Variance-Reduced Post-Training via Action-Level Behavior Policy Optimization

Jingkai Huang
New York University
jh9959@nyu.edu

Qinxun Bai
Isara Laboratories
jerry@isara.io

Zhengyuan Zhou
Arena Physica & New York University
zzhou@stern.nyu.edu

Abstract

Policy gradient methods may suffer from high variance, which limits their sample efficiency. Existing variance-reduction techniques largely focus on estimator design under a fixed sampling policy. In contrast, we ask whether the action-sampling distribution itself, often left at the on-policy default, can be optimized to reduce policy gradient variance. We show that this default is in fact not variance-optimal even when the state distribution is held on-policy. To exploit this, we introduce *Action-Level Behavior Policy Optimization* (AL-BPO), a framework that treats the behavior policy as an action-level proposal distribution at each on-policy state, while keeping the state distribution itself on-policy. For a one-step importance-sampled policy gradient estimator, we sharply decompose its variance into an irreducible state term and a controllable action term, and show that the latter can be minimized in closed form via AL-BPO. Under exact critic and sampling oracles, this yields a strictly tighter nonconvex SGD bound than on-policy sampling whenever the score-weighted critic is action-dependent. We then extend the analysis to approximate critics and parametrized behavior policies updated by a single stochastic step, and verify the resulting sample-efficiency gains empirically. These results identify AL-BPO as a principled mechanism for improving the sample efficiency of policy gradient methods. We also apply AL-BPO to post-train Qwen3-4B-Instruct for one-shot radio-frequency (RF) antenna design using Arena Physica's EM foundation model as a physics reward verifier.

1 Introduction

Policy gradient (PG) methods [Williams, 1992, Sutton et al., 1999] are widely used in reinforcement learning, and their sample efficiency is largely determined by the variance of the stochastic gradient estimator. A natural and widely studied choice is to estimate the gradient by sampling actions on-policy from the target policy π_θ, but the resulting estimator, while unbiased, can have a large variance and thus produces noisy actor updates that lead to slow convergence. The cause is a *sampling-frequency mismatch*: on-policy sampling weights actions by the policy's visit frequency rather than by their gradient contribution, leaving the highest-signal actions systematically undersampled.

Most existing variance reduction techniques accept this on-policy sampling distribution as the default and improve the estimator through subtracting baselines, replacing returns with advantage estimates, blending TD bootstraps via generalized advantage estimation, or constructing control variates. However, the action sampling distribution itself remains underexplored.

When actions are instead drawn from a behavior policy μ and corrected by importance sampling (IS), the choice of μ directly determines the variance of the resulting gradient. The classical off-policy PG literature [Precup et al., 2000, Degris et al., 2012, Liu et al., 2019] treats μ as given, typically a logger or a heuristic exploration wrapper around π_θ, and does not focus on optimizing μ itself. Two recent works move toward optimizing μ, but each in a direction different from ours. Hanna et al. [2024]

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

optimize μ for off-policy evaluation, targeting the mean squared error of a value estimator rather than the variance of the policy gradient. Papini et al. [2024] target the policy gradient directly and study the question at the *trajectory level*: they collect whole rollouts under μ, derive a trajectory optimum behavior, which is generally not realizable and has to be fitted through cross-entropy projection rather than evaluated in closed form.

In this paper we approach the goal of variance reduction for PG from a different angle: Rather than collecting full trajectories from μ and correcting the induced trajectory distribution shift through trajectory-level IS, we obtain on-policy state samples through a geometric stopping rollout under π_θ [Kakade, 2003, Metelli et al., 2023] and then take a single off-policy action query at the returned state. This setup uses two standard primitives in actor-critic analysis: a resettable simulator that can be advanced under π_θ [Kakade, 2003] and a critic oracle that can be queried at the returned state. In exchange, μ never enters the transition kernel, the importance ratio does not accumulate over horizon, and the total variance of the resulting one-step estimator decomposes into a state term invariant to μ and an action term that μ alone controls. The result is a separable per-state BPO problem with structurally smaller variance than the trajectory-level construction.

Within this setting the per-state optimum is a closed-form distribution on the action space alone, expressed in the score and the critic that any actor-critic algorithm already maintains, in contrast to the trajectory-space optimum that requires cross-entropy projection. The variance gap against on-policy sampling can now be given by an exact per-state formula that makes precise both when AL-BPO helps and how much. Under exact critic and sampling oracles, this gap plugs into the standard nonconvex SGD analysis and yields a strictly tighter convergence bound than on-policy sampling whenever the score-weighted critic is action-dependent. Conceptually, AL-BPO supplies a new lever for PG variance reduction, distinct from estimator-side techniques and trajectory-level active IS, and implies the on-policy default provably suboptimal even when the state distribution is held on-policy. Further, when the critic is approximate and the behavior policy is restricted to a parametric family, the resulting convergence bound separates the critic bias and the behavior-side error. We complement these theoretical results with experiments on synthetic environments confirming the predicted variance and sample-efficiency gains over standard on-policy PG baselines.

Contributions Our contributions form a three-step arc: strictly improved action-level variance, leading to a tighter convergence bound than default on-policy PG, delivered by a practical single-step algorithm with theoretical guarantees and empirical validation. We also apply AL-BPO to post-train Qwen3-4B-Instruct for one-shot radio-frequency (RF) antenna design using Arena Physica's EM foundation model as a physics reward verifier. [1]

(1) Closed-form AL-BPO. For every state s, the conditional second moment of the one-step importance-sampled PG estimator is minimized by $\mu^\star(a \mid s) \propto \pi_\theta(a \mid s)\sqrt{\psi(s,a)}$ with $\psi(s,a) := \|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 Q^{\pi_\theta}(s,a)^2$ (Lemma 1). The per-state gap against on-policy sampling makes precise both *when AL-BPO helps and how much*: the gain over on-policy sampling grows with the dispersion of $\sqrt{\psi(s,a)}$ across actions.

(2) Exact-oracle convergence with strictly improved variance. Under exact observation oracles, AL-BPO plugs into the standard nonconvex SGD analysis and *strictly tightens its variance constant* whenever the score-weighted critic is action-dependent (Theorem 3 and Corollary 4).

(3) Practical algorithm and approximate-oracle convergence. A single stochastic step on a parametric μ_β gives an implementable AL-BPO actor-critic (Algorithm 2). Its convergence bound (Theorem 5) decomposes into a critic-induced bias that is independent of μ, a behavior-side variance term, and a tracking error against the stage-wise optimum behavior policy (Proposition 6).

The rest of the paper is organized as follows. Section 2 provides the model setup. Section 3 derives the optimal behavior policy. Sections 4 and 5 give convergence guarantees under exact and approximate oracles. Section 6 presents experiments on synthetic environments and LLM post-training, and Section 7 discusses limitations and open directions. Appendices A and B contain related work and proofs; Appendix C presents practical AL-BPO implementations, and Appendices D and E provide details of the synthetic and continuous-control experiments and the LLM post-training experiments.

[1] Additional implementation details and results omitted from the main text due to space constraints are available at `https://www.arenaphysica.com/publications/one-shot-antenna-design`.

2 Model Setup

We work with a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$ with general state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $\mathbb{P}$, bounded reward $|r| \le r_{\max}$ and discount factor $\gamma \in (0,1)$. The target policy $\pi_\theta(a \mid s)$ is parameterized by $\theta \in \mathbb{R}^d$. For a policy π_θ, the state-value function V^{π_θ} and action-value function Q^{π_θ} are

$$V^{\pi_\theta}(s) \;:=\; \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t \cdot r(s_t,a_t) \;\middle|\; s_0 = s\right], \quad Q^{\pi_\theta}(s,a) \;:=\; r(s,a) + \gamma\, \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}[V^{\pi_\theta}(s')]\,,$$

where the expectation $\mathbb{E}^{\pi_\theta}$ is taken over the trajectory generated by $a_t \sim \pi_\theta(\cdot \mid s_t)$ and $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$. Fix an initial-state distribution $\rho \in \Delta(\mathcal{S})$, the policy-optimization objective is $J(\theta) := \mathbb{E}_{s_0\sim\rho}[V^{\pi_\theta}(s_0)]$ and we denote $J^\star := \sup_\theta J(\theta)$. The discounted state visitation distribution is

$$d_\rho^{\pi_\theta}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \mathbb{P}(s_t = s \mid s_0 \sim \rho, \pi_\theta),$$

The policy gradient (PG) theorem [Sutton et al., 1999] gives

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}}\left[\int_{\mathcal{A}} \pi_\theta(a \mid s)\nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a)\,\mathrm{d}a\right]. \tag{1}$$

Observation model The learner interacts with a resettable simulator augmented with local oracles on visited states. **(O1) Restart:** draw $s_0 \sim \rho$ at any time. **(O2) On-policy forward simulation:** at any visited state s and any parameter θ, draw the environment action $a^{\mathrm{env}} \sim \pi_\theta(\cdot \mid s)$ and query $s' \sim \mathbb{P}(\cdot \mid s, a^{\mathrm{env}})$. **(O3) Local critic oracle:** at any parameter θ and any visited state s, evaluate $Q^{\pi_\theta}(s,a)$ for any queried action $a \in \mathcal{A}$. **(O4) Local behavior oracle:** at any visited state s, draw one policy-gradient action $a^{\mathrm{pg}} \sim \mu(\cdot \mid s)$ for policy gradient estimator.

Together, **(O1)**–**(O2)** support a geometric-stopping rollout that returns a sample $\tilde{s} \sim d_\rho^{\pi_\theta}$ through: draw $s_0 \sim \rho$, repeatedly stop with probability $1-\gamma$, and otherwise advance under π_θ. This requires $O((1-\gamma)^{-1})$ simulator steps in expectation.

The reason for restricting μ to a single action query at the returned state, rather than letting μ drive the entire rollout as in standard fully off-policy data collection, is to keep the state distribution exactly $d_\rho^{\pi_\theta}$. If μ were used to advance the simulator, the induced rollout state distribution would deviate from $d_\rho^{\pi_\theta}$, and the resulting policy gradient estimator would suffer from a state-distribution mismatch whose unbiased correction requires the discounted state-density ratio $d_\rho^{\pi_\theta}/d_\rho^{\mu}$, an object that is generally unknown and notoriously hard to estimate [Zhang, 2022]. Carrying such a correction through the analysis would entangle the variance contribution of μ with the bias and variance of the density-ratio estimator and obscure the action-level variance benefit of AL-BPO. The current observation model isolates this effect by absorbing the state distribution shift into the rollout, which is standard in the PG sample complexity literature [Kakade, 2003, Metelli et al., 2023] for theoretical analysis.

Besides, we note that the rollout actions a_t^{env} are used only to generate the returned state sample $\tilde{s}$ and are not reused for policy-gradient estimation. By contrast, the action a^{pg} sampled in **(O4)** is used only for constructing the importance-sampled policy-gradient estimator at the returned state and does not affect the state transition. In Section 5, **(O3)** will be weakened to an ξ-approximate critic oracle and **(O4)** will be replaced by a parametrized behavior oracle.

Assumptions Throughout, we impose the following standard regularity conditions. **(A1) Coverage:** every behavior policy μ used by the algorithm and every (s,a), $\pi_\theta(a \mid s) > 0$ implies $\mu(a \mid s) > 0$. **(A2) Score bound:** there exists a constant $G < \infty$ such that $\|\nabla_\theta \log \pi_\theta(a \mid s)\|_2 \le G$ for all (θ, s, a). **(A3) Smoothness:** the objective $J(\theta)$ is $\widetilde{L}$-smooth for some constant $\widetilde{L} < \infty$.

Assumptions **(A1)**–**(A3)** are standard in IS-based off-policy policy gradient analysis and ensure that the estimator is well-defined and the optimization landscape is regular enough for nonconvex SGD arguments.

3 Variance Reduction via Behavior Policy Optimization

Policy gradient estimator Given state $\tilde{s} \sim d_\rho^{\pi_\theta}$, we construct the one-step IS gradient estimator

$$\hat{g}(\tilde{s}, a^{\mathrm{pg}}) \;:=\; \frac{1}{1-\gamma} \cdot \frac{\pi_\theta(a^{\mathrm{pg}} \mid \tilde{s})}{\mu(a^{\mathrm{pg}} \mid \tilde{s})} \cdot \nabla_\theta \log \pi_\theta(a^{\mathrm{pg}} \mid \tilde{s}) \cdot Q^{\pi_\theta}(\tilde{s}, a^{\mathrm{pg}}), \qquad a^{\mathrm{pg}} \sim \mu(\cdot \mid \tilde{s}). \tag{2}$$

This estimator is action-only off-policy: the state distribution is on-policy and needs no correction. For any fixed s,

$$\mathbb{E}_{a\sim\mu}[\hat{g}(s,a) \mid s] = \frac{1}{1-\gamma}\mathbb{E}_{a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a)\right]. \tag{3}$$

Taking expectation over $\tilde{s} \sim d_\rho^{\pi_\theta}$ in (2) recovers the (exact) policy gradient in (1).

Variance decomposition The (co)variance (trace) of the estimator decomposes as

$$\mathrm{Var}(\hat{g}(s,a)) = \underbrace{\mathbb{E}_{s\sim d_\rho^{\pi_\theta}}[\mathrm{Var}_{a\sim\mu}(\hat{g}(s,a) \mid s)]}_{\sigma^2_{\mathrm{action}}(\mu)} + \underbrace{\mathrm{Var}_{s\sim d_\rho^{\pi_\theta}}\left(\tfrac{1}{1-\gamma} \cdot \mathbb{E}_{a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a)\right]\right)}_{\sigma^2_{\mathrm{state}}}.$$

Note that the state variance term $\sigma^2_{\mathrm{state}}$ is independent of μ (due to the change of measure technique according to (3)), thus behavior policy optimization can only reduce the action-level variance $\sigma^2_{\mathrm{action}}(\mu)$.

Action-level objective For any state s, define the conditional second moment

$$\mathcal{V}_\mu(s) := \frac{1}{(1-\gamma)^2}\int \frac{\pi_\theta(a \mid s)^2}{\mu(a \mid s)} \cdot \psi(s,a)\,\mathrm{d}a, \quad \psi(s,a) := \|\nabla \log \pi_\theta(a \mid s)\|^2 Q^{\pi_\theta}(s,a)^2.$$

Let $h(\theta, s) := \mathbb{E}_{a\sim\mu}[\hat{g}(s,a) \mid s]$. Since $\mathrm{Var}(\hat{g}(s,a) \mid s) = \mathcal{V}_\mu(s) - \|h(\theta,s)\|^2$ and $h(\theta, s)$ does not depend on μ, minimizing the action-level variance is *equivalent to* minimizing $\mathcal{V}_\mu(s)$. The following lemma provides the closed-form of the optimal behavior policy based on the Cauchy-Schwarz inequality according to standard importance sampling literature.

Lemma 1 (Optimal behavior policy)**.** *Under **(A1)**, for each state s the optimal behavior policy $\mu^\star$ minimizing $\mathcal{V}_\mu(s)$ is*

$$\mu^\star(a \mid s) \;=\; \frac{\pi_\theta(a \mid s)\sqrt{\psi(s,a)}}{Z(s)}, \qquad Z(s) \;:=\; \int_{\mathcal{A}} \pi_\theta(a \mid s)\sqrt{\psi(s,a)}\,\mathrm{d}a.$$

Equivalently, $\mu^\star(a \mid s) \propto \pi_\theta(a \mid s) \cdot \|\nabla_\theta \log \pi_\theta(a \mid s)\| \cdot |Q^{\pi_\theta}(s,a)|$. The per-state variance reduction against on-policy is

$$\mathcal{V}_{\pi_\theta}(s) - \mathcal{V}_{\mu^\star}(s) \;=\; \frac{1}{(1-\gamma)^2}\,\mathrm{Var}_{a\sim\pi_\theta}\left[\sqrt{\psi(s,a)}\right] \;\geq\; 0. \tag{4}$$

with strict inequality whenever $\psi(s,\cdot)$ is non-constant on its support.

The optimal behavior policy reweighs actions according to their contribution to the gradient variance, as captured by $\psi(s,a) = \|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 Q^{\pi_\theta}(s,a)^2$, thereby prioritizing actions that both *strongly influence the policy* (large score norm) and have *large value magnitude*. This highlights that policy gradient variance depends jointly on the *state-action value* and *policy sensitivity*.

4 Convergence under Exact Oracles

We first analyze the ideal benchmark in which the critic oracle returns the exact value $Q^{\pi_{\theta_i}}$, the behavior oracle can be sampled directly from the optimal $\mu_i^\star$ at each iteration $i = 1, 2, \cdots, N$. The purpose of this section is to isolate the variance-reduction effect of AL-BPO without any approximation error to better illustrate how AL-BPO benefits the convergence speed.

We have the following property holds for the estimator under Algorithm 1.

Algorithm 1 Policy Gradient with AL-BPO (exact oracles).

Require: Initial θ_1; step size η; number of iterations N; distribution ρ.

1: **for** $i = 1, 2, \cdots, N$ **do**

2: **Phase 1 (State sampling):** Obtain $\tilde{s}_i \sim d_\rho^{\pi_{\theta_i}}$ via trajectory rollout by first initialize $t = 0$ and then:

(1) Draw $s_t \sim \rho$.

(2) Repeat: with prob. $1-\gamma$ stop, return $\tilde{s}_i = s_t$; else: draw $a_t^{\text{env}} \sim \pi_{\theta_i}(\cdot|s_t)$, query $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t^{\text{env}})$, update $t \leftarrow t+1$.

3: **Phase 2 (AL-BPO action sampling):** At state $\tilde{s}_i$,

(1) Compute $\psi_i(\tilde{s}_i, a) = \|\nabla_\theta \log \pi_{\theta_i}(a \mid \tilde{s}_i)\|^2 \cdot Q^{\pi_{\theta_i}}(\tilde{s}_i, a)^2$ via the exact critic oracle **(O3)**.

(2) Compute $\mu_i^\star(a \mid \tilde{s}_i) \propto \pi_{\theta_i}(a \mid \tilde{s}_i)\sqrt{\psi_i(\tilde{s}_i, a)}$ via the exact behavior oracle **(O4)**.

(3) Sample $a_i^{\text{pg}} \sim \mu_i^\star(\cdot \mid \tilde{s}_i)$.

4: **Phase 3 (Gradient estimate and update):**

$$\hat{g}_i(\tilde{s}_i, a_i^{\text{pg}}) = \frac{1}{1-\gamma} \cdot \frac{\pi_{\theta_i}(a_i^{\text{pg}} \mid \tilde{s}_i)}{\mu_i^\star(a_i^{\text{pg}} \mid \tilde{s}_i)} \cdot \nabla_\theta \log \pi_{\theta_i}(a_i^{\text{pg}} \mid \tilde{s}_i) \cdot Q^{\pi_{\theta_i}}(\tilde{s}_i, a_i^{\text{pg}}).$$

5: Update

$$\theta_{i+1} = \theta_i + \eta \cdot \hat{g}_i(\tilde{s}_i, a_i^{\text{pg}}).$$

6: **end for**

7: **return** θ_{N+1}

Proposition 2. *Under **(O1)–(O4)** and **(A1)–(A3)**, for each iteration i: (i) the policy gradient estimator is unbiased, i.e., $\mathbb{E}[\hat{g}_i(\tilde{s}_i, a_i^{\text{pg}}) \mid \theta_i] = \nabla J(\theta_i)$; (ii) the behavior policy $\mu_i^\star$ minimizes the conditional second moment $\mathbb{E}[\|\hat{g}_i(\tilde{s}_i, a_i^{\text{pg}})\|^2 \mid \tilde{s}_i, \theta_i]$.*

Thus Algorithm 1 is simply stochastic gradient ascent with an unbiased gradient estimator and a variance-optimal action-sampling distribution.

For any behavior policy sequence $\boldsymbol{\mu} := \{\mu_i\}_{i=1}^N$, let $\hat{g}_i^{(\mu_i)}$ denote the policy gradient estimator at iteration i constructed from $\tilde{s}_i \sim d_\rho^{\pi_{\theta_i}}$ and $a_i^{\text{pg}} \sim \mu_i(\cdot \mid \tilde{s}_i)$. Define the conditional variance

$$\sigma^2(\mu_i, \theta_i) := \mathbb{E}\left[\left\|\hat{g}_i^{(\mu_i)} - \nabla_\theta J(\theta_i)\right\|^2 \,\middle|\, \theta_i\right] = \mathbb{E}[\|\hat{g}_i^{(\mu_i)}\|^2 \mid \theta_i] - \|\nabla_\theta J(\theta_i)\|^2, \quad \bar{\sigma}_N^2(\boldsymbol{\mu}) := \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\sigma^2(\mu_i, \theta_i)],$$

where the equality holds by the unbiasedness of $\hat{g}_i^{(\mu_i)}$. We have the following convergence holds.

Theorem 3 (Convergence under exact oracles)**.** *Suppose **(O1)–(O4)** hold. Under **(A1)–(A3)**, assume $\bar{\sigma}_N^2(\mu^\star) < \infty$. If $\eta \le 1/(2\widetilde{L})$, then Algorithm 1 satisfies*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\le\; \frac{2(J^\star - J(\theta_1))}{\eta N} + \widetilde{L}\,\eta\,\bar{\sigma}_N^2(\boldsymbol{\mu}^\star). \tag{5}$$

Corollary 4 (Improvement over on-policy sampling)**.** *Let $\boldsymbol{\pi} := \{\pi_{\theta_i}\}_{i=1}^N$ denote the on-policy behavior sequence. Then*

$$\bar{\sigma}_N^2(\boldsymbol{\pi}) - \bar{\sigma}_N^2(\boldsymbol{\mu}^\star) = \frac{1}{(1-\gamma)^2 N}\sum_{i=1}^N \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_i}}}\left[\mathrm{Var}_{a\sim\pi_{\theta_i}(\cdot|s)}\sqrt{\psi_i(s,a)}\right] \ge 0.$$

Hence the variance term in Theorem 3 is always no greater than the on-policy policy gradient.

In the exact-oracle benchmark, AL-BPO is therefore not merely a reparameterization of policy gradient: it strictly improves the SGD constant by optimally reducing the action-level variance at the returned state.

We note that the per-state gap (4) is sharp: by Cauchy–Schwarz, no behavior policy satisfying coverage **(A1)** attains a smaller conditional second moment than $\mu_i^\star$, so $\bar{\sigma}_N^2(\boldsymbol{\mu}^\star)$ is the smallest

variance constant achievable in Theorem 3 over all admissible behavior sequences. Since Theorem 3 differs from the standard nonconvex SGD analysis [Ghadimi and Lan, 2013] only through this variance constant, AL-BPO retains the optimal $\mathcal{O}(1/\sqrt{N})$ rate at $\eta = \Theta(1/\sqrt{N})$ and improves only the leading constant, which isolates a clean variance-reduction effect rather than a rate improvement.

When does AL-BPO help most? The variance gap $\mathrm{Var}_{a\sim\pi_\theta}[\sqrt{\psi(s,a)}]$ in Lemma 1 is large whenever $\psi(s,a) = \|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 Q^{\pi_\theta}(s,a)^2$ varies strongly across actions, which arises in three representative regimes:

- *Heavy-tailed critic.* When $Q^{\pi_\theta}(s,\cdot)$ has high spread across actions, a small fraction of actions carries most of the gradient signal; AL-BPO concentrates samples on these informative actions, whereas on-policy sampling allocates queries in proportion to π_θ.
- *Action-dependent score.* When the score $\|\nabla_\theta \log \pi_\theta(a \mid s)\|$ varies strongly with a, some actions carry much more of the gradient signal than others. AL-BPO samples more frequently from these high-leverage actions, whereas on-policy sampling allocates queries in proportion to π_θ. The Gaussian policy $\pi_\theta = \mathcal{N}(\theta^\top s, \sigma^2)$ is a canonical instance of this regime, with $\|\nabla_\theta \log \pi_\theta(a \mid s)\| \propto |a - \theta^\top s|$ scaling linearly in the deviation from the mean.
- *Heavy-tailed rewards.* Reward functions that penalize suboptimal actions superlinearly amplify the spread of both the critic and the score. Section 6 and Appendix D confirms a monotonically growing AL-BPO advantage with the reward tail heaviness.

Conversely, AL-BPO offers little benefit when $\psi(s,\cdot)$ is approximately constant on the support of π_θ.

5 Convergence with Approximate Oracles

In practice, both **(O3)** and **(O4)** of Section 4 may fail, since critics are learned from finite data, and $\mu_i^\star$ may have no closed-form normalization and would need to be re-fitted as $\theta_i, \hat{Q}_i$ evolve. We therefore replace them by a ξ-approximate critic and a parametric behavior family defined below.

Approximate oracles We replace **(O3)** by the approximate critic oracle

$$\textbf{(O3}')\qquad \sup_{a\in\mathcal{A}} \left|\hat{Q}_i(s,a) - Q^{\pi_{\theta_i}}(s,a)\right| \le \xi \quad \text{for every visited state } s.$$

Here we treat the approximate critic as a given oracle to isolate the effect of behavior-policy optimization in the theoretical analysis. Appendix B.2 discusses the tracking assumptions, and Appendix B.5 further gives a score-weighted mean-squared relaxation of critic accuracy. Later in Appendix C we provide practical AL-BPO implementations with a TD critic for MDPs and with group-relative advantages in GRPO for LLM post-training. We also replace **(O4)** by a parametric behavior oracle

$$\textbf{(O4}')\qquad \text{for any visited state } s \text{ and parameter } \beta,\ \text{sample } a^{\mathrm{pg}} \sim \mu_\beta(\cdot \mid s).$$

Since coverage **(A1)** depends on the choice of μ_β and may fail in practice, we enforce it by deploying:

$$\bar{\mu}_{\beta,\theta}(a \mid s) = (1 - \varepsilon_{\mathrm{mix}}) \cdot \mu_\beta(a \mid s) + \varepsilon_{\mathrm{mix}} \cdot \pi_\theta(a \mid s), \qquad \varepsilon_{\mathrm{mix}} \in (0,1], \tag{6}$$

thus **(A1)** is always satisfied under the constructed behavior policy $\bar{\mu}_{\beta,\theta}$. The mixture retains at least a $1 - \varepsilon_{\mathrm{mix}}$ fraction of any improvement from a coverage-satisfying behavior policy (Appendix B.4). Appendix D.5 will empirically quantify the losses due to the coverage mixture, family restrictions, and behavior optimization.

Practical objective and best-in family behavior Define the plug-in quantity $\hat{\psi}_i(s,a) := \|\nabla_\theta \log \pi_{\theta_i}(a \mid s)\|^2 \, \hat{Q}_i(s,a)^2$. For any coverage-satisfying behavior policy μ, define the plug-in per-state second moment $\hat{\mathcal{V}}_\mu^{(i)}(s) := \dfrac{1}{(1-\gamma)^2} \int_{\mathcal{A}} \frac{\pi_{\theta_i}(a|s)^2}{\mu(a|s)} \hat{\psi}_i(s,a)\, \mathrm{d}a$. As in Section 4, once $\theta_i, \hat{Q}_i$ are fixed, the behavior policy affects only the action-level variance, so minimizing action-level variance reduces to minimizing $\hat{\mathcal{V}}_\mu^{(i)}(s)$.

When the behavior policy is restricted to the parametric family $\{\mu_\beta : \beta \in \mathcal{B}\}$, and coverage is enforced through the deployed mixture $\bar{\mu}_{\beta,\theta_i}$, this induces the stage-wise objective

$$F_i(\beta) := \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_i}}}\left[\hat{\mathcal{V}}^{(i)}_{\bar{\mu}_{\beta,\theta_i}}(s)\right] = \frac{1}{(1-\gamma)^2}\, \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_i}},\, a\sim\bar{\mu}_{\beta,\theta_i}(\cdot|s)} \left[\left(\frac{\pi_{\theta_i}(a \mid s)}{\bar{\mu}_{\beta,\theta_i}(a \mid s)}\right)^2 \hat{\psi}_i(s,a)\right].$$

Thus, $F_i(\beta)$ is exactly the plug-in analogue of the action-level variance objective from the exact-oracle setting restricted to the deployed parametric behavior family. Let $\beta_i^\star \in \arg\min_{\beta\in\mathcal{B}} F_i(\beta)$ denote the best behavior policy within the chosen parametric family.

Single-step behavior update The best response $\beta_i^\star$ is generally not available in closed form, and solving $\min_{\beta\in\mathcal{B}} F_i(\beta)$ exactly at every outer iteration would introduce an expensive inner loop. Instead, we freeze $(\theta_i, \hat{Q}_i)$ and take a single stochastic gradient step on $F_i(\beta)$.

Specifically, after drawing $\tilde{s}_i \sim d_\rho^{\pi_{\theta_i}}$ and sampling an action $a_i^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid \tilde{s}_i)$, we define

$$\hat{d}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) := -\frac{1}{(1-\gamma)^2}\left(\frac{\pi_{\theta_i}(a_i^{\mathrm{pg}} \mid \tilde{s}_i)}{\bar{\mu}_{\beta_i,\theta_i}(a_i^{\mathrm{pg}} \mid \tilde{s}_i)}\right)^2 \hat{\psi}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) \cdot \nabla_\beta \log \bar{\mu}_{\beta_i,\theta_i}(a_i^{\mathrm{pg}} \mid \tilde{s}_i). \tag{7}$$

$\hat{d}_i(\tilde{s}_i, a_i^{\mathrm{pg}})$ is an unbiased gradient estimator, i.e., $\mathbb{E}[\hat{d}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) \mid \theta_i, \beta_i] = \nabla_\beta F_i(\beta_i)$, and for given step size $\eta_\beta > 0$, the behavior parameter is updated by

$$\beta_{i+1} = \beta_i - \eta_\beta \cdot \hat{d}_i(\tilde{s}_i, a_i^{\mathrm{pg}}). \tag{8}$$

Thus the behavior policy does not solve the stage-wise BPO problem exactly at iteration i; instead, it tracks the moving minimizer $\beta_i^\star$ over time. To quantify this optimization gap, we define the average tracking error $\varepsilon_{\mathrm{track},N} := \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[F_i(\beta_i) - F_i(\beta_i^\star)\big]$. This term measures the variance overhead from using a single-step behavior update per iteration. Let $m_i := \mathbb{E}[\hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) \mid \theta_i, \beta_i]$ and $\hat{\sigma}_i^2(\beta) := \mathbb{E}[\|\hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) - m_i\|^2 \mid \theta_i, \beta_i = \beta]$. Since m_i does not depend on β, $F_i(\beta) = \hat{\sigma}_i^2(\beta) + \|m_i\|^2$, so $\varepsilon_{\mathrm{track},N}$ is exactly the average excess action-level variance from using β_i instead of $\beta_i^\star$.

Actor and behavior updates At iteration i, the actor uses the currently deployed behavior policy $\bar{\mu}_{\beta_i,\theta_i}$ to sample the action $a_i^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid \tilde{s}_i)$ for policy gradient estimator and forms

$$\hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) = \frac{1}{1-\gamma}\frac{\pi_{\theta_i}(a_i^{\mathrm{pg}} \mid \tilde{s}_i)}{\bar{\mu}_{\beta_i,\theta_i}(a_i^{\mathrm{pg}} \mid \tilde{s}_i)} \nabla_\theta \log \pi_{\theta_i}(a_i^{\mathrm{pg}} \mid \tilde{s}_i)\, \hat{Q}_i(\tilde{s}_i, a_i^{\mathrm{pg}}), \tag{9}$$

followed by the actor update $\theta_{i+1} = \theta_i + \eta \cdot \hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}})$ with step size η. In the meantime, the behavior policy takes a single stochastic gradient step on F_i following the same $(\tilde{s}, a_i^{\mathrm{pg}})$. We calculate $\hat{d}_i(\tilde{s}, a_i^{\mathrm{pg}})$ as in (7), and update the behavior with step size η_β via (8). Thus, the algorithm with approximate oracles is given as follows in Algorithm 2.

Algorithm 2 Policy Gradient with AL-BPO (approximate oracles)

Require: Initial (θ_1, β_1); actor step size η; behavior step size η_β; iterations N; initial distribution ρ.

1: **for** $i = 1, 2, \cdots, N$ **do**
2: **Phase 1 (State sampling):** Draw $\tilde{s}_i \sim d_\rho^{\pi_{\theta_i}}$ by the geometric-stopping rollout using **(O1)**–**(O2)** as that in Algorithm 1 Phase 1.
3: **Phase 2 (AL-BPO action sampling):** At state $\tilde{s}_i$, sample action $a_i^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid \tilde{s}_i)$.
4: **Phase 3 (Gradient estimate and update):** Compute the actor's gradient estimator $\hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}})$ by (9), and the behavior gradient estimator $\hat{d}_i(\tilde{s}_i, a_i^{\mathrm{pg}})$ by (7).
5: Update
$$\theta_{i+1} = \theta_i + \eta \cdot \hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}}), \qquad \beta_{i+1} = \beta_i - \eta_\beta \cdot \hat{d}_i(\tilde{s}_i, a_i^{\mathrm{pg}}).$$
6: **end for**
7: **return** θ_{N+1}

Theorem 5 (Convergence under approximate oracles)**.** *Suppose **(O1)–(O2)**, **(O3′)**, and **(O4′)** hold, and under **(A1)–(A3)**. Let $B := \frac{G\xi}{1-\gamma}$. For $\boldsymbol{\beta}^\star := \{\beta_i^\star\}_{i=1}^N$, define the average action-level variance under the stage-wise optimal behavior policy $\beta_i^\star$ by $\bar{\hat{\sigma}}_N^2(\boldsymbol{\beta}^\star) := \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\hat{\sigma}_i^2(\beta_i^\star)\big]$. If the actor step size satisfies $\eta \le 1/(2\widetilde{L})$ and $\bar{\hat{\sigma}}_N^2(\boldsymbol{\beta}^\star) < \infty$, then Algorithm 2 satisfies*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\le\; \frac{2(J^\star - J(\theta_1))}{\eta N} \;+\; \widetilde{L}\,\eta\,\bar{\hat{\sigma}}_N^2(\boldsymbol{\beta}^\star) \;+\; \widetilde{L}\,\eta\,\varepsilon_{\mathrm{track},N} \;+\; B^2. \tag{10}$$

When $\xi = 0$ and $\varepsilon_{\mathrm{track},N} = 0$, the bound reduces to the exact-oracle form of Theorem 3 with the variance constant $\bar{\sigma}_N^2(\boldsymbol{\mu}^\star)$ replaced by $\bar{\tilde{\sigma}}_N^2(\boldsymbol{\beta}^\star) \geq \bar{\sigma}_N^2(\boldsymbol{\mu}^\star)$. The non-negative difference is the *expressiveness gap* of the parametric family $\{\bar{\mu}_{\beta,\theta}\}$: it vanishes when the family contains $\mu_i^\star$ at every iteration and grows when the family is too restrictive.

Theorem 5 isolates the effect of the single-step behavior update through the residual term $\varepsilon_{\mathrm{track},N}$, which cleanly separates the actor-side convergence argument from the behavior-side optimization argument. To understand when the practical algorithm nearly matches the stage-wise optimal behavior policy benchmark, it remains to bound $\varepsilon_{\mathrm{track},N}$. To make this term explicit, we impose additional regularity conditions on the stage-wise behavior objective $F_i(\beta)$. The next proposition shows that the tracking error is controllable whenever the behavior-side optimization problem is well-conditioned.

Proposition 6 (Tracking-error bound)**.** *Suppose conditions **(B1)–(B4)** (stated in Appendix B) hold for the behavior-side optimization, and the actor and behavior step sizes satisfy $\eta \leq 1/(2\widetilde{L})$ and $\eta_\beta \leq 1/(2L_\beta)$. Then*

$$\varepsilon_{\mathrm{track},N} \lesssim \frac{\mathbb{E}\|\beta_1 - \beta_1^\star\|_2^2}{\eta_\beta\, N} + \eta_\beta + \frac{\eta^2}{\eta_\beta^2}, \tag{11}$$

*where $\lesssim$ hides the numerical constants related with **(B1)–(B4)** (see Appendix B for details).*

Each term in (11) has a clear interpretation: the first decays as $1/(\eta_\beta N)$ from geometric forgetting of the behavior initialization; the second scales linearly with η_β and is the steady-state noise floor of constant stepsize SGD on F_i (with the gradient-noise variance of $\hat{d}_i$ absorbed into $\lesssim$); the third η^2/η_β^2 captures the drift of the moving target $\beta_i^\star$ induced by the actor update. This drift term reveals an inherent *two-timescale* structure, and making it vanish requires $\eta \ll \eta_\beta$, so that the actor evolves slowly enough for the behavior parameter to track its stage-wise optimum. The behavior step size η_β here plays the same fast-timescale role as the critic step size in standard two-timescale actor-critic analyses: behavior and critic updates run on the fast timescale at comparable order, both faster than the actor on the slow timescale. Combining Proposition 6 with Theorem 5, choosing $\eta = \Theta(N^{-1/2})$ together with $\eta_\beta = \Theta(N^{-1/3})$ optimally balances the latter two contributions of $\varepsilon_{\mathrm{track},N}$ at $\Theta(N^{-1/3})$, so $\eta\,\varepsilon_{\mathrm{track},N} = O(N^{-5/6})$ is dominated by the $\Theta(N^{-1/2})$ actor-side terms in (10). The algorithm therefore attains the standard nonconvex-SGD stationarity rate $O(N^{-1/2})$ up to the constant critic bias floor B^2. AL-BPO's benefit lies in the variance constant $\bar{\tilde{\sigma}}_N^2(\boldsymbol{\beta}^\star)$: including π_θ in the family makes this benchmark no larger than its on-policy counterpart, with strict improvement when the family contains a strictly better behavior policy.

6 Experiments

We first use synthetic environments to examine the variance reduction and convergence benefits of AL-BPO. We then apply AL-BPO to GRPO post-training for RF antenna design and evaluate the resulting designs with full-wave electromagnetic simulations.

6.1 Synthetic Study

Environments We consider (i) a continuous-action contextual bandit ($\gamma = 0$) with state $s \sim \mathcal{N}(0, I_d)$ for $d = 4$; action $a \in \mathbb{R}$ and reward $r(s,a) = -(a - w^\top s)^4/4$, where $w \in \mathbb{R}^d$ is a fixed weight vector drawn from $\mathcal{N}(0, I_d/d)$. The optimal action at state s is thus $w^\top s$. Setting $\gamma = 0$ collapses $d_\rho^{\pi_\theta}$ to $\rho = \mathcal{N}(0, I_d)$ and the critic $\hat{Q}_i = r$ (thus $\xi = 0$ and the B^2 term in (10) vanishes); and (ii) a continuous-action MDP with $\gamma = 0.9$ with the same per-step reward and action-independent linear-Gaussian dynamics $s_{t+1} = As_t + \varepsilon_t$, where $A \in \mathbb{R}^{d\times d}$ has spectral radius 0.9 to make the system dynamics stable yet non-trivially correlated, and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_d)$ with $\sigma_\varepsilon = 0.5$. The critic $\hat{V}$ is a linear function on quadratic state features, fit by ridge regression (see Appendix D for details). We vary the reward to $r(s,a) = -(a - w^\top s)^p/p$ with $p \in \{2, 4, 6\}$ to study the effect of reward tail heaviness and shows AL-BPO's advantage over on-policy grows monotonically with p (Appendix D).

Algorithms The actor policy is $\pi_\theta(\cdot \mid s) = \mathcal{N}(\theta^\top s, 1)$. We compare four choices of the behavior: **B1 (On-policy)** $\mu_\beta = \pi_\theta$; **B2 (Fixed Uniform Behavior)** $\mu_\beta \equiv \mathrm{Unif}[-10, 10]$; **B3 (Fixed Gaussian Behavior)** $\mu_\beta \equiv \mathcal{N}(0, 25)$; **B4 (Our AL-BPO under Algorithm 2)** $\mu_\beta(\cdot \mid s) = \mathcal{N}(\theta^\top s,\, \sigma_\beta^2)$ with a single learned scalar σ_β (initialized to 1) updated following the behavior gradient (7).

All the above algorithms use the same actor update and apply the same mixture following (6) with $\varepsilon_{\text{mix}} = 0.1$. Each run uses $N = 10000$ iterations for $\gamma = 0$ and $N = 20000$ for $\gamma = 0.9$, repeated over 20 random seeds. Actor and behavior step sizes per environment are listed in Appendix Table 2. The results are presented in Figure 1. We additionally ablate $\varepsilon_{\text{mix}} \in \{0, 0.05, 0.1, 0.2, 0.5\}$ and find AL-BPO robust for $\varepsilon_{\text{mix}} \in [0, 0.2]$, degrading only as $\varepsilon_{\text{mix}} \to 0.5$ (Appendix D).

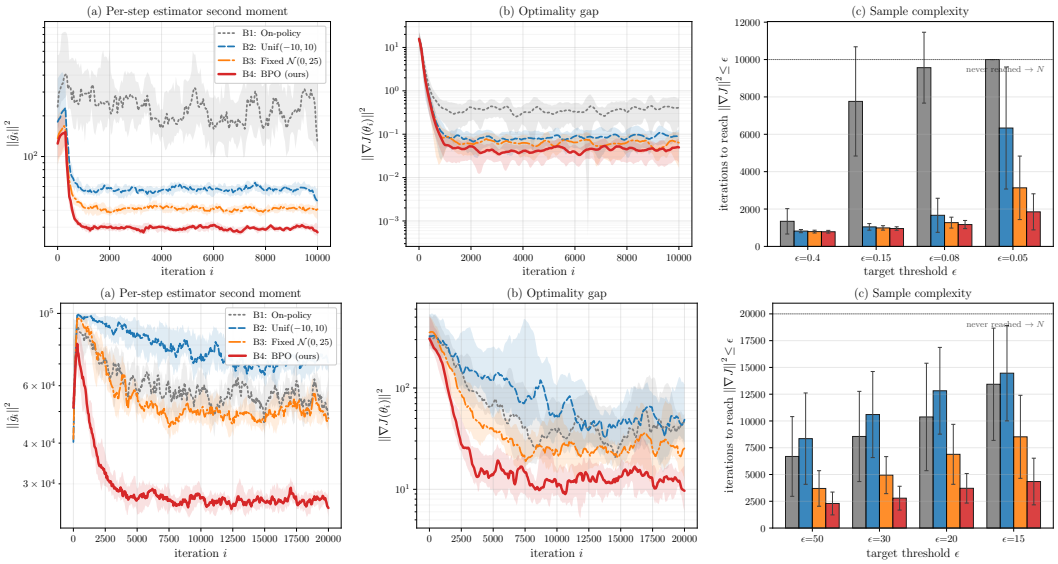


Figure 1: Top for contextual bandit; bottom for MDP with $\gamma = 0.9$ (bands cover 25–75 percentiles in panels (a)-(b), see Appendix D). **(a)** Second moment $\|\hat{g}_i\|^2$ of the policy gradient estimator (9). **(b)** Squared gradient norm $\|\nabla J(\theta_i)\|^2$ versus iteration. **(c)** Iterations needed to reach $\|\nabla J(\theta_i)\|^2 \le \epsilon$.

Figure 1 confirms the advantage of AL-BPO illustrated before. Panel (a) visualizes the variance constant entering Theorem 5: **B4** attains the smallest empirical second moment $\|\hat{g}_i\|^2$; Panel (b) shows the variance ordering carries over to the gradient norm $\|\nabla J(\theta_i)\|^2$: **B4** reaches the lowest level; Panel (c) translates this into sample complexity, where **B4** requires the minimum iterations.

Robustness to on-policy observation model The experiments above sample $\tilde{s} \sim d_\rho^{\pi_\theta}$ via geometric stopping. To verify that AL-BPO's benefit survives a more practical setup, we replace the geometric rollout with the standard online on-policy observation model: at each iteration, sample one step under π_θ to obtain a single transition, use 1-step TD bootstrapping to update the linear critic $\hat{V}$. Keeping all other components of Environment (ii) unchanged, we compare a vanilla A2C-style on-policy actor-critic (with batch size 1) against the same algorithm equipped with AL-BPO. Figure 2 shows that AL-BPO's variance-reduction and convergence advantage carry over to this practical setup.

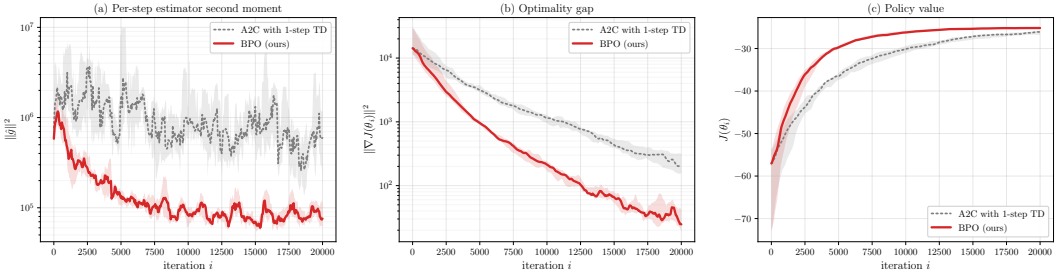


Figure 2: Comparison under the standard online on-policy observation model. Panels (a)–(b) follow the structure of Figure 1; Panel (c) plots the policy value $J(\theta_i)$ versus iteration.

Besides, we also evaluate AL-BPO on Pendulum-v1 using neural actor, critic, and behavior networks, with the results summarized in Appendix D.6.

6.2 LLM Post-Training

To examine AL-BPO in LLM post-training, we integrate it into GRPO and apply the combined procedure to Qwen3-4B-Instruct for one-shot radio-frequency (RF) antenna design. Given a specification of the frequency band, impedance match, and board requirements, the model generates a complete printed circuit board geometry. Training rewards come from finetuned Heaviside-v0 Bryant et al. [2026b] augmented with a far-field prediction head. Starting from a supervised fine-tuning (SFT) checkpoint, we train a rank-16 LoRA adapter for 500 updates, each using 32 prompts and 16 responses per prompt. GRPO uses group-relative rewards without a learned critic. [2]

AL-BPO changes the sampling distribution at one token in one response per group. After a random starting position, we select the first prefix whose largest next-token probability is below 0.99. The behavior mixture combines the actor's top-20 distribution at a learned temperature with the actor in equal proportions, with all other tokens sampled from the actor. A single importance weight, bounded by 2, corrects the group objective, and the behavior temperature is updated using the squared group-gradient norm. Appendix C.2 provides an illustration of the algorithm.

We evaluate with full-wave Palace simulations on the 470 of 501 held-out prompts with simulation evidence for all three policies. Invalid designs receive zero reward. Success requires the requested RF criteria, with total efficiency allowed up to 15 percentage points below the target. Appendix E gives the full protocol and example designs.

Table 1 compares Base, SFT, and SFT + AL-BPO, where the last denotes GRPO post-training with AL-BPO. Relative to SFT, mean Palace reward rises from 0.256 to 0.305 (19.1%), success increases by 4.9 percentage points, and the fraction with reward above 0.5 rises from 28.5% to 35.1%. Invalid designs fall from 5.1% to 3.8%. These gains, measured with a full-wave simulator separate from the training reward model, demonstrate the use of AL-BPO within GRPO post-training and the improvement of the combined procedure over SFT.

Table 1: RF design results for Qwen3-4B-Instruct on 470 held-out prompts, evaluated with Palace. Brackets give 95% prompt-bootstrap intervals. Reward > 0.5 reports the fraction scoring above 0.5.

Training	Mean reward	Success (%)	Reward > 0.5 (%)	Invalid (%)
Base	0.017	0.0	0.6	61.1
	[0.010, 0.025]	[0.0, 0.0]	[0.0, 1.5]	[56.6, 65.3]
SFT	0.256	17.9	28.5	5.1
	[0.227, 0.285]	[14.9, 21.1]	[25.1, 32.1]	[3.2, 7.0]
SFT + AL-BPO	0.305	22.8	35.1	3.8
	[0.276, 0.335]	[19.4, 26.2]	[31.5, 38.9]	[2.1, 5.7]

7 Limitations and open questions

Our theoretical guarantees assume access to a resettable simulator that can be advanced under π_θ, together with a critic oracle that can be queried for any action at each sampled state. These assumptions go beyond access to purely logged data. How AL-BPO performs under standard off-policy data collection remains an open direction. A further direction is to integrate AL-BPO into natural policy gradient, TRPO, and PPO, and investigate whether it improves sample efficiency in theory and practice.

Acknowledgments and Disclosure of Funding

This work is supported by NSF (CCF-2312205, ECCS-2419564), ONR-13983263 and 2027 New York University Center for Global Economy and Business grant. We thank Berkin Cilingiroglu and

[2] More details on the task, the physics reward, and the resulting antenna designs at `https://www.arenaphysica.com/publications/one-shot-antenna-design`

Tommaso Dreossi from Arena Physica for implementing AL-BPO in the LLM post-training pipeline for our RF antenna design experiments and the anonymous reviewers for constructive feedback. We also thank Pratap Ranade for his visionary insights into the transformative power and opportunity of physical foundation model as both tools and reward verifiers, without which the work would not be possible.

References

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Christopher Bryant, Tommaso Dreossi, Hao Liu, Nathan Mirman, Noah Kessler, Kaustubh Pabba, Usman Farman, Daniel Pedraza, Trevor Beaton, Akshay Kumar, Ben Gelb, Michael Frei, Roberto Riganti, Arun Natarajan, and Harish Krishnaswamy. Introducing Atlas RF Studio: Toward a foundation model for electromagnetics. Arena Physica Blog, mar 2026a. URL `https://www.arenaphysica.com/publications/rf-studio`. Published March 30, 2026.

Christopher M. Bryant, Hao Liu, Nathan Mirman, Tommaso Dreossi, Roberto Riganti, Ruichen Zhao, Noah Kessler, Boyuan Zhang, Trevor Beaton, Michael Frei, and Pratap Ranade. Introducing heaviside-1. Arena Physica, September 2026b. URL `https://www.arenaphysica.com/publications/heaviside-1`.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36:7017–7049, 2023.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Thinh T Doan, Lam M Nguyen, Nhan H Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under markov randomness. *arXiv preprint arXiv:2003.10973*, 2020.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Alex W Goodall, Francesco Belardinelli, et al. Behaviour policy optimization: Provably lower variance return estimates for off-policy reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21324–21332, 2026.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov):1471–1530, 2004.

Josiah P Hanna, Yash Chandak, Philip S Thomas, Martha White, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. *Journal of Machine Learning Research*, 25(313):1–58, 2024.

Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *International Conference on Machine Learning*, pages 4434–4443. PMLR, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

Shuze Daniel Liu, Claire Chen, and Shangtong Zhang. Doubly optimal policy evaluation for reinforcement learning. *arXiv preprint arXiv:2410.02226*, 2024.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. A tale of sampling and estimation in discounted reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 4575–4601. PMLR, 2023.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

Subhojyoti Mukherjee, Josiah P Hanna, and Robert D Nowak. Revar: Strengthening policy evaluation via reduced variance sampling. In *Uncertainty in Artificial Intelligence*, pages 1413–1422. PMLR, 2022.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pages 4026–4035. Pmlr, 2018.

Matteo Papini, Giorgio Manganini, Alberto Maria Metelli, and Marcello Restelli. Policy gradient with active importance sampling. *arXiv preprint arXiv:2405.05630*, 2024.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

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

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pages 11581–11591. PMLR, 2021.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *International Conference on Machine Learning*, pages 26713–26749. PMLR, 2022.

Shangtong Zhang. *Breaking the deadly triad in reinforcement learning*. PhD thesis, University of Oxford, 2022.

A Related Work

Our setting synthesizes four lines of related work, each of which supplies one ingredient but none of which, on its own, addresses the action-level behavior policy optimization problem we study.

Off-policy policy gradient REINFORCE [Williams, 1992], the PG theorem [Sutton et al., 1999], and early importance-sampled off-policy estimators [Precup et al., 2000] established the structural form of PG estimators and the validity of IS correction. Degris et al. [2012] made this incremental and online, and Liu et al. [2019] clarified that correcting only for action ratios is, in general, insufficient. In all of these the behavior policy is *given*; our contribution is to *optimize* it.

Variance reduction for policy gradient A long line of work reduces PG variance under fixed on-policy sampling via state-dependent baselines [Greensmith et al., 2004], advantage estimation [Mnih et al., 2016, Schulman et al., 2015], and action-dependent control variates [Liu et al., 2024]. These techniques modify the gradient *estimator* for a given sampling distribution, while we note that AL-BPO is orthogonal in that it modifies the sampling distribution itself and is composable with any of them. SVRPG [Papini et al., 2018] uses a reference-policy control variate, while SRVR-PG [Xu et al., 2019] recursively updates gradient estimates using importance-corrected differences between successive policies. These methods reduce variance across updates, while AL-BPO optimizes the action-sampling distribution within each update. The mechanisms are therefore complementary.

Doubly robust and OPE-to-PG bridges Jiang and Li [2016], Kallus and Uehara [2020], Zhang et al. [2022] study efficient off-policy evaluation via nuisance functions, and Huang and Jiang [2020], Xu et al. [2021] explicitly port this technology to policy gradients. The role of our critic oracle is exactly that of a nuisance object in this literature; what is new is that the critic feeds into the design of μ, not only into a control variate.

Behavior-policy optimization and active importance sampling Papini et al. [2024] address the trajectory-level version of this question: they derive the minimum-variance trajectory proposal and fit it by weighted cross-entropy with defensive IS, handling state distribution shift through trajectory-level importance correction. We instead obtain on-policy state samples through geometric-stopping rollouts and optimize the action-level proposal at the returned state, which yields a per-state closed-form optimum and a single-step IS ratio. Goodall et al. [2026] also study behavior policy optimization for off-policy RL but target lower-variance return estimation under standard off-policy collection rather than the action-level proposal at on-policy returned states. The remaining works target policy evaluation rather than the gradient: Hanna et al. [2024] optimize μ against the MSE of a value estimator, Mukherjee et al. [2022] develop oracle reduced-variance data collection in structured MDPs, and Liu et al. [2024] jointly design an optimal estimator together with an optimal behavior policy for evaluation.

Stochastic approximation for actor-critic Finite-time analyses of actor-critic under dependent sampling [Wu et al., 2020, Chen and Zhao, 2023] and the broader Markovian-SGD program [Doan et al., 2020, Borkar and Meyn, 2000] provide the convergence templates that our proofs adapt. We deliberately work in the generative / trajectory-rollout regime [Kakade, 2003, Metelli et al., 2023], which eliminates mixing-time factors and lets the pure AL-BPO benefit be seen in isolation.

B Proofs and Additional Theoretical Analysis

B.1 Proofs of Main Results

Proof of Lemma 1. By Cauchy-Schwarz inequality and under Assumption (**A1**) (which makes the integrals well-defined), we have

$$\int_{\mathcal{A}} \pi_\theta^2(a \mid s) \cdot \psi(s,a)/\mu(a \mid s)\,\mathrm{d}a \cdot \int_{\mathcal{A}} \mu(a \mid s)\,\mathrm{d}a \geq \left(\int_{\mathcal{A}} \pi_\theta(a \mid s)\sqrt{\psi(s,a)}\,\mathrm{d}a\right)^2 = Z(s)^2,$$

with equality holds iff $\mu(a \mid s) \propto \pi_\theta(a \mid s)\sqrt{\psi(s,a)}$.

For (4), $\mathcal{V}_{\pi_\theta}(s) = \mathbb{E}_{\pi_\theta}[\psi(s,a)]/(1-\gamma)^2$ and $\mathcal{V}_{\mu^\star}(s) = (\mathbb{E}_{\pi_\theta}[\sqrt{\psi(s,a)}])^2/(1-\gamma)^2$, and their difference is exactly $\mathrm{Var}_{\pi_\theta}[\sqrt{\psi(s,a)}]/(1-\gamma)^2$. □

Proof of Proposition 2. Fix iteration i and condition on θ_i, by Assumption **(A1)** we have the importance ratio $\pi_{\theta_i}/\mu_i^\star$ is well-defined on the support of $\pi_{\theta_i}(\cdot \mid \tilde{s}_i)$ for every visited $\tilde{s}_i$.

Unbiasedness. For any $\tilde{s} \in \mathcal{S}$ and any μ satisfying **(A1)**,

$$\begin{aligned}\mathbb{E}_{a\sim\mu}\big[\hat{g}_i(\tilde{s},a) \,\big|\, \tilde{s},\theta_i\big] &= \frac{1}{1-\gamma}\int_{\mathcal{A}} \mu(a \mid \tilde{s}) \cdot \frac{\pi_{\theta_i}(a \mid \tilde{s})}{\mu(a \mid \tilde{s})} \cdot \nabla_\theta \log \pi_{\theta_i}(a \mid \tilde{s}) \cdot Q^{\pi_{\theta_i}}(\tilde{s},a)\,\mathrm{d}a \\ &= \frac{1}{1-\gamma}\int_{\mathcal{A}} \pi_{\theta_i}(a \mid \tilde{s}) \nabla_\theta \log \pi_{\theta_i}(a \mid \tilde{s})\, Q^{\pi_{\theta_i}}(\tilde{s},a)\,\mathrm{d}a,\end{aligned}$$

where the second equality uses **(A1)** to cancel the importance ratio. Taking the expectation over $\tilde{s} \sim d_\rho^{\pi_{\theta_i}}$ then recovers the policy gradient theorem (1), so $\mathbb{E}[\hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}}) \mid \theta_i] = \nabla_\theta J(\theta_i)$.

Variance-optimality. Conditional on $\tilde{s}_i$ and θ_i, for any μ satisfying **(A1)**,

$$\mathbb{E}_{a\sim\mu}\big[\|\hat{g}_i(\tilde{s}_i,a)\|^2 \,\big|\, \tilde{s}_i,\theta_i\big] = \frac{1}{(1-\gamma)^2}\int_{\mathcal{A}} \frac{\pi_{\theta_i}(a \mid \tilde{s}_i)^2}{\mu(a \mid \tilde{s}_i)}\,\psi_i(\tilde{s}_i,a)\,\mathrm{d}a = \mathcal{V}_\mu(\tilde{s}_i),$$

which is exactly the per-state second moment of Lemma 1 (applied at $\theta = \theta_i$). By Lemma 1, the minimizer over μ satisfying **(A1)** is $\mu_i^\star(\cdot \mid \tilde{s}_i) \propto \pi_{\theta_i}(\cdot \mid \tilde{s}_i)\sqrt{\psi_i(\tilde{s}_i,\cdot)}$, which is the behavior policy used in Algorithm 1. □

Proof of Theorem 3. The proof follows the standard nonconvex SGD framework of Ghadimi and Lan [2013], instantiated with the unbiasedness and conditional second-moment characterization from Proposition 2. In the remaining proof, we will abbreviate $\hat{g}_i(\tilde{s}_i, a_i^{\mathrm{pg}})$ as $\hat{g}_i$ for simplicity.

Step 1 (descent inequality). By the $\widetilde{L}$-smoothness of J in **(A3)** and the update $\theta_{i+1} = \theta_i + \eta\,\hat{g}_i$,

$$J(\theta_{i+1}) \;\geq\; J(\theta_i) + \eta\,\langle \nabla_\theta J(\theta_i),\, \hat{g}_i\rangle - \frac{\widetilde{L}\,\eta^2}{2}\|\hat{g}_i\|^2.$$

Taking the conditional expectation $\mathbb{E}[\cdot \mid \theta_i]$ and applying the unbiasedness $\mathbb{E}[\hat{g}_i \mid \theta_i] = \nabla_\theta J(\theta_i)$ from Proposition 2,

$$\mathbb{E}[J(\theta_{i+1}) \mid \theta_i] \;\geq\; J(\theta_i) + \eta\,\|\nabla_\theta J(\theta_i)\|^2 - \frac{\widetilde{L}\,\eta^2}{2}\,\mathbb{E}[\|\hat{g}_i\|^2 \mid \theta_i]. \tag{12}$$

Step 2 (second-moment decomposition). Unbiasedness also gives the exact decomposition of $\mathbb{E}[\|\hat{g}_i\|^2 \mid \theta_i] = \|\nabla_\theta J(\theta_i)\|^2 + \sigma^2(\mu_i^\star, \theta_i)$. Taking it back to (12), we get

$$\mathbb{E}[J(\theta_{i+1}) \mid \theta_i] \;\geq\; J(\theta_i) + \eta\Big(1 - \frac{\widetilde{L}\,\eta}{2}\Big)\cdot\|\nabla_\theta J(\theta_i)\|^2 - \frac{\widetilde{L}\,\eta^2}{2}\,\sigma^2(\mu_i^\star,\theta_i). \tag{13}$$

Step 3 (telescoping). Taking total expectation and rearranging of the terms in (13), we have

$$\eta\Big(1 - \frac{\widetilde{L}\,\eta}{2}\Big)\cdot\mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\leq\; \mathbb{E}[J(\theta_{i+1}) - J(\theta_i)] + \frac{\widetilde{L}\,\eta^2}{2}\,\mathbb{E}[\sigma^2(\mu_i^\star,\theta_i)].$$

Summing $i = 1, \cdots, N$, we have the J summation terms cancel out. Using the fact that $\mathbb{E}[J(\theta_{N+1})] \le J^\star$ and $\sum_{i=1}^N \mathbb{E}[\sigma^2(\mu_i^\star, \theta_i)] = N\,\bar{\sigma}_N^2(\boldsymbol{\mu}^\star)$ by its definition,

$$\eta\Big(1 - \frac{\widetilde{L}\,\eta}{2}\Big) \cdot \sum_{i=1}^N \mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\le\; J^\star - J(\theta_1) + \frac{\widetilde{L}\,\eta^2\,N}{2}\,\bar{\sigma}_N^2(\boldsymbol{\mu}^\star).$$

Step 4 (step-size condition). Since $\eta \le 1/(2\widetilde{L})$, we have $\widetilde{L}\,\eta/2 \le 1/4$, so $1 - \widetilde{L}\eta/2 > 1/2$. Dividing both sides by $\eta(1 - \widetilde{L}\eta/2)\,N \ge \eta\,N/2$ yields

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\le\; \frac{2(J^\star - J(\theta_1))}{\eta\,N} + \widetilde{L}\,\eta\,\bar{\sigma}_N^2(\boldsymbol{\mu}^\star),$$

which is our desired result (5). □

Proof of Corollary 4. Fix iteration i and θ_i. By Proposition 2 and the unbiasedness of $\hat{g}_i^{(\mu_i)}$ for any μ_i satisfying **(A1)**,

$$\sigma^2(\mu_i, \theta_i) \;=\; \mathbb{E}[\|\hat{g}_i^{(\mu_i)}\|^2 \mid \theta_i] - \|\nabla_\theta J(\theta_i)\|^2 \;=\; \mathbb{E}_{\tilde{s}\sim d_\rho^{\pi_{\theta_i}}}[\mathcal{V}_{\mu_i}(\tilde{s})] - \|\nabla_\theta J(\theta_i)\|^2,$$

where the second equality uses the per-state second moment computed in the proof of Proposition 2. Subtracting the two cases of $\mu_i = \pi_{\theta_i}$ and $\mu_i = \mu_i^\star$, the $\|\nabla_\theta J(\theta_i)\|^2$ terms cancel and Lemma 1 gives

$$\begin{aligned}\sigma^2(\pi_{\theta_i}, \theta_i) - \sigma^2(\mu_i^\star, \theta_i) \;&=\; \mathbb{E}_{\tilde{s}\sim d_\rho^{\pi_{\theta_i}}}\Big[\mathcal{V}_{\pi_{\theta_i}}(\tilde{s}) - \mathcal{V}_{\mu_i^\star}(\tilde{s})\Big] \\ &=\; \frac{1}{(1-\gamma)^2}\,\mathbb{E}_{\tilde{s}\sim d_\rho^{\pi_{\theta_i}}}\Big[\mathrm{Var}_{a\sim\pi_{\theta_i}(\cdot|\tilde{s})}\sqrt{\psi_i(\tilde{s}, a)}\Big].\end{aligned}$$

Taking $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\cdot]$ on both sides recovers the claimed identity. Non-negativity follows from $\mathrm{Var} \ge 0$, with strict inequality holds whenever there exists some iteration i such that $\psi_i(s, \cdot)$ is non-constant on the support of $\pi_{\theta_i}(\cdot \mid s)$. □

Proof of Theorem 5. The proof extends the argument of Theorem 3 in two places: the gradient estimator $\hat{g}_i$ is now biased (because the critic $\hat{Q}_i$ is approximate), and the deployed behavior $\bar{\mu}_{\beta_i,\theta_i}$ may differ from the stage-wise optimum $\bar{\mu}_{\beta_i^\star,\theta_i}$ within the parametric family. We will handle the two effects in the following separate steps.

Step 1 (bias bound). Fix iteration i and condition on (θ_i, β_i). The deployed behavior $\bar{\mu}_{\beta_i,\theta_i}$ satisfies coverage **(A1)** by construction (6), so the importance ratio in $\hat{g}_i$ cancels under the conditional expectation over $a^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid \tilde{s})$ and the conditional expectation of the gradient estimator can be expressed as:

$$m_i \;=\; \mathbb{E}[\hat{g}_i \mid \theta_i, \beta_i] \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{\tilde{s}\sim d_\rho^{\pi_{\theta_i}}}\Big[\int_{\mathcal{A}} \pi_{\theta_i}(a \mid \tilde{s})\,\nabla_\theta \log \pi_{\theta_i}(a \mid \tilde{s})\,\hat{Q}_i(\tilde{s}, a)\,\mathrm{d}a\Big].$$

For fixed θ_i and $\hat{Q}_i$, importance weighting makes the conditional mean independent of the behavior parameter. Rewriting the inner integral as an expectation over $a \sim \pi_{\theta_i}(\cdot \mid \tilde{s})$,

$$m_i \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{\tilde{s}\sim d_\rho^{\pi_{\theta_i}},\, a\sim\pi_{\theta_i}(\cdot|\tilde{s})}\Big[\nabla_\theta \log \pi_{\theta_i}(a \mid \tilde{s})\,\hat{Q}_i(\tilde{s}, a)\Big]. \tag{14}$$

The policy gradient (1) has the same form on the same sampling measure, with $\hat{Q}_i$ replaced by $Q^{\pi_{\theta_i}}$. Subtracting (14) from (1) by linearity of expectation,

$$m_i - \nabla_\theta J(\theta_i) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{\tilde{s}\sim d_\rho^{\pi_{\theta_i}},\, a\sim\pi_{\theta_i}(\cdot|\tilde{s})}\Big[\nabla_\theta \log \pi_{\theta_i}(a \mid \tilde{s}) \cdot \big(\hat{Q}_i(\tilde{s}, a) - Q^{\pi_{\theta_i}}(\tilde{s}, a)\big)\Big].$$

By taking the norm and applying **(A2)** ($\|\nabla_\theta \log \pi_{\theta_i}\| \le G$) and **(O3′)** ($|\hat{Q}_i - Q^{\pi_{\theta_i}}| \le \xi$), we have

$$\|m_i - \nabla_\theta J(\theta_i)\| \;\le\; \frac{G\xi}{1-\gamma} \;=\; B.$$

Step 2 (descent inequality). By $\widetilde{L}$-smoothness in **(A3)** and the actor update $\theta_{i+1} = \theta_i + \eta \cdot \hat{g}_i$,

$$J(\theta_{i+1}) \;\geq\; J(\theta_i) + \eta \,\langle \nabla_\theta J(\theta_i),\, \hat{g}_i \rangle - \frac{\widetilde{L}\,\eta^2}{2}\|\hat{g}_i\|^2.$$

Taking $\mathbb{E}[\,\cdot \mid \theta_i, \beta_i]$ and applying (i) $\langle \nabla_\theta J(\theta_i), m_i \rangle = \frac{1}{2}\big(\|\nabla_\theta J(\theta_i)\|^2 + \|m_i\|^2 - \|m_i - \nabla_\theta J(\theta_i)\|^2\big)$ and (ii) the second-moment decomposition $\mathbb{E}[\|\hat{g}_i\|^2 \mid \theta_i, \beta_i] = \|m_i\|^2 + \hat{\sigma}_i^2(\beta_i)$, we have

$$\mathbb{E}[J(\theta_{i+1}) \mid \theta_i, \beta_i] \geq J(\theta_i) + \frac{\eta}{2}\|\nabla_\theta J(\theta_i)\|^2 + \frac{\eta}{2}(1 - \widetilde{L}\eta) \cdot \|m_i\|^2 - \frac{\eta}{2}\|m_i - \nabla_\theta J(\theta_i)\|^2 - \frac{\widetilde{L}\eta^2}{2}\hat{\sigma}_i^2(\beta_i).$$

Step 3 (apply the step-size condition and bias bound). Since $\eta \leq 1/(2\widetilde{L})$, $1 - \widetilde{L}\eta \geq 1/2$, so the $\|m_i\|^2$ term (which is a constant w.r.t. the behavior policy and will not influence our variance) is non-negative and can be dropped from the lower bound. Combining with the bias bound $\|m_i - \nabla_\theta J(\theta_i)\|^2 \leq B^2$ from Step 1,

$$\mathbb{E}[J(\theta_{i+1}) \mid \theta_i, \beta_i] \;\geq\; J(\theta_i) + \frac{\eta}{2}\|\nabla_\theta J(\theta_i)\|^2 - \frac{\eta B^2}{2} - \frac{\widetilde{L}\eta^2}{2}\hat{\sigma}_i^2(\beta_i). \tag{15}$$

Step 4 (tracking decomposition). The mean–variance relation of $F_i(\beta) = \hat{\sigma}_i^2(\beta) + \|m_i\|^2$ gives $F_i(\beta) = \hat{\sigma}_i^2(\beta) + \|m_i\|^2$, and since $\|m_i\|^2$ is independent of β, we have

$$\hat{\sigma}_i^2(\beta_i) \;=\; \hat{\sigma}_i^2(\beta_i^\star) + \big(F_i(\beta_i) - F_i(\beta_i^\star)\big).$$

Substituting it into the bound of (15) and rearranging the terms, we can get

$$\frac{\eta}{2}\|\nabla_\theta J(\theta_i)\|^2 \;\leq\; \mathbb{E}[J(\theta_{i+1}) - J(\theta_i) \mid \theta_i, \beta_i] + \frac{\eta B^2}{2} + \frac{\widetilde{L}\eta^2}{2}\hat{\sigma}_i^2(\beta_i^\star) + \frac{\widetilde{L}\eta^2}{2}\big(F_i(\beta_i) - F_i(\beta_i^\star)\big).$$

Step 5 (telescoping). Take total expectation and sum $i = 1, \cdots, N$, we have the J summation terms cancel out. Using $\mathbb{E}[J(\theta_{N+1})] \leq J^\star$, the definitions of $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\hat{\sigma}_i^2(\beta_i^\star)] = \bar{\hat{\sigma}}_N^2(\boldsymbol{\beta}^\star)$ and $\frac{1}{N}\sum_{i=1}^N \mathbb{E}[F_i(\beta_i) - F_i(\beta_i^\star)] = \varepsilon_{\mathrm{track},N}$, we have

$$\frac{\eta}{2}\sum_{i=1}^N \mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\leq\; J^\star - J(\theta_1) + \frac{\eta N B^2}{2} + \frac{\widetilde{L}\eta^2 N}{2}\bar{\hat{\sigma}}_N^2(\boldsymbol{\beta}^\star) + \frac{\widetilde{L}\eta^2 N}{2}\varepsilon_{\mathrm{track},N}.$$

Dividing both sides by $\eta N/2$ yields

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\nabla_\theta J(\theta_i)\|^2 \;\leq\; \frac{2(J^\star - J(\theta_1))}{\eta N} + B^2 + \widetilde{L}\eta\,\bar{\hat{\sigma}}_N^2(\boldsymbol{\beta}^\star) + \widetilde{L}\eta\,\varepsilon_{\mathrm{track},N},$$

which is our desired result (10). □

Proof of Proposition 6. We first state the regularity conditions on the behavior-side optimization. For each iteration i and all $\beta, \beta' \in \mathcal{B}$, the stage-wise objective $F_i(\beta)$ satisfies:

- **(B1)** *Convexity and smoothness*: F_i is convex in β and L_β-smooth, i.e., $\|\nabla_\beta F_i(\beta) - \nabla_\beta F_i(\beta')\| \leq L_\beta\|\beta - \beta'\|$.
- **(B2)** *Quadratic growth*: $F_i(\beta) - F_i(\beta_i^\star) \geq \frac{m_\beta}{2}\|\beta - \beta_i^\star\|^2$ for all β and some $m_\beta \leq L_\beta$.
- **(B3)** *Bounded gradient noise*: $\hat{d}_i$ in (7) is unbiased, $\mathbb{E}[\hat{d}_i \mid \theta_i, \beta_i] = \nabla_\beta F_i(\beta_i)$, and $\mathbb{E}[\|\hat{d}_i - \nabla_\beta F_i(\beta_i)\|^2 \mid \theta_i, \beta_i] \leq \sigma_\beta^2$.
- **(B4)** *Lipschitzness of optimal behavior*: the map $\theta \mapsto \beta^\star(\theta)$ satisfies $\|\beta^\star(\theta') - \beta^\star(\theta)\| \leq L_\star\|\theta' - \theta\|$ for all θ, θ'.

We adopt the convention $\beta_{N+1}^\star := \beta_N^\star$. Let $V_i := \|\beta_i - \beta_i^\star\|^2$ and $e_i := \mathbb{E}V_i$. We proceed in three steps:

Step 1 (one-step distance contraction). By expanding the behavior update (8), we get

$$\|\beta_{i+1} - \beta_i^\star\|^2 \;=\; V_i \;-\; 2\eta_\beta \langle \hat{d}_i,\, \beta_i - \beta_i^\star \rangle \;+\; \eta_\beta^2 \|\hat{d}_i\|^2. \tag{16}$$

Take $\mathbb{E}[\cdot \mid \theta_i, \beta_i]$ and use unbiasedness from **(B3)** on the cross term, by convexity **(B1)**, $\langle \nabla_\beta F_i(\beta_i), \beta_i - \beta_i^\star \rangle \geq F_i(\beta_i) - F_i(\beta_i^\star)$. Smoothness **(B1)** together with $\nabla_\beta F_i(\beta_i^\star) = 0$ gives $\|\nabla_\beta F_i(\beta_i)\|^2 \leq 2L_\beta(F_i(\beta_i) - F_i(\beta_i^\star))$, so combining with the variance bound in **(B3)**, we derive

$$\mathbb{E}[\|\hat{d}_i\|^2 \mid \theta_i, \beta_i] \;\leq\; 2L_\beta(F_i(\beta_i) - F_i(\beta_i^\star)) + \sigma_\beta^2. \tag{17}$$

By taking (17) back to (16), we can get

$$\mathbb{E}[\|\beta_{i+1} - \beta_i^\star\|^2 \mid \theta_i, \beta_i] \;\leq\; V_i \;-\; 2\eta_\beta(1 - L_\beta\eta_\beta) \cdot (F_i(\beta_i) - F_i(\beta_i^\star)) \;+\; \eta_\beta^2\sigma_\beta^2.$$

Since $\eta_\beta \leq 1/(2L_\beta)$, we have $1 - L_\beta\eta_\beta \geq 1/2$, and:

$$\mathbb{E}[\|\beta_{i+1} - \beta_i^\star\|^2 \mid \theta_i, \beta_i] \;\leq\; V_i \;-\; \eta_\beta(F_i(\beta_i) - F_i(\beta_i^\star)) \;+\; \eta_\beta^2\sigma_\beta^2. \tag{18}$$

By **(B2)**, $F_i(\beta_i) - F_i(\beta_i^\star) \geq \frac{m_\beta}{2} V_i$. Splitting $\eta_\beta(F_i - F_i^\star)$ in two equal halves and applying Quadratic growth property to one half,

$$-\eta_\beta(F_i(\beta_i) - F_i(\beta_i^\star)) \;\leq\; -\frac{\eta_\beta(F_i(\beta_i) - F_i(\beta_i^\star))}{2} \;-\; \frac{\eta_\beta m_\beta}{4} V_i.$$

Substituting it into (18), we get

$$\mathbb{E}[\|\beta_{i+1} - \beta_i^\star\|^2 \mid \theta_i, \beta_i] \;\leq\; \Big(1 - \frac{\eta_\beta m_\beta}{4}\Big) \cdot V_i \;-\; \frac{\eta_\beta(F_i(\beta_i) - F_i(\beta_i^\star))}{2} \;+\; \eta_\beta^2\sigma_\beta^2. \tag{19}$$

Step 2 (upper bound V_i). For any $\rho > 0$, the elementary inequality gives $\|a + b\|^2 \leq (1 + \rho) \cdot \|a\|^2 + (1 + 1/\rho) \cdot \|b\|^2$. Taking $a = \beta_{i+1} - \beta_i^\star$ and $b = \beta_i^\star - \beta_{i+1}^\star$ gives

$$V_{i+1} \;\leq\; (1 + \rho) \cdot \|\beta_{i+1} - \beta_i^\star\|^2 \;+\; (1 + 1/\rho) \cdot \|\beta_{i+1}^\star - \beta_i^\star\|^2. \tag{20}$$

Choose $\rho := \eta_\beta m_\beta/8$. Since $m_\beta \leq L_\beta$ and $\eta_\beta \leq 1/(2L_\beta)$, we have $\rho \leq 1/16$. Hence $1 + \rho \leq 9/8$ and $1 + 1/\rho \leq 9/(\eta_\beta m_\beta)$. Notice that

$$(1 + \rho) \cdot (1 - \frac{\eta_\beta m_\beta}{4}) \;=\; 1 - \frac{\eta_\beta m_\beta}{8} - \frac{(\eta_\beta m_\beta)^2}{32} \;\leq\; 1 - \frac{\eta_\beta m_\beta}{8}.$$

Multiplying (19) by $(1 + \rho) \geq 1$ and dropping the non-positive descent term and taking the results back to (20),

$$\mathbb{E}[V_{i+1} \mid \theta_i, \beta_i] \;\leq\; (1 - \frac{\eta_\beta m_\beta}{8}) V_i \;+\; \frac{9}{8}\eta_\beta^2\sigma_\beta^2 \;+\; \frac{9}{\eta_\beta m_\beta}\|\beta_{i+1}^\star - \beta_i^\star\|^2. \tag{21}$$

Step 3 (telescope). Take total expectation in (21) and let $r := 1 - \eta_\beta m_\beta/8 \in (0, 1)$ and $b_i := \frac{9}{8}\eta_\beta^2\sigma_\beta^2 + \frac{9}{\eta_\beta m_\beta}\mathbb{E}\|\beta_{i+1}^\star - \beta_i^\star\|^2$, so $e_{i+1} \leq r\, e_i + b_i$. Summing $i = 1, \cdots, N$ and using $\sum_{i=1}^N e_{i+1} = \sum_{i=1}^N e_i + e_{N+1} - e_1$,

$$(1 - r)\sum_{i=1}^N e_i \;\leq\; e_1 - e_{N+1} + \sum_{i=1}^N b_i \;\leq\; e_1 + \sum_{i=1}^N b_i. \tag{22}$$

Then, by **(B4)** and the actor update formula $\theta_{i+1} = \theta_i + \eta\, \hat{g}_i$,

$$\|\beta_{i+1}^\star - \beta_i^\star\|^2 \;\leq\; L_\star^2\, \|\theta_{i+1} - \theta_i\|^2 \;=\; L_\star^2\, \eta^2\, \|\hat{g}_i\|^2.$$

Under **(A1)**–**(A3)** and **(O3′)**, the importance ratio $\pi_{\theta_i}/\bar{\mu}_{\beta_i,\theta_i}$ is bounded by $1/\varepsilon_{\text{mix}}$ via the coverage mixture (6); the score is bounded by G via **(A2)**, and $|\hat{Q}_i| \leq V_{\max} := r_{\max}/(1 - \gamma) + \xi$ from bounded reward and **(O3′)**. Hence

$$\|\hat{g}_i\| \;\leq\; \frac{1}{1 - \gamma} \cdot \frac{1}{\varepsilon_{\text{mix}}} \cdot G \cdot V_{\max} \;=:\; M.$$

Thus $\mathbb{E}\|\beta^\star_{i+1}-\beta^\star_i\|^2 \le L_\star^2\,\eta^2\,M^2$ and $\sum_{i=1}^N \mathbb{E}\|\beta^\star_{i+1}-\beta^\star_i\|^2 \le N\,L_\star^2\,\eta^2\,M^2$ (by using the convention $\beta^\star_{N+1} := \beta^\star_N$). Therefore

$$\sum_{i=1}^N b_i \;\le\; \frac{9}{8}N\eta_\beta^2\sigma_\beta^2 + \frac{9NL_\star^2M^2\eta^2}{\eta_\beta m_\beta}.$$

Taking the result back to (22) and dividing it by $(1-r)N = (\eta_\beta m_\beta/8)N$,

$$\frac{1}{N}\sum_{i=1}^N e_i \;\le\; \frac{8\,\mathbb{E}V_1}{\eta_\beta m_\beta\,N} \;+\; \frac{9\,\eta_\beta\,\sigma_\beta^2}{m_\beta} \;+\; \frac{72\,L_\star^2\,M^2\,\eta^2}{\eta_\beta^2 m_\beta^2}.$$

Finally, by smoothness **(B1)** and $\nabla_\beta F_i(\beta_i^\star)=0$, we have $F_i(\beta_i)-F_i(\beta_i^\star) \le \frac{L_\beta}{2}V_i$. Thus

$$\varepsilon_{\mathrm{track},N} \;\le\; \frac{L_\beta}{2}\cdot\frac{1}{N}\sum_{i=1}^N e_i \;\le\; \frac{4L_\beta\,\mathbb{E}V_1}{\eta_\beta m_\beta\,N} \;+\; \frac{9L_\beta\,\eta_\beta\,\sigma_\beta^2}{2\,m_\beta} \;+\; \frac{36\,L_\beta\,L_\star^2\,M^2\,\eta^2}{\eta_\beta^2\,m_\beta^2}.$$

Hiding the numerical constants and the problem-dependent factors $L_\beta, L_\star, 1/m_\beta, M^2$ inside $\lesssim$ recovers (11).

□

B.2 Discussion of Assumptions (B1)–(B4).

We note that Assumptions **(B1)**–**(B4)** are sufficient for the explicit tracking rate in Proposition 6; they are not required for the closed-form optimal behavior policy or Theorem 5, which leaves the tracking error explicit. Here we discuss how restrictive these assumptions are below:

(B1) Convexity and smoothness. Smoothness of $F_i(\beta)$ holds whenever $\bar\mu_{\beta,\theta_i}$ is twice continuously differentiable in β with bounded first- and second-order log-density derivatives. Since F_i is a composition of these with the bounded importance ratio $\pi/\bar\mu \le 1/\varepsilon_{\mathrm{mix}}$ (from the coverage mixture (6)) and the bounded plug-in $\hat\psi_i$ (from **(A2)**, **(O3′)**, and bounded reward), smoothness is satisfied by standard parametric families (e.g., Gaussian, softmax). Convexity is more restrictive. For a Gaussian behavior $\mu_\beta(\cdot \mid s) = \mathcal{N}(\theta_i^\top s, \sigma_\beta^2)$ *without* the coverage mixture, $1/\mu_\beta(a;\sigma_\beta)$ is pointwise convex in σ_β, so F_i is convex in σ_β. With the deployed mixture $\bar\mu_{\beta,\theta_i}$, or for richer parametric families (e.g., deep neural networks), global convexity is a structural assumption rather than a derived property and has to be checked case by case. Extending the analysis to nonconvex F_i via the PL condition alone is a natural direction for future work.

(B2) Quadratic growth. Quadratic growth requires $F_i(\beta)-F_i(\beta_i^\star) \ge \frac{m_\beta}{2}\|\beta-\beta_i^\star\|^2$, i.e., the function value grows at least quadratically away from its minimum. This is strictly weaker than strong convexity, which requires the Hessian to be uniformly positive definite Karimi et al. [2016]. QG holds automatically near any non-degenerate minimum by Taylor expansion, so the substantive content of **(B2)** is requiring this quadratic growth to hold uniformly in β rather than only locally. For our 1-D experimental setting it can be verified numerically.

(B3) Unbiased and variance-bounded gradient. Unbiasedness $\mathbb{E}[\hat d_i \mid \theta_i, \beta_i] = \nabla_\beta F_i(\beta_i)$ holds by construction of the estimator (7). Bounded conditional variance $\sigma_\beta^2 < \infty$ follows from (i) the coverage mixture bounding the IS ratio $\pi_{\theta_i}/\bar\mu_{\beta_i,\theta_i}$ by $1/\varepsilon_{\mathrm{mix}}$; (ii) the bounded first-order log-density derivative of $\bar\mu_{\beta,\theta_i}$ already required by **(B1)** smoothness, which yields a uniform bound $\|\nabla_\beta \log \bar\mu_{\beta,\theta_i}\| \le G_\beta$; and (iii) the bounded plug-in $\hat\psi_i$, which follows from **(A2)** (score on the actor side, $\|\nabla_\theta \log \pi_{\theta_i}\| \le G$), **(O3′)**, and bounded reward. Thus no additional structural assumption beyond is needed.

(B4) Lipschitzness of optimal behavior. Lipschitz is the only assumption tied to the joint (θ,β) structure. Assume cross smoothness holds such that $\|\nabla_\beta F_i(\beta;\theta) - \nabla_\beta F_i(\beta;\theta')\| \le L_{F,\theta}\|\theta-\theta'\|$ for some constant $L_{F,\theta} > 0$. Then evaluating $\nabla_\beta F_i$ at $\beta = \beta_i^\star(\theta')$ yields two competing bounds,

$$\underbrace{\frac{m_\beta}{2}\|\beta_i^\star(\theta') - \beta_i^\star(\theta)\|}_{\text{convex+QG lower bound}} \;\le\; \|\nabla_\beta F_i(\beta_i^\star(\theta');\theta)\| \;\le\; \underbrace{L_{F,\theta}\|\theta-\theta'\|}_{\text{cross-smoothness upper bound}},$$

which leads to $L_\star \le 2L_{F,\theta}/m_\beta$. We note that this requires a uniform cross-gradient bound involving both the actor and critic. When the critic is learned, the stage-wise behavior optimum can move as either the actor or critic changes, so applying the tracking rate requires controlling both sources of drift.

B.3 Sampling costs, rollout reuse, and batching

Cost units. For $T \sim \mathrm{Geom}(1-\gamma)$ on $\{0,1,\cdots\}$, $\mathbb{E}[T] = \gamma/(1-\gamma)$ and $\mathbb{E}[T+1] = 1/(1-\gamma)$. Thus N occupancy-state samples require $N\gamma/(1-\gamma)$ expected forward transitions, in addition to N sampled proposal actions and their critic evaluations. The same state-generation cost applies to an on-policy comparator using this sampler. These counts exclude data used to fit a critic and any extra queries for behavior fitting. Consequently, the iteration bounds are oracle bounds, not end-to-end environment-interaction bounds. The sampling budgets, update counts, and query requirements of the additional experiments are described in Appendices D.4 and D.6.

Reusing intermediate transitions. Fix the actor during a rollout and write $z_\theta(s,a) = \nabla_\theta \log \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a)$. Retain the on-policy actions A_t^{env} for $t < T$, and draw a separate terminal action $\widetilde{A}_T \sim \mu(\cdot \mid S_T)$. An unbiased reuse estimator is

$$g_{\mathrm{reuse}} = \sum_{t=0}^{T-1} z_\theta(S_t, A_t^{\mathrm{env}}) + \frac{\pi_\theta(\widetilde{A}_T \mid S_T)}{\mu(\widetilde{A}_T \mid S_T)} z_\theta(S_T, \widetilde{A}_T).$$

Indeed, the prefix and terminal contributions at time t have weights $\mathbb{P}(T > t) = \gamma^{t+1}$ and $\mathbb{P}(T = t) = (1-\gamma)\gamma^t$, whose sum is γ^t. The policy-gradient identity therefore gives $\mathbb{E}[g_{\mathrm{reuse}}] = \nabla J(\theta)$. This estimator needs critic evaluations at the retained transitions. Temporal correlations remain, so unbiasedness alone does not establish a variance improvement. Continuing trajectories are also useful in practice (Appendix C.1), but their unweighted state samples need not follow the discounted occupancy distribution.

Batching. For K conditionally independent actions drawn from the behavior policy at one sampled state, let g_K average their importance-weighted gradients. With $\mathrm{Var}(X) := \mathbb{E}\|X - \mathbb{E}X\|^2$, the conditional-variance identity gives

$$\mathrm{Var}(g_K) = \sigma^2_{\mathrm{state}} + \frac{\sigma^2_{\mathrm{action}}}{K}, \qquad \sigma^2_{\mathrm{state}} = \mathrm{Var}_S(\mathbb{E}[g \mid S]), \quad \sigma^2_{\mathrm{action}} = \mathbb{E}_S[\mathrm{Var}(g \mid S)].$$

This reduces action variance using K local critic queries and one state rollout. Averaging K independent rollout estimators instead divides the full variance by K, at K times the rollout cost.

B.4 The cost of the coverage mixture

Fix the actor, critic, and state. Write $\mathcal{V}_\mu$ for the corresponding second moment and $\bar{\mu} = (1-\varepsilon)\mu + \varepsilon\pi$. Convexity of $x \mapsto 1/x$ gives

$$\mathcal{V}_{\bar{\mu}} \le (1-\varepsilon)\mathcal{V}_\mu + \varepsilon\mathcal{V}_\pi, \qquad \mathcal{V}_\pi - \mathcal{V}_{\bar{\mu}} \ge (1-\varepsilon)(\mathcal{V}_\pi - \mathcal{V}_\mu).$$

The mean gradient does not depend on the behavior policy, so the same statement holds for conditional variance reduction. At $\varepsilon = 0.1$, the mixture retains at least 90% of the reduction achieved by an improving behavior policy while bounding $\pi/\bar{\mu}$ by 10.

B.5 Relaxing the critic accuracy assumption

Uniform action-wise critic accuracy is sufficient but stronger than needed for the actor-bias bound. Conditional on the current actor and critic, define

$$\Delta_i = \hat{Q}_i - Q^{\pi_{\theta_i}}, \qquad \widetilde{\xi}_i^2 = \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_i}},\, a\sim\pi_{\theta_i}} \left[\|\nabla_\theta \log \pi_{\theta_i}(a \mid s)\|^2 \Delta_i(s,a)^2\right].$$

If b_i is the critic-induced mean-gradient error and e_i is the difference between the approximate- and exact-critic estimators on the same state-action sample, then

$$\|b_i\|^2 \le \frac{\widetilde{\xi}_i^2}{(1-\gamma)^2}, \qquad \mathbb{E}\|e_i\|^2 \le \frac{\widetilde{\xi}_i^2}{\varepsilon_{\mathrm{mix}}(1-\gamma)^2}.$$

The first bound follows from Jensen's inequality after importance cancellation; the second follows by changing measure to π_{θ_i} and using $\pi_{\theta_i}/\bar{\mu}_i \le 1/\varepsilon_{\mathrm{mix}}$. Keeping the other actor-analysis conditions, the proof of Theorem 5 therefore replaces B^2 by $N^{-1}\sum_i \mathbb{E}[\widetilde{\xi}_i^2/(1-\gamma)^2]$. This condition weights action errors by their policy probability and score, rather than requiring a uniform bound on every action. It is still an assumption on critic quality, not a guarantee supplied by TD fitting. In particular, controlling the variance of the behavior gradient requires additional moments of the squared critic and behavior score.

C Practical Gradient Estimation for AL-BPO

This appendix presents two practical implementations of AL-BPO: one using a TD critic to estimate action values in discounted MDPs, and the other using group-relative reward advantages in GRPO for LLM post-training without a critic. We specify the corresponding gradient estimators and actor and behavior-policy updates for each setting.

C.1 AL-BPO with a TD critic for discounted MDPs

We keep the actor as the rollout policy and use the behavior policy only for action queries at visited states. We learn an action-value critic $\hat{Q}_\omega$ by one-step TD and alternate critic fitting with actor and behavior updates.

At iteration i, freeze (θ_i, β_i) and copy $\bar{\omega}_i = \omega_i$. For an actor transition $(s, a^{\mathrm{env}}, r, s')$, sample $a' \sim \pi_{\theta_i}(\cdot \mid s')$ if s' is nonterminal and set

$$G_i^{\mathrm{TD}} = \begin{cases} r, & \text{if } s' \text{ is terminal,} \\ r + \gamma \hat{Q}_{\bar{\omega}_i}(s', a'), & \text{otherwise.} \end{cases} \tag{23}$$

A rollout cutoff alone does not count as termination. Store each $(s, a^{\mathrm{env}}, G_i^{\mathrm{TD}})$ in $\mathcal{D}_i$ and fit the critic by minimizing

$$\mathcal{L}_i(\omega) = \frac{1}{|\mathcal{D}_i|} \sum_{(s,a,G)\in\mathcal{D}_i} \big(\hat{Q}_\omega(s,a) - G\big)^2, \qquad \hat{Q}_i = \hat{Q}_{\omega_{i+1}}. \tag{24}$$

Keep the targets fixed and differentiate only $\hat{Q}_\omega(s,a)$. Starting from ω_i, take K_c minibatch gradient steps to obtain ω_{i+1}, then freeze $\hat{Q}_i$ for the actor and behavior updates.

At a visited state s, draw $a^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid s)$ using (6). With $w_i = \pi_{\theta_i}/\bar{\mu}_{\beta_i,\theta_i}$, the local estimates are

$$\hat{g}_i(s,a) = w_i(s,a)\nabla_\theta \log \pi_{\theta_i}(a \mid s)\hat{Q}_i(s,a), \qquad \hat{d}_i(s,a) = -\|\hat{g}_i(s,a)\|^2 \nabla_\beta \log \bar{\mu}_{\beta_i,\theta_i}(a \mid s). \tag{25}$$

The squared gradient norm is held fixed when computing the behavior score. These estimates are averaged over rollout states, rather than scaled as independent discounted-occupancy samples.

C.2 AL-BPO with GRPO updates

For each prompt, generate $M \geq 2$ responses as in Algorithm 4: only one token in the first response uses the behavior mixture; all other tokens use the actor. For the completed group $\mathcal{Y}$, all sampling factors except that token cancel, giving

$$c_i = \frac{\pi_{\theta_i}(a_i^{\mathrm{pg}} \mid s_\tau)}{\bar{\mu}_{\beta_i,\theta_i}(a_i^{\mathrm{pg}} \mid s_\tau)} = \frac{\mathbb{P}_i^{\mathrm{actor}}(\mathcal{Y} \mid x, \tau)}{\mathbb{P}_i^{\text{AL-BPO}}(\mathcal{Y} \mid x, \tau)}. \tag{26}$$

Set $c_i = 1$ if no intervention occurs. Let $\hat{A}_{i,j} = (R_j - \bar{R}_i)/(\sigma_i + \epsilon_A)$ be GRPO's group-normalized reward advantage [Shao et al., 2024], with $\bar{R}_i, \sigma_i$ the group mean and population standard deviation. For valid response lengths L_j (including EOS), define

$$S_i^{\mathrm{std}}(\theta) = \frac{1}{M}\sum_{j=1}^{M} \frac{1}{L_j} \sum_{t=0}^{L_j-1} \Big[\min\{u_{j,t}\hat{A}_{i,j}, \mathrm{clip}(u_{j,t}, 1-\epsilon_c, 1+\epsilon_c)\hat{A}_{i,j}\} - \lambda_{\mathrm{KL}}\kappa_{j,t}\Big], \tag{27}$$

$$u_{j,t}(\theta) = \frac{\pi_\theta(a_{j,t} \mid s_{j,t})}{\pi_{\theta_i}(a_{j,t} \mid s_{j,t})}, \quad \kappa_{j,t} = q_{j,t} - \log q_{j,t} - 1, \quad q_{j,t}(\theta) = \frac{\pi_{\mathrm{ref}}(a_{j,t} \mid s_{j,t})}{\pi_\theta(a_{j,t} \mid s_{j,t})}.$$

The corrected objective is $S_i = c_i S_i^{\mathrm{std}}$, including its KL term. Hold c_i, advantages, lengths, the old actor, and reference fixed during actor differentiation; keep c_i outside clipping. The updates use

$$\hat{g}_i = c_i \nabla_\theta S_i^{\mathrm{std}}(\theta_i), \qquad \hat{d}_i = -\|\hat{g}_i\|^2 \nabla_\beta \log \bar{\mu}_{\beta_i,\theta_i}(a_i^{\mathrm{pg}} \mid s_\tau), \tag{28}$$

with $\hat{d}_i = 0$ if no intervention occurs. Importance correction preserves the expected standard-GRPO surrogate gradient. Differentiating its second moment gives $\hat{d}_i$: the sampling density contributes one

Algorithm 3 AL-BPO for MDPs with a TD critic and continuing actor rollouts

Require: Initial actor θ_1, behavior β_1, critic ω_1; $\rho, N, B_c, B_g, K_c, \gamma, \varepsilon_{\mathrm{mix}}$; step sizes $\eta, \eta_\beta, \eta_\omega$.

Initialize the rollout state $s \sim \rho$; reset $s \sim \rho$ at episode boundaries throughout the algorithm.
for $i = 1, 2, \cdots, N$ **do**
  Phase 1 (Critic data collection): Freeze (θ_i, β_i), copy $\bar{\omega}_i = \omega_i$, and initialize $\mathcal{D}_i = \varnothing$.
  for $j = 1, 2, \cdots, B_c$ **do**
    Execute $a^{\mathrm{env}} \sim \pi_{\theta_i}(\cdot \mid s)$; observe reward r and $s' \sim \mathbb{P}(\cdot \mid s, a^{\mathrm{env}})$.
    If s' is nonterminal, sample $a' \sim \pi_{\theta_i}(\cdot \mid s')$ for the TD target.
    Compute $G_{i,j}^{\mathrm{TD}}$ by (23); append $(s, a^{\mathrm{env}}, G_{i,j}^{\mathrm{TD}})$ to $\mathcal{D}_i$.
    Advance the rollout to $s \leftarrow s'$ (or reset at an episode boundary).
  end for
  Phase 2 (TD critic update): Initialize $\omega_{i,1} = \omega_i$; keep $\bar{\omega}_i$ and all stored TD targets fixed.
  for $k = 1, 2, \cdots, K_c$ **do**
    Sample a nonempty minibatch $\mathcal{B}_{i,k}$ uniformly from $\mathcal{D}_i$; define $\mathcal{L}_{i,k}$ by (24) with $\mathcal{D}_i$ replaced by $\mathcal{B}_{i,k}$.
    Update

$$\omega_{i,k+1} = \omega_{i,k} - \eta_\omega \cdot \nabla_\omega \mathcal{L}_{i,k}(\omega_{i,k}).$$

  end for
  Set $\omega_{i+1} = \omega_{i,K_c+1}$ and freeze $\hat{Q}_i = \hat{Q}_{\omega_{i+1}}$.
  Phase 3 (On-policy state collection): Continue the rollout under the frozen actor π_{θ_i}.
  for $j = 1, 2, \cdots, B_g$ **do**
    Store the current state as $s_j = s$; execute $a^{\mathrm{env}} \sim \pi_{\theta_i}(\cdot \mid s_j)$.
    Observe $s' \sim \mathbb{P}(\cdot \mid s_j, a^{\mathrm{env}})$ and set $s \leftarrow s'$ (or reset at an episode boundary).
  end for
  Phase 4 (Behavior action queries):
  for $j = 1, \cdots, B_g$ **do**
    Draw $a_j^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid s_j)$ without advancing the rollout.
    Query $\hat{Q}_i(s_j, a_j^{\mathrm{pg}})$ and compute $\hat{g}_{i,j}, \hat{d}_{i,j}$ by (25).
  end for
  Phase 5 (Actor and behavior updates):
  Update

$$\theta_{i+1} = \theta_i + \eta \cdot \frac{1}{B_g} \sum_{j=1}^{B_g} \hat{g}_{i,j}, \qquad \beta_{i+1} = \beta_i - \eta_\beta \cdot \frac{1}{B_g} \sum_{j=1}^{B_g} \hat{d}_{i,j}.$$

end for
return θ_{N+1}

behavior score, and the squared weight contributes minus twice that score. All valid tokens enter the loss, so there is no factor H. Note that setting $\varepsilon_{\mathrm{mix}} = 1$ just recovers the standard GRPO.

Algorithm 4 AL-BPO with GRPO: one intervention per response group

Require: Initial SFT actor θ_1, behavior β_1, fixed π_{ref}; $\rho, H, N, M, \varepsilon_{\mathrm{mix}}, \epsilon_A, \epsilon_c, \lambda_{\mathrm{KL}}$; step sizes η, η_β.

for $i = 1, 2, \cdots, N$ **do**
  Freeze (θ_i, β_i) during Phases 1–4; keep π_{ref} fixed throughout training.
  Phase 1 (State sampling): Draw a prompt $x \sim \rho$ and independent $\tau \sim \mathrm{Uniform}\{0, \cdots, H-1\}$.
  Generate the first τ tokens under π_{θ_i} to obtain s_τ; stop at EOS and treat subsequent states as absorbing.
  Phase 2 (AL-BPO action sampling and response completion):
  if s_τ is absorbing **then**
    Retain the completed actor response as $y^{(1)}$; set $c_i = 1$.
  else
    Draw $a_i^{\mathrm{pg}} \sim \bar{\mu}_{\beta_i,\theta_i}(\cdot \mid s_\tau)$ using (6) and append it to the prefix.
    Complete $y^{(1)}$ under π_{θ_i} until EOS or total response length H; compute c_i by (26).
  end if
  Generate $y^{(2)}, \cdots, y^{(M)}$ independently from the same prompt x under π_{θ_i}, each until EOS or H tokens.
  Phase 3 (Group objective): Evaluate $R_j = R(x, y^{(j)})$ for each $j = 1, \cdots, M$; record valid lengths L_j and token pairs $(s_{j,t}, a_{j,t})$.
  Compute the group-normalized advantages $\hat{A}_{i,j}$ from the rewards as defined above.
  Form the surrogate in (27) over all valid response tokens; multiply the entire objective, including the KL term, by c_i.
  Phase 4 (Gradient estimation): Compute $\hat{g}_i$ by (28), holding c_i, advantages, lengths, the old actor, and reference fixed.
  if no intervention occurred or $\varepsilon_{\mathrm{mix}} = 1$ **then**
    Set $\hat{d}_i = 0$.
  else
    Compute $\hat{d}_i$ by (28), holding the whole-group norm $\|\hat{g}_i\|^2$ fixed.
  end if
  Phase 5 (Actor and behavior updates):
  Update

$$\theta_{i+1} = \theta_i + \eta \cdot \hat{g}_i, \qquad \beta_{i+1} = \beta_i - \eta_\beta \cdot \hat{d}_i.$$

end for
return θ_{N+1}

D Experiment Details for Synthetic and Continuous-Control Benchmarks

All experiments for synthetic environments were run on a local laptop with an Apple M2 chip and 8GB of memory. Each experiment took less than 10 minutes to complete.

D.1 Environments and closed-form gradients

Contextual bandit ($\gamma = 0$). State $s \sim \mathcal{N}(0, I_d)$ with $d = 4$, action $a \in \mathbb{R}$, and reward $r(s, a) = -\frac{1}{4}(a - w^\top s)^4$ with $w \sim \mathcal{N}(0, I_d/d)$ drawn at the beginning of the experiment and held fixed across all random seeds. The optimal policy is $\theta^\star = (0, w)$ with $J^\star = -\frac{3}{4}$.[3] Under the Gaussian policy $\pi_\theta(a \mid s) = \mathcal{N}(\theta_0 + \theta_1^\top s, 1)$, $u := a - w^\top s$ is Gaussian with mean θ_0 and variance $\|\theta_1 - w\|^2 + 1$, giving

$$J(\theta) = -\frac{1}{4}(\mu_u^4 + 6\mu_u^2\nu_u + 3\nu_u^2), \qquad \mu_u = \theta_0,\ \nu_u = \|\theta_1 - w\|^2 + 1,$$

and a closed-form ∇J can be calculated exactly by the chain rule. We use this exact gradient to evaluate the optimality gap $\|\nabla J(\theta_i)\|^2$ for every 50 iterations.

Multi-step MDP ($\gamma > 0$). State $s_t \in \mathbb{R}^d$ with $d = 4$, $s_0 \sim \mathcal{N}(0, I_d)$, linear-Gaussian dynamics $s_{t+1} = As_t + \varepsilon_t$ with $A \in \mathbb{R}^{d\times d}$ a randomly sampled matrix of spectral radius 0.9 and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_d)$ with $\sigma_\varepsilon = 0.5$. The per-step reward is the same quartic $r(s, a) = -\frac{1}{4}(a - w^\top s)^4$, and the horizon is infinite.

For this MDP, $J(\theta)$ admits a closed form because conditioned on s_t, u_t (defined before) is Gaussian and the discounted state covariance $\Sigma_t = \mathbb{E}[s_t s_t^\top]$ obeys the recursion of $\Sigma_{t+1} = A\Sigma_t A^\top + \sigma_\varepsilon^2 I_d$ and we can conclude:

$$J(\theta) \;=\; -\frac{1}{4}\sum_{t\geq 0}\gamma^t\big(\mu_u^4 + 6\mu_u^2\nu_u^{(t)} + 3(\nu_u^{(t)})^2\big), \qquad \nu_u^{(t)} = (\theta_1 - w)^\top\Sigma_t(\theta_1 - w) + 1.$$

We truncate the total reward summation at $t = 200$ and use this as the optimality-gap reference, identical in role to the bandit case.

D.2 Algorithm and implementation details

All four behaviors **B1**–**B4** share the actor update and the deployed mixture $\bar{\mu}_{\beta,\theta} = (1 - \varepsilon_{\text{mix}}) \cdot \mu_\beta + \varepsilon_{\text{mix}} \cdot \pi_\theta$ from (6) with $\varepsilon_{\text{mix}} = 0.1$. For **B4**, the behavior policy is $\mu_\beta(\cdot \mid s) = \mathcal{N}(\theta^\top s, \exp(2\log\sigma_\beta))$ and only the scalar $\log\sigma_\beta$ is updated, by the single-sample behavior gradient (7).

Critic (for $\gamma > 0$ only). We construct the estimator of the value function through $\hat{V}_\omega(s) = \omega^\top\phi(s)$ on quadratic features $\phi(s) = (1, s, \text{vec}(ss^\top)) \in \mathbb{R}^{1+d+d^2}$, fit (the linear coefficient ω) by ridge regression (with penalty coefficient 10^{-3}) on 300 Monte Carlo on-policy returns, and re-fit for every 400 actor iterations.

Stability guarantees. We clip the actor gradient norm at 200 (for $\gamma = 0$) or 500 (for $\gamma = 0.9$), the behavior gradient $\rho^2 \cdot \psi$ at 10^3, and clamp $\log\sigma_\beta$ to $[-1.5, 1.5]$. Note that the experiments show that AL-BPO does not hit these boundaries in its steady state.

Hyperparameters. Table 2 lists the actor and behavior step sizes used in each setting. The step sizes for $\gamma = 0.9$ are $30\times$ smaller than those for $\gamma = 0$ to compensate the $1/(1 - \gamma) = 10$ factor in the policy gradient estimator. Each experiment is repeated over 20 seeds.

Omitted details in the figures. In Figure 1(a) plot $\|\hat{g}_i\|^2$ as a proxy for the per-step estimator second moment. Because $\hat{g}$ is heavy-tailed (its per-step standard deviation is roughly twice its mean in our settings), a single-sample snapshot every 50 iterations would give a misleadingly noisy trace. We therefore record the running average of $\|\hat{g}\|^2$ over each 50-iteration window. For display, we additionally smooth across seeds with a window of 10 logged points, plot the median and shade the 25–75% percentile band.

[3] We introduce an intercept term as does in standard contextual bandit literature.

Table 2: Hyperparameters for the experiments in Figure 1

Setting	N	η (actor)	η_β (behavior)	ε_{mix}	seeds
Bandit ($\gamma = 0$)	10000	10^{-3}	4×10^{-3}	0.1	20
MDP ($\gamma = 0.9$)	20000	3×10^{-5}	1.2×10^{-4}	0.1	20

D.3 Supplementary experiments

Different reward functions. We extend the bandit setting with reward $r(s, a) = -|a - w^\top s|^p/p$ for $p \in \{2, 4, 6\}$ (so $p = 4$ recovers the quartic setting in Figure 1). Hyperparameters are tuned per p to keep updates well-scaled: actor step sizes $(\eta_2, \eta_4, \eta_6) = (3 \times 10^{-3}, 10^{-3}, 2 \times 10^{-4})$, with $\eta_\beta = 4\eta$. We compare on-policy **(B1)** against AL-BPO **(B4)** over 20 seeds.

Figure 3 shows AL-BPO's advantage, measured as the ratio of the on-policy final $\|\nabla J\|^2$ to the AL-BPO final $\|\nabla J\|^2$, which grows monotonically with p: $3.3\times$ at $p = 2$, $7.9\times$ at $p = 4$, and $10.6\times$ at $p = 6$. The learned σ_β also increases with p (from ≈ 2.1 to ≈ 3.0), consistent with the prediction that heavy-tailed rewards requires wider behavior support to control $\rho^2\psi$.

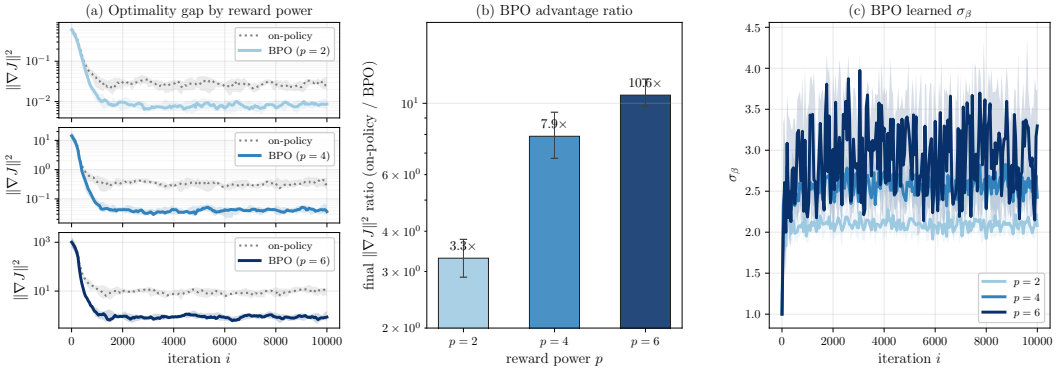


Figure 3: AL-BPO advantage in the contextual bandit setting as a function of reward power $p \in \{2, 4, 6\}$.

Different mixture parameters. Holding all other components of B4 fixed in the Environment (ii) ($\gamma = 0.9$ setting), we vary $\varepsilon_{\text{mix}} \in \{0, 0.05, 0.1, 0.2, 0.5\}$ over 10 seeds. Figure 4 shows the final $\|\nabla J\|^2$ is essentially constant (≈ 11–13) for $\varepsilon_{\text{mix}} \in [0, 0.2]$ and only degrades to ≈ 16 at $\varepsilon_{\text{mix}} = 0.5$. Notably, the $\varepsilon_{\text{mix}} = 0$ case, where the finite-variance proof technically does not apply, did not diverge in any of the 10 seeds. By gradient clipping and BPO's tendency to learn a moderate σ_β provide enough stability in practice. The learned σ_β shifts upward with ε_{mix}, as the on-policy part in the mixture biases the single sample behavior gradient toward larger σ_β.

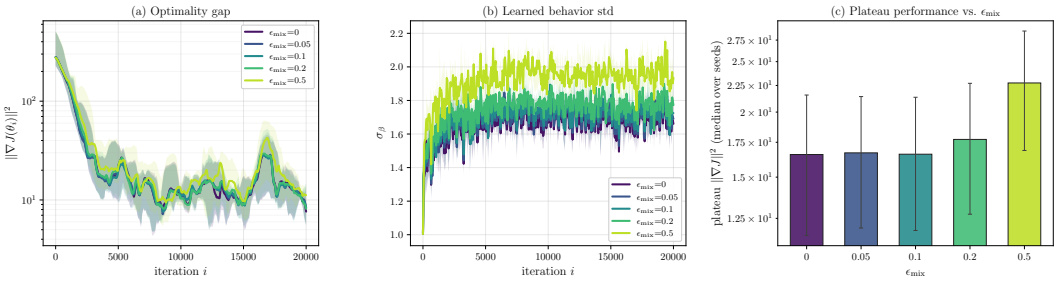


Figure 4: Ablation over the coverage mixture weight $\varepsilon_{\text{mix}} \in \{0, 0.05, 0.1, 0.2, 0.5\}$ for **B4** in Environment (ii) with 10 seeds.

D.4 Comparison with recursive variance reduction

Scope and rates. SVRPG [Papini et al., 2018] and SRVR-PG [Xu et al., 2019] reduce estimation error across updates, whereas AL-BPO optimizes action sampling within an update. For an averaged expected squared gradient norm at most δ, our exact-oracle bound gives $O(\delta^{-2})$ occupancy samples; SRVR-PG gives $O(\delta^{-3/2})$ trajectories under its assumptions. AL-BPO improves the variance constant, not this worst-case exponent. A trajectory costs multiple transitions, while our estimator also requires a local critic query. Different horizons, oracle assumptions, and hidden constants preclude an end-to-end efficiency ranking from these rates alone.

Controlled comparison. We use the quartic, action-independent MDP of Environment (ii), with $d = 4$, $\gamma = 0.9$, and the same local critic for all methods, evaluated from the linear-Gaussian model with a 200-step future sum. Unlike the fitted-critic experiment in Section 6.1, this comparison isolates gradient estimation. Analytic returns and gradients use a 240-step sum. Geometric state rollouts are capped at 256 steps. Hyperparameters were selected using pilot seeds 0–4 and then frozen for 20 seeds 100–119. All methods use actor step size 0.0003 and a budget cap of 20000 state samples. Ordinary updates use batches of 8; recursive methods use a snapshot batch of 128, followed by 16 inner updates with batches of 8. AL-BPO uses behavior step size 0.001 and coverage 0.1. Actor and behavior gradients are clipped at norm 500, the behavior objective's sample contribution at 10000, and the behavior log standard deviation to $[-1.5, 1.5]$.

For the combination, action-independent dynamics make $d_\rho^{\pi_\theta}$ independent of θ. Let $G_\theta^\mu(s, a)$ denote the gradient estimator formed from this local critic with denominator $\mu(a \mid s)$. Between snapshot refreshes, use

$$v_t = v_{t-1} + \frac{1}{b}\sum_{j=1}^{b}\left[G_{\theta_t}^{\bar{\mu}_t}(s_j, a_j) - G_{\theta_{t-1}}^{\bar{\mu}_t}(s_j, a_j)\right], \qquad a_j \sim \bar{\mu}_t(\cdot \mid s_j).$$

Both terms use the current sampling denominator. This is a controlled combination for action-independent transitions, not a general-MDP convergence result.

Table 3: Controlled comparison. Each entry is the median across seeds. Direction error is the squared error of the unclipped update estimator relative to the analytic gradient.

Method	$\|\nabla J\|^2$	$J^\star - J$	Direction error
On-policy	30.20	0.4905	7458.2
AL-BPO	11.96	0.1950	3293.1
SVRPG	34.67	0.5646	508.5
SRVR-PG	37.70	0.6155	395.4
AL-BPO + SRVR-PG	11.70	0.2059	168.5

Relative to SRVR-PG, the combined method has paired median ratios of 0.415 [0.391, 0.478] for direction error and 0.336 [0.266, 0.495] for return gap (95% percentile intervals from 20000 paired-seed bootstrap resamples). These support complementarity in this setting. Its return-gap ratio relative to AL-BPO alone is 1.077 [0.768, 1.321], so a further optimization improvement is not resolved. AL-BPO alone also has larger direction error than the recursive methods; lower error and faster optimization need not rank methods identically.

Within the 20000-state budget cap, on-policy PG and AL-BPO perform 2500 actor updates, while the snapshot and recursive batches allow 1326 updates for SVRPG, SRVR-PG, and AL-BPO + SRVR-PG. Each recursive inner sample requires critic evaluations at both the current and reference policies, whereas an ordinary sample requires one. These analytic critic queries require no additional simulator transitions or critic fitting.

D.5 Coverage, family restriction, and behavior optimization

This experiment examines the gap between the optimal behavior policy and the behavior policy learned by AL-BPO. With the actor fixed, we compare the effects of the coverage mixture, the restriction to a Gaussian behavior-policy family, and behavior optimization on the gradient estimator's second moment.

We fix a Gaussian actor and 256 Gaussian contexts in the four-dimensional quartic bandit, and evaluate the action integrals by 64-point Gauss–Hermite quadrature. The actor has unit standard deviation; the fixed problem uses seed 42, with $w \sim \mathcal{N}(0, I_4)/2$ and actor parameters drawn from $\mathcal{N}(0, 0.25^2 I_5)$. The behavior family is Gaussian, centered at the actor mean, with one shared log standard deviation. Its reference optimum is obtained by a grid search followed by golden-section refinement. Table 4 normalizes the on-policy second moment to one and uses coverage 0.1. The learned behavior policy uses 2000 single-sample updates at step size 0.004 over 20 seeds (0–19), with scalar gradient clipping at 200 and log standard-deviation ratio constrained to $[-2, 2]$.

Table 4: Comparison of behavior policies with a fixed actor. The covered oracle is the mixture of the unconstrained oracle with the actor. The learned entry is the median over seeds.

Behavior policy	Normalized second moment
On-policy	1.000000
Unconstrained oracle	0.089479
Covered oracle	0.094058
Best covered Gaussian	0.265677
Learned covered Gaussian	0.265926

The successive costs are 0.004579 for mixing the oracle, 0.171619 between that reference and the best covered Gaussian, and 0.000249 for the final optimization residual. The learned behavior policy reduces the on-policy second moment by 73.4%, with its interquartile range $[0.265762, 0.266770]$. We can see that family restriction dominates in this example.

D.6 Continuous control on Pendulum-v1

Setup. We further evaluate AL-BPO on Pendulum-v1 with neural actor, value, action-value, and behavior networks, each using two hidden layers of 64 tanh units. The actor is a tanh-squashed Gaussian with a learned log standard deviation. We compare an A2C-style baseline using observed actions and GAE advantages, an on-policy action-query control (OnPolicy-Q), and AL-BPO. All methods collect 100000 online transitions under the actor and perform 1563 actor updates and 7815 critic updates. The query-based methods draw one additional, unexecuted action per collected state for gradient estimation; only its distribution differs between OnPolicy-Q and AL-BPO.

Approximate critic and training. These sampled actions are evaluated by 10 transitions of the public Pendulum dynamics, followed by a learned terminal value. After the queried first action, the rollout uses the actor's mean action. We note that this is a model-rollout approximate critic. The learned value and action-value networks fit GAE-derived bootstrapped returns, with $\gamma = 0.99$, $\lambda = 0.95$, and reward scale 0.1. Rollout batches contain 64 transitions; critic fitting uses five epochs and minibatches of 64. Adam learning rates for actor, critic, and behavior are 0.0003, 0.001, and 0.001. The query-based updates use importance-weighted advantages obtained by subtracting the learned value from the model-rollout estimate. Their centering and scaling statistics are computed from queries at the collected actor actions. The entropy coefficient is 0.001. AL-BPO takes one behavior update per batch using 8 states and 7-node quadrature, with coverage 0.1.

Evaluation and results. We use all ten prespecified seeds 200–209. Evaluation occurs about every 5000 online transitions; the final return averages five deterministic episodes at the final checkpoint. Intervals use 100000 percentile bootstrap resamples of seed statistics. Table 5 reports final returns. AL-BPO improves return over the A2C-style baseline by 120.3 $[61.7, 185.3]$, while the paired difference from OnPolicy-Q is 5.2 $[-23.8, 38.8]$. At common states with the same actor and critic, the conditional second-moment ratio of AL-BPO to on-policy action sampling is 0.747 $[0.720, 0.776]$, a 25.3% reduction. Thus the experiment supports action-level second-moment reduction with neural networks.

Table 5: Pendulum-v1 results over ten seeds. Returns are means with 95% seed-bootstrap intervals. All methods use 100000 online training transitions and 1563 actor updates.

Method	Final return [95% CI]
A2C-style	-1091.2 $[-1122.1, -1060.9]$
OnPolicy-Q	-976.1 $[-1021.2, -933.4]$
AL-BPO	-970.9 $[-1028.1, -913.4]$

E Experiment Details for LLM Post-Training

Task and data. We study one-shot antenna design: a prompt specifies the frequency band, impedance match, and board requirements, and the model returns a complete printed circuit board (PCB) geometry. The prompts cover monopole, patch, MIFA, and meander-line antennas with target frequencies from 1 to 20 GHz. Each prompt is derived from a reference design, with RF criteria reviewed by RF engineers. We use 6185 training examples and 500 validation examples for SFT, and 4560 training examples and 501 validation examples for RL. No reference design is shared across these four splits.

Training reward. Heaviside predicts electromagnetic performance from the generated geometry [Bryant et al., 2026a]. A design receives zero reward if the PCB is invalid, does not match the requested board, or has no resonance reaching -10 dB within 5% of the target frequency. Otherwise, the reward averages clipped piecewise-linear scores for resonance, bandwidth, realized gain, and total efficiency. These predictions supply the training reward, and the validation results below are computed with Palace.

Post-training setup. We fine-tune Qwen3-4B-Instruct with a rank-16 LoRA adapter, containing approximately 33M trainable parameters. SFT runs for 40 epochs. RL starts from the merged epoch-31 SFT checkpoint with a fresh rank-16 adapter and uses GRPO with AL-BPO. Each of the 500 RL updates uses 32 prompts and 16 responses per prompt, giving 512 responses per update and about 3.5 passes through the RL training set. Table 6 lists the training settings.

Table 6: Training settings for the RF antenna design experiment.

Setting	Value
LoRA rank / trainable parameters	16 / approximately 33M
SFT epochs / checkpoint used to initialize RL	40 / epoch 31
RL updates / prompts per update	500 / 32
Responses per prompt / per update	16 / 512
Actor learning rate / sampling temperature	10^{-6} / 1
Maximum response length	4096 tokens
KL penalty coefficient	10^{-3}
Behavior top-k / actor mixture weight	20 / $\varepsilon_{\mathrm{mix}} = 1/2$
Prefix-selection probability threshold	0.99
Behavior updates per batch	1
Training / rollout framework	FSDP / vLLM
Hardware	16× NVIDIA A10G GPUs
Reported training compute	1044 GPU-hours

AL-BPO implementation. Within each group of 16 responses, only one token in one response is sampled from the behavior mixture. Starting from a random position, we choose the first prefix where the actor's largest next-token probability is below 0.99. The behavior policy μ_β is the actor's top-20 distribution at temperature e^β, mixed equally with the actor. All other tokens are sampled from the actor. Since the prefix-selection rule depends only on the actor, the group correction is the single-token ratio in (26), bounded by 2. We apply this weight to the group objective and take one behavior-parameter update per batch using the squared group-gradient norm, as in Algorithm 4.

Evaluation. We compare the base model, the SFT model, and SFT + AL-BPO (the SFT model after GRPO post-training with AL-BPO). Of the 501 held-out RL validation prompts, 470 have

simulation evidence for all three policies; every reported metric uses this common subset. Palace full-wave simulations provide the electromagnetic quantities used to compute validation rewards. Invalid designs receive zero reward. A design counts as successful if it meets the requested RF criteria, except that total efficiency may be up to 15 percentage points below the requested target. We also report the fraction of designs with reward above 0.5. The 95% intervals in Table 1 are obtained by bootstrapping validation prompts which describe uncertainty across prompts. The zero-width interval for base-model success reflects zero observed successes in this sample.

Results. Figure 5 shows three selected validation prompts and the designs generated by each policy, together with their Palace-simulated reflection responses. In the patch and MIFA examples, the post-trained designs reach -10 dB within the target frequency window, whereas the SFT designs do not. These examples illustrate the changes in geometry and resonance. The aggregate results are reported in Table 1 in the main context, which demonstrate the application of AL-BPO to GRPO-based LLM post-training and report the gains of the combined procedure over SFT.

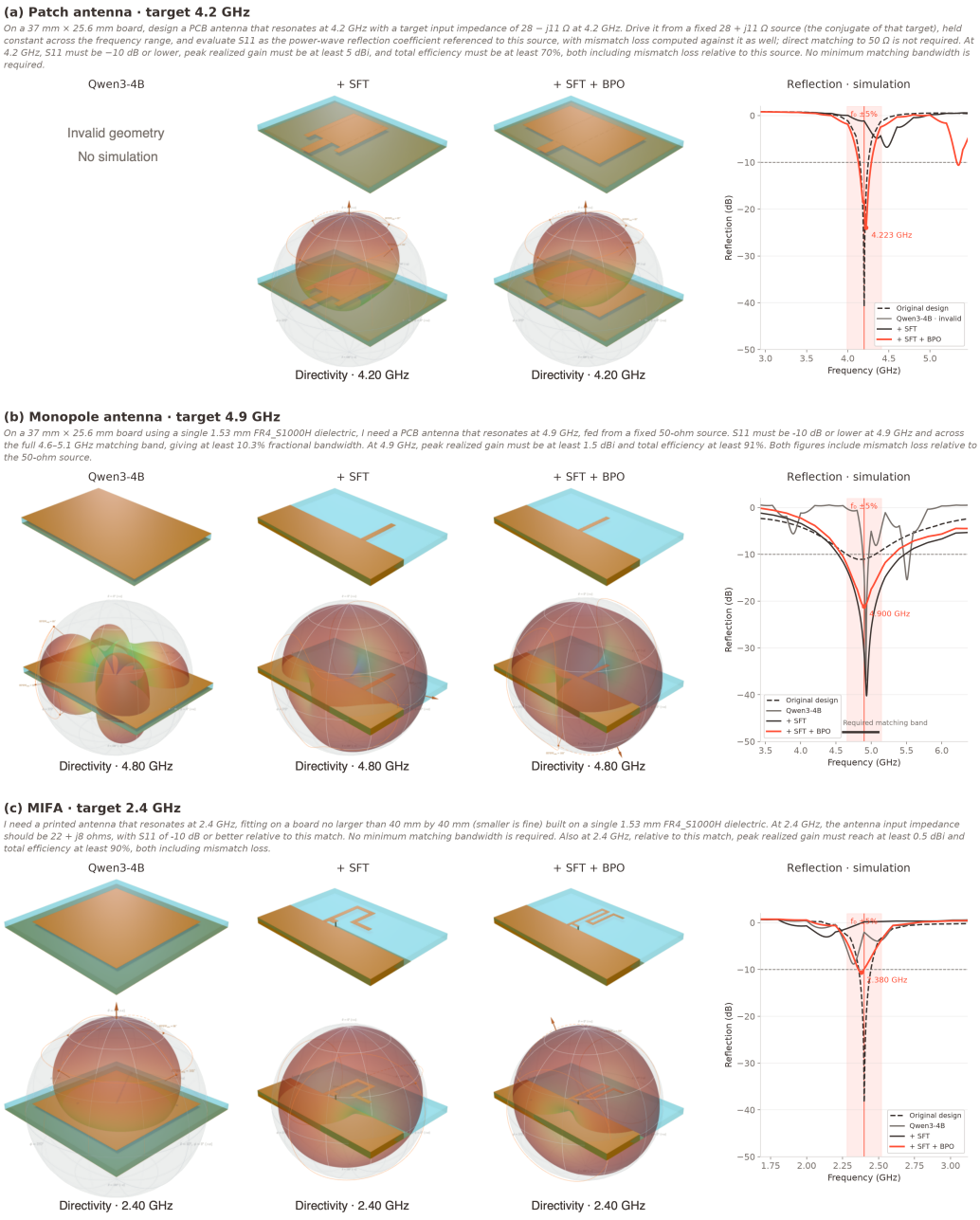


Figure 5: Three selected RF validation examples: (a) patch antenna, (b) monopole antenna, and (c) MIFA. Columns show the base, SFT, and post-trained designs, followed by Palace-simulated reflection curves. Reflection is referenced to the source specified in each prompt; directivity is shown at the frequency labeled below each board. Shaded bands mark $\pm 5\%$ of the target frequency, and the horizontal dashed line marks -10 dB.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions listed in the abstract and Section 1 are each supported by a formal result in Sections 3–5 and validated empirically in Section 6. The LLM case study reports the gains of the combined GRPO + AL-BPO post-training procedure over SFT.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Limitations are discussed in Section 7.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions used are stated in Sections 2 and 5. The proofs are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix D describes the synthetic environments and implementations. Appendix E describes the RF data splits, post-training configuration, and evaluation protocol, with the corresponding algorithms specified in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code for synthetic experiments is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 summarizes the experimental settings. Appendices D and E provide the training and evaluation details for the synthetic and RF experiments, respectively.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The error bars are provided in all the figures of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D reports the hardware and runtime for the synthetic experiments. Appendix E reports the GPU configuration and training compute for the LLM post-training experiment.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The experiments use synthetic environments and RF antenna-design specifications and do not involve research with human subjects or personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This theoretical paper on policy-gradient variance reduction has indirect societal impact on that it can reduce the compute needed for RL training, thus saving costs. Appendix E discusses the potential computational benefits of AL-BPO in LLM post-training.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.