rebalancing in AsymRE (Arnal et al., 2025); reference-relative preference optimization in DPO (Rafailov et al., 2023); and probability-aware tapering in TOPR (Le Roux et al., 2025). Unlikelihood and preference-based training also exploit failed or rejected behavior (Welleck et al., 2020; Ethayarajh et al., 2024; Duan et al., 2024; Zhang et al., 2024; Zhu et al., 2025; Song, 2025). These approaches control negative feedback through different weighting, balancing, or reference mechanisms; DRPO instead treats learner-relative remoteness as a reuse-induced dynamical state and attenuates only the far-field negative tail.

3 Problem Setup

3.1 Off-Policy Signed Actor Updates

We consider a historical update distribution ν over state–action pairs, which may represent an offline dataset, replay buffer, or trajectories from earlier or stale behavior policies. Let $\pi_\theta(a \mid s)$ denote the current policy and $\widehat{A}(s,a)$ the actor's advantage-like signal. During an actor step, the update distribution and advantage estimates are treated as fixed with respect to θ. The empirical signed actor field has score-function form (Williams, 1992; Sutton et al., 2000):

$$\mathbf{F}(\theta) = \mathbb{E}_{(s,a)\sim\nu}\left[\widehat{A}(s,a)\nabla_\theta \log \pi_\theta(a \mid s)\right], \tag{1}$$

with update $\theta_{t+1} = \theta_t + \eta\mathbf{F}(\theta_t)$. Repeated reuse evaluates the same historical samples and signed feedback under an evolving current policy; Section 4 studies the induced actor-side dynamics.

Write $\widehat{A}^+ = \max\{\widehat{A}, 0\}$ and $\widehat{A}^- = \max\{-\widehat{A}, 0\}$, giving the actor-field decomposition

$$\mathbf{F}(\theta) = \mathbb{E}_\nu \left[\widehat{A}^+(s,a) \nabla_\theta \log \pi_\theta(a \mid s) \right] - \mathbb{E}_\nu \left[\widehat{A}^-(s,a) \nabla_\theta \log \pi_\theta(a \mid s) \right]. \tag{2}$$

For a negative sample, the update magnitude factorizes as $\widehat{A}^-(s,a) \|\nabla_\theta \log \pi_\theta(a \mid s)\|$, separating disadvantage magnitude from learner-relative policy geometry. We call $\widehat{A}^-$ *coefficient magnitude*, $\|\nabla \log \pi\|$ *score response*, their product *update magnitude*, $R_u(a) = \|\nabla_u \log \pi_u(a)\|_2^2$ the *squared-score response*, and an expectation or sum of signed updates *aggregate force*.

Under the common update base measure ν, let

$$p = \mathbb{E}_\nu[\widehat{A}^+], \qquad q = \mathbb{E}_\nu[\widehat{A}^-], \qquad \rho = q/p \quad (p > 0). \tag{3}$$

As population expectations under ν, p and q already include positive and negative sample frequencies.

4 Far-Field Dynamics of Historical Policy Updates

4.1 Learner-Relative Remoteness and Update Strength

To isolate action-side geometry, fix context s and write u for its policy-output coordinates. For a historical action a, define learner-relative remoteness as $D_u(a) = -\log \pi_u(a)$. For categorical policies, $D_u(a)$ is selected-action surprisal; for Gaussian policies, it is a scale-dependent constant plus half the squared Mahalanobis distance from the policy mean. Thus, $D_u(a)$ measures how unlikely the historical action is under the current learner, not an action-space metric. Define the squared score norm by $R_u(a) = \|\nabla_u \log \pi_u(a)\|_2^2$. A sample with advantage estimate $\widehat{A}(s,a)$ contributes output-space update magnitude $|\widehat{A}(s,a)| \sqrt{R_u(a)}$.

Proposition 4.1 (Distance-dependent score-function magnitude)**.** *Fix a context and a historical action a.*

For a Gaussian policy with mean μ and fixed covariance $\Sigma \succ 0$, let $C_\Sigma = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|$ denote the Gaussian normalization constant, which is independent of a and μ. Then learner-relative remoteness satisfies $D_\mu(a) - C_\Sigma = \frac{1}{2}(a-\mu)^\top \Sigma^{-1}(a-\mu)$. Moreover, the squared mean-score norm $R_\mu(a) = \|\nabla_\mu \log \pi_\mu(a)\|_2^2$ is bounded above and below by positive constants times $D_\mu(a) - C_\Sigma$, with constants determined only by Σ. Consequently, as standardized distance increases, the Gaussian mean-score response grows without bound. When $\Sigma = \sigma^2 I$, $R_\mu(a) = 2(D_\mu(a) - C_\Sigma)/\sigma^2$, so remoteness and squared mean-score norm have the same global ordering.

For a categorical policy with logits z, learner-relative remoteness is the selected-action surprisal $D_z(a) = -\log \pi_z(a)$. The selected-logit score response is $\left|\frac{\partial}{\partial z_a} \log \pi_z(a)\right| = 1 - e^{-D_z(a)}$. Thus, categorical surprisal can diverge as the selected probability approaches zero, but the selected-logit response saturates at one. The full logit-score norm remains bounded as well.

(Proof: Appendix A.1.)

4.2 Self-Amplification under Historical Reuse

We now fix a historical action and reuse it as the learner changes. Write $D(u) = D_u(a), \qquad R(u) = \|\nabla_u D(u)\|_2^2$, and consider the repeated single-sample update

$$u_{t+1} = u_t + \eta \widehat{A} \nabla_u \log \pi_{u_t}(a) = u_t - \eta \widehat{A} \nabla_u D(u_t). \tag{4}$$

Action a and signed coefficient $\widehat{A}$ remain fixed over the reuse window.

Theorem 4.2 (Self-attenuation and self-amplification under reuse)**.** *Assume that $D(u)$ is differentiable and convex on a convex neighborhood containing the update segments, and define $D_t = D(u_t), \qquad R_t = R(u_t)$. If $\widehat{A} < 0$, then every reuse step satisfies $D_{t+1} \geq D_t + \eta|\widehat{A}|R_t, \qquad R_{t+1} \geq R_t$. Thus, negative reuse makes the historical action increasingly remote while preventing its score-function norm from decreasing.*

If $\widehat{A} > 0$, and D additionally has an L-Lipschitz gradient on the same neighborhood with $0 < \eta\widehat{A} \leq 1/L$, then $D_{t+1} \leq D_t - \frac{\eta\widehat{A}}{2} R_t, \qquad R_{t+1} \leq R_t$. Thus, positive reuse makes the historical action increasingly compatible with the learner while attenuating its subsequent score response.

For the exponential-family policy coordinates considered here, this convexity condition holds globally (Appendix A.2). Together with Proposition 4.1, the theorem distinguishes one-time distant updates from reuse feedback, yielding unbounded amplification for fixed-covariance Gaussian mean updates and bounded but persistent amplification for categorical logits. For an isolated isotropic-Gaussian negative sample, remoteness grows geometrically under uncontrolled reuse; Appendix A.3 gives the exact rate.

4.3 Aggregate Equilibria under Historical Reuse

The preceding analysis tracks one fixed historical action; aggregating positive and negative historical samples, we ask whether competing attraction and repulsion yield a finite stable equilibrium, persistent drift, or equilibrium instability. A common fixed-context exponential-family representation covers Gaussian and categorical policies; Appendix A.6 gives the coordinates and regularity assumptions. Using p, q, ρ from Equation (3), let $\mathbf{m}_+$ denote the positive-only mean-parameter target. When $q > 0$, let $\mathbf{m}_-$ denote the magnitude-weighted negative-sample moment (a repulsion moment, not an attraction target). Intuitively, $\mathbf{m}_+$ is the attraction center and $\mathbf{m}_-$ the repulsion center.

Theorem 4.3 (Aggregate attraction–repulsion regimes)**.** *Consider the continuous aggregate actor dynamics and the corresponding discrete update under the regular minimal exponential-family representation specified in Appendix A.6. Assume $p > 0$.*

Positive-only case *($\rho = 0$). If $\mathbf{m}_+$ lies in the interior of the attainable mean-parameter space, there is a unique finite equilibrium whose mean parameter is $\mathbf{m}_+$. It is locally asymptotically stable in continuous time and for sufficiently small discrete step sizes.*

Positive-dominant regime *($0 < \rho < 1$). Define the signed target*

$$\mathbf{m}^\star(\rho) = \mathbf{m}_+ + \frac{\rho}{1-\rho}\left(\mathbf{m}_+ - \mathbf{m}_-\right). \tag{5}$$

If $\mathbf{m}^\star(\rho)$ lies in the interior of the attainable mean-parameter space, there is a unique finite equilibrium whose mean parameter is $\mathbf{m}^\star(\rho)$. It is locally asymptotically stable in continuous time and for sufficiently small discrete step sizes. Holding $\mathbf{m}_+$ and $\mathbf{m}_-$ fixed, its displacement from $\mathbf{m}_+$ grows without bound as $\rho \to 1^-$, unless $\mathbf{m}_+ = \mathbf{m}_-$. If the signed target leaves the attainable mean-parameter space, no finite equilibrium exists.

Critical regime *($\rho = 1$). If $\mathbf{m}_+ \neq \mathbf{m}_-$, no finite equilibrium exists; if $\mathcal{H}$ is full or forward invariant in the drift direction, the natural parameter exhibits persistent linear drift. If $\mathbf{m}_+ = \mathbf{m}_-$, every point is stationary, but no point is an isolated asymptotically stable equilibrium.*

Negative-dominant regime ($\rho > 1$). Every finite stationary point, if one exists, is unstable in continuous time and under every discrete update with positive step size. Hence, this regime admits no locally asymptotically stable finite equilibrium.

Appendix A.6 gives the proof, actor field, and exact discrete-time step-size condition. Equation (5) shows that while positive feedback dominates, negative feedback displaces the equilibrium from $\mathbf{m}_+$ along $\mathbf{m}_+ - \mathbf{m}_-$, away from $\mathbf{m}_-$.

Corollary 4.4 (Stable over-extrapolation)**.** *Suppose $\mathbf{m}_+ \neq \mathbf{m}_-$, and let $\mathcal{K}$ be any fixed bounded target region in mean-parameter space. In the positive-dominant regime, sufficiently strong negative feedback either pushes the finite locally stable equilibrium outside $\mathcal{K}$, or, if the signed target becomes unattainable first, eliminates the finite equilibrium.*

Taking $\mathcal{K}$ as a bounded region representing historical data coverage or desirable task performance, sufficiently strong negative feedback can produce stable over-extrapolation before equilibrium instability.

4.4 Policy-Family Manifestations of Negative Dominance

Under negative dominance, no finite stable equilibrium exists, while score and support manifestations differ by policy family.

Theorem 4.5 (Policy-family score behavior under negative dominance)**.** *Assume $p < q$.*

Gaussian policy. *Consider a Gaussian policy $\pi_\mu = \mathcal{N}(\mu, \Sigma)$ with fixed $\Sigma \succ 0$. Its aggregate mean dynamics have a unique finite stationary point, which is repelling. Unless initialized exactly at that point, the distance from it diverges under both the continuous dynamics and every discrete trajectory with fixed step size $\alpha > 0$. Consequently, for every fixed historical action a, $\|\nabla_\mu \log \pi_\mu(a)\|_2 = \left\|\Sigma^{-1}(a-\mu)\right\|_2 \longrightarrow \infty$.*

Categorical policy. *Consider a categorical policy $\pi_z = \mathrm{softmax}(z)$ with at least two actions and the zero-sum gauge $\mathbf{1}^\top z = 0$. Any finite stationary point, when one exists, is unique and repelling. Unless initialized exactly at that point, the continuous dynamics and every discrete trajectory with fixed step size $\alpha > 0$ eventually leave every compact subset of the logit space. Along some subsequence, the corresponding probability vector approaches the boundary of the simplex. Nevertheless, for every action a and every finite z, $\|\nabla_z \log \pi_z(a)\|_2^2 = \|e_a - \pi_z\|_2^2 \le 2(1-\pi_z(a))^2 \le 2$.*

Appendix A.7 gives the proof; fixed covariance suffices for the Gaussian result.

Control implication. For fixed-covariance Gaussian policies with finitely many negative atoms, any remoteness-dependent taper satisfying $\omega_i(D_i) \to 0$ makes the far-field radial component eventually inward for every fixed finite $\rho = q/p$, though not uniformly in ρ. The negative contribution becomes negligible relative to the linear positive restoring force, yielding inward drift. Full details are in Appendix A.4. These conclusions assume frozen data, fixed advantages, and unbounded reuse; data refresh and actor–critic coupling alter the field.

5 Dynamic Remoteness-Aware Policy Optimization

The selective-tapering analysis motivates DRPO: it targets the remote negative source without uniformly weakening the negative branch.

5.1 Update and Policy-Family Instantiations

DRPO uses the thresholded exponential taper

$$\omega_{\mathrm{Exp}}(D) = \exp\left\{-\lambda\left[\frac{D-\tau}{c}\right]_+\right\}, \tag{6}$$

where τ is the threshold, $c > 0$ the remoteness scale, and $\lambda > 0$ the decay rate. The weight is exactly one for $D \le \tau$ and decays smoothly beyond it.

Using $D_\theta(s,a) = -\log \pi_\theta(a \mid s)$, define $\bar{D}_\theta(s,a) = \mathrm{sg}[D_\theta(s,a)]$, where $\mathrm{sg}[\cdot]$ denotes stop-gradient. DRPO changes only the negative branch: $\widetilde{A}_\theta(s,a) = \widehat{A}^+(s,a) - \omega_{\mathrm{Exp}}\big(\bar{D}_\theta(s,a)\big)\,\widehat{A}^-(s,a)$, with actor field

$$\mathbf{F}_{\mathrm{DRPO}}(\theta) = \mathbb{E}_{(s,a)\sim\nu}\left[\widetilde{A}_\theta(s,a)\nabla_\theta \log \pi_\theta(a \mid s)\right]. \tag{7}$$

DRPO recomputes and detaches current-policy remoteness before weighting, yielding a dynamic weighting rule rather than a differentiable distance regularizer.

Gaussian policies. For a Gaussian policy with mean $\mu_\theta(s)$ and fixed covariance Σ, define $D_{\mathrm{G},\theta}(s,a) = (a-\mu_\theta(s))^\top \Sigma^{-1}(a-\mu_\theta(s))$. Because $D_\theta - C_\Sigma = \frac{1}{2}D_{\mathrm{G},\theta}$, the constant and factor $1/2$ can be absorbed into the threshold and scale: $\omega_{\mathrm{G}}(s,a) = \exp\left\{-\lambda_{\mathrm{G}}\left[\frac{D_{\mathrm{G},\theta}(s,a)-\tau_{\mathrm{G}}}{c_{\mathrm{G}}}\right]_+\right\}$.

Categorical policies. For categorical policies, substituting selected-action surprisal into Equation (6) gives $\omega_{\mathrm{C}}(s,a) = \min\left\{1,\ (e^\tau \pi_\theta(a \mid s))^{\lambda/c}\right\}$.

5.2 Minimum-Order Far-Field Control

The preceding analysis establishes that remoteness-dependent attenuation can control far-field negative feedback. We now ask how rapidly the taper must decay to keep a growing far-field response bounded.

Proposition 5.1 (Minimum-order far-field control)**.** *Let $I(D) \geq 0$ and $I(D) = \Theta(D^m)$ as $D \to \infty$ for some finite $m \geq 0$. For any nonnegative taper $\omega(D)$,*

$$\omega(D)I(D) = O(1) \quad \Longleftrightarrow \quad \omega(D) = O(D^{-m}).$$

At the exact matching order $\omega(D) = \Theta(D^{-m})$, the weighted response remains $\Theta(1)$.

For continuous-time isolated negative reuse with fixed coefficient $\widehat{A} < 0$,

$$\dot{D} = |\widehat{A}|\, \omega(D) R(D), \qquad R(D) = \|\nabla_u D\|_2^2.$$

For fixed-covariance Gaussian means, $R(D) = \Theta(D)$ in the far field, so $\omega(D) = O(D^{-1})$ is the matching order for bounded remoteness growth. At exact matching, the parameter-force scale $\omega(D)\sqrt{R(D)}$ vanishes while $\omega(D)R(D) = \Theta(1)$. In standardized distance, this is the asymptotic order implemented by reciprocal-quadratic attenuation.

The exponential taper in Equation (6) satisfies this minimum-order requirement for every finite m, since its asymptotic decay is faster than D^{-m}. The proof is in Appendix A.8.

5.3 Redescending Far-Field Control

Corollary 4.4 shows that finite local stability does not preclude outward displacement. Section 5.2 further shows that exact matching can leave a constant-order contribution to remoteness growth even after the corresponding parameter force has vanished. We therefore seek the stronger redescending condition

$$\omega(D)I(D) \longrightarrow 0 \qquad \text{as } D \to \infty.$$

Proposition 5.2 (Vanishing finite-order far-field response)**.** *Let $I(D) \geq 0$ denote a far-field response and suppose*

$$I(D) \leq C(1 + D)^m$$

for some finite $m \geq 0$ and $C > 0$. Then the exponential taper in Equation (6) *satisfies*

$$\omega_{\mathrm{Exp}}(D)I(D) \longrightarrow 0 \qquad \textit{as } D \to \infty.$$

At each iterate, tapering replaces the negative mass and moment by

$$q_\omega = \mathbb{E}_\nu[\widehat{A}^- \omega], \qquad \mathbf{M}_{-,\omega} = \mathbb{E}_\nu[\widehat{A}^- \omega T(a)].$$

When $q_\omega < p$, the corresponding signed target is

$$\mathbf{m}_\omega^\star = \frac{p\mathbf{m}_+ - \mathbf{M}_{-,\omega}}{p - q_\omega}.$$

Corollary 5.3 (Far-field signed-target recovery)**.** *In the fixed-data setting with $p > 0$ and fixed $\mathbf{m}_+$, consider a fixed finite negative set with fixed coefficients. If every negative atom becomes asymptotically remote along a policy sequence and $\omega_i(D_i) \to 0$, then*

$$q_\omega \to 0, \qquad \mathbf{M}_{-,\omega} \to \mathbf{0}, \qquad \mathbf{m}_\omega^\star \to \mathbf{m}_+.$$

If $\mathbf{m}_+ \in \mathrm{int}(\mathcal{K})$, then $\mathbf{m}_\omega^\star \in \mathcal{K}$ eventually.

Vanishing weights recover the signed target in this limit; exponential tapering additionally makes every finite polynomial-order response vanish after weighting. Exponential tapering also yields Gaussian ultimate boundedness and polynomial probability decay under isolated categorical suppression (Appendices A.5 and A.11). Proofs of the results above are in Appendix A.10.

Where existing off-policy controls intervene. We organize off-policy controls by four questions: what data are used, what signal they carry, how much each sample contributes, and how policy is anchored to historical behavior or a reference. These questions classify mechanisms, not algorithms; methods may instantiate multiple mechanisms. DRPO acts at the sample-influence level, attenuating negative feedback according to learner-relative remoteness rather than a behavior-policy ratio. Appendix C gives the full taxonomy and comparisons.

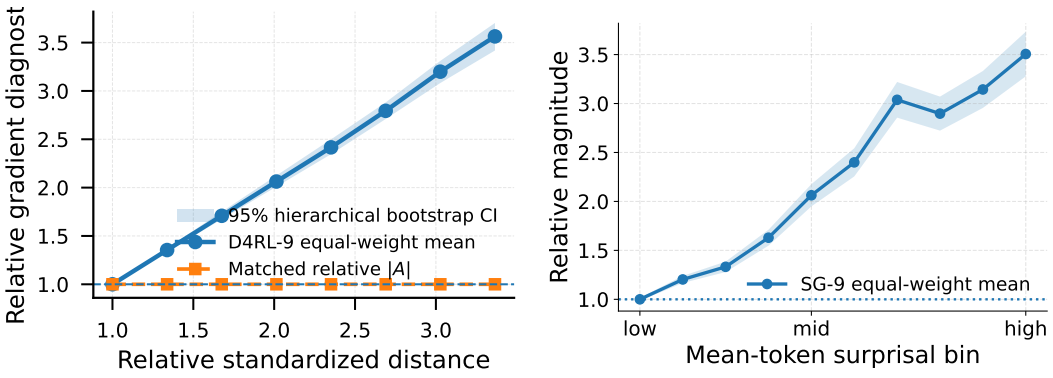


Figure 1: **External far-field diagnostics.** Left: D4RL locomotion aggregate of relative implemented actor-gradient versus policy-relative standardized distance. The dashed orange curve shows matched relative $|A|$, aggregated with equal dataset-cell weighting. Right: SG-9 aggregate over nine structured-generation tasks of relative implemented actor-gradient versus mean-token surprisal bin. Shading shows 95% confidence intervals.

6 Experiments

External probes test whether the predicted far-field pattern appears in realistic continuous-control and discrete autoregressive policies; controlled environments separate coefficient magnitude from policy geometry, trace updates into policy behavior, and compare negative-weighting rules under known near/far structure; task-level evaluations cover D4RL and structured generation. C-U1 (a Gaussian continuous-action contextual bandit) and D-U1 (an unordered categorical environment with shared representations) provide controlled environments with known ground truth for source isolation, policy dynamics, and held-out-context generalization. External diagnostics use nine D4RL locomotion cells (Fu et al., 2020) and nine structured-generation tasks— Countdown (Hutton, 2002; Pan et al., 2025), seven Reasoning Gym tasks (Stojanovski et al., 2025), and WikiSQL (Zhong et al., 2017). Controlled source-isolation reports policy-score response; neural probes report implemented actor-gradient magnitude to test whether the near-to-far pattern persists in trainable updates. Task-performance collapse, support/variance-boundary events, and NaN/Inf failures are reported separately. Full configurations and protocols are in Appendices D and E.

6.1 Far-Field Repulsion: External Diagnostics and Controlled Identification

We compare near- and far-field actor gradients using matched negative-signal magnitudes in D4RL and fixed negative coefficients in structured generation; Appendices D.1 and D.2 give the protocols. In both domains, implemented actor-gradient increases from near to far field (Figure 1). The approximately linear D4RL trend agrees with the Gaussian prediction; the fixed-coefficient structured-generation trend is consistent with persistent categorical suppression. These probes show the near-to-far pattern in realistic neural policies.

The external probes show the near-to-far gradient pattern with negative-signal magnitude approximately matched in D4RL, with learned-critic $|A|_{\text{far}}/|A|_{\text{near}} \approx 1$, and unit-valued by construction for verifier-labeled negatives in structured generation; controlled experiments then isolate its source and test its downstream training effects by independently varying coefficient magnitude and remoteness and intervening on near- versus far-field negative updates (Appendix E). C-U1 negative actions share a reward contour and fixed advantage but differ in standardized distance; the categorical direct-softmax reference fixes the negative coefficient while varying current-policy surprisal. For the causal intervention, C-U1 branches share data, fixed labels, initialization, optimizer, and horizon, differing only in near/far negative-update treatment (Appendix E.1).

The top panel of Figure 2 separates coefficient and policy-score terms: coefficient magnitude stays flat across remoteness bins while score response grows, so larger updates arise from policy-score geometry rather than larger advantages. Gaussian policies show amplitude amplification, whereas categorical direct-logit responses saturate while selected-action probability decreases toward zero.

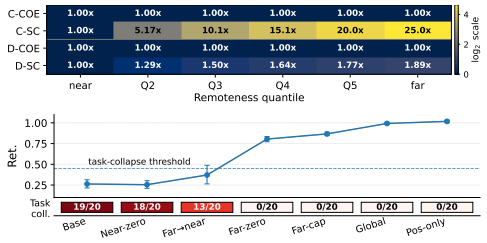


Figure 2: **Controlled source isolation and causal transmission. Top:** coefficient magnitude (COE) and policy-score response (SC) across remoteness bins, normalized to the near bin, for the C-U1 Gaussian protocol and direct-softmax categorical reference. **Bottom:** C-U1 interventions; Ret.: reward retention (20-seed means, bootstrap CIs); Task coll.: task-performance collapse counts.

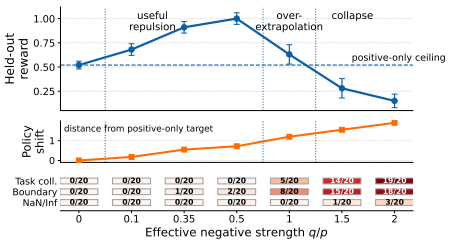


Figure 3: **Controlled negative-strength response in C-U1.** Held-out-context reward and normalized policy shift as the effective negative-to-positive strength q/p increases. The strip reports task-performance collapse, boundary events, and NaN/Inf failures separately.

This isolates the source of the update difference; the next intervention tests whether it propagates into drift and collapse.

The bottom panel of Figure 2 shows rescue: removing near-field negative updates leaves task-performance collapse largely unchanged, whereas removing or capping far-field updates restores reward retention and eliminates task-performance collapse. Global rescaling also rescues the run, supporting the interpretation of abnormal negative-update magnitude as a mediator. Restoring the uncontrolled negative-gradient norm through near-field updates yields only partial recovery, showing that gradient magnitude alone does not explain the rescue. These interventions identify far-field negative updates—not negative feedback generally—as a causal path to policy instability in C-U1.

6.2 From Useful Repulsion to DRPO Control

Causal rescue does not imply removing all negative feedback. C-U1 and D-U1 place a hidden optimum beyond the positive-only target and direct local repulsion toward it, while remote negatives add far-field pressure. This separates useful repulsion from harmful dominance and enables controlled taper comparisons.

Corollary 4.4 predicts that increasing negative feedback can drive stable over-extrapolation. We test this in C-U1 by varying only q/p. Figure 3 shows a non-monotone response: intermediate strength raises held-out-context reward above the Positive-only baseline ($q/p = 0$) while increasing policy shift, whereas stronger feedback further increases displacement as reward declines. Collapse, boundary events, and NaN/Inf failures are reported separately.

Positive-only removes both the harmful tail and beneficial intermediate regime; remoteness-aware control should preserve local repulsion while suppressing the far-field tail. Matched controls show that remoteness-aware tapering can preserve near-field repulsion while suppressing far-field influence; global scaling reduces both. Appendix E.5 reports Positive-only, Global-α, Reciprocal-Linear, Reciprocal-Quadratic, and exponential tapering under controls matched either by mean near-field weight or by pre-Adam negative-gradient norm.

We next test this remoteness-aware control principle externally. DRPO weights D4RL transitions by standardized remoteness under the current Gaussian policy and structured-generation responses by mean-token surprisal; protocol-matched methods share data, labels, initialization, optimizer, and budget. DRPO has the highest displayed aggregate score in both domains (Table 1).

Table 1: **Protocol-matched external transfer.** D4RL aggregate and nine-task structured-generation (SG-9) macro Pass@8.

	Pos.	G-α	Rec.-L	Rec.-Q	DRPO
D4RL	70.1	72.9	74.3	72.1	**77.7**
SG-9	67.9	66.7	68.1	68.2	**69.2**

Table 2: D4RL-9 normalized returns in (a) and structured-generation Pass@8 in (b).

(a) D4RL-9 normalized return

Dataset	CQL†	TD3+BC†	IQL†	**DRPO**
halfcheetah-medium	44.0	**48.3**	47.4	47.3
hopper-medium	58.5	59.3	**66.3**	57.3
walker2d-medium	72.5	83.7	78.3	**85.5**
halfcheetah-replay	**45.5**	44.6	44.2	43.3
hopper-replay	**95.0**	60.9	94.7	87.8
walker2d-replay	77.2	**81.8**	73.9	78.1
halfcheetah-expert	**91.6**	90.7	86.7	91.1
hopper-expert	**105.4**	98.0	91.5	98.0
walker2d-expert	108.8	110.1	109.6	**110.4**
D4RL-9 total	*698.5*	*677.4*	*692.6*	***698.8***

(b) Structured generation Pass@8

Task	PO	AsymRE	β-TOPR	DPO‡	**DRPO**
Countdown	13.98	14.38	14.41	8.48	**15.72**
Word Sorting	62.06	61.72	59.61	34.38	**63.38**
Spiral Matrix	**100.00**	99.92	**100.00**	**100.00**	99.94
Mini Sudoku	95.91	**97.58**	96.88	91.41	97.06
Maze	76.74	77.81	75.86	**79.69**	77.66
Word Ladder	10.68	12.50	11.41	4.69	**13.38**
Knights & Knaves	67.53	**70.16**	68.59	58.20	69.53
Graph Coloring	98.67	98.91	99.37	96.88	**99.88**
WikiSQL	85.65	85.78	82.89	78.91	**86.69**
Macro Avg.	*67.91*	*68.75*	*67.67*	*61.40*	***69.25***

Note. PO = Positive-only. **(a)** "replay"/"expert" abbreviate "medium-replay"/"medium-expert". †: published results, not protocol-matched reruns. Bold/underline: best/second-best among shown methods. DRPO reports the mean over 10 runs. **(b)** AsymRE, β-TOPR, and DRPO initialize LoRA directly from Qwen2.5-0.5B-Instruct; ‡DPO uses the same LoRA parameterization but initializes policy and frozen reference from the short SFT checkpoint (Appendix E.8). β-TOPR is the joint fitted-reference TOPR variant (Le Roux et al., 2025). Each λ point is summarized by mean Pass@8 over updates $800, 900, 1000, 1100$, and 1200. DRPO transfer scores are the maximum means over five adjacent λ values on the coefficient-response curve. Macro Avg. is the unweighted mean of nine task-level Pass@8 percentages.

6.3 Task Performance

Task-level transfer uses D4RL normalized return and structured-generation verifier success. Both domains use fixed data while recomputing learner-relative remoteness under the current policy; protocol-matched controls share data and compute budgets. Task-performance collapse, support/variance-boundary events, and NaN/Inf failures remain separate.

For continuous-control transfer, Table 2 evaluates DRPO on the D4RL locomotion suite (Fu et al., 2020) against published CQL (Kumar et al., 2020), TD3+BC (Fujimoto & Gu, 2021), and IQL (Kostrikov et al., 2021) results; Appendix E.7 gives protocol and implementation details. With the shared advantage-weighted actor form, DRPO remains competitive on the D4RL-9 total without an added behavior-policy constraint or conservative value penalty; its only method-specific component is a detached negative-branch remoteness taper.

For structured generation, we evaluate Qwen2.5-0.5B-Instruct (Yang et al., 2024) on these nine tasks using deterministic verifiers and Pass@8. For each prompt, we construct a fixed bank of verified correct and incorrect completions and reuse it throughout training, yielding a zero-refresh stale-negative-reuse setting (Appendices D.5 and E.8). We compare DRPO with AsymRE (Arnal et al., 2025), joint fitted-reference β-TOPR (Le Roux et al., 2025), and canonical DPO (Rafailov et al., 2023); Table 2 and Appendix E.8 report method-specific initialization.

Across the nine structured-generation tasks, reported DRPO scores exceed those of AsymRE on six, those of joint fitted-reference β-TOPR on eight, and the corresponding DPO entries on seven.

7 Conclusion

Repeated off-policy reuse can turn useful negative feedback into self-reinforcing far-field repulsion, linking per-sample repulsion to stable displacement, persistent drift, and instability. DRPO selectively attenuates remote negative updates while preserving local ones; matched diagnostics and interventions support this mechanism across continuous and discrete policies. Our guarantees analyze the zero-refresh limit for clean isolation, but the same learner-relative weighting principle applies when negative samples are already remote under the current policy, including replay and stale-policy settings; remoteness-aware control is not restricted to static offline datasets.

AI Use Statement

Generative AI tools were used in limited supporting roles during theoretical refinement, mathematical verification, implementation, literature retrieval, and manuscript preparation. The research questions, hypotheses, core conceptual framework, experimental methodology and design, and interpretation of

experimental results were developed by the authors. AI-assisted discussion was used only to refine limited aspects of the theoretical framework.

For mathematical analysis, AI tools were used primarily to check theorem statements and derivations, identify missing assumptions and boundary cases, suggest localized corrections, and assist with proof verification. The core mathematical claims and proof strategies were developed by the authors.

AI tools were also used for implementation, code review and debugging, experimental workflow support, and consistency checking. Their use in manuscript preparation was primarily for language polishing and local sentence-level revision. AI tools were used to a limited extent for literature search and retrieval; the selection, assessment, and use of related work were performed by the authors.

AI tools were not used to formulate the research questions or hypotheses, design the experiments or research methodology, or interpret the experimental results. All AI-assisted code, mathematical arguments, experimental procedures, and manuscript edits were reviewed and verified by the authors, who take responsibility for the final content of the work.

REFERENCES

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Remi Munos. Asymmetric REINFORCE for off-policy reinforcement learning: Balancing positive and negative rewards. *arXiv preprint arXiv:2506.20520*, 2025.

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

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Yan Liu, Zheng Liu, Tun Lu, Xing Xie, and Ning Gu. Negating negatives: Alignment with human negative samples via distributional dispreference optimization. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. arXiv:2403.03419.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning (ICML)*, 2018.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.01306.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning (ICML)*, pp. 2052–2062, 2019.

Graham Hutton. The countdown problem. *Journal of Functional Programming*, 12(6):609–616, 2002. doi: 10.1017/S0956796801004300.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Nicolas Le Roux, Marc G. Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy REINFORCE: Stable and efficient reinforcement learning for LLMs. *arXiv preprint arXiv:2503.14286*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. In *arXiv preprint arXiv:2006.09359*, 2020.

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. `https://github.com/Jiayi-Pan/TinyZero`, 2025. Minimal reproduction of DeepSeek R1-Zero on Countdown; accessed 2025.

Xue Bin Peng, Aviral Kumar, et al. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning (ICML)*, 2000.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mohamed Sana, Nicola Piovesan, Antonio De Domenico, Fadhel Ayed, and Haozhe Zhang. Hpo: Hysteretic policy optimization for stable and efficient training under sparse-reward regime. *arXiv preprint arXiv:2605.30201*, 2026.

Jun Song and Chaoyue Zhao. Optimistic distributionally robust policy optimization. *arXiv preprint arXiv:2006.07815*, 2020.

Meng Song. Good actions succeed, bad actions generalize: A case study on why RL generalizes better. *arXiv preprint arXiv:2503.15693*, 2025.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. REASONING GYM: Reasoning environments for reinforcement learning with verifiable rewards. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1913. Datasets and Benchmarks Track.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

Yunhao Tang, Taco Cohen, David W. Zhang, Gabriel Synnaeve, and Rémi Munos. Rl-finetuning llms from on- and off-policy data with a single algorithm. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 415–423. PMLR, 02–05 May 2026.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*, 2020.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024. arXiv:2412.15115.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.05868.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103, 2017.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. *arXiv preprint arXiv:2506.01347*, 2025.

A Proofs for the Theoretical Results

Proof map. The score and reuse results are proved in Sections A.1–A.3; the aggregate regimes and policy-family manifestations are proved in Sections A.6 and A.7. The selective-tapering bridge, Gaussian boundedness, control order, exponential tail, and categorical results appear in Sections A.4, A.5, A.8, A.10, and A.11–A.12, respectively.

A.1 Proof of Proposition 4.1

Proof. Let $\delta = a - \mu$ and $x = \Sigma^{-1/2}\delta$. For a fixed-covariance Gaussian policy,

$$D_\mu(a) - C_\Sigma = \frac{1}{2}\|x\|_2^2, \qquad \nabla_\mu \log \pi_\mu(a) = \Sigma^{-1}\delta, \qquad R_\mu(a) = x^\top \Sigma^{-1} x.$$

The Rayleigh quotient therefore gives

$$2\lambda_{\min}(\Sigma^{-1})\big(D_\mu(a) - C_\Sigma\big) \le R_\mu(a) \le 2\lambda_{\max}(\Sigma^{-1})\big(D_\mu(a) - C_\Sigma\big).$$

Hence the Gaussian mean-score response is linearly comparable to $D_\mu(a) - C_\Sigma$ and grows without bound with standardized distance. If $\Sigma = \sigma^2 I$, the bounds collapse to

$$R_\mu(a) = \frac{2}{\sigma^2}\big(D_\mu(a) - C_\Sigma\big).$$

For a categorical policy, the $K = 1$ case is degenerate: $\pi_z(a) = 1$, $D_z(a) = 0$, and the score is identically zero. Assume $K \ge 2$ below and let $p_a = \pi_z(a) = e^{-D_z(a)}$. Since

$$\nabla_z \log \pi_z(a) = e_a - \pi_z,$$

the selected-logit response is

$$\left|\frac{\partial}{\partial z_a} \log \pi_z(a)\right| = 1 - p_a = 1 - e^{-D_z(a)},$$

which increases with surprisal and saturates at one. Moreover,

$$R_z(a) = (1 - p_a)^2 + \sum_{j \ne a} \pi_z(j)^2.$$

The non-selected probabilities are nonnegative and sum to $1 - p_a$, so

$$\frac{(1 - p_a)^2}{K - 1} \le \sum_{j \ne a} \pi_z(j)^2 \le (1 - p_a)^2.$$

Therefore

$$\frac{K}{K - 1}(1 - p_a)^2 \le R_z(a) \le 2(1 - p_a)^2 \le 2,$$

which proves the claimed bounded categorical score behavior. □

A.2 Proof of Theorem 4.2

For a regular exponential-family policy in natural coordinates ξ,

$$\pi_\xi(a) = h(a) \exp\{\xi^\top T(a) - \psi(\xi)\}.$$

Its learner-relative remoteness is

$$D_\xi(a) = -\log \pi_\xi(a) = \psi(\xi) - \xi^\top T(a) - \log h(a),$$

and therefore

$$\nabla_\xi^2 D_\xi(a) = \nabla^2 \psi(\xi) = \mathrm{Cov}_{\pi_\xi}[T(a)] \succeq 0.$$

Convexity is preserved under the affine policy-output coordinate maps used here. Fixed-covariance Gaussian means correspond to $\xi = \Sigma^{-1}\mu$, while categorical logits are natural coordinates up to the usual additive gauge. Hence both policy families satisfy the convexity assumption of Theorem 4.2.

Proof. Write

$$f = D, \qquad g_t = \nabla f(u_t), \qquad R_t = \|g_t\|_2^2.$$

If $\widehat{A} = -c < 0$, let $\alpha = \eta c$. Then $u_{t+1} = u_t + \alpha g_t$, and convexity gives

$$D_{t+1} \geq D_t + g_t^\top (u_{t+1} - u_t) = D_t + \alpha R_t.$$

Gradient monotonicity gives

$$(g_{t+1} - g_t)^\top (u_{t+1} - u_t) \geq 0.$$

Since $u_{t+1} - u_t = \alpha g_t$, $g_{t+1}^\top g_t \geq \|g_t\|_2^2$; Cauchy–Schwarz then yields $\|g_{t+1}\|_2 \geq \|g_t\|_2$, with the zero-gradient case immediate. Hence

$$D_{t+1} \geq D_t + \eta |\widehat{A}| R_t, \qquad R_{t+1} \geq R_t.$$

The second inequality is strict whenever the gradient changes nontrivially along a direction of positive curvature.

If $\widehat{A} = c > 0$, again let $\alpha = \eta c$, so $u_{t+1} = u_t - \alpha g_t$. The descent lemma gives

$$D_{t+1} \leq D_t - \alpha \left(1 - \frac{\alpha L}{2}\right) R_t \leq D_t - \frac{\alpha}{2} R_t$$

whenever $0 < \alpha \leq 1/L$. For a convex L-smooth function, cocoercivity gives

$$(g_t - g_{t+1})^\top (u_t - u_{t+1}) \geq \frac{1}{L} \|g_t - g_{t+1}\|_2^2.$$

Let $d = g_{t+1} - g_t$. Since $u_t - u_{t+1} = \alpha g_t$,

$$d^\top g_t \leq -\frac{\|d\|_2^2}{\alpha L},$$

and therefore

$$R_{t+1} = \|g_t + d\|_2^2 \leq R_t + \left(1 - \frac{2}{\alpha L}\right) \|d\|_2^2 \leq R_t.$$

Substituting $\alpha = \eta \widehat{A}$ proves the positive-reuse claim. □

A.3 Quantitative Gaussian Reuse Rates

Proposition A.1 (Quantitative Gaussian reuse rates)**.** *For an isotropic Gaussian mean policy with covariance $\sigma^2 I$, one fixed negative sample of mass $q > 0$, and $X_t = D_t - C_\Sigma > 0$, the uncontrolled update satisfies*

$$X_{t+1} = \left(1 + \eta q/\sigma^2\right)^2 X_t,$$

so remoteness grows geometrically. If the same sample is weighted by $\exp\{-a[D_t - \tau]_+\}$, $a > 0$, then its continuous-time remoteness is $O(a^{-1} \log t)$, and the discrete update obeys

$$X_t \leq a^{-1}\{\log t + \log\log t + O(1)\}.$$

This slowdown does not by itself prove aggregate boundedness; that requires the closed-loop attraction established in Theorem A.4.

Proof of Proposition A.1. Let $\delta_t = \mu_t - a$ and $X_t = \|\delta_t\|_2^2/(2\sigma^2) = D_t - C_\Sigma$. For an uncontrolled negative sample of mass q,

$$\delta_{t+1} = \left(1 + \frac{\eta q}{\sigma^2}\right) \delta_t,$$

which gives the stated exact geometric recurrence.

Now let $\omega_t = e^{-a[D_t - \tau]_+}$. Once $D_t > \tau$, write $b = \eta q/\sigma^2$ and $K_0 = e^{a(\tau - C_\Sigma)}$. Then

$$X_{t+1} = X_t \left(1 + bK_0 e^{-aX_t}\right)^2. \tag{8}$$

Consequently $\Delta X_t \geq 0$, and for all sufficiently large X_t,

$$0 \leq \Delta X_t \leq C X_t e^{-aX_t} \tag{9}$$

for a finite constant C. Put $Y_t = e^{aX_t}$. Since $aX_t e^{-aX_t} \leq 1/e$, Equation (9) also gives a uniform bound $a\Delta X_t \leq M$. Hence

$$\Delta Y_t = Y_t(e^{a\Delta X_t} - 1) \tag{10}$$

$$\leq C_1 Y_t \, a\Delta X_t \leq C_2 \log Y_t. \tag{11}$$

The standard comparison sequence $\overline{Y}_t = Kt\log t$, with K chosen larger than the constant in Equation (11), is eventually a supersolution. Therefore $Y_t = O(t\log t)$ and

$$X_t \le a^{-1}\{\log t + \log\log t + O(1)\}.$$

In continuous time, the far-branch equation is $\dot{X} = C_3 X e^{-aX}$. Thus $Y = e^{aX}$ satisfies $\dot{Y} = C_3 \log Y$, and the same comparison gives $X(t) = O(a^{-1}\log t)$. □

A.4 Far-Field Radial Bridge

Lemma A.2 (Far-field radial balance)**.** *For a fixed-covariance Gaussian field with finite negative atoms, $F(\mu) = p\Sigma^{-1}(m_+ - \mu) + \sum_i q_i\Sigma^{-1}(\mu - a_i), q = \sum_i q_i$, and $\mu = m_+ + rv$, $\|v\|_2 = 1$, $\lim_{r\to\infty} \frac{\langle v, F(\mu)\rangle}{r} = (q-p)v^\top\Sigma^{-1}v$.*

Corollary A.3 (Selective far-field attenuation)**.** *If each negative atom is multiplied by $\omega_i(D_i) \to 0$, the limit in Lemma A.2 becomes $-pv^\top\Sigma^{-1}v < 0$. Thus, for every fixed finite $\rho = q/p$, the far-field radial component is eventually inward; the onset radius is not uniform in ρ and depends on the tail rate of ω. If additionally $\omega(D)\sqrt{D} \to 0$, the absolute tapered negative force vanishes.*

Proofs of Lemma A.2 and Corollary A.3. Set $\mu = m_+ + rv$ with $\|v\|_2 = 1$. Direct expansion gives

$$\frac{\langle v, F(\mu)\rangle}{r} = -p\, v^\top\Sigma^{-1}v + \sum_i q_i v^\top\Sigma^{-1}v \tag{12}$$

$$+ \frac{1}{r}\sum_i q_i\langle v, \Sigma^{-1}(m_+ - a_i)\rangle. \tag{13}$$

The final term vanishes, proving the lemma.

With atom-specific weights, $D_\mu(a_i) \to \infty$ along every ray. For finitely many atoms, $\omega_i(D_i) \to 0$ therefore makes both the weighted coefficient of r and the weighted constant term vanish. The radial limit is $-pv^\top\Sigma^{-1}v$, which is strictly negative. This is a pointwise-in-q/p statement: the radius at which it becomes negative can grow with q/p. Moreover, $\|\Sigma^{-1}(\mu - a_i)\|_2 = O(\sqrt{D_i})$, so $\omega(D)\sqrt{D} \to 0$ removes the absolute negative force. □

A.5 Gaussian Closed-Loop Boundedness

Theorem A.4 (Gaussian DRPO boundedness)**.** *Let $p > 0$, $\Sigma \succ 0$, and*

$$F_\omega(\mu) = p\Sigma^{-1}(m_+ - \mu) + \sum_i q_i\omega_i(D_\mu(a_i))\Sigma^{-1}(\mu - a_i).$$

For finitely many negative atoms with exponential tapers, define

$$B_\omega = \sup_\mu \left\| \sum_i q_i\omega_i(D_i)\Sigma^{-1}(\mu - a_i) \right\|_2, \qquad R^* = \frac{B_\omega}{p\lambda_{\min}(\Sigma^{-1})}. \tag{14}$$

Every continuous trajectory satisfies $\limsup_{t\to\infty} \|\mu(t) - m_+\|_2 \le R^*$, and the field has at least one finite equilibrium. If an uncontrolled stable equilibrium places every negative atom strictly below its threshold, the controlled and original fields have the same local linearization in a neighborhood of that equilibrium.

The proof below gives an explicit branch-wise bound, including the case where the shifted threshold is negative and tapering is active from radius zero. Along every escaping sequence,

$$\rho_{\text{eff}}(\mu) = p^{-1}\sum_i q_i\omega_i(D_i) \longrightarrow 0;$$

this explains how the remote negative source disappears, but it is a diagnostic consequence rather than the boundedness argument.

Proof of Theorem A.4. Let $e = \mu - m_+$ and

$$N_\omega(\mu) = \sum_i q_i \omega_i(D_i)\Sigma^{-1}(\mu - a_i).$$

The finite-atom exponential-taper field is locally Lipschitz and has at most linear growth, so its solutions exist for all forward times. For $V(e) = \frac{1}{2}\|e\|_2^2$,

$$\dot{V} = -pe^\top \Sigma^{-1} e + e^\top N_\omega(\mu) \tag{15}$$

$$\leq -p\lambda_{\min}(\Sigma^{-1})\|e\|_2^2 + B_\omega \|e\|_2. \tag{16}$$

Thus $\dot{V} < 0$ outside every ball of radius $R^* + \varepsilon$. The standard ultimate-boundedness argument gives

$$\limsup_{t\to\infty} \|\mu(t) - m_+\|_2 \leq R^*.$$

On the boundary of any ball centered at m_+ with radius greater than R^*, Equation (16) is strictly negative. Continuity of the field and the Poincaré–Bohl inward-pointing theorem therefore imply at least one zero inside the ball. This is an existence result, not a uniqueness or global-convergence claim.

It remains to verify $B_\omega < \infty$ and the stated constant. Put $r_i = \|\Sigma^{-1/2}(\mu - a_i)\|_2$, so $D_i = C_\Sigma + r_i^2/2$, and

$$\|\Sigma^{-1}(\mu - a_i)\|_2 \leq \sqrt{\lambda_{\max}(\Sigma^{-1})}\, r_i.$$

For $a = \lambda/c$ and $\widetilde{\tau}_i = \tau_i - C_\Sigma$, maximize

$$g_i(r) = r\exp\{-a[r^2/2 - \widetilde{\tau}_i]_+\}, \qquad r \geq 0.$$

If $\widetilde{\tau}_i < 0$, the taper is active at zero and the maximum occurs at $r = a^{-1/2}$, giving $a^{-1/2}e^{a\widetilde{\tau}_i - 1/2}$. If $\widetilde{\tau}_i \geq 0$, the untapered plateau ends at $r_\tau = \sqrt{2\widetilde{\tau}_i}$. When $a^{-1/2} \leq r_\tau$, the maximum is r_τ; otherwise it is the same interior-tail value. These cases show that g_i has a finite maximum, and the triangle inequality gives $B_\omega < \infty$.

For near-field fidelity, let μ^* be an uncontrolled equilibrium with $D_{\mu^*}(a_i) < \tau_i$ for every negative atom. Strictness and continuity provide a neighborhood in which every $\omega_i = 1$. The controlled and uncontrolled fields, including their Jacobians, are therefore identical there.

Finally, along any sequence $\|\mu_n\|_2 \to \infty$, each fixed atom satisfies $D_{\mu_n}(a_i) \to \infty$, hence $\omega_i(D_{\mu_n}(a_i)) \to 0$. Finiteness of the atom set gives $\rho_{\text{eff}}(\mu_n) \to 0$. □

A.6 Derivation and Proof of Theorem 4.3

At a fixed context, use a regular minimal exponential-family representation

$$\pi_\xi(a) = h(a)\exp\{\xi^\top T(a) - \psi(\xi)\}, \qquad \xi \in \mathcal{H},$$

and let

$$\mathcal{M} = \{\nabla\psi(\xi) : \xi \in \mathcal{H}\}$$

be the attainable mean-parameter space. For a fixed-covariance Gaussian, $\xi = \Sigma^{-1}\mu$, $T(a) = a$, and $\nabla\psi(\xi) = \mu$. For a categorical policy, a reference-logit gauge gives the corresponding minimal $(K-1)$-dimensional representation. Regularity and minimality imply $\nabla^2\psi(\xi) \succ 0$ on $\mathcal{H}$ and make $\nabla\psi : \mathcal{H} \to \mathcal{M}$ one-to-one.

Write

$$\widehat{A} = \widehat{A}^+ - \widehat{A}^-, \qquad p = \mathbb{E}_\nu[\widehat{A}^+], \qquad q = \mathbb{E}_\nu[\widehat{A}^-],$$

and define the unnormalized moments

$$\bar{\mathbf{m}}_+ = \mathbb{E}_\nu[\widehat{A}^+ T(a)], \qquad \bar{\mathbf{m}}_- = \mathbb{E}_\nu[\widehat{A}^- T(a)].$$

Because $p > 0$, let $\mathbf{m}_+ = \bar{\mathbf{m}}_+/p$; when $q > 0$, let $\mathbf{m}_- = \bar{\mathbf{m}}_-/q$. If $q = 0$, use only $\bar{\mathbf{m}}_- = \mathbf{0}$.

Set

$$b = \bar{\mathbf{m}}_+ - \bar{\mathbf{m}}_-.$$

Up to a constant independent of ξ, the signed objective, field, and Hessian are

$$\mathcal{L}(\xi) = b^\top\xi - (p-q)\psi(\xi) + C, \qquad \mathbf{F}(\xi) = b - (p-q)\nabla\psi(\xi),$$
$$J_{\mathbf{F}}(\xi) = \nabla^2\mathcal{L}(\xi) = -(p-q)\nabla^2\psi(\xi).$$

We consider

$$\dot{\xi} = \mathbf{F}(\xi), \qquad \xi_{t+1} = \xi_t + \alpha \mathbf{F}(\xi_t), \qquad \alpha > 0.$$

Proof. **Positive-dominant regime, including positive-only** $(p > q \geq 0)$**.** Define

$$\mathbf{m}^\star = \frac{b}{p-q}.$$

A finite equilibrium is equivalent to

$$\nabla\psi(\xi^\star) = \mathbf{m}^\star.$$

Thus, if $\mathbf{m}^\star \in \mathcal{M}$, existence and uniqueness follow from the bijectivity of $\nabla\psi : \mathcal{H} \to \mathcal{M}$; if $\mathbf{m}^\star \notin \mathcal{M}$, no finite equilibrium exists. Because $p - q > 0$,

$$\nabla^2\mathcal{L}(\xi) = -(p-q)\nabla^2\psi(\xi) \prec 0,$$

so any finite equilibrium is the unique finite maximizer of $\mathcal{L}$. Its continuous-time Jacobian is negative definite, hence it is locally asymptotically stable. The discrete local Jacobian is

$$J_{\mathrm{disc}}(\xi^\star) = I - \alpha(p-q)\nabla^2\psi(\xi^\star),$$

whose eigenvalues lie strictly inside the unit disk whenever

$$0 < \alpha < \frac{2}{(p-q)\lambda_{\max}(\nabla^2\psi(\xi^\star))}.$$

When $q = 0$, $\bar{\mathbf{m}}_- = \mathbf{0}$ and $\mathbf{m}^\star = \mathbf{m}_+$, recovering the positive-only case. When $q > 0$, with $\rho = q/p$,

$$\mathbf{m}^\star = \frac{p\mathbf{m}_+ - q\mathbf{m}_-}{p-q} = \mathbf{m}_+ + \frac{\rho}{1-\rho}(\mathbf{m}_+ - \mathbf{m}_-),$$

so

$$\mathbf{m}^\star - \mathbf{m}_+ = \frac{\rho}{1-\rho}(\mathbf{m}_+ - \mathbf{m}_-).$$

For fixed unequal $\mathbf{m}_+$ and $\mathbf{m}_-$, this algebraic signed-target displacement grows without bound as $\rho \to 1^-$; a finite equilibrium exists only while the target remains in $\mathcal{M}$.

If a sequence of reachable targets $\mathbf{m}_n^\star \in \mathcal{M}$ leaves every compact subset of $\mathcal{M}$, then the associated $\xi_n^\star = (\nabla\psi)^{-1}(\mathbf{m}_n^\star)$ leaves every compact subset of $\mathcal{H}$. Otherwise some compact $K \subset \mathcal{H}$ would contain a subsequence of $\xi_n^\star$, while its continuous image $\nabla\psi(K)$ would be compact in $\mathcal{M}$ and contain the corresponding targets, a contradiction.

Critical regime $(p = q > 0)$**.** Here

$$\mathbf{F}(\xi) = \bar{\mathbf{m}}_+ - \bar{\mathbf{m}}_- = p(\mathbf{m}_+ - \mathbf{m}_-)$$

is constant. If $\mathbf{m}_+ \neq \mathbf{m}_-$,

$$\xi(t) = \xi(0) + tp(\mathbf{m}_+ - \mathbf{m}_-), \qquad \xi_t = \xi_0 + t\alpha p(\mathbf{m}_+ - \mathbf{m}_-),$$

so both dynamics exhibit linear drift and have no finite equilibrium; the drift persists for all forward time when $\mathcal{H}$ is full or forward invariant along this direction. If $\mathbf{m}_+ = \mathbf{m}_-$, the field vanishes identically. Every point is stationary, but no point is an isolated locally asymptotically stable equilibrium.

Negative-dominant regime $(p < q)$**.** Now

$$\nabla^2\mathcal{L}(\xi) = (q-p)\nabla^2\psi(\xi) \succ 0.$$

Thus $\mathcal{L}$ is strictly convex. If a finite stationary point $\xi^\star$ exists, its continuous field Jacobian is positive definite, so it is repelling. The discrete local Jacobian is

$$I + \alpha(q-p)\nabla^2\psi(\xi^\star),$$

whose eigenvalues are $1 + \alpha(q-p)\lambda_i > 1$ for every $\alpha > 0$. Hence every finite stationary point is unstable in both time forms; if no stationary point exists, there is no finite stable equilibrium. □

Proof of Corollary 4.4. Because $\mathcal{K}$ is bounded, define

$$R_{\mathcal{K}} = \sup_{\mathbf{m}\in\mathcal{K}} \|\mathbf{m} - \mathbf{m}_+\|_2 < \infty, \qquad \delta = \|\mathbf{m}_+ - \mathbf{m}_-\|_2 > 0.$$

Equation (5) gives

$$\|\mathbf{m}^\star(\rho) - \mathbf{m}_+\|_2 = \frac{\rho}{1-\rho}\delta.$$

Therefore, with

$$\rho_{\mathcal{K}} = \frac{R_{\mathcal{K}}}{R_{\mathcal{K}} + \delta},$$

every $\rho > \rho_{\mathcal{K}}$ satisfies $\mathbf{m}^\star(\rho) \notin \mathcal{K}$. If the signed target remains in the interior of the attainable mean-parameter space, Theorem 4.3 gives a finite locally stable equilibrium outside $\mathcal{K}$; if attainability is lost first, no finite equilibrium remains. □

A.7 Proof of Theorem 4.5

Proof. Let $r = q - p > 0$.

Gaussian policy. Define the unnormalized signed moments

$$\bar{\mu}_+ = \mathbb{E}_\nu[\widehat{A}^+(a)a], \qquad \bar{\mu}_- = \mathbb{E}_\nu[\widehat{A}^-(a)a].$$

For a fixed-covariance Gaussian,

$$\nabla_\mu \log \pi_\mu(a) = \Sigma^{-1}(a - \mu),$$

so

$$\mathbf{F}_\mu(\mu) = \Sigma^{-1}\big[\bar{\mu}_+ - \bar{\mu}_- - (p - q)\mu\big].$$

Define

$$\mu^\dagger = \frac{\bar{\mu}_- - \bar{\mu}_+}{r}.$$

Then

$$\mathbf{F}_\mu(\mu) = r\Sigma^{-1}(\mu - \mu^\dagger).$$

Writing $\delta = \mu - \mu^\dagger$, the continuous and discrete solutions are

$$\delta(t) = e^{r\Sigma^{-1}t}\delta(0), \qquad \delta_t = \left(I + \alpha r\Sigma^{-1}\right)^t \delta_0.$$

Since $\Sigma^{-1} \succ 0$,

$$\|\delta(t)\|_2 \geq e^{r\lambda_{\min}(\Sigma^{-1})t}\|\delta(0)\|_2,$$

and

$$\|\delta_t\|_2 \geq \left[1 + \alpha r\lambda_{\min}(\Sigma^{-1})\right]^t \|\delta_0\|_2.$$

Thus $\mu^\dagger$ is repelling and every trajectory not initialized exactly there diverges. For any fixed historical action a,

$$\|\nabla_\mu \log \pi_\mu(a)\|_2 = \|\Sigma^{-1}(a - \mu)\|_2 \geq \lambda_{\min}(\Sigma^{-1})\|a - \mu\|_2 \longrightarrow \infty.$$

Categorical policy. Work on the zero-sum gauge

$$\mathcal{H}_0 = \{z \in \mathbb{R}^K : \mathbf{1}^\top z = 0\}.$$

Let the fixed signed action-mass vector be

$$b = \mathbb{E}_\nu[\widehat{A}(a)e_a] = \mathbb{E}_\nu[\widehat{A}^+(a)e_a] - \mathbb{E}_\nu[\widehat{A}^-(a)e_a].$$

Then

$$\mathcal{L}(z) = b^\top z + r\log\sum_{j=1}^K e^{z_j} + C,$$

with gradient and Hessian

$$g(z) = \nabla\mathcal{L}(z) = b + r\pi_z, \qquad H(z) = r\big(\mathrm{Diag}(\pi_z) - \pi_z\pi_z^\top\big).$$

Here g is the aggregate objective gradient, not the single-action score whose norm is bounded in Proposition 4.1. Because $\mathbf{1}^\top b = p - q = -r$, $\mathbf{1}^\top g(z) = 0$, so both continuous and discrete dynamics remain in $\mathcal{H}_0$. For every nonzero $v \in \mathcal{H}_0$,

$$v^\top H(z)v = r\ \mathrm{Var}_{A\sim\pi_z}(v_A) > 0.$$

Hence $\mathcal{L}$ is strictly convex on $\mathcal{H}_0$, has at most one finite stationary point, and any such point is repelling: the continuous Jacobian restricted to $\mathcal{H}_0$ is positive definite, while the discrete Jacobian restricted to $\mathcal{H}_0$ has eigenvalues strictly larger than one for every $\alpha > 0$.

It remains to prove escape for every nonstationary initialization. If $g(z_0) = 0$, both the continuous and discrete dynamics remain at the stationary point. Otherwise, since $g(z) = b + r\pi_z$ is bounded, the continuous flow $\dot{z} = g(z)$ exists for all forward time, and

$$c_0 = \|g(z_0)\|_2^2 > 0.$$

Along the flow,

$$\frac{d}{dt}\|g(z(t))\|_2^2 = 2g(z(t))^\top H(z(t))g(z(t)) \geq 0,$$

so $\|g(z(t))\|_2^2 \geq c_0$, and therefore

$$\frac{d}{dt}\mathcal{L}(z(t)) = \|g(z(t))\|_2^2 \geq c_0.$$

Thus $\mathcal{L}(z(t)) \geq \mathcal{L}(z_0) + c_0 t$.

For the discrete update

$$z_{t+1} = z_t + \alpha g_t, \qquad g_t = g(z_t),$$

gradient monotonicity gives

$$(g_{t+1} - g_t)^\top (z_{t+1} - z_t) \geq 0.$$

Since $z_{t+1} - z_t = \alpha g_t$, $g_{t+1}^\top g_t \geq \|g_t\|_2^2$, hence $\|g_{t+1}\|_2 \geq \|g_t\|_2 \geq \sqrt{c_0}$. Convexity then gives

$$\mathcal{L}(z_{t+1}) - \mathcal{L}(z_t) \geq g_t^\top (z_{t+1} - z_t) = \alpha\|g_t\|_2^2 \geq \alpha c_0.$$

Thus the discrete objective also grows at least linearly.

A continuous function is bounded above on every compact subset of $\mathcal{H}_0$. The preceding linear-growth bounds therefore imply that every nonstationary continuous or discrete trajectory eventually leaves every compact subset of $\mathcal{H}_0$.

Finally, on $\mathcal{H}_0$,

$$z_i = \log \pi_z(i) - \frac{1}{K}\sum_{j=1}^{K} \log \pi_z(j).$$

If the probability vectors stayed in a compact subset of the simplex interior, all coordinates would be bounded below by some $\varepsilon > 0$, and the inverse-softmax formula would keep the logits bounded. Escape in $\mathcal{H}_0$ therefore implies a subsequence along which at least one action probability approaches zero. Nevertheless, for every action a,

$$\|\nabla_z \log \pi_z(a)\|_2^2 = (1 - \pi_z(a))^2 + \sum_{j \neq a} \pi_z(j)^2 \leq 2(1 - \pi_z(a))^2 \leq 2,$$

as established in Proposition 4.1. □

A.8 Proof of Minimum-Order Control

If $I(D) = \Theta(D^m)$, there exist constants $0 < c_1 \leq c_2 < \infty$ and D_0 such that, for all $D \geq D_0$,

$$c_1 D^m \leq I(D) \leq c_2 D^m.$$

For any nonnegative taper $\omega(D)$,

$$c_1 D^m \omega(D) \leq \omega(D) I(D) \leq c_2 D^m \omega(D).$$

Hence $\omega(D)I(D) = O(1)$ if and only if $\omega(D) = O(D^{-m})$. At the exact matching order $\omega(D) = \Theta(D^{-m})$, the same bounds give $\omega(D)I(D) = \Theta(1)$.

A.9 Empirical Negative-Update Utility

Let $\widehat{J}_{\mathcal{B}}(\theta)$ be an empirical task-performance estimate on evaluation batch $\mathcal{B}$. For a negative sample $z = (s, a)$, define

$$g_\theta^-(z) = -\widehat{A}^-(z)\nabla_\theta \log \pi_\theta(a \mid s), \qquad \Delta\theta(z; w) = \eta w g_\theta^-(z), \qquad 0 \leq w \leq 1.$$

Its empirical and population utilities are

$$\widehat{\mathcal{U}}_{\mathcal{B}}(z; w \mid \theta) = \widehat{J}_{\mathcal{B}}\big(\theta + \Delta\theta(z; w)\big) - \widehat{J}_{\mathcal{B}}(\theta),$$
$$\mathcal{U}_\theta(z; w) = J\big(\theta + \eta w g_\theta^-(z)\big) - J(\theta).$$

The empirical utility-optimal weight is

$$\widehat{w}^{\star}_{\mathcal{U}}(z \mid \theta) \in \arg\max_{w\in[0,1]} \widehat{\mathcal{U}}_{\mathcal{B}}(z; w \mid \theta).$$

If $\widehat{J}_{\mathcal{B}}$ is unbiased for J, then for fixed θ, z, w,

$$\mathbb{E}_{\mathcal{B}}\left[\widehat{\mathcal{U}}_{\mathcal{B}}(z; w \mid \theta)\right] = \mathcal{U}_{\theta}(z; w).$$

Thus empirical utility is evaluation-batch dependent rather than an intrinsic deterministic sample label.

Lemma A.5 (Far-field shrinkage of the utility-optimal step)**.** *Let D index learner-relative remoteness and let $g(D) \neq 0$ be the corresponding negative-update direction. Define*

$$I(D) = \|g(D)\|_2, \qquad v(D) = \frac{g(D)}{I(D)}.$$

Suppose the local update space admits an orthogonal decomposition

$$\mathcal{T} = \mathcal{T}_{\mathrm{task}} \oplus \mathcal{T}_{\mathrm{far}},$$

with projection P_{task}, and suppose that:

1. $\nabla J(\theta) \in \mathcal{T}_{\mathrm{task}}$*;*
2. $\|P_{\mathrm{task}} v(D)\|_2 \to 0$ *as* $D \to \infty$*;*
3. *there exist $\rho_0 > 0$, $\kappa > 0$, and D_0 such that, for every $D \geq D_0$ and every $\rho \in [0, \rho_0]$,*
$$v(D)^{\top} \nabla^2 J\big(\theta + \rho v(D)\big) v(D) \leq -\kappa.$$

Define

$$\mathcal{V}_D(\rho) = J\big(\theta + \rho v(D)\big) - J(\theta), \qquad 0 \leq \rho \leq \rho_0.$$

Continuity of J implies that $\mathcal{V}_D$ attains a maximum on the compact interval $[0, \rho_0]$; let

$$\rho^{\star}_{\mathcal{U}}(D) \in \arg\max_{\rho\in[0,\rho_0]} \mathcal{V}_D(\rho).$$

Then

$$\rho^{\star}_{\mathcal{U}}(D) \to 0 \qquad \text{as } D \to \infty.$$

Moreover, for $D \geq D_0$, if $\langle \nabla J(\theta), v(D) \rangle < 0$, then

$$\mathcal{V}_D(\rho) < 0 \qquad \text{for every } \rho \in (0, \rho_0].$$

Proof. Let

$$a(D) = \langle \nabla J(\theta), v(D) \rangle.$$

Because $\nabla J(\theta) \in \mathcal{T}_{\mathrm{task}}$,

$$|a(D)| = |\langle \nabla J(\theta), P_{\mathrm{task}} v(D) \rangle| \leq \|\nabla J(\theta)\|_2 \|P_{\mathrm{task}} v(D)\|_2 \longrightarrow 0.$$

For $D \geq D_0$ and $\rho \in (0, \rho_0]$, Taylor's theorem gives some $\xi \in (0, \rho)$, and the curvature assumption yields

$$\mathcal{V}_D(\rho) \leq \rho a(D) - \frac{\kappa}{2}\rho^2.$$

The same bound is trivial at $\rho = 0$. Since $\mathcal{V}_D(0) = 0$, a maximizing step must have nonnegative utility. The right-hand side is strictly negative whenever

$$\rho > \frac{2[a(D)]_+}{\kappa},$$

so

$$0 \leq \rho^{\star}_{\mathcal{U}}(D) \leq \frac{2[a(D)]_+}{\kappa} \longrightarrow 0.$$

If $a(D) < 0$, the same upper bound is strictly negative for every $\rho > 0$, proving the second claim. □

This is a sufficient local mechanism, not a universal monotonicity law. It applies when far-field negative updates become asymptotically misaligned with the task-improving direction and the evaluated objective is locally harmed along those directions. The curvature condition can represent

local losses caused by parameter drift or policy-support contraction, provided those effects reduce the evaluated objective J. If far-field growth remains aligned with $\nabla J(\theta)$, the conclusion need not hold.

A sample weight w produces effective step length

$$\rho(D; w) = \eta w I(D).$$

Because the lemma is local to $[0, \rho_0]$, define

$$w^\star_{\mathcal{U},\mathrm{loc}}(D) \in \arg \max_{\substack{0 \le w \le 1 \\ \eta w I(D) \le \rho_0}} \mathcal{V}_D\big(\eta w I(D)\big).$$

The same quadratic bound and feasibility of the zero step yield

$$0 \le \eta I(D) w^\star_{\mathcal{U},\mathrm{loc}}(D) \le \frac{2[a(D)]_+}{\kappa} \longrightarrow 0.$$

For fixed $\eta > 0$, if $\liminf_{D\to\infty} I(D) > 0$, then $w^\star_{\mathcal{U},\mathrm{loc}}(D) \to 0$ as well.

A.10 Proofs for Redescending Far-Field Control

Proof of Proposition 5.2. For $D > \tau$,

$$\omega_{\mathrm{Exp}}(D) = e^{-(\lambda/c)(D-\tau)}.$$

Hence

$$0 \le \omega_{\mathrm{Exp}}(D) I(D) \le C e^{(\lambda/c)\tau} e^{-(\lambda/c)D} (1+D)^m \longrightarrow 0,$$

since exponential decay dominates every finite-order polynomial. □

Proof of Corollary 5.3. For each fixed negative atom, asymptotic remoteness and $\omega_i(D_i) \to 0$ imply that its tapered weight vanishes. Because the negative set is finite and its coefficients and sufficient statistics are fixed,

$$q_\omega \longrightarrow 0, \qquad \mathbf{M}_{-,\omega} \longrightarrow \mathbf{0}.$$

Since $p > 0$, eventually $q_\omega < p$, and therefore

$$\mathbf{m}^\star_\omega = \frac{p\mathbf{m}_+ - \mathbf{M}_{-,\omega}}{p - q_\omega} \longrightarrow \mathbf{m}_+.$$

If $\mathbf{m}_+ \in \mathrm{int}(\mathcal{K})$, some neighborhood of $\mathbf{m}_+$ lies inside $\mathcal{K}$; convergence then implies $\mathbf{m}^\star_\omega \in \mathcal{K}$ eventually. □

A.11 Categorical One-vs-Rest Dynamics

Theorem A.6 (Categorical self-limiting suppression)**.** *For a repeatedly suppressed action i, define $y_i = z_i - \log \sum_{j\neq i} e^{z_j}$, so $\pi_i = \sigma(y_i)$. Under an uncontrolled negative atom of mass $q > 0$,*

$$\dot{y}_i = -q(1-\pi_i)c_K(t), \qquad \frac{K}{K-1} \le c_K(t) \le 2.$$

Hence y_i eventually decreases linearly and π_i exponentially. With $\omega_i = \min\{1, (e^{\tau_i}\pi_i)^\beta\}$, $\beta = \lambda/c$, the far branch instead satisfies

$$y_i(t) = -\beta^{-1} \log t + O(1), \qquad \pi_i(t) = \Theta(t^{-1/\beta}).$$

Thus tapering changes persistent suppression from exponential to polynomial probability decay; it does not freeze the action probability.

Proof of Theorem A.6. For a single negative atom i, the direct-logit field is

$$\dot{z} = -q(e_i - \pi), \qquad \dot{z}_i = -q(1-\pi_i), \qquad \dot{z}_j = q\pi_j \quad (j \neq i).$$

Define the normalized off-target probabilities $\bar{\pi}_j = \pi_j/(1-\pi_i)$. Differentiating $y_i = z_i - \log \sum_{j\neq i} e^{z_j}$ gives

$$\dot{y}_i = -q(1-\pi_i) - q\frac{\sum_{j\neq i}\pi_j^2}{1-\pi_i} \tag{17}$$

$$= -q(1-\pi_i)c_K(t), \qquad c_K(t) = 1 + \sum_{j\neq i} \bar{\pi}_j^2. \tag{18}$$

Because $\bar{\pi}$ is a distribution on $K-1$ actions,

$$\frac{1}{K-1} \le \sum_{j \neq i} \bar{\pi}_j^2 \le 1,$$

which proves the coefficient envelope without a concentration assumption.

We next justify entry into the far half-line before using its bounds. The log-odds is strictly decreasing. Until $\pi_i < 1/2$, monotonicity gives $1-\pi_i(t) \ge 1-\pi_i(0) > 0$, so Equation (18) has a strictly negative upper bound. The trajectory therefore crosses $\pi_i = 1/2$ in finite time. Thereafter $1-\pi_i \ge 1/2$, so $y_i \to -\infty$ at a linear rate and $\pi_i = \sigma(y_i)$ decays exponentially.

For the generic coefficient, suppose $j^* \neq i$ is the unique largest off-target logit initially. For every $k \neq i, j^*$,

$$\frac{d}{dt}(z_{j^*} - z_k) = q(\pi_{j^*} - \pi_k) > 0.$$

The initial ordering is therefore preserved. If a gap stayed bounded, then, once $\pi_i \to 0$, maximality of j^* and the positive initial gap would give a positive lower bound on $\pi_{j^*} - \pi_k$, a contradiction. Hence all such gaps diverge, off-target mass concentrates on j^*, and $c_K(t) \to 2$. Under complete off-target symmetry, symmetry is preserved and $\bar{\pi}_j = 1/(K-1)$, giving $c_K(t) \equiv K/(K-1)$.

Now apply the taper $\omega_i = \min\{1, (e^{\tau_i}\pi_i)^\beta\}$. Since the tapered log-odds is strictly decreasing with a rate bounded away from zero on every bounded log-odds interval, it reaches the far branch $\pi_i < \min\{1/2, e^{-\tau_i}\}$ in finite time. Thereafter let $U = e^{-\beta y_i}$. Since $\pi_i e^{-y_i} = 1-\pi_i$,

$$\dot{U} = -\beta e^{-\beta y_i}\dot{y}_i \tag{19}$$

$$= \beta q e^{\beta\tau_i} c_K(t)(1-\pi_i)^{\beta+1}. \tag{20}$$

Thereafter, the right-hand side is bounded above and below by positive constants. Thus $U = \Theta(t)$, which is equivalent to

$$y_i(t) = -\beta^{-1}\log t + O(1), \qquad \pi_i(t) = \Theta(t^{-1/\beta}).$$

□

A.12 Categorical Restoring-Force Balance

Write the one-dimensional restored dynamics as

$$\dot{y} = F(y) = E(y) - S_{\mathrm{tap}}(y),$$

where $S_{\mathrm{tap}} > 0$ is downward suppression and $E > 0$ is upward restoration.

Proposition A.7 (Finite restoring crossing)**.** *Assume E and S_{tap} are continuous and the zeros of $F = E - S_{\mathrm{tap}}$ in $[y_L, y_R]$ are isolated. If $F(y_L) > 0$ and $F(y_R) < 0$, then the interval contains a zero at which the phase line has a positive-to-negative crossing. Such a crossing is locally asymptotically stable; if it is hyperbolic, its condition is*

$$F'(y^*) < 0 \quad \Longleftrightarrow \quad [S_{\mathrm{tap}} - E]'(y^*) > 0.$$

Proof. The intermediate value theorem gives at least one zero. Because the zeros of F are isolated on a compact interval, only finitely many occur; as the endpoint signs differ, at least one has positive sign immediately to its left and negative sign immediately to its right. The flow then points toward that zero from both sides, proving phase-line stability. At a hyperbolic crossing this sign change is equivalent to $F'(y^*) < 0$, and the displayed derivative identity follows from $F = E - S_{\mathrm{tap}}$. □

In the binary or symmetric one-vs-rest tail, the tapered suppression has order $e^{\beta y}$, whereas entropy restoration has order $|y|e^y$ as $y \to -\infty$. Hence $\beta \ge 1$ supplies the required far-tail ordering (with the $|y|$ factor breaking the tie at $\beta = 1$), but it does not by itself verify the finite phase-line conditions of Proposition A.7.

B Finite-Measure I-Divergence Reweighting Interpretation

DRPO's exponential attenuation admits an optimistic finite-measure analogue of divergence-ball DRO under generalized I-divergence (Csiszár, 1975). Related optimistic distributional objectives

appear in likelihood approximation and policy optimization (Nguyen et al., 2019; Song & Zhao, 2020). Unlike normalized-distribution ambiguity sets, the finite negative-update measure need not preserve total mass, so reweighting may attenuate total negative-update mass. Let $d\xi^-(z) = \widehat{A}^-(z)d\nu(z)$ and $h(z) = [D(z) - \tau(z)]_+$. Consider

$$\min_{0\le w\le 1} \int wh\, d\xi^- \quad \text{s.t.} \quad \int (w\log w - w + 1)d\xi^- \le \delta. \tag{21}$$

Proposition B.1 (Finite-measure exponential reweighting)**.** *Let $\delta_0 = \xi^-(\{h > 0\})$. For $0 < \delta < \delta_0$, the constraint in Equation* (21) *is active and has a unique finite $\kappa > 0$ with $w_\kappa(z) = e^{-h(z)/\kappa}$. At $\delta = \delta_0$, the endpoint is $w_0 = \mathbf{1}\{h = 0\}$; for $\delta > \delta_0$, zero objective is already attainable and no finite one-to-one κ–δ relation remains. The later endpoint $\delta \ge |\xi^-|$ merely makes $w \equiv 0$ feasible.*

Proof of Proposition B.1. Use the convention $\phi(0) = 1$ for $\phi(w) = w\log w - w + 1$. For an active constraint, the Lagrangian with multiplier $\kappa > 0$ is

$$\mathcal{L}(w, \kappa) = \int \{wh + \kappa\phi(w)\}\, d\xi^- - \kappa\delta.$$

Pointwise stationarity on $0 < w \le 1$ gives $h + \kappa\log w = 0$, hence

$$w_\kappa(z) = e^{-h(z)/\kappa}.$$

For $h = 0$, this equals one; for $h > 0$, it lies strictly between zero and one. Define $\mathcal{D}(\kappa) = \int \phi(w_\kappa)d\xi^-$. Dominated convergence gives

$$\lim_{\kappa\to\infty} \mathcal{D}(\kappa) = 0, \qquad \lim_{\kappa\downarrow 0} \mathcal{D}(\kappa) = \xi^-(\{h > 0\}) = \delta_0.$$

On a set of positive h, increasing κ strictly increases w_κ toward one and strictly decreases $\phi(w_\kappa)$. Thus $\mathcal{D}(\kappa)$ is continuous and strictly decreasing from δ_0 to zero, proving existence and uniqueness for $0 < \delta < \delta_0$.

At $\delta = \delta_0$, the limit is $w_0 = \mathbf{1}\{h = 0\}$, which has zero objective and divergence δ_0. For $\delta > \delta_0$, the same zero-objective point is strictly feasible, so the constraint need not be active and there is no finite one-to-one multiplier relation. Finally, $\int \phi(0)d\xi^- = |\xi^-|$, so $w \equiv 0$ becomes feasible only at the later endpoint $\delta \ge |\xi^-|$. □

C Where Off-Policy Controls Intervene

A weighted actor direction has the schematic form

$$\mathbf{g}(\theta) \propto \mathbb{E}_{z\sim\nu}\left[w_\theta(z)\,\widehat{A}(z)\,\nabla_\theta \log \pi_\theta(a \mid s)\right].$$

We use an operational distinction: data controls alter the base measure ν; signal controls transform the signed feedback or apply sign-branch-wide scaling; sample-influence controls hold the sampled data and signed feedback fixed while conditioning per-sample influence on policy-dependent quantities. Policy anchoring instead uses dataset actions as fitting targets or a reference policy to define policy optimization.

The taxonomy classifies intervention points rather than algorithms; a method may instantiate more than one mechanism.

Sample influence. Canonical TOPR attenuates negative feedback using the current-to-behavior clipped ratio

$$w^-_{\mathrm{TOPR}}(s, a) = \min\left\{\frac{\pi_\theta(a \mid s)}{\mu(a \mid s)}, 1\right\},$$

so its control coordinate is behavior-relative mismatch (Le Roux et al., 2025). DRPO instead conditions negative-sample influence on learner-relative remoteness under the current policy, with a designed near–far boundary and decay law. Both intervene at the sample-influence level, but they condition influence on different policy-dependent coordinates.

Policy anchoring. Policy-anchoring mechanisms use dataset actions as fitting targets or a reference policy to define policy optimization. TD3+BC adds behavior-cloning pressure to the actor objective

Question	Core control logic	Representative methods	Primary intervention
Which data are used?	Data selection / resampling	Filtering, curation, resampling	Changes the sampling distribution ν.
What learning signal do they carry?	Signed-feedback shaping / global negative scaling	AsymRE; HPO-style fixed hysteretic weighting; Positive-only / Global-α controls	Changes signed feedback or the global positive–negative balance.
How much does each sample contribute?	Sample-level weighting	Importance weighting; TOPR; DRPO	Modulates sample influence using behavior-relative mismatch or learner-relative remoteness.
How is the policy anchored?	Behavior / reference anchoring	TD3+BC; AGRO; DPO	Uses dataset actions as fitting targets or defines policy optimization relative to a persistent reference policy.

Table 3: Four intervention questions for organizing off-policy controls.

(Fujimoto & Gu, 2021). AGRO uses a response-consistency-based objective derived from KL-regularized policy optimization to learn from on- and off-policy data (Tang et al., 2026). DPO defines preference optimization through reference-relative log-probability ratios (Rafailov et al., 2023).

D Experimental Details

D.1 D4RL Locomotion Gradient-Diagnostic Protocol

Critic fitting and frozen advantages. For each of the nine D4RL locomotion task–dataset combinations, we train one value critic, then compute, standardize, and freeze the transition-level learned-critic advantages. All actor seeds for a given task–dataset combination use the same frozen critic outputs.

Positive-only reference actors. For each seed, we train a Gaussian reference actor using only transitions with positive frozen advantage. The terminal actor checkpoint is fixed before the gradient diagnostic and serves only to define learner-relative distance and the full-parameter gradient measurement.

Matched negative-transition probe. The probe population contains transitions with negative frozen advantage. We form near and far pools by standardized distance, match them by absolute advantage, and evaluate the full actor-parameter gradient on the matched probe pairs. The probe performs no optimizer update.

Aggregation. Every far point is normalized by its original matched near partner. Curves are first aggregated within seed and task–dataset combination; the main panel then assigns equal weight to the nine task–dataset combinations on their shared distance support. The 95% confidence interval is obtained by hierarchical bootstrap over task–dataset combinations and seeds.

D.2 Structured-Generation Gradient-Diagnostic Protocol

Probe definition. For each structured-generation task, learner-relative remoteness is measured by the current-policy mean completion-token surprisal,

$$D_{\mathrm{SG}}(x, y) = -\frac{1}{T}\sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t}),$$

where only completion tokens are included and prompt and padding positions are excluded. Define

$$\bar{\ell}_\theta(x, y) = \frac{1}{T}\sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t}),$$

and measure

$$G_{\mathrm{SG}}(x, y) = \left\| \nabla_{\theta_{\mathrm{train}}} \bar{\ell}_\theta(x, y) \right\|_2 ,$$

Table 4: Configuration of the D4RL locomotion gradient diagnostic.

Component	Configuration
Datasets	HalfCheetah, Hopper, Walker2d × medium, medium-replay, medium-expert
Critic	One learned value critic per dataset; 100,000 optimizer updates
Advantage	Transition-level learned-critic advantage; standardized and frozen
Reference actor	Positive-only Gaussian actor; 100,000 optimizer updates
Number of seeds	10
Probe population	Negative frozen-advantage transitions
Near/far pools	Standardized-distance quantiles 0.25 / 0.75
Advantage matching	20 bins; 5% relative tolerance
Matched pool	256 pairs per seed
Full-gradient subset	64 pairs per seed
Distance summary	Seven far-distance bins with original-pair normalization
Probe updates	None
Cross-dataset summary	Within-seed/task–dataset normalization; equal task–dataset weighting
Uncertainty	Hierarchical bootstrap over task–dataset combinations and seeds

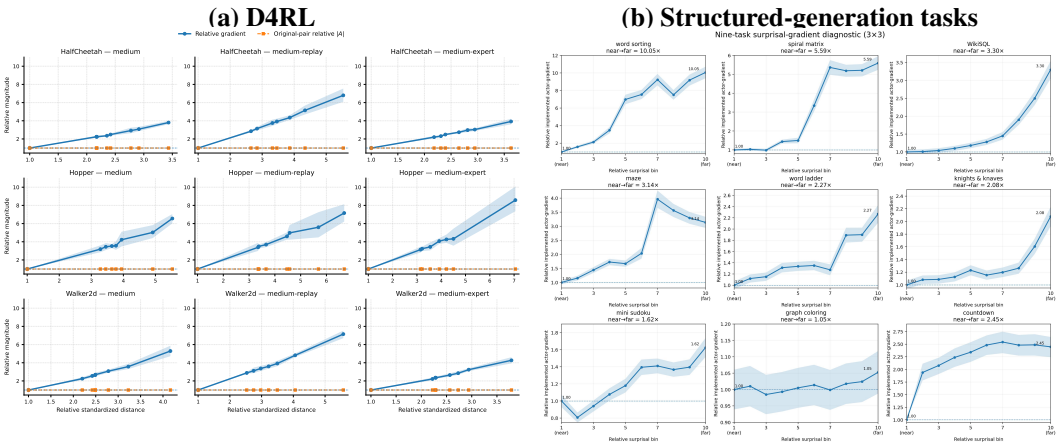


Figure 4: **Disaggregated external far-field diagnostics.** (a) Nine D4RL locomotion task–dataset combinations. (b) Nine structured-generation tasks.

where the gradient is taken only with respect to trainable policy parameters at the fixed probe checkpoint, and no optimizer update is applied during the probe.

Fixed-coefficient gradient measurement. All probed negative responses use the same absolute negative coefficient. Thus, negative-signal magnitude is held fixed while learner-relative surprisal varies. This measures how the full-parameter gradient changes with surprisal at fixed negative-signal magnitude.

Within-task normalization and aggregation. Within each task, responses are ordered by current-policy surprisal and partitioned into ten equal-count bins. The mean gradient norm in each bin is normalized by the lowest-surprisal bin of the same task. The task-level normalized curves are then aggregated with equal task weight; the corresponding task-level curves are reported separately in Figure 4. Confidence intervals are obtained by cluster bootstrap within each structured-generation task, using that task's evaluation instance as the resampling unit, followed by equal-weight aggregation across tasks where applicable.

D.3 Unified Continuous Environment C-U1

C-U1 is the shared controlled continuous environment used by all continuous mechanism experiments. A context $s \in \mathbb{R}^6$ is sampled from $\mathcal{N}(0, I_6)$. For each seed, we independently sample 4096 training

contexts and 4096 test contexts from the same distribution. Test contexts never participate in optimization and are used only to measure held-out-context generalization.

State-dependent task geometry. Each context determines a two-dimensional positive-support center $a_+(s)$ and a unit task direction $u(s)$. We define

$$\begin{aligned}[a_+(s)]_1 &= 0.70\tanh\big(0.85s_1 - 0.30s_2s_3 + 0.20\sin(1.6s_4)\big),\\ [a_+(s)]_2 &= 0.65\tanh\big(-0.50s_2 + 0.35\cos(1.1s_5) + 0.22s_1s_6\big).\end{aligned} \tag{22}$$

The direction angle is

$$\begin{aligned}\varphi(s) = 1.15\tanh\big(&0.75s_1 + 0.50s_3 - 0.30s_6\big)\\ &+ 0.30\sin(1.35s_2),\end{aligned} \tag{23}$$

with

$$u(s) = \begin{bmatrix}\cos\varphi(s)\\ \sin\varphi(s)\end{bmatrix}, \qquad v(s) = \begin{bmatrix}-u_2(s)\\ u_1(s)\end{bmatrix}. \tag{24}$$

The unobserved task optimum and the directionally useful negative anchor are

$$\begin{aligned}a_\star(s) &= a_+(s) + 0.70u(s),\\ a_-(s) &= a_+(s) - 0.50u(s).\end{aligned} \tag{25}$$

Positive and negative actions. Each context is associated with four positive actions and eight negative actions. The positive actions lie on a reward contour of radius $r_+ = 0.75$ centered at $a_\star(s)$:

$$a_{\mathrm{pos},k}(s) = a_\star(s) + r_+\big(\cos\phi_k\, u(s) + \sin\phi_k\, v(s)\big), \tag{26}$$

where

$$\begin{aligned}\{\phi_k\}_{k=1}^4 &= \{\pi - \theta_1,\, \pi + \theta_1,\, \pi - \theta_2,\, \pi + \theta_2\},\\ \theta_1 = 0.20, &\qquad \cos\theta_2 = \frac{2(0.70)}{0.75} - \cos\theta_1.\end{aligned} \tag{27}$$

This construction makes the four actions equal-reward while ensuring that their centroid is exactly $a_+(s)$. It also leaves nonzero conditional action spread, preventing deterministic maximum-likelihood variance contraction from being confused with the far-field mechanism.

The eight negative actions lie on a second reward contour of radius $r_- = 1.20$:

$$a_{\mathrm{neg},j}(s) = a_\star(s) + r_-\big(\cos\psi_j\, u(s) + \sin\psi_j\, v(s)\big), \tag{28}$$

with

$$\{\psi_j\}_{j=1}^8 = \left\{\pi,\, \frac{3\pi}{4},\, \frac{\pi}{2},\, \frac{\pi}{4},\, 0,\, -\frac{\pi}{4},\, -\frac{\pi}{2},\, -\frac{3\pi}{4}\right\}. \tag{29}$$

The first negative action equals $a_-(s)$. All eight negative actions have the same distance to $a_\star(s)$ and therefore the same reward and fixed advantage, while having different distances from the current policy.

Reward and fixed advantages. The ground-truth reward is

$$R(s,a) = \exp\left(-\frac{\|a - a_\star(s)\|_2^2}{2(0.75)^2}\right). \tag{30}$$

Advantages are computed once using the fixed baseline 0.40:

$$\widehat{A}(s,a) = R(s,a) - 0.40, \tag{31}$$

and remain frozen throughout optimization. Consequently, all four positive actions have positive advantages, all eight negative actions have negative advantages, and negative-action quality is exactly matched within each context.

Policy parameterization. The policy is an isotropic state-conditioned Gaussian,

$$\pi_\theta(a \mid s) = \mathcal{N}\left(\mu_\theta(s), \sigma_\theta^2(s) I_2\right). \tag{32}$$

Both $\mu_\theta(s)$ and the scalar $\log\sigma_\theta(s)$ are produced by a shared two-layer ReLU MLP with 64 hidden units per layer and separate output heads. The initial standard deviation is 0.60, and the learnable-variance experiments use no artificial variance clamp. The continuous mechanism results reported in the main text use the fixed-variance branch unless explicitly labeled otherwise. A finite crossing of

$\log \sigma_\theta(s) < -12$ is recorded as a support or variance-boundary event, whereas nonfinite parameters or outputs are reported separately as NaN/Inf numerical failure.

Training minibatches are sampled by context, and all actions associated with a selected context are retrieved together. Thus, the replicated actions within one context are not treated as independent contexts. Positive and negative objectives are averaged separately before their relative coefficient is applied, preventing the four-to-eight action count ratio from mechanically changing the total negative-update mass.

Use across continuous experiments. All C-U1 comparisons use the same contexts, reward geometry, fixed advantages, and policy architecture. When near–far membership is used, it is recomputed from the current standardized distance

$$d_\theta(s,a) = \frac{\|a - \mu_\theta(s)\|_2}{\sigma_\theta(s)}, \tag{33}$$

with a fixed near–far threshold of 5.0.

D.4 Controlled Categorical Environment D-U1

D-U1 denotes the controlled categorical environment family used to study persistent negative suppression, probability-support boundaries, and shared-semantic generalization. Its shared-semantic instantiation uses a fixed catalogue of 64 unordered actions, each associated with a four-dimensional unit semantic vector. Action identifiers are randomly permuted for every seed, so numerical action indices carry no geometric ordering.

For each seed, we independently sample 2048 training contexts and 2048 test contexts from $s \sim \mathcal{N}(0, I_6)$. The two splits follow the same context distribution, and test contexts never participate in optimization. Results on this split are therefore reported as held-out-context or unseen-state generalization rather than OOD generalization.

Semantic action catalogue. Let $\{e_i\}_{i=1}^K \subset \mathbb{R}^4$, with $K = 64$, denote the unit reward embeddings of the categorical actions. They are sampled once per seed and randomly assigned to action identifiers. For each context, two fixed linear maps generate a demonstrated semantic target and a candidate improvement direction:

$$\begin{aligned} t_+(s) &= \mathrm{unit}\big(W_+^\top s\big), \\ \widetilde{u}(s) &= W_u^\top s - \big\langle W_u^\top s, t_+(s)\big\rangle\, t_+(s), \\ u(s) &= \mathrm{unit}(\widetilde{u}(s)). \end{aligned} \tag{34}$$

The hidden optimal target and the local negative target are

$$\begin{aligned} t_\star(s) &= \mathrm{unit}(t_+(s) + \delta u(s)), \\ t_-(s) &= \mathrm{unit}(t_+(s) - \delta u(s)), \end{aligned} \qquad \delta = 0.45. \tag{35}$$

Reward and action roles. The ground-truth reward of action i is determined by its semantic alignment with the hidden target:

$$R(s,i) = \frac{1}{2}\left(1 + t_\star(s)^\top e_i\right). \tag{36}$$

The hidden optimal action is

$$i_\star(s) = \arg\max_i t_\star(s)^\top e_i. \tag{37}$$

It is excluded from all positive demonstrations.

For each context, the four positive actions are the available actions with the largest similarity to $t_+(s)$. After excluding the hidden optimum and positive actions, the local negative action is the action most aligned with $t_-(s)$. The four far-negative actions are selected by their alignment with $-t_+(s)$. Thus,

$$\begin{aligned} \mathcal{P}(s) &= \mathrm{Top4}_{i \neq i_\star(s)}\ t_+(s)^\top e_i, \\ i_{\mathrm{local}}(s) &= \arg\max_{i \notin \{i_\star(s)\} \cup \mathcal{P}(s)} t_-(s)^\top e_i, \\ \mathcal{F}(s) &= \mathrm{Top4}_{i \notin \{i_\star(s), i_{\mathrm{local}}(s)\} \cup \mathcal{P}(s)} \left[-t_+(s)^\top e_i\right]. \end{aligned} \tag{38}$$

The positive and negative advantages are fixed throughout training:

$$\widehat{A}(s,i) = \begin{cases} +1, & i \in \mathcal{P}(s), \\ -1, & i = i_{\text{local}}(s) \text{ or } i \in \mathcal{F}(s). \end{cases} \tag{39}$$

Consequently, differences between local and far-negative updates cannot be attributed to different advantage magnitudes.

Shared-semantic categorical policy. The policy uses a two-layer tanh MLP with 64 hidden units per layer to map the context to a unit semantic direction $q_\theta(s) \in \mathbb{R}^4$. Let $\widetilde{e}_i$ denote the action embedding used by the policy. The categorical logits and probabilities are

$$\begin{aligned} \ell_{\theta,i}(s) &= \kappa_\theta(s) q_\theta(s)^\top \widetilde{e}_i, \\ \pi_\theta(i \mid s) &= \frac{\exp(\ell_{\theta,i}(s))}{\sum_{j=1}^{K} \exp(\ell_{\theta,j}(s))}. \end{aligned} \tag{40}$$

The concentration is either fixed at $\kappa_\theta(s) = 8$ or learned as

$$\kappa_\theta(s) = \text{softplus}(c_\theta(s)) + 0.05, \tag{41}$$

with initial value 8 and no upper clamp.

In the semantically aligned condition, $\widetilde{e}_i = e_i$. In the shuffled control, the same policy embeddings are randomly permuted across action identifiers while the reward embeddings and training samples remain unchanged. This intervention preserves the catalogue size and categorical parameterization but breaks the correspondence between policy-side similarity and ground-truth semantic utility.

Evaluation and failure events. We evaluate expected semantic reward, probability assigned to the hidden optimal action, entropy, effective support, and normalized semantic extrapolation on both training and held-out contexts. Task-performance collapse, effective-support or concentration-boundary events, and NaN/Inf numerical failure are reported separately. A policy can therefore retain high reward while still approaching a support boundary; the two outcomes are not treated as equivalent.

Use across categorical experiments. D-U1 uses fixed concentration for negative-strength sweeps, learnable concentration for support-dynamics interventions, aligned/shuffled embeddings for semantic-transfer controls, and a direct-softmax diagnostic for repeated suppression.

D.5 Structured-Generation Tasks and Dataset Construction

The structured-generation evaluation contains nine tasks: Countdown, Word Sorting, Spiral Matrix, Mini Sudoku, Maze, Word Ladder, Knights & Knaves, Graph Coloring, and WikiSQL. Countdown follows the arithmetic task of Hutton (2002) and the verifier-guided LLM formulation used by TinyZero (Pan et al., 2025); Word Sorting, Spiral Matrix, Mini Sudoku, Maze, Word Ladder, Knights & Knaves, and Graph Coloring are generated from Reasoning Gym (Stojanovski et al., 2025); WikiSQL follows Zhong et al. (2017). The versions of the task sources and generators are fixed before data generation.

All tasks use a common prompt-indexed replay bank containing the task identity, prompt, verifier-confirmed oracle completion, verifier outcomes, model-independent negative completions, and task metadata. Learner-relative quantities are excluded from bank construction: current-policy probability, surprisal, near/far labels, and taper weights are recomputed from the learner when required rather than stored with the data. The replay banks remain frozen during optimization.

Each structured-generation task uses 5,000 training prompts and 500 validation prompts under the common structured-generation protocol. Reserved test data are not used for coefficient tuning. Each training prompt uses 16 negative responses; the replay data remain fixed during optimization, while learner-relative mean completion-token surprisal is recomputed from the current policy. This zero-refresh construction isolates stale-negative reuse without storing learner-relative remoteness in the bank.

Table 5: Structured-generation tasks and verified outputs.

Task	Source	Verified output
Countdown	Generated	Arithmetic expression
Word Sorting	Reasoning Gym	Word permutation
Spiral Matrix	Reasoning Gym	Integer sequence
Mini Sudoku	Reasoning Gym	4×4 grid
Maze	Reasoning Gym	Shortest-path length
Word Ladder	Reasoning Gym	Word path
Knights & Knaves	Reasoning Gym	Role assignment
Graph Coloring	Reasoning Gym	Vertex-color map
WikiSQL	WikiSQL	SQL logical form

E Controlled Experimental Protocols

The following protocols define the controlled source-isolation, causal-transmission, negative-strength, and taper-comparison experiments. C-U1 and D-U1 are defined in Appendix D; experiment-specific training and evaluation details are given with the corresponding protocol.

E.1 Shared Notation and Controls

For a negative sample $z = (s, a)$, define its raw repulsive update as

$$g_\theta^-(z) = -\widehat{A}^-(z)\nabla_\theta \log \pi_\theta(a \mid s), \qquad \widehat{A}^-(z) = \max\{-\widehat{A}(z), 0\}. \tag{42}$$

Learner-relative remoteness is measured by

$$D_\theta(z) = -\log \pi_\theta(a \mid s), \tag{43}$$

or by its equivalent standardized Gaussian component when additive density constants are irrelevant.

Within each controlled comparison, all branches use the same contexts, samples, fixed rewards or advantage labels, initialization, optimizer, minibatch order, training horizon, and evaluation schedule. Only the intervention variable changes. When multiple actions share a context, statistical aggregation is performed at the context level rather than treating those actions as independent contexts.

E.2 Source-Isolation Protocol

Purpose. The source-isolation experiment asks whether policy-relative remoteness changes negative-update magnitude when sample quality is held fixed. It identifies the source of a per-sample difference; the causal-transmission protocol below tests its downstream effect on training.

Continuous product-manifold construction. For each C-U1 context, we select multiple negative actions from a common reward contour:

$$R(s, a_i) = R(s, a_j), \qquad \widehat{A}(s, a_i) = \widehat{A}(s, a_j) < 0, \tag{44}$$

while ensuring that

$$D_\theta(s, a_i) \neq D_\theta(s, a_j). \tag{45}$$

The reward coordinate and the policy-distance coordinate are therefore independently controlled. This exact counterfactual comparison is the reason for using a synthetic product-manifold construction: an external offline dataset generally cannot provide two actions with precisely matched task quality and independently prescribed distance from the same current policy.

The comparison is performed both at policy initialization and after positive-only pretraining. No negative sample used as a probe is applied as an optimizer update during this protocol. Let $u_\theta(s)$ denote the policy-output coordinate used by the controlled probe, e.g., fixed-covariance Gaussian mean coordinates in C-U1 and direct logits in D-U1. For every probe, we record

$$\begin{aligned} I_{\text{score}}(z) &= \|\nabla_u \log \pi_u(a \mid s)\|_2 , \\ I_{\text{sample}}(z) &= |\widehat{A}(z)|\, I_{\text{score}}(z). \end{aligned} \tag{46}$$

Because $|\widehat{A}|$ is identical across the matched actions, differences in I_{sample} arise from policy-score geometry rather than advantage magnitude.

For neural policies, we also measure the full parameter-gradient norm and the aggregate gradient obtained from samples in the same remoteness range. These full-parameter actor-gradient diagnostics assess transfer in trainable networks, complementing the output-coordinate score measures used for controlled source isolation.

We report near-to-far ratios of remoteness, score norm, single-sample gradient norm, aggregate gradient norm, and gradient-direction coherence. Ratios are first computed within each context and seed and are then aggregated across seeds.

Categorical counterpart. D-U1 uses unordered categorical actions with identical fixed negative labels. The comparison varies the current probability assigned to the negative action while preserving its reward role and advantage:

$$\widehat{A}(s,i) = \widehat{A}(s,j) < 0, \qquad \pi_\theta(i \mid s) \neq \pi_\theta(j \mid s). \tag{47}$$

For a direct categorical logit parameterization, the score norm is bounded and need not grow without limit as probability decreases. The relevant far-field effect is instead persistence: the negative update continues to increase the selected action's logit gap even after its probability has become small. We therefore report probability, surprisal, logit gap, score norm, and cumulative suppression over repeated updates. This distinguishes categorical support suppression from the unbounded Gaussian-distance amplification measured in C-U1.

E.3 Causal-Transmission Protocol

Purpose. The source-isolation protocol identifies the source of abnormal update magnitude but not its effect on task behavior. Causal transmission is tested through near–far interventions in C-U1 and through the categorical support reconstruction in D-U1. The primary C-U1 causal result in the main text uses the fixed-variance branch to test mean drift and task-performance collapse. The learnable-variance branch provides complementary evidence on support contraction and variance-boundary dynamics.

C-U1 dynamic near–far partition. At each intervention step, negative samples are partitioned using their current-policy remoteness. Let $\mathcal{N}_\theta$ and $\mathcal{F}_\theta$ denote the prespecified near- and far-field sets. Their membership is recomputed as the policy evolves rather than frozen at initialization. The thresholds or quantiles are fixed before training and specified with the corresponding environment.

C-U1 interventions. For a negative sample z, the following multipliers are applied to $g_\theta^-(z)$:

$$\begin{aligned} w_{\text{uncontrolled}}(z) &= 1, \\ w_{\text{positive}}(z) &= 0, \\ w_{\text{near-zero}}(z) &= \mathbf{1}[z \notin \mathcal{N}_\theta], \\ w_{\text{far-zero}}(z) &= \mathbf{1}[z \notin \mathcal{F}_\theta]. \end{aligned} \tag{48}$$

Let G_N and G_F denote the aggregate near- and far-field negative gradients on the current minibatch, respectively, and let $G_{\text{raw}} = G_N + G_F$.

The Far-cap intervention preserves the near-field component and scales the far-field aggregate gradient to at most 5% of the near-field magnitude:

$$\gamma_{\text{cap}} = \min\left\{1,\ 0.05\,\frac{\|G_N\|_2}{\|G_F\|_2}\right\}, \qquad G_{\text{far-cap}} = G_N + \gamma_{\text{cap}} G_F, \tag{49}$$

with $\gamma_{\text{cap}} = 1$ when $G_F = 0$.

To distinguish selective intervention from a global rescaling of negative updates, the global-matched control applies one scalar to the full negative gradient so that its aggregate magnitude matches Far-cap:

$$\alpha_{\text{match}} = \frac{\|G_{\text{far-cap}}\|_2}{\|G_{\text{raw}}\|_2}, \qquad G_{\text{global}} = \alpha_{\text{match}} G_{\text{raw}}, \tag{50}$$

for $G_{\text{raw}} \neq 0$; when $G_{\text{raw}} = 0$, the matched global negative gradient is zero.

The Far-to-near control retains the capped far-field component and rescales the near-field aggregate gradient to restore the uncontrolled aggregate-gradient norm. It chooses the larger nonnegative solution c_N of

$$\|c_N G_N + \gamma_{\text{cap}} G_F\|_2 = \|G_N + G_F\|_2 \,, \tag{51}$$

and uses $G_{\text{far}\to\text{near}} = c_N G_N + \gamma_{\text{cap}} G_F$. All matching coefficients are treated as detached during the policy update.

Causal contrasts. The primary causal contrast is between removing near-field and far-field negative updates. If Near-zero leaves the dynamics unchanged while Far-zero or Far-cap alters them, the effect cannot be attributed to negative feedback in general. The global-matched branch quantifies how much of the change follows from globally rescaling negative updates and how much requires selecting samples by remoteness.

D-U1 categorical counterpart. D-U1 applies analogous interventions to the fixed local and far-negative action groups defined in Appendix D.4. The action-group roles are fixed by the environment rather than recomputed from current-policy remoteness. We compare Positive-only and the uncontrolled baseline with Near-zero, Far-zero, Far-cap, and a non-selective Global-scale control. Near-zero removes the fixed local-negative component, whereas Far-zero removes the fixed far-negative component; Far-cap preserves local negative feedback while limiting the far-negative component, and Global-scale attenuates both components without selecting by remoteness. All branches keep the fixed negative labels, sample roles, data, initialization, optimizer, and training horizon unchanged. We track task reward, negative-action probability and surprisal, logit gap, effective support, and the first task-performance, support-boundary, or NaN/Inf event.

Temporal and terminal measurements. We record policy displacement, reward, negative-gradient magnitude, mean dynamics and, for learnable-variance runs, scale dynamics, entropy or effective support, and the first time at which each prespecified event occurs. The analysis checks whether far-field update magnitude rises before policy drift and whether drift precedes task-performance loss.

Terminal outcomes are classified separately as:

1. **task-performance collapse**: a sustained loss of the prespecified task metric while optimization remains finite;
2. **support or variance-boundary event**: categorical effective support or Gaussian variance crosses its prespecified boundary without requiring NaN or Inf;
3. **numerical collapse**: a parameter, loss, gradient, or model output becomes NaN or Inf.

For C-U1, task-performance collapse is recorded when held-out-context reward falls below 45% of the branch-initial reference reward for three consecutive evaluations. Boundary events and task-performance collapse are reported separately.

D-U1 categorical causal result. The uncontrolled baseline and Near-zero each produce task-performance collapse and a support-boundary event in all 20 seeds, whereas Far-zero and Far-cap produce neither event in any seed. Global-scale prevents task-performance collapse but does not prevent the support-boundary event, which occurs in all 20 seeds. Positive-only produces neither event. No NaN/Inf numerical failure occurs in any of the 120 runs. Thus, selectively suppressing the far-negative component interrupts persistent support suppression, whereas removing the local-negative component does not; non-selective global attenuation can preserve task performance while still allowing support degeneration.

E.4 Controlled Negative-Strength Sweep

The controlled negative-strength sweep varies only the effective negative-repulsion strength q/p, where p denotes the aggregate positive-attraction budget and q denotes the aggregate negative-repulsion budget. All sweep arms share the same data, fixed labels, initialization, optimizer, training horizon, and event thresholds. The positive-only baseline corresponds to

Table 6: D-U1 categorical causal interventions over 20 seeds.

Intervention	Task collapse	Support boundary	NaN/Inf
Positive-only	0/20	0/20	0/20
Uncontrolled	20/20	20/20	0/20
Near-zero	20/20	20/20	0/20
Far-zero	0/20	0/20	0/20
Far-cap	0/20	0/20	0/20
Global-scale	0/20	20/20	0/20

$q/p = 0$. For the fixed-variance C-U1 sweep, the registered strength points are $q/p \in \{0,\ 0.148,\ 0.295,\ 0.443,\ 0.591,\ 0.738,\ 0.886,\ 1.033\}$.

Held-out-context reward evaluates unseen-state generalization within the C-U1 controlled protocol. Reward retention is terminal held-out-context reward normalized by the corresponding branch-initial reward. Normalized policy shift is the displacement of the policy mean from the Positive-only target along the hidden-optimum direction, normalized by the 0.70 gap between $a_+(s)$ and $a_\star(s)$:

$$\Delta_{\text{policy}} = \mathbb{E}_s \left[\frac{\langle \mu_\theta(s) - a_+(s),\, u(s) \rangle}{0.70} \right].$$

Task-performance collapse, support or variance-boundary events, and NaN/Inf numerical failures are reported as separate event types.

E.5 Controlled Taper Comparison

Controlled utility construction. In C-U1, the hidden optimum $a_\star(s)$ lies beyond the center of the positive demonstrations $a_+(s)$. A local negative anchor is placed on the opposite side of $a_+(s)$, so repulsion from this anchor points toward $a_\star(s)$. Additional negative actions are placed on matched reward contours at larger policy-relative distances. Increasing negative strength therefore produces three observable regimes: positive-only under-extrapolation, useful local extrapolation, and far-field over-extrapolation or instability.

D-U1 implements the same logic with unordered semantic actions. The hidden optimal target lies beyond the demonstrated positive direction, while the local negative target lies in the opposite semantic direction. Repelling the local negative action can therefore improve the shared representation toward the hidden optimal action. Additional remote negative actions create support pressure without changing their fixed negative label. Evaluation on independently sampled contexts is reported as held-out-context generalization.

Let

$$x(D) = \left[\frac{D - \tau}{c} \right]_+, \qquad r_x(D) = \sqrt{x(D)}.$$

The Linear and Quadratic names refer to polynomial order in the radialized coordinate $r_\theta = \sqrt{x_\theta}$, rather than in the squared-remoteness coordinate x_θ.

Weighting rules. For negative-sample remoteness D, the compared rules are

$$\begin{aligned} \omega_{\text{positive}}(D) &= 0, \\ \omega_{\text{uncontrolled}}(D) &= 1, \\ \omega_{\text{global}}(D) &= \alpha, \\ \omega_{\text{Rec-L}}(D) &= \frac{1}{1 + \lambda r_x(D)}, \\ \omega_{\text{Rec-Q}}(D) &= \frac{1}{1 + \lambda r_x(D)^2}, \\ \omega_{\text{exp}}(D) &= \exp\{-\lambda r_x(D)^2\}. \end{aligned} \tag{52}$$

All weighting values are computed from detached remoteness statistics; the policy cannot reduce its loss by differentiating through the taper.

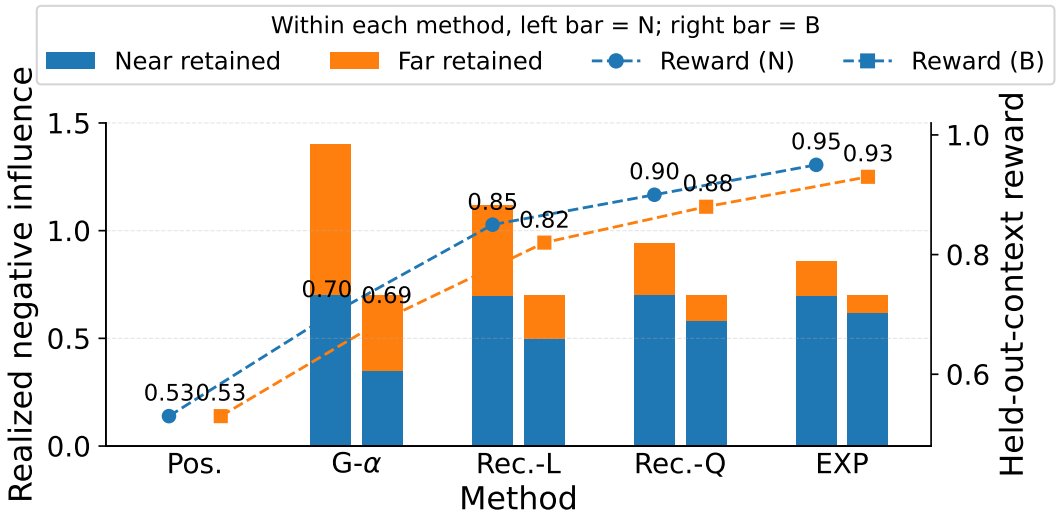


Figure 5: **Controlled taper-shape comparison.** Matched operating points (left: mean near-field weight matched; right: stepwise pre-Adam negative-gradient L_2 matched). Bars summarize near- and far-field negative influence; dashed traces show held-out-context reward.

Near-field-retention matching. Let θ_{ref} denote the fixed Positive-only reference policy used for calibration and define the prespecified near set

$$\mathcal{N}_{\mathrm{ref}} = \{z : d_{\theta_{\mathrm{ref}}}(s,a) \le 5.0\}.$$

For a C-U1 taper $w_c(d)$, define its mean retained near-field weight as

$$\rho_{\mathrm{near}}(w_c) = \frac{1}{|\mathcal{N}_{\mathrm{ref}}|} \sum_{z \in \mathcal{N}_{\mathrm{ref}}} w_c(d_{\theta_{\mathrm{ref}}}(s,a)) \,. \tag{53}$$

Taper parameters are calibrated on the pooled training negatives so that $\rho_{\mathrm{near}} = 0.75$, and are then frozen for the matched comparison. This isolates far-field taper shape after matching the average weight retained on the same prespecified near region.

Aggregate-gradient matching. At update t, let $G^-_{m,t}$ denote method m's aggregate negative gradient after applying its taper rule but before the matching scalar and the Adam update. Reciprocal-Linear supplies the reference norm $\|G^-_{\mathrm{Rec\text{-}L},t}\|_2$. For a nonzero candidate gradient, each matched method uses the detached scalar

$$s_{m,t} = \frac{\|G^-_{\mathrm{Rec\text{-}L},t}\|_2}{\|G^-_{m,t}\|_2}, \qquad \widetilde{G}^-_{m,t} = s_{m,t} G^-_{m,t}. \tag{54}$$

The non-selective Global control uses the unweighted negative-gradient direction and the same target norm. Thus, the matched methods share the same pre-Adam negative-gradient L_2 norm at each update while differing in how that gradient is distributed across near and far samples. The Adam parameter-update norm is not the matching target.

Measurements. Each method is evaluated over the full training horizon. We report task performance, held-out-context generalization, distance from the hidden optimum, policy displacement, entropy or variance, effective support, near-field retention, pre-Adam negative-gradient norm, and all three terminal-event categories. Method comparisons use the same seeds and paired statistics.

We compare positive-only learning, global negative scaling, reciprocal-linear tapering, reciprocal-quadratic tapering, and exponential tapering while changing only the negative weighting rule. The near-field comparison matches average retained weight on the prespecified near set, whereas the aggregate-gradient comparison matches the pre-Adam negative-gradient L_2 norm at every update. Together they separate taper shape from a non-selective global rescaling at the same pre-Adam gradient scale.

Under both matched comparisons, exponential tapering retains near-field repulsion while suppressing the far-field tail more rapidly than polynomial tapers. Positive-only removes both components, whereas global scaling reduces them indiscriminately.

E.6 Shared Objectives and Negative-Update Controls

This section defines the protocol controls and remoteness-aware weighting rules shared by the D4RL and structured-generation experiments. Their task-specific remoteness signals, data construction, and optimization protocols are described separately in Appendices E.7 and E.8.

Shared signed objective. Let z denote an offline transition or response, let $A(z)$ denote its fixed signed learning signal during an actor update, and let $D_\theta(z)$ denote its learner-relative remoteness under the current policy. We write

$$A^+(z) = \max\{A(z), 0\}, \qquad A^-(z) = \max\{-A(z), 0\}.$$

The shared family of actor objectives is

$$\mathcal{J}_\omega(\theta) = \mathbb{E}_{z\sim\mathcal{B}}\left[A^+(z)\log\pi_\theta(z) - \omega(D_\theta(z))\,A^-(z)\log\pi_\theta(z)\right],$$

where positive updates retain unit weight and $\omega(D)$ controls only the negative branch. During each actor update, the advantage and remoteness-derived weight are treated as fixed coefficients; gradients are not propagated through the weighting rule.

For the remoteness-aware methods, define the normalized excess remoteness

$$x_\theta(z) = \left[\frac{D_\theta(z) - \tau}{c}\right]_+,$$

where τ is the near-field threshold, $c > 0$ is the task-specific scale, and $[u]_+ = \max\{u, 0\}$. Consequently, samples inside the near-field region $D_\theta(z) \le \tau$ retain their full negative weight.

Positive-only. Positive-only removes all negative updates:

$$\omega_{\text{pos}}(D) = 0.$$

It therefore provides an extreme no-repulsion reference. In D4RL it retains only transitions with positive actor advantages; in structured-generation tasks it retains only verifier-confirmed positive responses under the corresponding task signal.

Global-α. Global-α applies the same attenuation to every negative sample:

$$\omega_{\text{global}}(D) = \alpha, \qquad 0 \le \alpha \le 1.$$

This control tests whether performance improvements can be explained by a global reduction in negative-update magnitude, without using learner-relative remoteness.

Reciprocal-Linear and Reciprocal-Quadratic. Using the normalized excess remoteness defined above, let

$$r_\theta(z) = \sqrt{x_\theta(z)}.$$

For Gaussian policies, r_θ is asymptotically proportional to standardized distance in the far field; for sequence policies, it is a surprisal-derived radial coordinate rather than a geometric distance in the discrete action space. The two reciprocal controls use the same remoteness signal and near-field threshold as DRPO, but differ in their far-field decay:

$$\omega_{\text{Rec-L}}(D) = \frac{1}{1 + \lambda r_\theta(z)},$$

and

$$\omega_{\text{Rec-Q}}(D) = \frac{1}{1 + \lambda r_\theta(z)^2} = \frac{1}{1 + \lambda x_\theta(z)}.$$

Reciprocal-Linear has the slowest far-field attenuation. The quadratic variant decays more rapidly but retains a polynomial tail. These controls isolate whether the benefit comes merely from using a remoteness signal or specifically from the decay profile selected by DRPO.

DRPO. DRPO uses an exponential remoteness taper:

$$\omega_{\text{DRPO}}(D) = \exp\left\{-\lambda x_\theta(z)\right\}.$$

The weight equals one throughout the near-field region and then decays exponentially once remoteness exceeds τ. Thus, DRPO preserves local negative feedback while suppressing the remote negative tail more rapidly than the reciprocal alternatives.

Hyperparameter selection. Tuned hyperparameters are selected using validation data only; fixed settings are specified with each experimental protocol. Test-task performance is not used for hyperparameter selection.

E.7 D4RL Baselines and Experimental Protocol

Positive-only, Global-α, Reciprocal-Linear, Reciprocal-Quadratic, and DRPO follow Appendix E.6 and are evaluated under the same D4RL actor-training protocol, differing only in their negative-update weighting rule. This section specifies the D4RL-specific baselines, remoteness construction, training protocol, and evaluation procedure.

Tasks and datasets. We evaluate HalfCheetah, Hopper, and Walker2d on the medium, medium-replay, and medium-expert D4RL datasets (Fu et al., 2020), producing the nine dataset cells reported in Table 2. The task-level comparison and external gradient diagnostic in Appendix D.1 cover the same nine cells. The same fixed offline dataset is used by every protocol-matched method within a task.

Published offline-RL baselines. CQL, TD3+BC, and IQL are reported using their published task-level means. Because they were not rerun under our actor-training protocol, they are marked by † in Table 2.

Gaussian-policy remoteness. For an action $a \in \mathbb{R}^{d_a}$ and current Gaussian policy with mean $\mu_\theta(s)$ and standard deviation $\sigma_\theta(s)$, we use the following standardized remoteness statistic

$$D_{\text{D4RL}}(s, a) = \frac{1}{d_a} \sum_{j=1}^{d_a} \left(\frac{a_j - \mu_{\theta,j}(s)}{\sigma_{\theta,j}(s)} \right)^2.$$

This remoteness definition is used throughout the D4RL comparisons. When the actor uses a transformed action distribution, this quantity is computed in the policy's native Gaussian coordinates. The resulting remoteness is used only to determine the negative-update coefficient; the distance term is detached during the actor update.

Table 7: Task-specific DRPO hyperparameters selected on the D4RL development protocol.

Dataset	s	c
halfcheetah-medium	1.0	0.0416667
hopper-medium	1.0	0.15
walker2d-medium	0.003	0.15
halfcheetah-medium-replay	1.0	0.0046875
hopper-medium-replay	1.0	0.122667
walker2d-medium-replay	1.0	0.0096
halfcheetah-medium-expert	1.0	2.3
hopper-medium-expert	0.01	0.425
walker2d-medium-expert	1.0	0.64

Hyperparameters. Hyperparameters are selected on the development protocol, and held-out evaluation data are not used for selection. For each tuned method and D4RL task, the setting with the highest late-window mean normalized return on the development runs is selected and held fixed for evaluation.

The policy-update budget is $1{,}000{,}000$ updates, with evaluation every $50{,}000$ updates using 10 episodes. The near-field threshold is $\tau = 0$.

For DRPO, Table 7 reports the selected task-specific negative scale s and exponential coefficient c.

Evaluation. Task performance is reported using the standard D4RL normalized return. For every protocol-matched method, the main table reports the mean across seeds. The D4RL-9 total is the sum of the nine task-level means.

Task-performance collapse is assessed from normalized return and training trajectories. Variance or support-boundary events are reported separately from task performance. NaN/Inf numerical failures are reported as a third outcome and are not merged with either low return or policy-boundary events.

E.8 Structured-Generation Baselines and Experimental Protocol

All structured-generation experiments use Qwen2.5-0.5B-Instruct (Yang et al., 2024) with LoRA adaptation and frozen prompt-indexed replay data. Learner-relative remoteness for DRPO is the current-policy mean completion-token surprisal and is recomputed during training rather than stored in the replay bank. In the far-field gradient diagnostic, verifier-incorrect completions use a unit negative coefficient throughout. Within each task, the compared methods use the same task data and verifier labels; their shared training and evaluation settings are specified below.

AsymRE. Asymmetric REINFORCE defines the response-level advantage using a tunable reward baseline V:

$$A_{\mathrm{AsymRE}}(x, y) = r(x, y) - V,$$

and optimizes

$$\mathcal{J}_{\mathrm{AsymRE}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{B}}\left[\big(r(x, y) - V\big) \log \pi_\theta(y \mid x)\right].$$

For a binary verifier reward, lowering V emphasizes successful responses, whereas increasing V strengthens suppression of failed responses. With branch-balanced signed rewards, $r \in \{-1, +1\}$ gives the empirical baseline $\widehat{V} = 0$. The swept offset therefore determines the objective baseline as $V = \widehat{V} + \delta_v = \delta_v$. The baseline is selected using the validation protocol and is held fixed within each run.

Global-α and AsymRE remain separate comparisons because Global-α scales the negative branch of the common signed advantage, whereas AsymRE constructs its signed signal from the verifier reward

relative to V. Under binary rewards and simplified normalization they are closely related, but their objectives and tuning parameters remain distinct.

β-TOPR (joint fitted reference). Let $\mathcal{X}$ be the training-prompt set. For each $x \in \mathcal{X}$, let y_x^+ be the verified oracle solution and let $\mathcal{N}_x$ be the set of unique negative completions after canonical-completion deduplication. Define the mean completion-token log probability

$$\bar{\ell}_\theta(x, y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}),$$

and the full-completion summed log probability used only for the ratio,

$$s_\theta(x, y) = |y|\, \bar{\ell}_\theta(x, y) = \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}).$$

For the jointly fitted reference adapter μ_{ref}, define

$$s_{\mathrm{ref}}(x, y) = \sum_{t=1}^{|y|} \log \mu_{\mathrm{ref}}(y_t \mid x, y_{<t}),$$

$$\rho_\theta(x, y) = \exp(s_\theta(x, y) - s_{\mathrm{ref}}(x, y)),$$

and

$$\begin{aligned} w_\beta(x, y) &= \exp(\beta \min\{s_\theta(x, y) - s_{\mathrm{ref}}(x, y), 0\}) \\ &= \min\{\rho_\theta(x, y)^\beta, 1\}. \end{aligned}$$

The policy objective is

$$\mathcal{J}_{\beta\text{-TOPR}}(\theta) = \mathbb{E}_{x \sim \mathrm{Unif}(\mathcal{X})} \left[\bar{\ell}_\theta(x, y_x^+) - \frac{1}{|\mathcal{N}_x|} \sum_{y \in \mathcal{N}_x} \mathrm{sg}[w_\beta(x, y)] \bar{\ell}_\theta(x, y) \right].$$

with actor update

$$g_{\beta\text{-TOPR}}(\theta) = \nabla_\theta \mathcal{J}_{\beta\text{-TOPR}}(\theta).$$

This objective samples prompts uniformly, assigns unit total coefficient to the positive branch and unit total coefficient to the negative branch, and averages unique negatives within each prompt. It is not a simple uniform average over all response rows. The ratio-derived weight is detached from the policy gradient.

The reference adapter is fitted with

$$\mathcal{J}_{\mathrm{ref}} = \mathbb{E}_{x \sim \mathrm{Unif}(\mathcal{X})} \left[\frac{1}{2} \bar{\ell}_{\mathrm{ref}}(x, y_x^+) + \frac{1}{2|\mathcal{N}_x|} \sum_{y \in \mathcal{N}_x} \bar{\ell}_{\mathrm{ref}}(x, y) \right].$$

The reference adapter is updated once per policy step. Reference outputs and ratio-derived weights are detached from the policy gradient. The 0.5/0.5 masses belong only to the reference-fitting objective, not to the policy objective. This variant does not use behavior-policy log probabilities; its ratio denominator is supplied by the jointly fitted reference adapter.

DPO. Canonical DPO (Rafailov et al., 2023) is trained with the same LoRA parameterization as the other structured-generation methods. Its policy and frozen reference are initialized from a task-specific SFT checkpoint trained to predict the verifier-confirmed correct completion for each training prompt; this difference is marked by ‡ in Table 2. For each prompt x, the verified oracle solution y_x^+ is paired with every unique verifier-incorrect completion $y \in \mathcal{N}_x$. Using the summed completion log probability s_θ defined above, the loss is

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{x\sim\mathrm{Unif}(\mathcal{X})}\left[\frac{1}{|\mathcal{N}_x|}\sum_{y\in\mathcal{N}_x}\log\sigma\Big(\beta\big[s_\theta(x,y_x^+) - s_{\mathrm{ref}}(x,y_x^+) - s_\theta(x,y) + s_{\mathrm{ref}}(x,y)\big]\Big)\right].$$

The loss is averaged over unique negatives within each prompt and then uniformly across prompts. The reference remains frozen throughout the fixed 1,200-step continuation, and β is selected using the validation sweep.

Method-specific remoteness. For DRPO and the other remoteness-taper controls, the learner-relative coordinate for a prompt x and completion $y = (y_1, \ldots, y_T)$ is the current mean completion-token surprisal

$$D_{\mathrm{SG}}(x,y) = -\frac{1}{T}\sum_{t=1}^{T}\log\pi_\theta\left(y_t \mid x, y_{<t}\right).$$

Only completion tokens are included; prompt tokens and padding positions are excluded. Mean rather than summed surprisal is used so that response length does not mechanically determine DRPO remoteness. The remoteness score is detached before applying the shared taper functions. AsymRE does not use remoteness and applies its signed reward directly. The Joint Fitted-Reference β-TOPR variant uses the dynamic full-completion policy/reference log-ratio defined above. Canonical DPO instead optimizes the frozen-reference pairwise log-ratio in its own objective; neither baseline uses DRPO's mean-token remoteness.

Training protocol. All protocol-matched methods use Qwen2.5-0.5B-Instruct (Yang et al., 2024) with LoRA rank 32, alpha 64, and dropout 0.05. The Joint Fitted-Reference β-TOPR baseline additionally initializes a reference LoRA from the initial policy adapter. Within each structured-generation task, the methods share the AdamW optimizer (Loshchilov & Hutter, 2019), learning-rate schedule, minibatch construction, 1,200 optimizer updates, and evaluation schedule, with no early stopping.

Hyperparameters are selected on the validation split, and reserved test data are not used for selection. Pass@8 is the primary sampled-decoding metric. The nine tasks use the same training protocol. For coefficient-response experiments, late-window Pass@8 is the arithmetic mean of evaluations at updates $800, 900, 1000, 1100$, and 1200.

Verifier-based evaluation. Greedy verifier success is the fraction of held-out instances solved by greedy decoding. Pass@k is the fraction solved by at least one of the k sampled responses. Task-specific output-validity diagnostics are reported separately from verifier success where applicable.

Task-performance degradation, token- or sequence-support boundary events, and NaN/Inf numerical failures are reported separately.

Baseline coefficient selection. For each structured-generation transfer task, Global-α scans $\alpha \in \{1/128, 1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1\}$, AsymRE scans $\delta_v \in \{-1.0, -0.9, -0.75, -0.5, 0.0\}$, Joint Fitted-Reference β-TOPR scans $\beta \in \{0.25, 0.5, 1.0\}$, and canonical DPO scans $\beta \in \{0, 0.05, 0.1, 0.2, 0.5, 1.0\}$. For DPO, $\beta = 0$ is the SFT-only initialization control with no subsequent DPO optimizer updates. For each task, coefficients are selected on the validation metric.

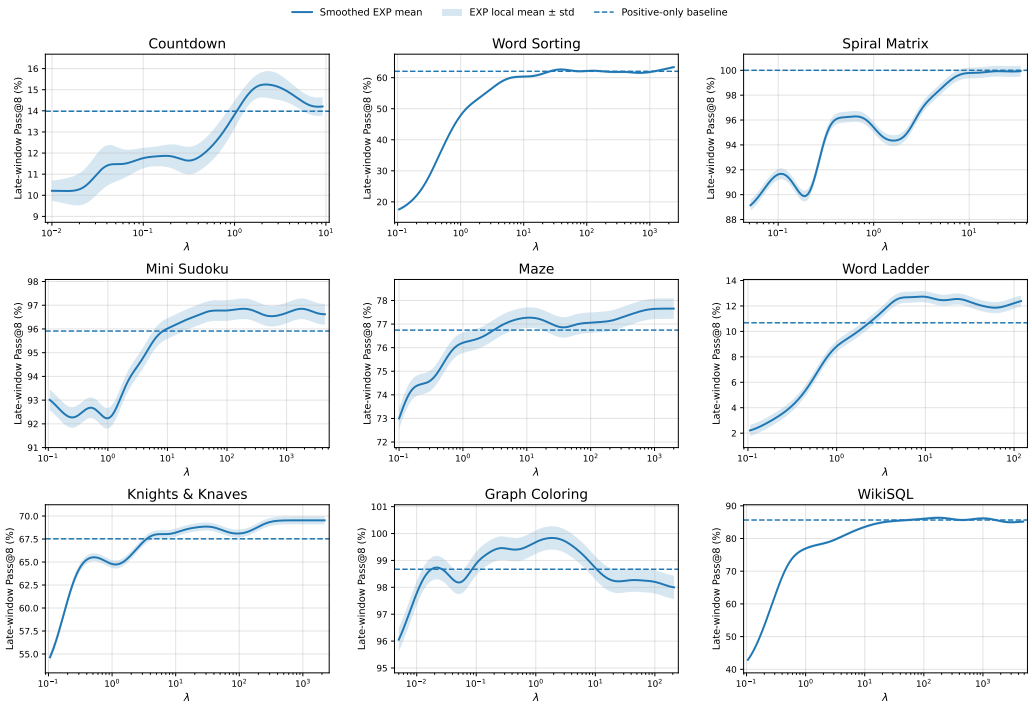


Figure 6: **DRPO coefficient response across structured-generation tasks.** Each panel reports late-window Pass@8 versus the evaluated taper coefficient λ; dashed horizontal lines denote the corresponding Positive-only references. Task-specific vertical scales are used to make within-task response shape visible.

E.9 Structured-Generation DRPO Coefficient Response

Figure 6 reports the DRPO coefficient-response curves across the structured-generation tasks. Each panel plots late-window Pass@8 against the evaluated taper coefficient λ, with the corresponding Positive-only result shown as a horizontal reference.

Across tasks, the response is task dependent rather than a monotone "larger is better" relationship. Countdown, Word Ladder, and Knights & Knaves show substantial gains over Positive-only over part of the evaluated coefficient range, while Mini Sudoku and Maze show smaller improvements. Spiral Matrix is already near the task ceiling. Word Sorting and WikiSQL improve as λ increases but the observed sweeps do not establish a closed interior optimum. For Graph Coloring, small λ is insufficient, an intermediate range is effective, and the benefit declines as λ becomes large, with high-λ performance returning toward the Positive-only level. These curves characterize coefficient sensitivity and task-dependent operating ranges over the evaluated training horizon.