ZEROTH-ORDER MOMENTUM AND ADAPTIVE METHODS FOR NONSMOOTH POLICY OPTIMIZATION FROM HUMAN PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning from human feedback (RLHF) commonly relies on reward models to convert human preferences into optimization signals, but reward inference introduces an additional modeling step and may suffer from reward misspecification. To avoid reward inference, direct policy optimization methods learn policies directly from human preferences without explicitly fitting a reward model, for which zeroth-order optimization provides a natural optimization framework. Existing zeroth-order policy methods such as Zhang & Ying (2025) rely on smooth policy objectives, leaving the nonsmooth setting less understood. Meanwhile, momentum has shown favorable theoretical and empirical properties, motivating its study in preference-based zeroth-order policy optimization. In this work, we study zeroth-order policy optimization from human preferences for general objectives that may be nonsmooth and nonconvex. We establish a refined characterization of the preference oracle induced by pairwise human feedback for reward-model-free RLHF. We further establish a unified convergence framework for momentum-based and adaptive-gradient zeroth-order nonsmooth policy optimization methods, called ZOPGM and ZOPGAM. Both methods achieve a (μ, ϵ)-expected stationarity (a natural relaxation of (δ, ϵ)-Goldstein stationarity) outer-loop complexity of $O(d\sqrt{\mu}\epsilon^{-4})$. This rate matches the parameter dependence obtained by translating the best-known zeroth-order Goldstein-stationarity bound to expected stationarity. Accounting for preference-oracle sampling yields a total human preference query complexity of $O(d^3\sqrt{\mu}\epsilon^{-10})$. Experiments in controlled preference-based policy optimization environments demonstrate consistent improvements in preference-query efficiency over zeroth-order baselines across different preference-feedback budgets, supporting the view that momentum helps exploit information accumulated from historical preference-induced updates.

1 INTRODUCTION

Over the past decade, reinforcement learning (RL) has become a central framework for sequential decision-making because it optimizes long-term behavior through interaction with an environment. Its development has been accelerated by deep neural networks, parallel sampling, self-play, and advances in policy optimization (Mnih et al., 2015; 2016; Espeholt et al., 2018; Silver et al., 2016; Schulman et al., 2017; Haarnoja et al., 2018), leading to notable successes in game playing and robotic control (Mnih et al., 2015; Silver et al., 2016; Levine et al., 2016). Many real-world objectives, however, are difficult to express through explicit numerical rewards. Reinforcement learning from human feedback (RLHF) addresses this difficulty by using human judgments to specify desired behavior and has become an important approach to aligning large language models with human intentions. Its effectiveness has been demonstrated in language generation, summarization, instruction following, and the development of helpful and harmless assistants (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022).

Many RLHF formulations treat generation as a single-step contextual-bandit problem (Shani et al., 2024; Zhang & Ying, 2025). Applications such as multi-turn dialogue and interactive agents instead require a sequential formulation, since early actions can affect future states and the final outcome (Shani et al., 2024). We therefore consider a finite-horizon episodic environment with horizon K,

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

potentially continuous state and action spaces, and a policy π_x parameterized by $x \in \mathbb{R}^d$. Executing π_x induces a trajectory distribution $\mathbb{P}_x$, and each trajectory τ has a latent return $r(\tau)$. Our objective is

$$\max_{x \in \mathbb{R}^d} J(x) := \mathbb{E}_{\tau \sim \mathbb{P}_x}[r(\tau)]. \tag{1}$$

This setting includes contextual bandits as the special case $K = 1$, while allowing stochastic dynamics and long-term effects of earlier actions.

The latent return is generally unobserved, and feedback is instead collected through pairwise trajectory comparisons. Conventional RLHF first learns a surrogate reward model from these comparisons and then optimizes the policy against the inferred reward (Christiano et al., 2017; Ouyang et al., 2022). This additional learning problem may suffer from model misspecification, overfitting, and distribution shift (Zhang & Ying, 2025). Moreover, optimizing an imperfect proxy too aggressively may increase its predicted reward without improving the underlying objective or actual human preference (Gao et al., 2023).

These limitations motivate methods that optimize a policy directly from preferences without learning a global reward model. DPO avoids training an explicit reward model in response-level settings (Rafailov et al., 2023), while episodic reward-inference-free methods have also been studied for sequential decision problems (Zhang et al., 2024; Zhang & Ying, 2025). Beyond specific tabular or closed-form policy classes, it is desirable to formulate preference-based policy optimization directly in terms of general policy parameters. Such a formulation accommodates a broad range of differentiable policy representations and provides a natural interface for zeroth-order optimization: by perturbing the policy parameters and comparing trajectories generated by nearby policies, preference feedback can be used to estimate local policy-value differences (Zhang & Ying, 2025). Randomized zeroth-order estimators can then convert these differences into gradient-free policy updates (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017). However, these works remain largely restricted to smooth settings, leaving a substantial gap between current theory and the nonsmooth objectives that can arise in practical RLHF.

At the same time, recent work has begun to show that momentum and similar mechanisms can preserve favorable complexity guarantees by accumulating information across optimization steps (Ji & Yuan, 2026; Kornowski & Shamir, 2024; Cutkosky et al., 2023). Momentum-based optimizers also remain effective in modern deep learning practice (Chen et al., 2023b; Liang et al., 2026). Under preference feedback, however, their effect is more subtle: temporal averaging can attenuate centered sampling noise, while systematic bias introduced by estimating and inverting preference probabilities may persist. We therefore study ZOPGM and ZOPGAM, with ZOPGAM using a scalar second-moment accumulator based on squared gradient norms.

Reward models explicitly encode historical preference knowledge and learn preference structure, while smoothness helps preserve the relevance of past optimization signals across nearby iterates. In the absence of both, momentum provides an implicit mechanism for reusing preference-induced optimization signals across iterations. Motivated by this intuition, we ask the following question:

> **Can zeroth-order RLHF with nonsmooth policy objectives retain and exploit useful information from past human preferences through momentum, without learning an explicit reward model?**

1.1 Contributions

We answer the question above in the affirmative. Our main contributions are as follows.

- **A sharp characterization of the preference-based zeroth-order oracle.** We separate the systematic bias caused by finite preference estimation from the stochastic variation caused by trajectory and label sampling. The resulting bounds make explicit their dependence on the policy dimension d, perturbation and smoothing radius ν, number of trajectory pairs N, and number of labels per pair M. Moreover, matching constructions show that these bounds are sharp in the leading dimension and sampling exponents for the estimator in our work. A sharp characterization of the preference-based zeroth-order oracle yields a more precise per-iteration preference cost.

- **A unified convergence theory for preference-based zeroth-order momentum.** We develop a common analysis for momentum and adaptive gradient methods driven by biased stochastic estimators. The bound in Theorem 2 separates optimization error, persistent oracle bias, and the stochastic component reduced through momentum averaging, providing a theoretical explanation for momentum, whose temporal aggregation reduces the stochastic component of the oracle error while preserving accumulated optimization signals.
- **Provably efficient momentum for nonsmooth RLHF.** For policy objectives that may be nonsmooth and nonconvex, our methods both attain $O(d\sqrt{\mu}\,\epsilon^{-4})$ outer-iteration complexity for (μ, ϵ)-expected stationarity (See Defintion 2). A concise comparison with related work is provided in Section 6. Moreover, our method attains optimal dimension dependence in the nonsmooth setting without introducing any additional algorithmic mechanisms beyond momentum, unlike the specially designed schemes discussed in Section A. On the analysis side, we avoid the additional dimension dependence induced by the smoothness constant of the smoothed objective. As a result, the dimension dependence in our bounds arises solely from the zeroth-order gradient estimator. Combining this result with the preference-oracle sampling requirements gives $O(d^3\sqrt{\mu}\,\epsilon^{-10})$ human-preference queries. Our experiments further demonstrate that momentum improves preference efficiency and optimization stability across different preference budgets and noise levels. We show that momentum can efficiently aggregate noisy preference-derived zeroth-order updates without an explicit reward model, and establish the corresponding convergence complexity.

Notation. We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ for the Euclidean norm and inner product, respectively. The symbols $\mathbb{E}$, $\mathbb{P}$, and Var denote expectation, probability, and variance. For a differentiable function h, $\nabla h(x)$ denotes its gradient; for a locally Lipschitz function h, $\partial h(x)$ denotes its Clarke subdifferential, the convex hull of limits of gradients at differentiable points approaching x. We write $\mathrm{conv}\, A$ for the convex hull of a set A and $\mathrm{dist}(0, A) = \inf_{a\in A}\|a\|$ for its distance from the origin. The sets $\mathbb{B}^d$ and $\mathbb{S}^{d-1}$ denote the Euclidean unit ball and unit sphere, and $\mathrm{Unif}(A)$ denotes the uniform distribution on A. We write $\mathbf{1}_A$ for the indicator of an event A, $(a)_+ = \max\{a, 0\}$ for the positive part, and $\mathrm{sign}(a)$ for the sign function, with $\mathrm{sign}(0) = 0$. Clipping to an interval is denoted by $\mathrm{clip}_{[a,b]}(p) = \min\{b, \max\{a, p\}\}$. $O(\cdot)$, $\Theta(\cdot)$, and $\widetilde{O}(\cdot)$ denote asymptotic upper bounds, two-sided bounds, and asymptotic upper bounds up to logarithmic factors, respectively.

2 PRELIMINARIES

Preference-based policy optimization. We consider a finite-horizon episodic MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, horizon K, transition kernels $\{P_h\}_{h=0}^{K-1}$, and initial-state distribution ρ_0. A parameter $x \in \mathbb{R}^d$ specifies a possibly time-dependent stochastic policy $\pi_x = \{\pi_{x,h}\}_{h=0}^{K-1}$, where $\pi_{x,h}(\cdot \mid s)$ is the action distribution at state s and step h. Starting from $s_0 \sim \rho_0$, the agent samples $a_h \sim \pi_{x,h}(\cdot \mid s_h)$, and the environment transitions to $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ for $h = 0, \ldots, K-1$. This interaction generates a trajectory $\tau = (s_0, a_0, \ldots, s_{K-1}, a_{K-1}, s_K) \in \mathcal{T}$, where $\mathcal{T}$ is the set of possible trajectories. We write $\mathbb{P}_x$ for its distribution, which includes randomness from the initial state, the policy, and the environment transitions.

Each trajectory has a latent return $r(\tau)$ that measures its overall quality and need not decompose into per-step rewards. The learner can execute policies and collect preference feedback, but does not observe this return. For an initial state s, let $V^{\pi_x}(s) = \mathbb{E}[r(\tau) \mid s_0 = s]$. The policy objective is

$$J(x) = \mathbb{E}_{s_0\sim\rho_0}[V^{\pi_x}(s_0)] = \mathbb{E}_{\tau\sim\mathbb{P}_x}[r(\tau)]. \tag{2}$$

To use the standard minimization convention, we write $F = -J$ and analyze $\min_x F(x)$. With zero initial momentum and the same random samples, changing the sign of the gradient estimates changes the sign of the momentum and leaves the squared-gradient accumulators unchanged. The resulting descent updates on F therefore coincide with ascent updates on J.

Preference feedback. For an ordered pair (τ^+, τ^-), the feedback is a binary label $o \in \{0, 1\}$, with $o = 1$ indicating that the first trajectory is preferred. Following Zhang & Ying (2025), a known link function σ maps the latent return difference to its preference probability:

$$\mathbb{P}(o = 1 \mid \tau^+, \tau^-) = \sigma\big(r(\tau^+) - r(\tau^-)\big). \tag{3}$$

The link specifies how differences in return affect feedback; the observed label is a random outcome from this model. Conditional on the trajectories, labels are independent across repetitions and pairs. Even when (τ^+, τ^-) is fixed, repeated labels can disagree.

From comparisons to policy values. For two policy parameters x^+ and x^-, draw independent trajectories $\tau^+ \sim \mathbb{P}_{x^+}$ and $\tau^- \sim \mathbb{P}_{x^-}$, and write $p = \mathbb{P}(o = 1 \mid \tau^+, \tau^-)$. Inverting the conditional probability recovers the return difference. Averaging over the trajectories then gives

$$\mathbb{E}[\sigma^{-1}(p)] = \mathbb{E}[r(\tau^+) - r(\tau^-)] = J(x^+) - J(x^-). \tag{4}$$

Here p varies with the sampled pair. Inversion occurs before averaging over pairs, so the order of operations matters. Inverting a probability pooled across different pairs generally does not give the same policy-value difference, because σ^{-1} need not be linear.

Smoothing and zeroth-order estimation. Since $F = -J$ may be nonsmooth, we analyze its uniform-ball smoothing using the same radius $\nu > 0$ as in the policy perturbations:

$$f(x) := F_\nu(x) = \mathbb{E}_{v \sim \mathrm{Unif}(\mathbb{B}^d)}[F(x + \nu v)]. \tag{5}$$

The smoothed function remains G-Lipschitz and is continuously differentiable, with a ρ-Lipschitz gradient for $\rho = C\sqrt{d}\,\frac{G}{\nu}$, where $C > 0$ is universal (Yousefian et al., 2012). In particular,

$$|f(x) - F(x)| \le \nu G, \qquad f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle - \frac{\rho}{2}\|y - x\|^2. \tag{6}$$

Thus, the weak convexity used in our analysis follows from smoothing rather than an additional assumption on F. A smaller ν improves the approximation to F but increases the smoothness bound.

The connection between smoothing and the algorithm is the following identity for $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$:

$$\nabla f(x) = \mathbb{E}_U\left[\frac{d}{2\nu}\big(F(x + \nu U) - F(x - \nu U)\big)U\right].$$

Although smoothing averages over a ball, its gradient can therefore be estimated through symmetric perturbations along a sphere-sampled direction (Shamir, 2017). Our preference oracle replaces the unavailable function-value difference with an estimate obtained from trajectory comparisons. Smoothing is only an analytical device: the algorithm requires neither exact policy values nor evaluations of the smoothed function f.

Stationarity. We assess approximate stationarity because the objectives may be nonconvex and nonsmooth. The following definitions apply to a locally Lipschitz function h.

Definition 1 (Goldstein stationarity (Goldstein, 1977)). *For $\delta > 0$, define $\partial_\delta h(x) = \mathrm{conv} \bigcup_{\|y-x\| \le \delta} \partial h(y)$. A point x is (δ, ϵ)-stationary if $\mathrm{dist}(0, \partial_\delta h(x)) \le \epsilon$.*

To analyze momentum, we also use a criterion that controls the average subgradient together with the mean squared distance of the points contributing to that average.

Definition 2 (Expected stationarity). *Let X be a random variable in $\mathbb{R}^d$ whose distribution is not fixed. For a locally Lipschitz function h and $\mu > 0$, define*

$$S_\mu^h(x)^2 := \inf_{X,g} \left\{ \|\mathbb{E}g\|^2 + \mu\mathbb{E}\|X - x\|^2 : g \in \partial h(X) \text{ almost surely} \right\}, \tag{7}$$

where g is a random subgradient at X and the infimum ranges over all joint distributions of (X, g) with $\mathbb{E}\|X - x\|^2 < \infty$ and $\mathbb{E}\|g\| < \infty$. A point is (μ, ϵ)-expected stationary if $S_\mu^h(x) \le \epsilon$.

In (7), X is a random parameter used to assess stationarity around x, and g is an associated subgradient. The term $\|\mathbb{E}g\|^2$ measures how well the subgradients cancel, while $\mu\mathbb{E}\|X - x\|^2$ penalizes using points far from x; increasing μ therefore enforces greater locality. Goldstein stationarity uses a fixed neighborhood of radius δ, whereas expected stationarity controls distance through this penalty. For the differentiable function f, the constraint forces $g = \nabla f(X)$, so the criterion matches the need to control averages of gradients evaluated at the past iterates contributing to momentum.

We prove $\mathbb{E}S_\mu^f(x)^2 \le \epsilon^2$ and transfer this guarantee to $\mathbb{E}\,\mathrm{dist}(0, \partial_\delta F(x)) \le \epsilon$ in Section 4.3. For a detailed discussion of the relationship between Goldstein stationarity and expected stationarity, please refer to Appendix B.

Algorithm 1 Preference-Based ZO-RLHF with Momentum

Require: Initial parameter x_1, iterations T, smoothing radius ν, sample sizes N, M, stepsizes $\{\alpha_t\}_{t=1}^{T}$, and an optimizer with its hyperparameters

1: Initialize the momentum as $m_0 \leftarrow 0$; for ZOPGAM, also set $v_0 \leftarrow 0$ and $\widehat{v}_0 \leftarrow 0$.
2: **for** $t = 1, \dots, T$ **do**
3: Sample $U_t \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and form $x_t^{\pm} \leftarrow x_t \pm \nu U_t$.
4: **for** $n = 1, \dots, N$ **do**
5: Sample independent trajectories $\tau_{t,n}^{\pm} \sim \mathbb{P}_{x_t^{\pm}}$ and query M labels $\{o_{t,n,q}\}_{q=1}^{M}$.
6: Compute the empirical preference frequency: $\widehat{p}_{t,n} \leftarrow M^{-1} \sum_{q=1}^{M} o_{t,n,q}$.
7: **end for**
8: Construct the estimate: $\widehat{g}_t \leftarrow -\frac{d}{2\nu N} \sum_{n=1}^{N} \sigma^{-1}(\mathrm{clip}_{I_\sigma}(\widehat{p}_{t,n}))U_t$.
9: Update the parameter and optimizer states using Algorithm 2 or 3 with $\widehat{g}_t$ and α_t.
10: **end for**
11: Set $\beta_{\mathrm{out}} \leftarrow \beta$ for ZOPGM and $\beta_{\mathrm{out}} \leftarrow \beta_1$ for ZOPGAM.
12: Independently sample $s \in \{1, \dots, T\}$ with $\mathbb{P}(s = t) \propto [1 - (1 - \beta_{\mathrm{out}})^t]^2$.
13: **return** The policy parameter x_s.

Algorithm 2 ZOPGM Step

Require: Parameter x_t, momentum m_{t-1}, estimate $\widehat{g}_t$, stepsize α_t, and momentum parameter β

1: Update the momentum: $m_t \leftarrow (1 - \beta)m_{t-1} + \beta\widehat{g}_t$.
2: Update the policy parameter: $x_{t+1} \leftarrow x_t - \alpha_t m_t$.
3: **return** (x_{t+1}, m_t).

3 Algorithm

Algorithm 1 follows the preference-based estimation framework of Zhang & Ying (2025): trajectory preferences are converted into policy-update directions without fitting an explicit reward model. Their ZPG uses the current estimate to update the policy, while ZBCPG uses block-coordinate perturbation directions. Our methods instead focus on how preference-derived estimates are used across iterations. Symmetric perturbations are formed in line 3, and the resulting estimate is passed to an optimizer that carries information from earlier preference batches into subsequent updates (line 9).

To clarify what these comparisons estimate, consider the ideal direction that would be available if policy values were known:

$$g_t^\star = -\frac{d}{2\nu}\big(J(x_t^+) - J(x_t^-)\big)U_t, \qquad \mathbb{E}[g_t^\star \mid \mathcal{F}_t] = \nabla F_\nu(x_t),$$

where $F = -J$ and $\mathcal{F}_t$ is the history before iteration t. The algorithm does not evaluate J: it samples trajectory pairs and their preference labels in line 5, computes each empirical preference frequency in line 6, and inverts each frequency before averaging in line 8. Thus $g_t^\star$ is the ideal counterpart of the implemented estimate, not an additional algorithm step.

The principal change is in the policy update. ZOPGM combines the current preference-derived estimate with momentum accumulated from previous iterations (Algorithm 2, lines 1–2). ZOPGAM retains this first-moment state and additionally adjusts the overall update scale using a scalar second-moment state (Algorithm 3, lines 1–4).

4 Theoretical framework and complexity

Our analysis proceeds in three stages. We first characterize the preference oracle and separate its persistent inversion bias from centered sampling noise, then we establish convergence under momentum and adaptive updates in the nonsmooth setting, and finally we translate the resulting expected-stationarity guarantees into Goldstein stationarity and total preference-query complexity.

We impose the following conditions on the policy objective, latent returns, and preference model.

Assumption 1. *The policy objective J is G-Lipschitz for some $G > 0$, i.e., $|J(x) - J(x')| \le G\|x - x'\|, \forall x, x' \in \mathbb{R}^d$.*

Algorithm 3 ZOPGAM Step

Require: Parameter x_t, states $(m_{t-1}, v_{t-1}, \widehat{v}_{t-1})$, estimate $\widehat{g}_t$, stepsize α_t, and parameters $\beta_1, \beta_2, \varepsilon_{\mathrm{ad}}$

1: Update the first moment: $m_t \leftarrow (1-\beta_1)m_{t-1} + \beta_1 \widehat{g}_t$.
2: Update the second moment: $v_t \leftarrow (1-\beta_2)v_{t-1} + \beta_2 \|\widehat{g}_t\|^2$.
3: Retain its historical maximum: $\widehat{v}_t \leftarrow \max\{\widehat{v}_{t-1}, v_t\}$.
4: Update the policy parameter: $x_{t+1} \leftarrow x_t - \frac{\alpha_t}{\sqrt{\widehat{v}_t}+\varepsilon_{\mathrm{ad}}} m_t$.
5: **return** $(x_{t+1}, m_t, v_t, \widehat{v}_t)$.

Assumption 2. *For some $H > 0$, $r(\tau) \in [0, H]$ for every $\tau \in \mathcal{T}$.*

Assumption 3. *Let $I_\sigma = [\sigma(-H), \sigma(H)]$. The known function $\sigma : \mathbb{R} \to (0,1)$ restricts to a strictly increasing bijection from $[-H, H]$ onto I_σ, and $|\sigma^{-1}(p) - \sigma^{-1}(q)| \le L_\sigma |p - q|$ for all $p, q \in I_\sigma$.*

The first assumption allows nonsmooth objectives, while the second bounds return differences by H. The third ensures that preference inversion on I_σ is well-conditioned through the Lipschitz constant L_σ, without revealing the latent returns or introducing a reward model.

An example is the Bradley–Terry link $\sigma(z) = (1 + e^{-z})^{-1}$. It satisfies Assumption 3, as does the normalized Weibull link $\sigma(z) = 2^{-2^{-z}}$ used in the experiments. Both links have a continuous, strictly positive derivative on $[-H, H]$, so their inverses are Lipschitz on I_σ.

4.1 Preference-oracle accuracy

We first characterize the accuracy of the preference-based zeroth-order oracle. The result separates the bias induced by finite preference estimation from the stochastic variation due to trajectory and label sampling, and provides the oracle quantities needed for the subsequent convergence analysis.

Theorem 1. *Suppose Assumptions 1–3 hold. For any $x \in \mathbb{R}^d$, $\nu > 0$, and integers $N, M \ge 1$, let $\widehat{g}$ be the estimator in Algorithm 1 evaluated at x. Then*

$$\|\mathbb{E}\widehat{g} - \nabla f(x)\| \le \frac{L_\sigma \sqrt{d}}{8\nu\sqrt{M}}, \qquad \mathbb{E}\|\widehat{g}\|^2 \le C\left(dG^2 + \frac{d^2H^2}{\nu^2 N} + \frac{d^2 L_\sigma^2}{\nu^2 M}\right), \tag{8}$$

where $C > 0$ is universal.

The first bound controls the bias from estimating each pair's preference probability. The second bounds the magnitude of the update and separates trajectory sampling from finite-label error. For the remainder of the paper, let B and Q^2 denote the right-hand sides of the bias and second-moment bounds in (8), with the universal constant chosen sufficiently large. Since these bounds hold uniformly in x, fresh sampling gives $\|\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] - \nabla f(x_t)\| \le B$ and $\mathbb{E}[\|\widehat{g}_t\|^2 \mid \mathcal{F}_t] \le Q^2$, where $\mathcal{F}_t$ is the history before the samples at iteration t are drawn. The momentum theorem below uses these two bounds. The AMSGrad specialization also uses the conditional exponential-moment bound established for the same oracle in Lemma 9. Appendix C proves the bounds. Compared with Zhang & Ying (2025), our bound is more refined; Appendix D provides examples matching this upper bound.

4.2 Momentum and adaptive updates

We next analyze the momentum update for a general estimator satisfying the preceding conditional bounds. This allows the same result to be used for the implemented algorithms. The weak-convexity parameter ρ and initial gap Δ are those of the smoothed objective defined in the preliminaries.

Theorem 2. *Use the update of Algorithm 2 with $m_0 = 0$, a possibly adaptive stepsize γ_t in place of α_t. Suppose $0 < \beta \le 1$, $\beta T \ge 2$, and the stepsizes are almost surely nonincreasing with deterministic bounds $0 < \gamma_{\min} \le \gamma_t \le \gamma_{\max}$ and $\rho\gamma_{\max} \le \frac{1}{2}$. For any $\mu > 0$, the independent output rule in Algorithm 1, with momentum parameter β, gives*

$$\mathbb{E}S_\mu^f(x_s)^2 \le C\left(1 + \frac{\mu\gamma_{\max}^2}{\beta^2}\right)\left(\frac{\Delta}{\gamma_{\min} T} + \frac{\gamma_{\max}}{\gamma_{\min}}(B^2 + \beta Q^2)\right), \tag{9}$$

where $C > 0$ is universal.

The proof appear in Appendix E. For ZOPGM in Algorithm 2, taking a constant stepsize gives $\gamma_{\min} = \gamma_{\max}$. For ZOPGAM in Algorithm 3, a constant base stepsize and the running maximum give a nonincreasing effective stepsize. Similar guarantees can be derived for other varying step-size.

Remark 1. *A notable feature of Theorem 2 is that the optimization bound itself is dimension-free, the ambient dimension enters only through the oracle quantities B and Q^2. In particular, if the zeroth-order oracle were replaced by a first-order oracle whose bias and second moment are independent of d, the same analysis would introduce no dimension dependence. Thus, the dimension cost in our final rate is attributable to zeroth-order gradient estimation rather than to the momentum dynamics.*

This differs from Kornowski & Shamir (2024), where the additional dimension dependence induced by the smoothness of the smoothed objective is avoided through the choice of the outer optimization procedure. Our analysis instead keeps the optimization dynamics dimension-free from the outset, so that the final dimension dependence is inherited directly from the zeroth-order oracle.

Corollary 3 (ZOPGM)**.** *Under Assumptions 1–3, run Algorithm 1 with Algorithm 2, $\alpha_t \equiv \eta = \frac{1}{\mu^{\frac{1}{4}}\sqrt{dT}}$, and $\beta = \frac{\mu^{\frac{1}{4}}}{\sqrt{dT}}$. If $0 < \beta \le 1$, $\beta T \ge 2$, and $\rho\eta \le \frac{1}{2}$, then*

$$\mathbb{E}S_\mu^f(x_s)^2 \le C\left[\frac{\sqrt{d}\,\mu^{\frac{1}{4}}}{\sqrt{T}}\left(\Delta + \frac{Q^2}{d}\right) + B^2\right]. \tag{10}$$

Building on ZOPGM, we consider ZOPGAM, which adaptively rescales preference-derived momentum using a running second-moment estimate, introducing a history-dependent random denominator into the analysis.

Corollary 4 (ZOPGAM)**.** *Under Assumptions 1–3, let $R^2 = (\log 2 + 4)C_0Q^2$, where C_0 is the universal constant in Lemma 9. Run Algorithm 1 with Algorithm 3, $\varepsilon_{\mathrm{ad}} = R$, $\alpha_t \equiv \eta = \frac{1}{\mu^{\frac{1}{4}}\sqrt{T}}$, $\beta_1 = \frac{\mu^{\frac{1}{4}}}{R\sqrt{T}}$, and $\beta_2 = \frac{\log T}{T}$. If $T \ge 2$, $0 < \beta_1 \le 1$, $\beta_1 T \ge 2$, and $\frac{\rho\eta}{R} \le \frac{1}{2}$, then*

$$\mathbb{E}S_\mu^f(x_s)^2 \le C\left[\frac{R(\Delta+1)\mu^{\frac{1}{4}}}{\sqrt{T}} + B^2\right] + G^2T^{-3}. \tag{11}$$

Both proofs appear in Appendix F. To make the bias sufficiently small and keep Q^2 of order dG^2, choose, for $0 < \epsilon \le G$,

$$N = \left\lceil C\frac{dH^2}{\nu^2G^2}\right\rceil, \qquad M = \left\lceil C\frac{dL_\sigma^2}{\nu^2\epsilon^2}\right\rceil. \tag{12}$$

Here C is sufficiently large. These choices give $B^2 \le c\epsilon^2$ for a sufficiently small numerical c and $Q^2 = \Theta(dG^2)$. With fixed problem constants, the leading outer-iteration term in both corollaries is $O(d\sqrt{\mu}\,\epsilon^{-4})$. Notably, both methods achieve a linear dependence on the policy dimension d, matching the best known dimension dependence for zeroth-order methods in nonsmooth and even smooth stochastic optimization. The corresponding leading label count is $O(\frac{d^3\sqrt{\mu}}{\nu^4\epsilon^6})$. If one additionally sets $\nu = \Theta(\epsilon)$, this becomes $O(d^3\sqrt{\mu}\,\epsilon^{-10})$. Benefiting from our refined characterization of the preference-based zeroth-order oracle, the resulting human preference query complexity improves the dimension dependence of Zhang & Ying (2025) by a factor of d^5.

4.3 Goldstein stationarity and total complexity

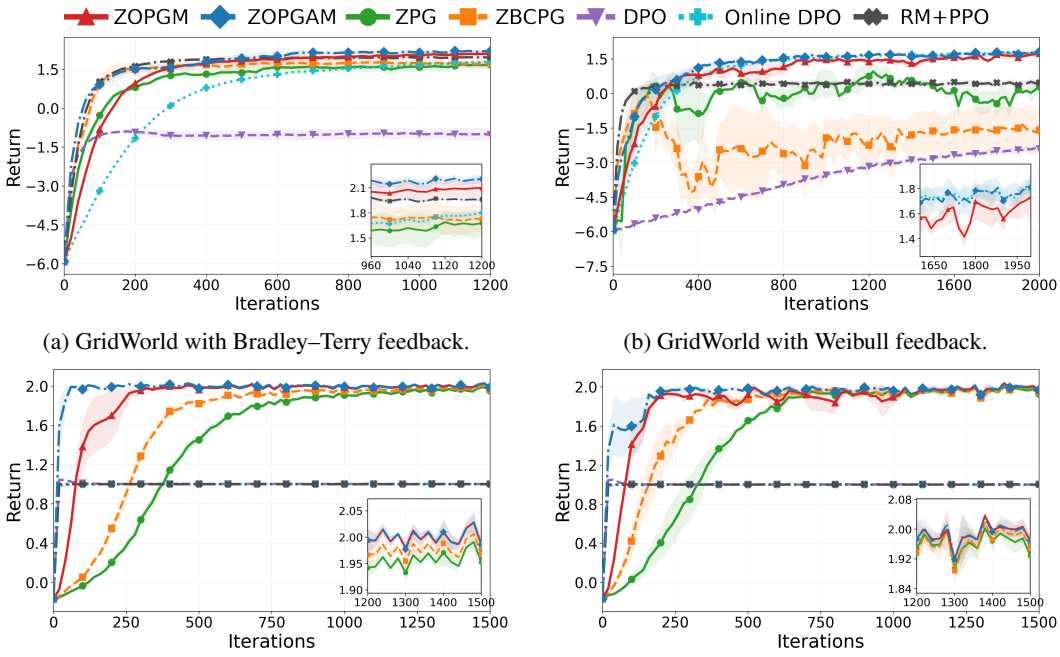
The preceding bounds concern the smoothed objective. We next choose the smoothing radius and distance penalty so that the same output satisfies a guarantee for the original objective F.

Theorem 5. *Suppose Assumptions 1–3 hold. Fix $\delta > 0$ and $0 < \epsilon \le G$, set $\nu = \frac{\delta}{2}$ and $\mu = \frac{4G\epsilon}{\delta^2}$, and use the budgets in* (12)*. Run either method with the parameters of its corollary and the sufficient horizon in* (18)*. Its output satisfies*

$$\mathbb{E}\,\mathrm{dist}(0, \partial_\delta F(x_s)) \le \epsilon. \tag{13}$$

The exact costs are T updates, $2NT$ trajectories, and NMT preference labels. In the small-accuracy regime $0 < \delta \le 1$ and $0 < \epsilon \le 1$, sufficiently large choices have the orders

$$T = O\left(\frac{d}{\delta\epsilon^{\frac{7}{2}}}\right), \qquad 2NT = O\left(\frac{d^2}{\delta^3\epsilon^{\frac{7}{2}}}\right), \qquad NMT = O\left(\frac{d^3}{\delta^5\epsilon^{\frac{11}{2}}}\right), \tag{14}$$



(a) GridWorld with Bradley–Terry feedback.

(b) GridWorld with Weibull feedback.

(c) Risky environment with Bradley–Terry feedback.

(d) Risky environment with Weibull feedback.

Figure 1: Comparison of policy-optimization methods

where fixed problem constants are suppressed.

Appendix G gives the bounds with ceilings and all admissibility terms for general $\delta > 0$.

Remark 2. *Although the resulting Goldstein complexity does not match the best known rate, the gap arises entirely from the conversion from expected stationarity to Goldstein stationarity. In particular, the current optimal Goldstein complexity scales as $O(d\delta^{-1}\epsilon^{-3})$, then our complexity for the expected-stationarity criterion matches the corresponding bound before this conversion. We further emphasize that this guarantee is obtained without introducing additional algorithmic mechanisms beyond momentum. In contrast, sharper Goldstein rates often rely on extra devices such as periodic momentum resets with epoch averaging (Ji & Yuan, 2026), or variance reduction (Chen et al., 2023a). Such mechanisms can also be incorporated into our framework. Importantly, even without such additional mechanisms, our basic method already improves upon the existing complexity guarantees for plain ZO-SGD and ZO-Adam.*

5 Experiments

We first modify the experiment of Zhang & Ying (2025) using nonsmooth policy objectives to examine whether the methods converge to a better policy when an imperfect reward model prevents the agent from recovering the optimal policy. We then consider a more challenging setting in which majority preference is systematically misaligned with expected return, so that both direct-preference and reward-model-based methods may favor a suboptimal policy.

Detailed environment specifications are provided in Appendix I. Briefly, GridWorld is a stochastic 5×5 environment with horizon $K = 20$, randomly generated cell rewards, and a 0.4 probability of reversing the chosen action. Risky Bandit contains one safe action with reward 1, one risky action yielding 10 with probability 0.2 and 0 otherwise, and ten actions with reward -0.5; the risky action has the highest expected return but loses to the safe action in most realized comparisons.

Figure 1 compares our momentum methods with zeroth-order, direct-preference, and reward-model-based baselines across four environment–feedback settings. ZOPGAM achieves the highest area under the return curve in all cases. In GridWorld, where DPO performs poorly under stochastic tran-

Table 1: Comparison of zeroth-order complexities. NS indicates whether nonsmooth objectives are supported. We use a common accuracy ϵ: nonsmooth methods target expected (δ, ϵ)-Goldstein stationarity, with our guarantees transferred to this criterion, while smooth methods target gradient-norm accuracy ϵ. Fixed problem-dependent constants are omitted; $\widetilde{O}$ also hides logarithmic factors.

Algorithm	NS	Outer iterations	Preference cost per iteration	Total preference labels
SGFM (Lin et al., 2022)	✓	$O(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$	–	–
ZO-AdaMM (Chen et al., 2019)	×	$\widetilde{O}(d^2\epsilon^{-4})$	–	–
ZPG/ZBCPG (Zhang & Ying, 2025)	×	$O(d\epsilon^{-2})$	$\widetilde{O}(d^7\epsilon^{-8})$	$\widetilde{O}(d^8\epsilon^{-10})$
ZOPGM (ours)	✓	$O(d\delta^{-1}\epsilon^{-\frac{7}{2}})$	$O(d^2\delta^{-4}\epsilon^{-2})$	$O(d^3\delta^{-5}\epsilon^{-\frac{11}{2}})$
ZOPGAM (ours)	✓	$O(d\delta^{-1}\epsilon^{-\frac{7}{2}})$	$O(d^2\delta^{-4}\epsilon^{-2})$	$O(d^3\delta^{-5}\epsilon^{-\frac{11}{2}})$

sitions, both ZOPGM and ZOPGAM outperform the zeroth-order baselines at convergence, while in Risky Bandit they reach the optimal-return region faster than ZPG and ZBCPG; DPO, Online DPO, and RM+PPO remain near a return of 1 under both preference links.

These observations reflect two complementary mechanisms: momentum aggregates preference-induced zeroth-order signals across iterations, reducing centered trajectory and preference-sampling noise, while inversion before averaging preserves expected-return differences rather than majority preference. This explains why the zeroth-order methods recover the risky action and why the momentum variants reach it faster, yielding consistently better optimization efficiency across the evaluated settings. Additional experiments on Acrobot, as well as ablations on fixed-budget performance and preference-query efficiency, are provided in Appendix I.

6 Comparisons

As summarized in Table 1, our guarantee can be transferred to a (δ, ϵ)-Goldstein stationary point of the original objective in expectation, with an outer-iteration complexity of $O(d\delta^{-1}\epsilon^{-\frac{7}{2}})$. This conversion incurs an additional factor of $\epsilon^{-\frac{1}{2}}$. Under the same oracle accounting, even in the nonsmooth setting, this rate still improves upon the best known results for vanilla zeroth-order SGD (SGFM) and zeroth-order Adam (ZO-AdaMM) as shown in Table 1. Regarding ZPG/ZBCPG, their lower outer-iteration complexity, scaling as $O(d\epsilon^{-2})$, is partly due to their smoothness assumption and stationarity criterion, but also reflects a different allocation of complexity: the outer iterations T control only the zeroth-order optimization error, while the preference-estimation and trajectory-sampling errors are reduced separately through the per-iteration sample sizes M and N. Consequently, a lower dependence on ϵ at the outer-iteration level does not directly translate into a lower overall oracle cost. When the full preference-query complexity TMN is considered, our results exhibit a substantially milder dependence on the policy dimension d.

7 Conclusion

We study momentum-based policy optimization from trajectory preferences without reward inference. Our oracle analysis separates finite-label inversion bias from trajectory and preference-sampling variation, with matching constructions showing sharp leading exponents. Building on these bounds, we establish expected-stationarity and Goldstein guarantees for ZOPGM and ZOPGAM with explicit iteration and query complexities. The analysis shows that momentum aggregates preference-derived update directions and reduces stochastic error, while finite-label bias persists. Since momentum does not explicitly model the structure of past preferences, richer mechanisms for reusing historical preference information may further improve reward-model-free RLHF.

AI USE STATEMENT

Generative AI tools were used at several stages of manuscript preparation. For drafting, they assisted with restructuring selected passages of the motivation, related work, descriptions of theoretical results, and experimental setup. For writing assistance, they were used for language polishing, improving clarity and conciseness, content condensation, and formatting. For technical checking, they were used to identify potential errors in proofs and to review the manuscript for completeness and logical consistency; all mathematical arguments and conclusions were independently checked by the authors. For retrieval and discovery, they were used to suggest potentially relevant related work and references beyond the primary comparison methods, which were subsequently examined and verified by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work does not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

The algorithms, assumptions, parameter choices, and complete proofs are provided in the main text and appendices. The experimental environments, preference-feedback models, baselines, hyperparameter selection, and evaluation protocols are also described in detail in the appendix.

REFERENCES

Gaon An, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song. Direct preference-based policy optimization without reward modeling. *Advances in Neural Information Processing Systems*, 36:70247–70266, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Taha El Bakkali, El Mahdi Chayti, and Omar Saadi. Nonsmooth optimization with zeroth order comparison feedback. *arXiv preprint arXiv:2602.05622*, 2026.

Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, pp. 5219–5233. PMLR, 2023a.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023b.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670. PMLR, 2023.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

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

Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13 (1):14–22, 1977.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Fanfan Ji and Xiao-Tong Yuan. Derandomized online-to-non-convex conversion for stochastic weakly convex optimization. *International Conference on Learning Representations*, 2026.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Kaizhao Liang, Lizhang Chen, Bo Liu, and Qiang Liu. Cautious optimizers: Improving training with one line of code. In *International Conference on Learning Representations*, volume 2026, pp. 106538–106563, 2026.

Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35: 26160–26175, 2022.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623. PMLR, 2016. URL `https://proceedings.mlr.press/v48/martins16.html`.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

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*, pp. 1928–1937. PmLR, 2016.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. Multi-turn reinforcement learning with preference human feedback. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

Qining Zhang and Lei Ying. Zeroth-order policy gradient for reinforcement learning from human feedback without reward inference. In *The Thirteenth International Conference on Learning Representations*, 2025.

Qining Zhang, Honghao Wei, and Lei Ying. Reinforcement learning from human feedback without reward inference: Model-free algorithm and instance-dependent analysis. *arXiv preprint arXiv:2406.07455*, 2024.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Forty-first International Conference on Machine Learning (ICML)*, pp. 58780 – 58799. PMLR, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Related Work

Related work relevant to our setting has developed along three complementary directions. First, preference-based policy optimization has evolved from reward-model-based RLHF toward direct and reward-inference-free approaches. Second, zeroth-order optimization has been extended from smooth objectives to nonsmooth nonconvex settings, with increasing attention to stationarity guarantees and dimension dependence. Third, momentum and adaptive methods have been studied as mechanisms for aggregating noisy optimization signals across iterations, while their role under biased preference-based zeroth-order feedback remains less understood. We review these three lines of work below and position our approach at their intersection.

Policy optimization from preferences. Reward-model-based RLHF learns a reward model before optimizing the policy (Christiano et al., 2017; Ouyang et al., 2022). DPO removes this intermediate stage for response-level preferences (Rafailov et al., 2023), although its standard derivation does not directly extend to general stochastic MDPs. Direct optimization from preferences has also been studied in offline RL (An et al., 2023). For trajectory-level feedback, Zhang & Ying (2025) develop zeroth-order policy updates under a known preference link and establish convergence for smooth policy objectives. Their updates use the current preference estimate, leaving the role of estimates retained across iterations open. More recently, Bakkali et al. (2026) obtain nonsmooth stationarity guarantees from noisy comparisons by recovering objective-value differences with an unbiased, randomized estimator. Their comparison oracle acts directly on objective values, rather than on returns of sampled trajectories.

Zeroth-order nonsmooth optimization. Random-direction, two-point estimators are a basic tool in zeroth-order optimization (Shamir, 2017; Nesterov & Spokoiny, 2017). For nonsmooth nonconvex objectives, Lin et al. (2022) use randomized smoothing to obtain an $O(d^{3/2}\delta^{-1}\epsilon^{-4})$ Goldstein-stationarity guarantee. Kornowski & Shamir (2024) subsequently improve this complexity to $O(d\delta^{-1}\epsilon^{-3})$. However, this rate depends on an algorithm-specific property (see Remark 1) and additional machinery, including an online-to-nonconvex reduction and auxiliary updates.

Momentum and adaptive methods. Alongside progress on zeroth-order oracle complexity, another line of work studies how past estimates can inform later updates. In value-based black-box optimization, ZO-AdaMM combines zeroth-order estimates with adaptive momentum (Chen et al., 2019). For nonsmooth nonconvex objectives, Zhang & Cutkosky (2024) analyze momentum with randomized scaling, while Ji & Yuan (2026) obtain a derandomized result under weak convexity. Together, these results clarify the role of momentum in nonsmooth optimization and provide rigorous convergence guarantees for momentum-based methods. Compared with our work, these analyses use stochastic gradients or function-value estimates. With finite-label preference inversion, momentum carries the estimation bias across iterations even as it averages centered noise. Their guarantees do not account for these two effects occurring together under preference feedback.

B Properties of Expected Stationarity

We first relate the stationarity notions in Definitions 1 and 2, then transfer the guarantees from the smoothed objective $f = F_\nu$ to the original objective F.

Proposition 6. *Let h be G-Lipschitz, and let $\mu, \delta > 0$ and $\eta \geq 0$.*

(a) If x is (δ, η)-Goldstein stationary for h, then it is $(\mu, \sqrt{\eta^2 + \mu\delta^2})$-expected stationary for h.

(b) If x is (μ, η)-expected stationary for h, then it is $(\delta, \eta + \frac{2G\eta^2}{\mu\delta^2})$-Goldstein stationary for h.

Proof. For (a), represent an element of $\partial_\delta h(x)$ with norm at most η as a convex combination of subgradients $g_i \in \partial h(y_i)$ at points $\|y_i - x\| \leq \delta$. Using the same weights to sample $(X, g) = (y_i, g_i)$ gives $\|\mathbb{E}g\| \leq \eta$ and $\mathbb{E}\|X - x\|^2 \leq \delta^2$.

For (b), given $\zeta > 0$, choose a witness (X, g) with $g \in \partial h(X)$ almost surely and $\|\mathbb{E}g\|^2 + \mu\mathbb{E}\|X - x\|^2 \leq (\eta + \zeta)^2$. Replace g outside the radius-δ ball by any $g_0 \in \partial h(x)$. The resulting average

belongs to $\partial_\delta h(x)$. Since subgradients have norm at most G, Markov's inequality gives

$$\operatorname{dist}(0, \partial_\delta h(x)) \le \|\mathbb{E} g\| + 2G\mathbb{P}(\|X - x\| > \delta) \le \eta + \zeta + \frac{2G(\eta+\zeta)^2}{\mu\delta^2}.$$

Let $\zeta \downarrow 0$. □

Choices of μ. For $\eta > 0$, the two directions specialize to

$$\mu = \frac{\eta^2}{\delta^2}: \quad (\delta,\eta)\text{-Goldstein} \implies (\mu, \sqrt{2}\eta)\text{-expected},$$
$$\mu = \frac{2G\eta}{\delta^2}: \quad (\mu,\eta)\text{-expected} \implies (\delta, 2\eta)\text{-Goldstein}.$$

The two implications use different values of μ.

Proposition 7. *Suppose F is G-Lipschitz, $f = F_\nu$ is its ball smoothing, and $\mu, \nu > 0$. If x is (μ,η)-expected stationary for f, then:*

(a) x is $(\mu, \sqrt{\eta^2 + \frac{\mu\nu^2 d}{d+2}})$-expected stationary for F.

(b) If ∇F is L-Lipschitz, then x is $(\mu, \eta + \frac{L\nu d}{d+1})$-expected stationary for F and

$$\|\nabla F(x)\| \le \sqrt{1 + \frac{L^2}{\mu}}\,\eta + \frac{L\nu d}{d+1}.$$

Thus x is also $(\delta, \sqrt{1+\frac{L^2}{\mu}}\,\eta + \frac{L\nu d}{d+1})$-Goldstein stationary for F for every $\delta > 0$. Conversely, if x is (μ,η)-expected stationary for F, it is $(\mu, \eta + \frac{L\nu d}{d+1})$-expected stationary for f.

(c) Without assuming F smooth, for every $\delta > \nu$, x is $(\delta, \eta + \frac{2G\eta^2}{\mu(\delta-\nu)^2})$-Goldstein stationary for F.

Proof. For (a), take an arbitrarily accurate witness X for $S_\mu^f(x)$, draw $V \sim \mathrm{Unif}(\mathbb{B}^d)$ independently, and set $Z = X + \nu V$. A Lipschitz function is differentiable almost everywhere, so $g = \nabla F(Z) \in \partial F(Z)$ almost surely. The ball-smoothing identity gives $\mathbb{E}[g \mid X] = \nabla f(X)$, while $\mathbb{E}V = 0$ and $\mathbb{E}\|V\|^2 = \frac{d}{d+2}$ give $\mathbb{E}\|Z - x\|^2 = \mathbb{E}\|X - x\|^2 + \frac{\nu^2 d}{d+2}$. Hence $S_\mu^F(x)^2 \le S_\mu^f(x)^2 + \frac{\mu\nu^2 d}{d+2}$.

For (b), the same smoothing identity and $\mathbb{E}\|V\| = \frac{d}{d+1}$ imply $\|\nabla f(y) - \nabla F(y)\| \le b_\nu := \frac{L\nu d}{d+1}$ for all y. Comparing the expected-stationarity objectives for the same witness in both directions yields $|S_\mu^F(x) - S_\mu^f(x)| \le b_\nu$. Moreover, for any witness X of $S_\mu^f(x)$, smoothness and Cauchy–Schwarz yield

$$\|\nabla F(x)\| \le \|\mathbb{E}\nabla f(X)\| + L\sqrt{\mathbb{E}\|X - x\|^2} + b_\nu.$$

Applying Cauchy–Schwarz to the first two terms and taking the infimum over witnesses gives the stated bound. Finally, $\nabla F(x) \in \partial_\delta F(x)$.

For (c), apply Proposition 6(b) to f with radius $r = \delta - \nu$. The inclusion $\partial_r f(x) \subseteq \partial_{r+\nu} F(x) = \partial_\delta F(x)$ from Lemma 11 completes the proof. □

Smoothing-to-Goldstein choice of μ. For $\eta > 0$ and $\delta > \nu$, part (c) yields

$$\mu = \frac{2G\eta}{(\delta-\nu)^2}, \quad S_\mu^{F_\nu}(x) \le \eta \quad \implies \quad \operatorname{dist}(0, \partial_\delta F(x)) \le 2\eta.$$

In particular, taking $\nu = \frac{\delta}{2}$ and $\eta = \frac{\epsilon}{2}$ gives $\mu = \frac{4G\epsilon}{\delta^2}$ and a (δ,ϵ)-Goldstein guarantee. Under this choice, part (a) gives only

$$S_\mu^F(x)^2 \le \frac{\epsilon^2}{4} + \frac{G\epsilon d}{d+2},$$

which need not be at most ϵ^2. Thus this choice guarantees Goldstein stationarity for F, not ϵ-level expected stationarity for F. The premise must hold at the stated value of μ.

Corollary 8. *If a random output x_s satisfies $\mathbb{E}S_\mu^f(x_s)^2 \le \eta^2$, then*

$$\mathbb{E}S_\mu^F(x_s)^2 \le \eta^2 + \frac{\mu\nu^2 d}{d+2}.$$

If F has an L-Lipschitz gradient, then $\mathbb{E}\|\nabla F(x_s)\| \le \sqrt{1+\frac{L^2}{\mu}}\,\eta + \frac{L\nu d}{d+1}$. If F is only G-Lipschitz, then for $\delta > \nu$,

$$\mathbb{E}\operatorname{dist}(0, \partial_\delta F(x_s)) \le \eta + \frac{2G\eta^2}{\mu(\delta-\nu)^2}.$$

In particular, $\nu = \frac{\delta}{2}$, $\mu = \frac{4G\epsilon}{\delta^2}$, and $\eta = \frac{\epsilon}{2}$ give $\mathbb{E}\operatorname{dist}(0, \partial_\delta F(x_s)) \le \epsilon$.

Proof. Apply Proposition 7 pointwise, take expectations, and use $\mathbb{E}S_\mu^f(x_s) \le \sqrt{\mathbb{E}S_\mu^f(x_s)^2} \le \eta$. □

C Preference-oracle bounds

C.1 Proof of Theorem 1

Proof. Fix x and condition first on one trajectory pair. Write $Z = r(\tau^+) - r(\tau^-)$, $p = \sigma(Z)$, and $Y = \sigma^{-1}(\operatorname{clip}_{I_\sigma}(\widehat{p}))$, where $M\widehat{p}$ is conditionally binomial with parameters M, p. Here $\widehat{p}$ is the raw empirical frequency, as in Algorithm 1. For a fixed $Z = z$, monotonicity and Lipschitz continuity give

$$\begin{aligned}
|\mathbb{E}[Y \mid Z=z] - z| &= |\mathbb{E}[(Y-z)_+ \mid Z=z] - \mathbb{E}[(z-Y)_+ \mid Z=z]| \\
&\le \max\{\mathbb{E}[(Y-z)_+ \mid Z=z], \mathbb{E}[(z-Y)_+ \mid Z=z]\} \\
&\le L_\sigma \mathbb{E}[(\widehat{p}-p)_+ \mid Z=z] = \frac{L_\sigma}{2}\mathbb{E}[|\widehat{p}-p| \mid Z=z] \\
&\le \frac{L_\sigma}{2}\sqrt{\frac{p(1-p)}{M}} \le \frac{L_\sigma}{4\sqrt{M}}.
\end{aligned}$$

The first equality decomposes the error into positive and negative parts. The first inequality bounds the absolute difference of two nonnegative numbers by their maximum. The second inequality uses the increasing, L_σ-Lipschitz map $\sigma^{-1} \circ \operatorname{clip}_{I_\sigma}$ and $\mathbb{E}(\widehat{p}-p)_+ = \mathbb{E}(p-\widehat{p})_+$, since $\mathbb{E}\widehat{p} = p$. The second equality uses the same identity. The third inequality is Cauchy–Schwarz together with the binomial variance, and the last uses $p(1-p) \le \frac{1}{4}$.

The conditional variance satisfies

$$\begin{aligned}
\operatorname{Var}(Y \mid Z=z) &\le \mathbb{E}[(Y-z)^2 \mid Z=z] \\
&\le L_\sigma^2 \mathbb{E}[(\widehat{p}-p)^2 \mid Z=z] = \frac{L_\sigma^2 p(1-p)}{M} \le \frac{L_\sigma^2}{4M}.
\end{aligned}$$

The first inequality uses the minimizing property of the mean in squared error. The second follows from the Lipschitz property. The equality is the binomial variance formula, and the final inequality again uses $p(1-p) \le \frac{1}{4}$.

For this proof, write $b(z) = \mathbb{E}[Y \mid Z = z] - z$. Conditional on U, the pairs and their label samples are independent and identically distributed. Since the conditional law of Y depends on the trajectories only through Z,

$$\begin{aligned}
\operatorname{Var}(Y \mid U) &= \mathbb{E}[\operatorname{Var}(Y \mid Z) \mid U] + \operatorname{Var}(Z + b(Z) \mid U) \\
&\le \frac{L_\sigma^2}{4M} + 2\operatorname{Var}(Z \mid U) + 2\operatorname{Var}(b(Z) \mid U) \\
&\le 2H^2 + \frac{3L_\sigma^2}{8M}.
\end{aligned}$$

The equality is conditional total variance. The first inequality uses the preceding bound and $\operatorname{Var}(A+B) \le 2\operatorname{Var}(A) + 2\operatorname{Var}(B)$. The second uses $|Z| \le H$ and $|b(Z)| \le \frac{L_\sigma}{4\sqrt{M}}$.

For $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, let $\bar{g}(x;U) := \frac{d}{2\nu}\big(F(x+\nu U) - F(x-\nu U)\big)U$ denote the ideal two-point estimator based on exact objective values.

The conditional return difference satisfies $\mathbb{E}[Z \mid U] = J(x+\nu U) - J(x-\nu U)$, so the ideal estimator from the preliminaries is $\bar{g}(x;U) = -d(2\nu)^{-1}\mathbb{E}[Z \mid U]U$. Consequently,

$$\begin{aligned}\|\mathbb{E}\widehat{g} - \nabla f(x)\| &= \frac{d}{2\nu}\sup_{\|w\|=1}|\mathbb{E}[b(Z)\langle U, w\rangle]| \\ &\le \frac{dL_\sigma}{8\nu\sqrt{M}}\sup_{\|w\|=1}\mathbb{E}|\langle U, w\rangle| \\ &= \frac{dL_\sigma}{8\nu\sqrt{M}}\mathbb{E}|U_1| \le \frac{L_\sigma\sqrt{d}}{8\nu\sqrt{M}}.\end{aligned}$$

The first equality uses the smoothing gradient identity and Euclidean norm duality. The first inequality applies the uniform pair-level bias bound. The second equality follows from rotational symmetry. The final inequality uses $\mathbb{E}|U_1| \le \sqrt{\mathbb{E}U_1^2} = d^{-\frac{1}{2}}$.

Conditional independence also yields

$$\begin{aligned}\mathbb{E}\|\widehat{g}\|^2 &= \frac{d^2}{4\nu^2}\mathbb{E}\left[(\mathbb{E}[Y \mid U])^2 + \frac{1}{N}\operatorname{Var}(Y \mid U)\right] \\ &\le \left(\sqrt{\mathbb{E}\|\bar{g}(x;U)\|^2} + \frac{dL_\sigma}{8\nu\sqrt{M}}\right)^2 + \frac{d^2}{4\nu^2 N}\left(2H^2 + \frac{3L_\sigma^2}{8M}\right) \\ &\le C\left(dG^2 + \frac{d^2L_\sigma^2}{\nu^2 M} + \frac{d^2H^2}{\nu^2 N} + \frac{d^2L_\sigma^2}{\nu^2 NM}\right) \\ &\le C\left(dG^2 + \frac{d^2H^2}{\nu^2 N} + \frac{d^2L_\sigma^2}{\nu^2 M}\right).\end{aligned}$$

The equality is the second-moment identity for a conditional sample mean. The first inequality uses the triangle inequality in mean square, the pair-level bias bound, and the conditional variance estimate. The second uses $\mathbb{E}\|\bar{g}(x;U)\|^2 \le CdG^2$ and $(a+b)^2 \le 2a^2 + 2b^2$. The last absorbs the mixed term using $N \ge 1$. □

C.2 Conditional tail control

Lemma 9 (Conditional exponential moment)**.** *Under the conditions of Theorem 1, with Q^2 chosen as specified after that theorem, a universal constant $C_0 > 0$ satisfies*

$$\mathbb{E}\left[\exp\left(\frac{\|\widehat{g}_t\|^2}{C_0 Q^2}\right)\middle|\mathcal{F}_t\right] \le 2. \tag{15}$$

The bound is uniform in the current iterate and holds for all integers $N, M \ge 1$.

Proof. Condition on $\mathcal{F}_t$ and write $x = x_t$; all calculations below retain this conditioning unless stated otherwise. The new direction is uniform on the sphere and independent of this history. For the ideal estimator, write $D(U) = J(x+\nu U) - J(x-\nu U)$. Then

$$\begin{aligned}D(-U) = -D(U), \qquad \mathbb{E}[D(U) \mid \mathcal{F}_t] = 0,\\ |D(U) - D(V)| \le |J(x+\nu U) - J(x+\nu V)| + |J(x-\nu U) - J(x-\nu V)|\\ \le 2\nu G\|U - V\|.\end{aligned}$$

The first equality exchanges the two perturbed parameters, and the second uses the symmetry of the uniform distribution. The first inequality is the triangle inequality, and the second applies the Lipschitz assumption to both differences.

Concentration of Lipschitz functions on the sphere, in the form stated in the proof of Lemma 9 of Shamir (2017), gives, for a universal $c > 0$,

$$\begin{aligned}\mathbb{P}(\|\bar{g}(x;U)\| > G\sqrt{\frac{du}{c}} \mid \mathcal{F}_t) &= \mathbb{P}(|D(U)| > 2\nu G\sqrt{\frac{u}{cd}} \mid \mathcal{F}_t) \\ &\le 2\exp\left(-\frac{cd}{(2\nu G)^2}\frac{4\nu^2G^2u}{cd}\right) = 2e^{-u}, \qquad u \ge 1.\end{aligned}$$

The first equality uses $\|\bar{g}(x;U)\| = \frac{d|D(U)|}{2\nu}$. The inequality applies sphere concentration with Lipschitz constant $2\nu G$ and mean zero.

Now condition also on U. With Y_n denoting the inverted estimate for pair n, the exact decomposition and the bound on its conditional-mean error are

$$\widehat{g}_t = \bar{g}(x;U) - \frac{d}{2\nu}\big(\mathbb{E}[Y \mid U, \mathcal{F}_t] - D(U)\big)U - \frac{d}{2\nu}\left(\frac{1}{N}\sum_{n=1}^{N} Y_n - \mathbb{E}[Y \mid U, \mathcal{F}_t]\right)U.$$

$$\frac{d}{2\nu}\left|\mathbb{E}[Y \mid U, \mathcal{F}_t] - D(U)\right| = \frac{d}{2\nu}\left|\mathbb{E}[b(Z) \mid U, \mathcal{F}_t]\right| \le \frac{d}{2\nu}\mathbb{E}[|b(Z)| \mid U, \mathcal{F}_t] \le \frac{dL_\sigma}{8\nu\sqrt{M}}.$$

The first equality adds and subtracts the conditional pair mean and the ideal return difference. The second equality uses $\mathbb{E}[Z \mid U, \mathcal{F}_t] = D(U)$. The first inequality is conditional Jensen, and the second uses the pair-level bias bound from the preceding proof.

Since $Y_1, \dots, Y_N$ are independent conditional on $(U, \mathcal{F}_t)$ and take values in $[-H, H]$, Hoeffding's inequality gives

$$\begin{aligned}
&\mathbb{P}\left(\frac{d}{2\nu}\left|\frac{1}{N}\sum_{n=1}^{N} Y_n - \mathbb{E}[Y \mid U, \mathcal{F}_t]\right| > \frac{dH\sqrt{u}}{\nu\sqrt{2N}} \,\middle|\, U, \mathcal{F}_t\right) \\
&\quad = \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} Y_n - \mathbb{E}[Y \mid U, \mathcal{F}_t]\right| > H\sqrt{\frac{2u}{N}} \,\middle|\, U, \mathcal{F}_t\right) \\
&\quad \le 2\exp\left(-\frac{2N}{(2H)^2}\frac{2H^2u}{N}\right) = 2e^{-u}, \qquad u \ge 1.
\end{aligned}$$

The first equality rescales the threshold. The inequality uses the interval length $2H$, and the second equality evaluates the exponent. Averaging over U preserves this bound.

By the definition of Q^2, the scales $G\sqrt{d}$, $\frac{dH}{\nu\sqrt{N}}$, and $\frac{dL_\sigma}{\nu\sqrt{M}}$ are each bounded by a universal multiple of Q. A sufficiently large numerical constant C_1 therefore gives, for $u \ge 1$,

$$\begin{aligned}
\mathbb{P}(\|\widehat{g}_t\| > C_1 Q\sqrt{u} \mid \mathcal{F}_t) &\le \mathbb{P}(\|\bar{g}(x;U)\| > C_1 Q\frac{\sqrt{u}}{3} \mid \mathcal{F}_t) \\
&\quad + \mathbb{P}\left(\frac{d}{2\nu}\left|\frac{1}{N}\sum_{n=1}^{N} Y_n - \mathbb{E}[Y \mid U, \mathcal{F}_t]\right| > C_1 Q\frac{\sqrt{u}}{3} \,\middle|\, \mathcal{F}_t\right) \\
&\le 2e^{-u} + 2e^{-u} = 4e^{-u}.
\end{aligned}$$

The first inequality uses the decomposition and a union bound; its deterministic middle term is at most $C_1 Q\frac{\sqrt{u}}{3}$. The second applies the two tail bounds, and the equality adds their contributions.

Choose a numerical $A \ge \max\{C_1^2, 4e\}$. For $v \ge C_1^2$, substitution of $u = \frac{v}{C_1^2}$ gives $4e^{-\frac{v}{C_1^2}} \le Ae^{-\frac{v}{A}}$. For $0 \le v < C_1^2$, the same bound follows from $Ae^{-\frac{v}{A}} \ge \frac{A}{e} \ge 1$. Thus the squared norm has tail at most $Ae^{-\frac{v}{A}}$ at threshold vQ^2 for every $v \ge 0$, and

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\|\widehat{g}_t\|^2}{C_0Q^2}\right)\middle|\mathcal{F}_t\right] &= 1 + \int_0^\infty e^u \mathbb{P}(\|\widehat{g}_t\|^2 > C_0Q^2u \mid \mathcal{F}_t)\, du \\
&\le 1 + A\int_0^\infty e^{-(\frac{C_0}{A}-1)u}\, du \\
&= 1 + \frac{A^2}{C_0 - A} \le 2.
\end{aligned}$$

The first equality follows by integrating $e^w - 1 = \int_0^w e^u\, du$ and applying conditional Tonelli. The first inequality uses the tail bound just established. The second equality evaluates the integral for $C_0 > A$, and the last inequality holds for $C_0 \ge A + A^2$. Every constant is independent of the iterate, the history, N, and M, as required. □

D MATCHING EXAMPLES FOR THE PREFERENCE ESTIMATOR

The following examples concern the clipped inverse-link estimator in Algorithm 1. They establish worst-case rates over the stated class of links and policy families.

Proposition 10. *For every dimension and radius $\nu > 0$, there is a one-step policy-return model satisfying the standing assumptions with $H = 1$, $L_\sigma = 4$, and constant J, such that at $x = 0$, for all sufficiently large M and every $N \geq 1$,*

$$\|\mathbb{E}\widehat{g} - \nabla f(0)\| \geq c\frac{\sqrt{d}}{\nu\sqrt{M}}, \qquad \mathbb{E}\|\widehat{g}\|^2 \geq c\frac{d^2}{\nu^2 M},$$

where $c > 0$ is universal. As $N \to \infty$, the second moment has order $\frac{d^2}{\nu^2 M}$.

Proof. On $I_\sigma = [\frac{1}{8}, \frac{7}{8}]$, choose the inverse link

$$\sigma^{-1}(p) = 2(p - \frac{1}{2}) + 2\,\mathrm{sign}(p - \frac{1}{2})(|p - \frac{1}{2}| - \frac{1}{4})_+.$$

Its slopes are 4, 2, and 4 on the three intervals separated by $\frac{1}{4}$ and $\frac{3}{4}$, respectively. It is continuous, strictly increasing, and 4-Lipschitz, with endpoint values -1 and 1. Thus its inverse defines σ on $[-1, 1]$. One possible extension is $\sigma(z) = \frac{e^{z+1}}{8}$ for $z < -1$ and $\sigma(z) = 1 - \frac{e^{1-z}}{8}$ for $z > 1$, which preserves strict monotonicity and takes values in $(0, 1)$.

Use a one-step model with three actions having deterministic returns 0, $\frac{1}{2}$, $\frac{3}{4}$. A policy whose first parameter coordinate is nonnegative chooses return $\frac{1}{2}$ deterministically; a policy whose first coordinate is negative chooses return 0 with probability $\frac{1}{3}$ and return $\frac{3}{4}$ with probability $\frac{2}{3}$. Both distributions have mean $\frac{1}{2}$, so J is constant and satisfies the objective assumption for every $G > 0$. In particular, $\nabla f(0) = 0$. At $x = 0$, conditional on $U_1 > 0$, the difference Z equals $\frac{1}{2}$ with probability $\frac{1}{3}$ and $-\frac{1}{4}$ with probability $\frac{2}{3}$; when $U_1 < 0$, its distribution is reflected. The event $U_1 = 0$ has probability zero, including when $d = 1$.

Write $b_M(z) = \mathbb{E}[Y \mid Z = z] - z$ in this proof. Reflection of a binomial frequency with parameter p gives one with parameter $1 - p$, while the clipped inverse changes sign under the same reflection. Therefore $b_M(-z) = -b_M(z)$.

At $z = \frac{1}{2}$, the preference probability is $p = \frac{3}{4}$. On the event $|\widehat{p} - \frac{3}{4}| \leq \frac{1}{16}$, clipping is inactive and the exact local formula is $Y - \frac{1}{2} = 2(\widehat{p} - \frac{3}{4}) + 2(\widehat{p} - \frac{3}{4})_+$. Outside this event, the absolute difference between the two sides is at most 4: indeed, $|Y - \frac{1}{2}| \leq \frac{3}{2}$, $|2(\widehat{p} - \frac{3}{4})| \leq \frac{3}{2}$, and $2(\widehat{p} - \frac{3}{4})_+ \leq \frac{1}{2}$. Consequently,

$$\begin{aligned}
\left| b_M(\frac{1}{2}) - 2\mathbb{E}[(\widehat{p} - \frac{3}{4})_+] \right| &= \left| \mathbb{E}\left[Y - \frac{1}{2} - 2\left(\widehat{p} - \frac{3}{4}\right) - 2\left(\widehat{p} - \frac{3}{4}\right)_+ \right] \right| \\
&\leq 4\mathbb{P}\left(\left| \widehat{p} - \frac{3}{4} \right| > \frac{1}{16} \right) \\
&\leq 8e^{-\frac{M}{128}}, \\
2\mathbb{E}[(\widehat{p} - \frac{3}{4})_+] &= \mathbb{E}|\widehat{p} - \frac{3}{4}|.
\end{aligned}$$

The first equality uses $\mathbb{E}\widehat{p} = \frac{3}{4}$. The first inequality uses the exact local formula and the uniform remainder bound. The second applies Hoeffding with deviation $\frac{1}{16}$, giving exponent $\frac{2M}{16^2} = \frac{M}{128}$. The last equality uses the equal expectations of the positive and negative parts of a centered variable.

At $z = -\frac{1}{4}$, the preference probability is $p = \frac{3}{8}$. On $|\widehat{p} - \frac{3}{8}| \leq \frac{1}{16}$, the inverse is affine and $Y + \frac{1}{4} = 2(\widehat{p} - \frac{3}{8})$. The absolute remainder is at most 3, since $|Y + \frac{1}{4}| \leq \frac{5}{4}$ and $|2(\widehat{p} - \frac{3}{8})| \leq \frac{5}{4}$. Hence

$$\begin{aligned}
|b_M(-\frac{1}{4})| &= \left| \mathbb{E}\left[Y + \frac{1}{4} - 2\left(\widehat{p} - \frac{3}{8}\right) \right] \right| \\
&\leq 3\mathbb{P}\left(\left| \widehat{p} - \frac{3}{8} \right| > \frac{1}{16} \right) \leq 6e^{-\frac{M}{128}}.
\end{aligned}$$

The equality uses the unbiasedness of $\widehat{p}$. The first inequality restricts the bounded remainder to the complement of the affine neighborhood, and the second again applies Hoeffding.

For an explicit lower bound on the absolute binomial error, let $W = \sum_{q=1}^{M}(o_q - \frac{3}{4})$ with independent Bernoulli labels of mean $\frac{3}{4}$. Its moments are

$$\mathbb{E}W^2 = \sum_{q=1}^{M}\mathbb{E}(o_q - \frac{3}{4})^2 = M\left[\frac{3}{4}\left(\frac{1}{4}\right)^2 + \frac{1}{4}\left(\frac{3}{4}\right)^2\right] = \frac{3M}{16},$$

$$\begin{aligned}\mathbb{E}W^4 &= \sum_{q=1}^{M}\mathbb{E}(o_q - \frac{3}{4})^4 + 6\sum_{q<j}\mathbb{E}(o_q - \frac{3}{4})^2\,\mathbb{E}(o_j - \frac{3}{4})^2 \\ &= \frac{21M}{256} + 3M(M-1)\left(\frac{3}{16}\right)^2 = \frac{27M^2 - 6M}{256} \le \frac{27M^2}{256}.\end{aligned}$$

The first equality in each expansion uses independence and centering: any product containing a label to the first power has zero expectation. The remaining equalities evaluate the Bernoulli moments and count the pairs of distinct indices. The inequality drops the negative linear term.

Hölder and Cauchy–Schwarz now give a two-sided estimate without a normal approximation:

$$\begin{aligned}\mathbb{E}W^2 &= \mathbb{E}\big[|W|^{\frac{2}{3}}|W|^{\frac{4}{3}}\big] \le (\mathbb{E}|W|)^{\frac{2}{3}}(\mathbb{E}W^4)^{\frac{1}{3}}, \\ \mathbb{E}|\widehat{p} - \frac{3}{4}| &= \frac{\mathbb{E}|W|}{M} \ge \frac{(\mathbb{E}W^2)^{\frac{3}{2}}}{M(\mathbb{E}W^4)^{\frac{1}{2}}} \ge \frac{1}{4\sqrt{M}}, \\ \mathbb{E}|\widehat{p} - \frac{3}{4}| &\le \sqrt{\mathbb{E}(\widehat{p} - \frac{3}{4})^2} = \frac{\sqrt{3}}{4\sqrt{M}}.\end{aligned}$$

The first equality factors $|W|^2$, and its inequality is Hölder with exponents $\frac{3}{2}$ and 3. The next equality rescales W; the first inequality rearranges Hölder, and the second substitutes the moment bounds. The last inequality is Cauchy–Schwarz, and its equality uses the binomial variance.

Set $a_M = \frac{b_M(\frac{1}{2})}{3} + \frac{2b_M(-\frac{1}{4})}{3}$. Combining the explicit remainder bounds gives

$$\begin{aligned}a_M &\ge \frac{1}{12\sqrt{M}} - \frac{20}{3}e^{-\frac{M}{128}} \ge \frac{1}{24\sqrt{M}}, \\ a_M &\le \frac{\sqrt{3}}{12\sqrt{M}} + \frac{20}{3}e^{-\frac{M}{128}} \le \frac{C}{\sqrt{M}},\end{aligned}$$

for all sufficiently large M. The first inequality on each line substitutes the bounds on the two biases. The second uses $\sqrt{M}e^{-\frac{M}{128}} \to 0$. In particular, a_M is positive and of order $M^{-\frac{1}{2}}$, with universal constants.

For completeness, the angular factor also has a dimension-uniform lower bound. Rotational symmetry and $\sum_{j=1}^{d} U_j^2 = 1$ give $\mathbb{E}U_1^2 = \frac{1}{d}$. For $d \ge 2$, rotation of $\frac{U_1+U_2}{\sqrt{2}}$ to the first coordinate, together with reflection symmetry, gives

$$\begin{aligned}\mathbb{E}U_1^4 &= \frac{1}{4}\mathbb{E}(U_1+U_2)^4 = \frac{1}{2}\mathbb{E}U_1^4 + \frac{3}{2}\mathbb{E}[U_1^2U_2^2], \\ 1 &= \mathbb{E}\left(\sum_{j=1}^{d}U_j^2\right)^2 = d\mathbb{E}U_1^4 + d(d-1)\mathbb{E}[U_1^2U_2^2], \\ \mathbb{E}[U_1^2U_2^2] &= \frac{1}{d(d+2)}, \qquad \mathbb{E}U_1^4 = \frac{3}{d(d+2)}.\end{aligned}$$

The first equality uses rotational invariance. The second expands the fourth power, removes odd cross terms by reflection, and uses identical marginal fourth moments. The next line squares the identity $\sum_j U_j^2 = 1$ and counts its diagonal and off-diagonal terms. The last two equalities solve these two linear equations. When $d = 1$, $\mathbb{E}U_1^4 = 1$ gives the same fourth-moment formula.

Applying the same Hölder interpolation as above to U_1 yields

$$\mathbb{E}|U_1| \geq \frac{(\mathbb{E}U_1^2)^{\frac{3}{2}}}{(\mathbb{E}U_1^4)^{\frac{1}{2}}} = \frac{\sqrt{d+2}}{\sqrt{3}\,d} \geq \frac{1}{\sqrt{3d}}.$$

The first inequality is the preceding interpolation inequality with U_1 in place of W. The equality substitutes the spherical moments, and the second inequality uses $d + 2 \geq d$. These moment identities also hold for $d = 1$, where U_1 takes the values -1 and 1.

Since the true conditional return difference has mean zero, $\mathbb{E}[Y \mid U] = a_M \operatorname{sign}(U_1)$. We therefore obtain

$$\begin{aligned}
\|\mathbb{E}\widehat{g} - \nabla f(0)\| &= \frac{da_M}{2\nu}\,\|\mathbb{E}[\operatorname{sign}(U_1)U]\| \\
&= \frac{da_M}{2\nu}\mathbb{E}|U_1| \geq \frac{\sqrt{d}}{48\sqrt{3}\,\nu\sqrt{M}}, \\
\mathbb{E}\|\widehat{g}\|^2 &= \frac{d^2}{4\nu^2}\left(a_M^2 + \frac{1}{N}\mathbb{E}\operatorname{Var}(Y \mid U)\right) \\
&\geq \frac{d^2 a_M^2}{4\nu^2} \geq \frac{d^2}{2304\nu^2 M}.
\end{aligned}$$

The first equality uses $\nabla f(0) = 0$ and the conditional mean. The second follows by reflecting every coordinate other than the first, which makes the corresponding expectations vanish. Its inequality uses the lower bounds on a_M and $\mathbb{E}|U_1|$. The next equality is the conditional sample-mean identity; the first inequality drops its nonnegative variance term, and the second substitutes $a_M \geq \frac{1}{24\sqrt{M}}$.

Finally, $|Y| \leq 1$ implies $\mathbb{E}\operatorname{Var}(Y \mid U) \leq 1$, so for each such M,

$$\lim_{N\to\infty} \mathbb{E}\|\widehat{g}\|^2 = \frac{d^2 a_M^2}{4\nu^2} = \Theta\left(\frac{d^2}{\nu^2 M}\right).$$

The first equality removes the variance term as N increases, and the second uses the two-sided estimate on a_M. Scaling all returns and the inverse link by a common factor scales H and L_σ by that factor, the bias by that factor, and the second moment by its square. □

The ideal-direction term. The dG^2 term is attained in the exact-label limit by a deterministic-return policy with $J(x) = \operatorname{clip}_{[0,H]}(\frac{H}{2} + Gx_1)$ at $x = 0$ and $\nu \leq \frac{H}{2G}$. In this neighborhood,

$$\begin{aligned}
J(\nu U) - J(-\nu U) = 2\nu G U_1, \quad \bar{g}(0;U) = -dGU_1 U, \\
\mathbb{E}\|\bar{g}(0;U)\|^2 = d^2G^2\mathbb{E}[U_1^2\|U\|^2] = d^2G^2\mathbb{E}U_1^2 = dG^2.
\end{aligned}$$

The first equality uses the affine part of the clipped objective, and the second substitutes this difference into the estimator. The equalities on the last line use its norm, $\|U\| = 1$, and $\mathbb{E}U_1^2 = \frac{1}{d}$, respectively. A one-step deterministic policy can realize this objective by selecting an action whose return is $J(x)$.

The trajectory-sampling term. Let every policy generate return 0 or H with equal probability, independently across trajectories. Then J is constant and the distribution of Z does not depend on U. For each fixed pair, the strong law and continuity of the clipped inverse give $Y \to Z$ almost surely as $M \to \infty$. Moreover, $|Y| \leq H$, so bounded convergence applies to the squared estimator. Thus

$$\begin{aligned}
\operatorname{Var}(Z) = \operatorname{Var}(r(\tau^+)) + \operatorname{Var}(r(\tau^-)) = \frac{H^2}{4} + \frac{H^2}{4} = \frac{H^2}{2}, \\
\lim_{M\to\infty} \mathbb{E}\|\widehat{g}\|^2 = \frac{d^2}{4\nu^2}\mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} Z_n\right)^2 \\
= \frac{d^2}{4\nu^2 N}\operatorname{Var}(Z) = \frac{d^2H^2}{8\nu^2 N}.
\end{aligned}$$

The first equality adds the variances of independent returns, and the next two evaluate them. The first equality on the second line is bounded convergence and $\|U\| = 1$. The next equality uses independent centered pair differences, and the last substitutes their variance.

The mixed label-sampling term. To display the unabsorbed $\frac{d^2L_\sigma^2}{\nu^2 NM}$ term explicitly, take $H = \frac{L_\sigma}{4}$, $I_\sigma = [\frac{1}{4}, \frac{3}{4}]$, and $\sigma^{-1}(p) = L_\sigma(p - \frac{1}{2})$ on I_σ, with a strictly increasing extension of σ outside $[-H, H]$. Let all trajectories have the same deterministic return $\frac{H}{2}$. Then $Z = 0$, $p = \frac{1}{2}$, and

$$Y = L_\sigma\big(\mathrm{clip}_{[\frac{1}{4},\frac{3}{4}]}(\widehat{p}) - \frac{1}{2}\big), \qquad M\widehat{p} \sim \mathrm{Bin}(M, \frac{1}{2}).$$

Reflection about $\frac{1}{2}$ gives $\mathbb{E}Y = 0$ exactly. Clipping agrees with the identity on $|\widehat{p} - \frac{1}{2}| \le \frac{1}{4}$ and reduces absolute deviations everywhere, so

$$\begin{aligned} 0 &\le \frac{L_\sigma^2}{4M} - \mathrm{Var}(Y) \\ &= L_\sigma^2 \mathbb{E}\left[(\widehat{p} - \frac{1}{2})^2 - \big(\mathrm{clip}_{[\frac{1}{4},\frac{3}{4}]}(\widehat{p}) - \frac{1}{2}\big)^2\right] \\ &\le \frac{L_\sigma^2}{4}\mathbb{P}(|\widehat{p} - \frac{1}{2}| > \frac{1}{4}) \le \frac{L_\sigma^2}{2} e^{-\frac{M}{8}}. \end{aligned}$$

The first inequality uses contraction toward $\frac{1}{2}$. The equality uses zero bias and the binomial variance. The second inequality bounds the nonzero difference by $\frac{1}{4}$ on the clipping event, and the third applies Hoeffding. Hence, for all sufficiently large M,

$$\begin{aligned} \mathbb{E}\|\widehat{g}\|^2 &= \frac{d^2}{4\nu^2 N}\,\mathrm{Var}(Y) \\ &= \frac{d^2 L_\sigma^2}{16\nu^2 NM} + O\left(\frac{d^2 L_\sigma^2}{\nu^2 N} e^{-\frac{M}{8}}\right) \\ &= \Theta\left(\frac{d^2 L_\sigma^2}{\nu^2 NM}\right). \end{aligned}$$

The first equality is conditional averaging with zero mean. The second substitutes the variance estimate, and the last uses $Me^{-\frac{M}{8}} \to 0$. The main theorem absorbs this mixed term into $\frac{d^2 L_\sigma^2}{\nu^2 M}$ because $N \ge 1$; the preceding constructions attain each of its three retained terms in the corresponding regime.

E Proof of the common momentum theorem

Proof of Theorem 2. Write $q = 1 - \beta$, set $x_0 = x_1$ and $\gamma_0 = \gamma_{\max}$, and decompose the oracle as $\widehat{g}_t = \nabla f(x_t) + b_t + \xi_t$, where $b_t = \mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] - \nabla f(x_t)$ and $\xi_t = \widehat{g}_t - \mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t]$. Thus $\|b_t\| \le B$, $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$, and $\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \le Q^2$. The noise bound follows from $\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] = \mathbb{E}[\|\widehat{g}_t\|^2 \mid \mathcal{F}_t] - \|\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t]\|^2 \le Q^2$.

Apply weak convexity at x_t, with the preceding iterate as the comparison point:

$$\begin{aligned} f(x_{t-1}) &\ge f(x_t) + \langle \nabla f(x_t), x_{t-1} - x_t\rangle - \frac{\rho}{2}\|x_{t-1} - x_t\|^2 \\ &= f(x_t) + \gamma_{t-1}\langle \nabla f(x_t), m_{t-1}\rangle - \frac{\rho\gamma_{t-1}^2}{2}\|m_{t-1}\|^2, \\ 2\gamma_{t-1}\langle \nabla f(x_t) + b_t, m_{t-1}\rangle &\le 2(f(x_{t-1}) - f(x_t)) + \rho\gamma_{t-1}^2\|m_{t-1}\|^2 \\ &\quad + \frac{\gamma_{t-1}}{2}\|m_{t-1}\|^2 + 2\gamma_{t-1}B^2. \end{aligned}$$

The first inequality is (6). The equality substitutes the preceding update; at $t = 1$ both displacement and momentum are zero. The second inequality rearranges the preceding line and applies $2\langle b_t, m_{t-1}\rangle \le \frac{\|m_{t-1}\|^2}{2} + 2B^2$.

Monotonicity of the stepsizes now gives

$$\begin{aligned}
\mathbb{E}[\gamma_t\|m_t\|^2] &\le \mathbb{E}[\gamma_{t-1}\|qm_{t-1}+\beta\widehat{g}_t\|^2] \\
&= q^2\mathbb{E}[\gamma_{t-1}\|m_{t-1}\|^2] + 2\beta q\mathbb{E}[\gamma_{t-1}\langle\nabla f(x_t)+b_t, m_{t-1}\rangle] \\
&\quad + \beta^2\mathbb{E}[\gamma_{t-1}\|\widehat{g}_t\|^2] \\
&\le q^2\mathbb{E}[\gamma_{t-1}\|m_{t-1}\|^2] + 2\beta q\mathbb{E}[f(x_{t-1})-f(x_t)] \\
&\quad + \beta q\rho\mathbb{E}[\gamma_{t-1}^2\|m_{t-1}\|^2] + \frac{\beta q}{2}\mathbb{E}[\gamma_{t-1}\|m_{t-1}\|^2] \\
&\quad + 2\beta q\gamma_{\max}B^2 + \beta^2\gamma_{\max}Q^2 \\
&\le q\mathbb{E}[\gamma_{t-1}\|m_{t-1}\|^2] + 2\beta q\mathbb{E}[f(x_{t-1})-f(x_t)] \\
&\quad + 2\beta q\gamma_{\max}B^2 + \beta^2\gamma_{\max}Q^2.
\end{aligned}$$

The first inequality uses $\gamma_t \le \gamma_{t-1}$. The equality expands the square and removes the martingale cross term because $\gamma_{t-1}m_{t-1}$ is $\mathcal{F}_t$-measurable. The second inequality applies weak convexity at x_t with comparison point x_{t-1}, uses $x_{t-1}-x_t=\gamma_{t-1}m_{t-1}$, bounds $2\langle b_t, m_{t-1}\rangle$ by $\frac{\|m_{t-1}\|^2}{2}+2B^2$, and applies the conditional second-moment bound. The last inequality uses $\rho\gamma_{\max}\le\frac{1}{2}$ and $q^2+\beta q=q$. In particular, the current stepsize need not be independent of the current oracle.

Summing the recursion gives

$$\begin{aligned}
\beta\gamma_{\min}\sum_{t=1}^{T}\mathbb{E}\|m_t\|^2 &\le \beta\sum_{t=1}^{T}\mathbb{E}[\gamma_t\|m_t\|^2] \\
&\le \sum_{t=1}^{T}\big(\mathbb{E}[\gamma_t\|m_t\|^2] - q\mathbb{E}[\gamma_{t-1}\|m_{t-1}\|^2]\big) \\
&\le 2\beta q\mathbb{E}[f(x_1)-f(x_T)] + 2\beta q\gamma_{\max}TB^2 + \beta^2\gamma_{\max}TQ^2 \\
&\le 2\beta\Delta + 2\beta\gamma_{\max}TB^2 + \beta^2\gamma_{\max}TQ^2.
\end{aligned}$$

The first inequality uses the lower stepsize bound. The second adds the nonnegative terminal term $q\mathbb{E}[\gamma_T\|m_T\|^2]$, using $m_0=0$. The third sums the preceding recursion and telescopes the objective differences. The last uses $q\le 1$ and $f(x_T)\ge\inf f$. Dividing by $\beta\gamma_{\min}T$ yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|m_t\|^2 \le \frac{2\Delta}{\gamma_{\min}T} + 2\frac{\gamma_{\max}}{\gamma_{\min}}B^2 + \frac{\gamma_{\max}}{\gamma_{\min}}\beta Q^2. \tag{16}$$

For this proof, set $a_t = 1-q^t$ and $z_t=\sum_{i=1}^{t}\beta q^{t-i}\xi_i$. For each fixed realization of the optimization history, choose a witness $X=x_i$ with probability $\beta\frac{q^{t-i}}{a_t}$ in Definition 2. These witness probabilities introduce randomness separate from the algorithm's samples. Then

$$\begin{aligned}
a_t^2 S_\mu^f(x_t)^2 &\le \left\|\sum_{i=1}^{t}\beta q^{t-i}\nabla f(x_i)\right\|^2 + \mu a_t\sum_{i=1}^{t}\beta q^{t-i}\|x_i-x_t\|^2 \\
&= \left\|m_t - \sum_{i=1}^{t}\beta q^{t-i}b_i - z_t\right\|^2 + \mu a_t\sum_{i=1}^{t}\beta q^{t-i}\|x_i-x_t\|^2 \\
&\le 3\|m_t\|^2 + 3a_t^2B^2 + 3\|z_t\|^2 + \mu a_t\sum_{i=1}^{t}\beta q^{t-i}\|x_i-x_t\|^2.
\end{aligned}$$

The first inequality evaluates the infimum at this witness. The equality unrolls the momentum recursion. The second inequality applies $\|u+v+w\|^2\le 3(\|u\|^2+\|v\|^2+\|w\|^2)$ and the bound on b_i. When $\beta=1$, the witness puts all its mass at x_t.

For $i<j$, ξ_i is $\mathcal{F}_j$-measurable. The cross terms consequently vanish:

$$\begin{aligned}
\mathbb{E}\langle\xi_i,\xi_j\rangle &= \mathbb{E}\big[\mathbb{E}[\langle\xi_i,\xi_j\rangle \mid \mathcal{F}_j]\big] \\
&= \mathbb{E}\langle\xi_i, \mathbb{E}[\xi_j \mid \mathcal{F}_j]\rangle = 0.
\end{aligned}$$

The first equality is the tower property, the second takes out the measurable factor, and the third uses conditional centering. This argument uses the original history, before any output event is imposed.

The noise and distance terms therefore satisfy

$$\mathbb{E}\|z_t\|^2 = \sum_{i=1}^{t} \beta^2 q^{2(t-i)} \mathbb{E}\|\xi_i\|^2 \le \beta^2 Q^2 \sum_{k=0}^{\infty} q^{2k} = \frac{\beta Q^2}{2-\beta} \le \beta Q^2,$$

$$\|x_i - x_t\|^2 = \left\| \sum_{j=i}^{t-1} \gamma_j m_j \right\|^2 \le (t-i) \sum_{j=i}^{t-1} \gamma_j^2 \|m_j\|^2 \le \gamma_{\max}^2 (t-i) \sum_{j=i}^{t-1} \|m_j\|^2.$$

On the first line, the first equality uses martingale orthogonality with deterministic coefficients, the first inequality uses the noise second moment and extends the sum, the second equality evaluates a geometric series, and the last inequality uses $\beta \le 1$. On the second line, the equality telescopes the updates, the first inequality is Cauchy–Schwarz, and the last uses the upper stepsize bound.

To sum the weighted distances, fix an increment index j and an interval length $k = t - i$. The conditions $i \le j < t = i + k$ imply $j - k + 1 \le i \le j$, so at most k intervals of length k contain that increment. Hence,

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \sum_{i=1}^{t} \beta q^{t-i} \|x_i - x_t\|^2 \\
&\quad \le \gamma_{\max}^2 \sum_{t=1}^{T} \sum_{i=1}^{t-1} \beta q^{t-i} (t-i) \sum_{j=i}^{t-1} \|m_j\|^2 \\
&\quad = \gamma_{\max}^2 \sum_{j=1}^{T-1} \|m_j\|^2 \sum_{1 \le i \le j < t \le T} \beta q^{t-i} (t-i) \\
&\quad \le \gamma_{\max}^2 \sum_{j=1}^{T-1} \|m_j\|^2 \sum_{k=1}^{\infty} \beta q^k k^2 \\
&\quad = \frac{\gamma_{\max}^2 q(1+q)}{\beta^2} \sum_{j=1}^{T-1} \|m_j\|^2 \le \frac{2\gamma_{\max}^2}{\beta^2} \sum_{j=1}^{T} \|m_j\|^2.
\end{aligned}$$

The first inequality uses $a_t \le 1$ and the distance bound. The first equality interchanges finite sums. The second inequality groups intervals by length, applies the preceding count, and extends a nonnegative series. The second equality uses $\sum_{k \ge 1} k^2 q^k = \frac{q(1+q)}{(1-q)^3}$ and $1 - q = \beta$. The final inequality uses $q(1+q) \le 2$ and adds a nonnegative terminal term. When $\beta = 1$, all weighted distances are zero, so the bound also covers that case.

The output normalization satisfies

$$\sum_{t=1}^{T} a_t^2 \ge \sum_{t=\lceil \frac{T}{2} \rceil}^{T} (1 - e^{-\beta t})^2 \ge \frac{T}{2} (1 - e^{-1})^2 \ge \frac{T}{8}.$$

The first inequality drops nonnegative terms and uses $(1-\beta)^t \le e^{-\beta t}$. The second uses $\beta T \ge 2$ and at least $\frac{T}{2}$ indices with $t \ge \frac{T}{2}$. The last uses $1 - e^{-1} \ge \frac{1}{2}$. Combining the witness, noise, and distance bounds gives

$$\begin{aligned}
\mathbb{E} S_\mu^f(x_s)^2 &= \frac{\sum_{t=1}^{T} a_t^2 \mathbb{E} S_\mu^f(x_t)^2}{\sum_{t=1}^{T} a_t^2} \\
&\le \frac{24 + \frac{16\mu\gamma_{\max}^2}{\beta^2}}{T} \sum_{t=1}^{T} \mathbb{E}\|m_t\|^2 + 24B^2 + 24\beta Q^2 \\
&\le C \left(1 + \frac{\mu\gamma_{\max}^2}{\beta^2}\right) \left(\frac{\Delta}{\gamma_{\min} T} + \frac{\gamma_{\max}}{\gamma_{\min}} (B^2 + \beta Q^2) \right).
\end{aligned}$$

The equality averages over the independent output index. The first inequality combines the preceding estimates and normalization. The second substitutes (16) and uses $\frac{\gamma_{\max}}{\gamma_{\min}} \ge 1$. □

F PROOFS FOR THE TWO OPTIMIZERS

F.1 PROOF OF COROLLARY 3

Proof. Take $\gamma_{\min} = \gamma_{\max} = \eta$ in Theorem 2. Then

$$\begin{aligned}\mathbb{E}S_\mu^f(x_s)^2 &\le C\left(1+\frac{\mu\eta^2}{\beta^2}\right)\left(\frac{\Delta}{\eta T}+B^2+\beta Q^2\right)\\ &= 2C\left(\frac{\sqrt{d}\,\mu^{\frac{1}{4}}\Delta}{\sqrt{T}}+B^2+\frac{\mu^{\frac{1}{4}}Q^2}{\sqrt{dT}}\right).\end{aligned}$$

The inequality is the common momentum bound. The equality substitutes the stepsize and momentum parameter, for which $\frac{\mu\eta^2}{\beta^2} = 1$. Absorbing the factor 2 into the universal constant proves (10). □

F.2 PROOF OF COROLLARY 4

Proof. Consider a virtual run, denoted by stars, with the same oracle and moment updates as Algorithm 3, but with denominator $\sqrt{\min\{\widehat{v}_t^*, R^2\}}+R$. This modification is used only in the proof. The running maximum is nondecreasing, and

$$0 \le \min\{\widehat{v}_t^*, R^2\} \le R^2, \qquad \frac{\eta}{2R} \le \frac{\eta}{\sqrt{\min\{\widehat{v}_t^*, R^2\}}+R} \le \frac{\eta}{R}.$$

The first bounds follow from nonnegativity and the cap. The stepsize bounds follow by adding R to a square root in $[0, R]$ and taking reciprocals. Consequently, the virtual effective stepsizes are nonincreasing and satisfy the hypotheses of Theorem 2 with $\gamma_{\min} = \frac{\eta}{2R}$ and $\gamma_{\max} = \frac{\eta}{R}$.

Let $\mathcal{F}_i^*$ denote the history of this virtual run before its ith oracle call. The oracle bounds hold with this history because they are uniform in the current parameter and each call uses fresh samples. For $0 \le a \le 1$, conditional Jensen and Lemma 9 yield

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{a\|\widehat{g}_i^*\|^2}{C_0Q^2}\right)\middle|\mathcal{F}_i^*\right] &= \mathbb{E}\left[\left(\exp\left(\frac{\|\widehat{g}_i^*\|^2}{C_0Q^2}\right)\right)^a\middle|\mathcal{F}_i^*\right]\\ &\le \left(\mathbb{E}\left[\exp\left(\frac{\|\widehat{g}_i^*\|^2}{C_0Q^2}\right)\middle|\mathcal{F}_i^*\right]\right)^a \le 2^a.\end{aligned}$$

The equality rewrites the exponential as a power. The first inequality uses concavity of $z \mapsto z^a$, and the second applies the conditional exponential-moment bound. At $a = 0$, both sides of the first inequality equal one.

Since $T \ge 2$, $0 < \beta_2 = \frac{\log T}{T} < 1$. Unrolling the scalar recursion and conditioning backwards gives

$$\begin{aligned}\mathbb{E}\exp\left(\frac{v_t^*}{\beta_2 C_0Q^2}\right) &= \mathbb{E}\prod_{i=1}^{t}\exp\left(\frac{(1-\beta_2)^{t-i}\|\widehat{g}_i^*\|^2}{C_0Q^2}\right)\\ &= \mathbb{E}\left[\prod_{i=1}^{t-1}\exp\left(\frac{(1-\beta_2)^{t-i}\|\widehat{g}_i^*\|^2}{C_0Q^2}\right)\mathbb{E}\left[\exp\left(\frac{\|\widehat{g}_t^*\|^2}{C_0Q^2}\right)\middle|\mathcal{F}_t^*\right]\right]\\ &\le 2\,\mathbb{E}\prod_{i=1}^{t-1}\exp\left(\frac{(1-\beta_2)^{t-i}\|\widehat{g}_i^*\|^2}{C_0Q^2}\right)\\ &\le 2^{\sum_{i=1}^{t}(1-\beta_2)^{t-i}}\\ &= \exp\left((\log 2)\frac{1-(1-\beta_2)^t}{\beta_2}\right) \le \exp\left(\frac{\log 2}{\beta_2}\right).\end{aligned}$$

The first equality uses $v_0^* = 0$. The second applies the tower property, since every preceding factor is $\mathcal{F}_t^*$-measurable. The first inequality applies the preceding conditional bound with exponent one. The second repeats this conditioning at $t-1, \dots, 1$, using exponents $(1-\beta_2)^{t-i}$; no independence

of the successive iterates or estimates is used. The third equality evaluates the geometric sum, and the final inequality bounds its numerator by one.

The cap-crossing probability therefore satisfies

$$\begin{aligned}\mathbb{P}(\widehat{v}_T^* > R^2) &= \mathbb{P}\left(\max_{t\le T} v_t^* > R^2\right)\\ &\le \sum_{t=1}^{T} \exp\left(-\frac{R^2}{\beta_2 C_0 Q^2}\right)\mathbb{E}\exp\left(\frac{v_t^*}{\beta_2 C_0 Q^2}\right)\\ &\le T\exp\left(-\frac{\log 2+4}{\beta_2}+\frac{\log 2}{\beta_2}\right)\\ &= T\exp\left(-\frac{4T}{\log T}\right)\le T^{-3}.\end{aligned} \tag{17}$$

The first equality is the definition of the running maximum. The first inequality is a union bound followed by exponential Markov. The second substitutes its exponential moment and $R^2 = (\log 2 + 4)C_0Q^2$. The second equality substitutes $\beta_2 = \frac{\log T}{T}$. The final inequality uses $(\log T)^2 \le T$ for $T \ge 2$, which follows from $\sup_{u\ge 1}\frac{(\log u)^2}{u} = \frac{4}{e^2} < 1$. This crossing estimate is valid for any positive constant base stepsize and any first-moment parameter, since only the uniform oracle tail and the second-moment recursion entered the argument.

For the parameters in the corollary,

$$\frac{\mu(\frac{\eta}{R})^2}{\beta_1^2} = \frac{\mu}{R^2\sqrt{\mu}\,T}\frac{R^2T}{\sqrt{\mu}} = 1, \qquad \frac{1}{\eta T} = \frac{\mu^{\frac{1}{4}}}{\sqrt{T}}, \qquad \frac{Q^2}{R^2} = \frac{1}{(\log 2+4)C_0}.$$

The first two equalities substitute η and β_1, and the last uses the definition of R. The common momentum bound now gives

$$\begin{aligned}\mathbb{E}S_\mu^f(x_s^*)^2 &\le C\left(1+\frac{\mu(\frac{\eta}{R})^2}{\beta_1^2}\right)\left(\frac{2R\Delta}{\eta T}+2B^2+2\beta_1Q^2\right)\\ &= 4C\left(\frac{R\Delta\mu^{\frac{1}{4}}}{\sqrt{T}}+B^2+\frac{\mu^{\frac{1}{4}}Q^2}{R\sqrt{T}}\right)\\ &\le C\left(\frac{R(\Delta+1)\mu^{\frac{1}{4}}}{\sqrt{T}}+B^2\right).\end{aligned}$$

The first inequality uses the deterministic virtual stepsize bounds. The equality substitutes the displayed identities. The second inequality absorbs the fixed ratio $\frac{Q^2}{R^2}$ into the universal constant.

Couple the implemented and virtual runs by using the same oracle randomness whenever their states agree and the same independent output index. Before the first crossing of R^2, their moment recursions and parameter updates coincide. At the first crossing, their running maxima are identical and both record a value greater than R^2. In particular, on the event $E = \{\widehat{v}_T^* \le R^2\}$, their entire paths and outputs coincide. Also, choosing the point-mass witness at x in Definition 2 gives $0 \le S_\mu^f(x)^2 \le \|\nabla f(x)\|^2 \le G^2$. Hence,

$$\begin{aligned}\mathbb{E}S_\mu^f(x_s)^2 &= \mathbb{E}[S_\mu^f(x_s)^2\mathbf{1}_E] + \mathbb{E}[S_\mu^f(x_s)^2\mathbf{1}_{E^c}]\\ &\le \mathbb{E}[S_\mu^f(x_s^*)^2\mathbf{1}_E] + G^2\mathbb{P}(E^c)\\ &\le \mathbb{E}S_\mu^f(x_s^*)^2 + G^2T^{-3}\\ &\le C\left(\frac{R(\Delta+1)\mu^{\frac{1}{4}}}{\sqrt{T}}+B^2\right)+G^2T^{-3}.\end{aligned}$$

The equality partitions the expectation. The first inequality uses equality of the paths on E and the uniform bound on its complement. The second removes the indicator from a nonnegative variable and applies (17). The third inserts the virtual-run estimate. All conditional oracle calculations were made before this final event decomposition. □

G GOLDSTEIN TRANSFER AND COMPLEXITY

G.1 A SUFFICIENT HORIZON FOR BOTH METHODS

The batch sizes in (12) control the bias and second moment separately. If the numerical multiplier in those budgets is at least one, then

$$\frac{d^2H^2}{\nu^2 N} \le dG^2, \qquad \frac{d^2L_\sigma^2}{\nu^2 M} \le d\epsilon^2 \le dG^2,$$
$$Q^2 = C\left(dG^2 + \frac{d^2H^2}{\nu^2 N} + \frac{d^2L_\sigma^2}{\nu^2 M}\right) \le 3CdG^2.$$

The first two inequalities substitute the respective lower bounds on N and M. The third uses $\epsilon \le G$. The equality is the previously fixed definition of Q^2, and its inequality substitutes the three bounds. The same definition gives $Q^2 \ge CdG^2$, so $Q^2 = \Theta(dG^2)$. For any desired numerical $c > 0$, choosing the label-budget multiplier sufficiently large ensures

$$M \ge \frac{dL_\sigma^2}{64c\nu^2\epsilon^2} \quad \Longrightarrow \quad B^2 = \frac{dL_\sigma^2}{64\nu^2 M} \le c\epsilon^2.$$

The equality squares the definition of B, and the inequality substitutes this lower bound on M.

Using R from Corollary 4, choose the common horizon

$$T = \left\lceil C\left[2 + \frac{\sqrt{\mu}}{\epsilon^4}\big(d(\Delta+G^2)^2 + R^2(\Delta+1)^2\big) + \frac{d+R^2}{\sqrt{\mu}} + \frac{\rho^2}{\sqrt{\mu}}\left(\frac{1}{d}+\frac{1}{R^2}\right) + \sqrt{\mu}\left(\frac{1}{d}+\frac{1}{R^2}\right) + \left(\frac{G}{\epsilon}\right)^{\frac{2}{3}}\right]\right\rceil. \tag{18}$$

Here and in (12), sufficiently large numerical constants may differ between occurrences. We verify separately the parameter restrictions and the final accuracy, so each term in this horizon has an explicit role.

For SGDM, the parameter conditions are exactly

$$\beta \le 1 \Longleftrightarrow T \ge \frac{\sqrt{\mu}}{d}, \quad \beta T \ge 2 \Longleftrightarrow T \ge \frac{4d}{\sqrt{\mu}}, \quad \rho\eta \le \frac{1}{2} \Longleftrightarrow T \ge \frac{4\rho^2}{d\sqrt{\mu}}.$$

Each equivalence substitutes $\beta = \frac{\mu^{\frac{1}{4}}}{\sqrt{dT}}$ and $\eta = \frac{1}{\mu^{\frac{1}{4}}\sqrt{dT}}$, then squares positive quantities. For AMSGrad, the corresponding restrictions are

$$\beta_1 \le 1 \Longleftrightarrow T \ge \frac{\sqrt{\mu}}{R^2}, \quad \beta_1 T \ge 2 \Longleftrightarrow T \ge \frac{4R^2}{\sqrt{\mu}}, \quad \frac{\rho\eta}{R} \le \frac{1}{2} \Longleftrightarrow T \ge \frac{4\rho^2}{R^2\sqrt{\mu}}.$$

These equivalences use $\beta_1 = \frac{\mu^{\frac{1}{4}}}{R\sqrt{T}}$ and $\eta = \frac{1}{\mu^{\frac{1}{4}}\sqrt{T}}$. All six lower bounds, as well as $T \ge 2$, are included in (18).

For the accuracy calculation, denote by C_2 a numerical constant large enough to cover both corollary bounds after $Q^2 = \Theta(dG^2)$ is substituted. Choose the budget multipliers so that $C_2B^2 \le \frac{\epsilon^2}{12}$. By increasing the horizon multiplier, its two leading terms and its final term ensure

$$\frac{C_2\sqrt{d}\,\mu^{\frac{1}{4}}(\Delta+G^2)}{\sqrt{T}} \le \frac{\epsilon^2}{12},$$
$$\frac{C_2R\mu^{\frac{1}{4}}(\Delta+1)}{\sqrt{T}} \le \frac{\epsilon^2}{12},$$
$$G^2T^{-3} \le \frac{\epsilon^2}{12}.$$

The first two inequalities follow by squaring and comparing with the respective accuracy terms in (18). The third is equivalent to $T \geq 12^{\frac{1}{3}}(\frac{G}{\epsilon})^{\frac{2}{3}}$. Consequently,

$$\mathbb{E}S_\mu^f(x_s)^2 \leq C_2\frac{\sqrt{d}\,\mu^{\frac{1}{4}}(\Delta+G^2)}{\sqrt{T}} + C_2B^2 \leq \frac{\epsilon^2}{6} \qquad \text{for SGDM,}$$

$$\mathbb{E}S_\mu^f(x_s)^2 \leq C_2\frac{R\mu^{\frac{1}{4}}(\Delta+1)}{\sqrt{T}} + C_2B^2 + G^2T^{-3} \leq \frac{\epsilon^2}{4} \qquad \text{for AMSGrad.}$$

The first inequality on each line is the corresponding corollary, and the second adds the allocated error terms. In particular, both methods satisfy

$$\mathbb{E}S_\mu^f(x_s)^2 \leq \frac{\epsilon^2}{4.} \tag{19}$$

For SGDM alone, the terms involving R and the final tail term in (18) can be omitted.

G.2 Transfer to the original objective

Lemma 11 (Expected-to-Goldstein transfer)**.** *For every $x \in \mathbb{R}^d$ and $r, \mu > 0$,*

$$\operatorname{dist}(0, \partial_r f(x)) \leq S_\mu^f(x) + \frac{2G}{\mu r^2}S_\mu^f(x)^2. \tag{20}$$

Moreover, $\partial_r f(x) \subseteq \partial_{r+\nu}F(x)$.

Proof. For any $\zeta > 0$, choose a witness X in Definition 2 with $\|\mathbb{E}\nabla f(X)\|^2 + \mu\mathbb{E}\|X-x\|^2 \leq S_\mu^f(x)^2+\zeta$. Replace the gradients outside the radius-r ball by $\nabla f(x)$. The resulting average belongs to $\partial_r f(x)$, and

$$\begin{aligned}
\operatorname{dist}(0, \partial_r f(x)) &\leq \left\|\mathbb{E}\left[\nabla f(X)\mathbf{1}_{\{\|X-x\|\leq r\}} + \nabla f(x)\mathbf{1}_{\{\|X-x\|>r\}}\right]\right\| \\
&= \left\|\mathbb{E}\nabla f(X) + \mathbb{E}\left[(\nabla f(x) - \nabla f(X))\mathbf{1}_{\{\|X-x\|>r\}}\right]\right\| \\
&\leq \|\mathbb{E}\nabla f(X)\| + 2G\mathbb{P}(\|X-x\| > r) \\
&\leq \|\mathbb{E}\nabla f(X)\| + \frac{2G}{r^2}\mathbb{E}\|X-x\|^2 \\
&\leq \sqrt{S_\mu^f(x)^2+\zeta} + \frac{2G}{\mu r^2}(S_\mu^f(x)^2+\zeta).
\end{aligned}$$

The first inequality uses membership in the convex hull defining $\partial_r f(x)$; that convex hull is compact because ∇f is continuous on a compact ball. The equality adds and subtracts the omitted gradients. The second inequality uses the triangle inequality and $\|\nabla f\| \leq G$. The third is Markov's inequality, and the fourth bounds the two nonnegative terms in the witness objective by their sum. Letting $\zeta \downarrow 0$ proves (20).

For the inclusion, the smoothing property $\nabla f(y) \in \partial_\nu F(y)$ (Kornowski & Shamir, 2024) and the triangle inequality give

$$\begin{aligned}
\partial_r f(x) &= \operatorname{conv} \bigcup_{\|y-x\|\leq r} \{\nabla f(y)\} \\
&\subseteq \operatorname{conv} \bigcup_{\|y-x\|\leq r} \partial_\nu F(y) \\
&= \operatorname{conv} \bigcup_{\|y-x\|\leq r} \left(\operatorname{conv} \bigcup_{\|z-y\|\leq \nu} \partial F(z)\right) \\
&\subseteq \operatorname{conv} \bigcup_{\|z-x\|\leq r+\nu} \partial F(z) = \partial_{r+\nu}F(x).
\end{aligned}$$

The first equality uses continuous differentiability of f, so its Clarke subdifferential is the singleton gradient. The first inclusion applies the smoothing property at every y. The second equality expands the Goldstein subdifferential of F. The second inclusion uses $\|z-x\| \leq \|z-y\| + \|y-x\| \leq \nu + r$ and the fact that taking a convex hull twice does not enlarge it further. The last equality is the Goldstein definition. □

G.3 Proof of Theorem 5

Proof. Take $r = \nu = \frac{\delta}{2}$. Lemma 11 and (19) yield

$$\begin{aligned}\mathbb{E}\,\mathrm{dist}(0, \partial_\delta F(x_s)) &\le \mathbb{E}\,\mathrm{dist}(0, \partial_{\frac{\delta}{2}} f(x_s)) \\ &\le \mathbb{E} S_\mu^f(x_s) + \frac{8G}{\mu\delta^2}\mathbb{E} S_\mu^f(x_s)^2 \\ &\le \sqrt{\mathbb{E} S_\mu^f(x_s)^2} + \frac{8G}{\mu\delta^2}\mathbb{E} S_\mu^f(x_s)^2 \\ &\le \frac{\epsilon}{2} + \frac{8G}{\mu\delta^2}\frac{\epsilon^2}{4} = \epsilon.\end{aligned}$$

The first inequality uses the smoothing inclusion. The second applies the transfer inequality. The third is Cauchy–Schwarz, the fourth substitutes (19), and the equality uses $\mu = \frac{4G\epsilon}{\delta^2}$.

For fixed G, H, L_σ and $\Delta \le H$, the individual terms in the sufficient horizon scale as follows:

$$\begin{aligned}\frac{\sqrt{\mu}}{\epsilon^4}\big(d(\Delta + G^2)^2 + R^2(\Delta+1)^2\big) &= O\left(\frac{d}{\delta\epsilon^{\frac{7}{2}}}\right), \\ \frac{d + R^2}{\sqrt{\mu}} &= O\left(\frac{d\delta}{\sqrt{\epsilon}}\right), \\ \frac{\rho^2}{\sqrt{\mu}}\left(\frac{1}{d} + \frac{1}{R^2}\right) &= O\left(\frac{1}{\delta\sqrt{\epsilon}}\right), \\ \sqrt{\mu}\left(\frac{1}{d} + \frac{1}{R^2}\right) &= O\left(\frac{\sqrt{\epsilon}}{d\delta}\right).\end{aligned}$$

Each equality substitutes $\sqrt{\mu} = \frac{2\sqrt{G\epsilon}}{\delta}$, $R^2 = \Theta(dG^2)$, and $\rho = O(\sqrt{d}\,\frac{G}{\delta})$; all suppressed quantities are fixed problem constants. Thus

$$\begin{aligned}T &= O\left(1 + \frac{d}{\delta\epsilon^{\frac{7}{2}}} + \frac{d\delta}{\sqrt{\epsilon}} + \frac{1}{\delta\sqrt{\epsilon}} + \frac{\sqrt{\epsilon}}{d\delta} + \epsilon^{-\frac{2}{3}}\right) \\ &= O\left(\frac{d}{\delta\epsilon^{\frac{7}{2}}} + \frac{d\delta}{\sqrt{\epsilon}}\right).\end{aligned}$$

The first equality adds the four terms above, the constant term, the tail term, and the ceiling. For the second equality, the ratios of the two discarded δ^{-1} terms to $\frac{d}{\delta\epsilon^{\frac{7}{2}}}$ are $\frac{\epsilon^3}{d} \le G^3$ and $\frac{\epsilon^4}{d^2} \le G^4$. The remaining absorptions follow from

$$\frac{d}{\delta\epsilon^{\frac{7}{2}}} + \frac{d\delta}{\sqrt{\epsilon}} \ge 2\sqrt{\frac{d}{\delta\epsilon^{\frac{7}{2}}}\frac{d\delta}{\sqrt{\epsilon}}} = \frac{2d}{\epsilon^2} \ge \frac{2}{\epsilon^2}, \qquad 1 \le G^2\epsilon^{-2}, \qquad \epsilon^{-\frac{2}{3}} \le G^{\frac{4}{3}}\epsilon^{-2}.$$

The first inequality is the arithmetic–geometric mean inequality. The equality simplifies the product, and the next inequality uses $d \ge 1$. The last two inequalities use $\epsilon \le G$. Hence both the ceiling contribution and the tail term are bounded by a fixed multiple of the two retained terms.

Keeping the batch ceilings, the costs satisfy

$$\begin{aligned}2NT &= O\left[\left(1 + \frac{d}{\delta^2}\right)\left(\frac{d}{\delta\epsilon^{\frac{7}{2}}} + \frac{d\delta}{\sqrt{\epsilon}}\right)\right], \\ NMT &= O\left[\left(1 + \frac{d}{\delta^2}\right)\left(1 + \frac{d}{\delta^2\epsilon^2}\right)\left(\frac{d}{\delta\epsilon^{\frac{7}{2}}} + \frac{d\delta}{\sqrt{\epsilon}}\right)\right].\end{aligned}$$

The first equality substitutes the trajectory-pair budget and horizon into the exact count $2NT$. The second additionally substitutes the label budget into the exact count NMT.

When $0 < \delta \le 1$ and $0 < \epsilon \le 1$, the nonconstant batch terms dominate, and $\frac{\frac{d\delta}{\sqrt{\epsilon}}}{\frac{d}{\delta\epsilon^{\frac{7}{2}}}} = \delta^2\epsilon^3 \le 1$. Thus,

$$T = O\left(\frac{d}{\delta\epsilon^{\frac{7}{2}}}\right), \qquad 2NT = O\left(\frac{d^2}{\delta^3\epsilon^{\frac{7}{2}}}\right), \qquad NMT = O\left(\frac{d^3}{\delta^5\epsilon^{\frac{11}{2}}}\right).$$

The first equality retains the dominant horizon term. The second and third multiply by the respective sampling budgets. □

H Comparison with ZO-SGD

For an explicit comparison under the same feedback model, take the update $x_{t+1} = x_t - \gamma \widehat{g}_t$ with the same conditional oracle bounds, and select the output index uniformly and independently. The ρ-Lipschitz gradient of f gives

$$\begin{aligned}\mathbb{E}[f(x_{t+1}) \mid \mathcal{F}_t] &\le f(x_t) - \gamma \langle \nabla f(x_t), \mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] \rangle + \frac{\rho\gamma^2}{2}\mathbb{E}[\|\widehat{g}_t\|^2 \mid \mathcal{F}_t] \\ &\le f(x_t) - \frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{\gamma}{2}B^2 + \frac{\rho\gamma^2}{2}Q^2.\end{aligned}$$

The first inequality is smoothness followed by conditional expectation. The second uses the bias bound and $-\langle a, b\rangle \le \frac{\|a\|^2+\|b\|^2}{2}$. Summing and using $f \ge \inf f$ gives

$$\mathbb{E}\|\nabla f(x_s)\|^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 \le \frac{2\Delta}{\gamma T} + B^2 + \rho\gamma Q^2.$$

The equality uses the uniform output rule, and the inequality telescopes the preceding descent bound. If $B^2 \le \frac{\epsilon^2}{4}$, choose $\gamma = \frac{\epsilon^2}{4\rho Q^2}$ and an integer $T \ge 16\frac{\rho Q^2\Delta}{\epsilon^4}$. Substitution gives

$$\begin{aligned}\mathbb{E}\|\nabla f(x_s)\|^2 &\le \frac{8\rho Q^2\Delta}{\epsilon^2 T} + \frac{\epsilon^2}{4} + \frac{\epsilon^2}{4} \\ &\le \frac{\epsilon^2}{2} + \frac{\epsilon^2}{4} + \frac{\epsilon^2}{4} = \epsilon^2.\end{aligned}$$

The first inequality substitutes γ and the bias budget into the preceding bound. The second uses the choice of T, and the equality adds the three error terms. Since $\nabla f(x_s) \in \partial_\nu F(x_s) \subseteq \partial_\delta F(x_s)$ when $\nu \le \delta$,

$$\mathbb{E}\,\mathrm{dist}(0, \partial_\delta F(x_s)) \le \mathbb{E}\|\nabla f(x_s)\| \le \sqrt{\mathbb{E}\|\nabla f(x_s)\|^2} \le \epsilon.$$

The first inequality evaluates the set distance at $\nabla f(x_s)$. The second is Cauchy–Schwarz, and the third uses the squared-gradient estimate.

With $\nu = \frac{\delta}{2}$ and the same budgets (12), $Q^2 = O(dG^2)$ and $\rho = O(\frac{\sqrt{d}G}{\delta})$ give the leading sufficient SGD iteration count $O(d^{\frac{3}{2}}\delta^{-1}\epsilon^{-4})$. This recovers the ϵ^{-4} dependence of the standard smoothing-and-SGD analysis, consistent with Lin et al. (2022). The corresponding momentum bound has exponent $\frac{7}{2}$ in the same expected Goldstein criterion. Each preference-based iteration still costs NM labels, so outer-iteration and human-query complexities must be distinguished.

I Experiments

GridWorld. We consider a stochastic 5×5 GridWorld with horizon $K = 20$. The agent starts at $(3, 3)$ and chooses among four actions: up, down, left, and right. The intended action is executed with probability 0.6 and reversed with 0.4. Each grid cell has a randomly generated reward.

Risky Bandit. We construct a one-step bandit with 12 actions to induce a mismatch between expected return and majority preference. One safe action deterministically yields reward 1, while one risky action yields 10 with probability 0.2 and 0 otherwise, giving expected return 2; the remaining ten actions yield -0.5. Thus, the risky action is optimal in expectation, while the uniform initial policy has expected return $-\frac{1}{6}$. However, the risky action loses to the safe action in most realized comparisons, testing whether preference-based zeroth-order methods recover expected-return differences rather than simply follow majority preference.

To examine a policy class that can yield nonsmooth objectives, we use a softmax–sparsemax (Martins & Astudillo, 2016) mixture in both environments. For each state s, let

$$\pi_x(\cdot \mid s) = (1-\lambda_{\mathrm{mix}})\,\mathrm{softmax}(x_s) + \lambda_{\mathrm{mix}}\,\mathrm{sparsemax}(x_s), \qquad \mathrm{sparsemax}(z) = \operatorname*{arg\,min}_{p \in \Delta_{\mathcal{A}}} \|p - z\|^2.$$

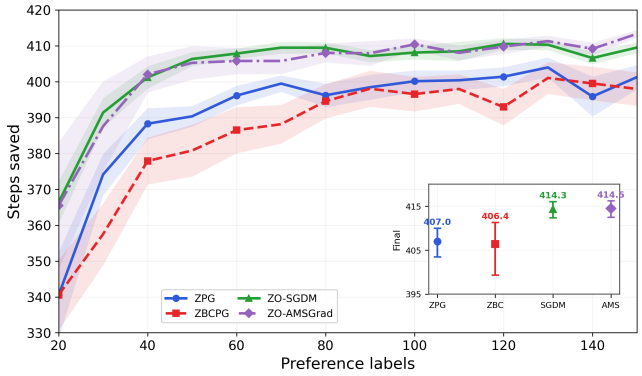


Figure 2: Acrobot-v1 with one preference label per update. The inset shows final performance after 300 labels; shading indicates standard error over 20 seeds.

For $\lambda_{\text{mix}} \in (0, 1)$, every action retains positive probability, while the policy remains within $2\lambda_{\text{mix}}$ of softmax in ℓ_1 distance. The sparsemax component is nondifferentiable when the action support changes. All compared methods use the same policy parameterization and are tuned separately.

Acrobot. We evaluate ZOPGM, ZOPGAM, ZPG, and ZBCPG on `Acrobot-v1` using a 6–16–3 ReLU policy with softmax output (163 parameters) and a 500-step episode limit. Returns are used only for evaluation and to generate labels according to $\Pr(\tau^+ \succ \tau^-) = \sigma\left(\frac{R(\tau^+)-R(\tau^-)}{25}\right)$; the optimizers observe no returns. Each method uses one pair and one label per update ($N = M = 1$), for 300 labels in total. All methods share an initialization. We evaluate every 10 updates over 50 episodes and average across 20 held-out seeds (10–29). Figure 2 reports steps saved (500 plus episode return), including final performance after 300 labels.

Query efficiency. In Bradley–Terry GridWorld, we vary the number of pairs N and labels per pair M over $(M, N) \in \{(1, 50), (5, 50), (20, 50), (5, 200), (20, 200), (100, 200)\}$. Each condition runs for 1500 updates on five seeds. Figure 3 plots exact expected return against cumulative label use tNM. At the same cost of 1000 labels per update, $(M, N) = (5, 200)$ yields higher final returns than $(20, 50)$ for all four methods, showing a greater benefit from sampling more pairs in this comparison, which is consistent with our theoretical results.

Fixed preference budget. We repeat these six conditions with $B = 1.5 \times 10^6$ labels per method. Setting $T = B/(MN)$ gives 30000, 6000, 1500, 1500, 375, and 75 updates, respectively. Figure 4 shows results over five seeds, evaluated at 15 equally spaced checkpoints after initialization. The $(100, 200)$ condition uses separately retuned learning rates; the other five retain the same per-method hyperparameters.

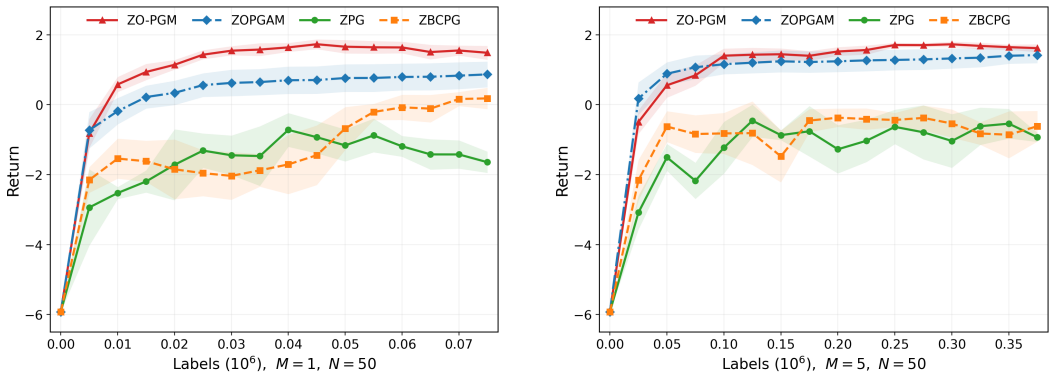
Hyperparameters. We carefully tuned the hyperparameters of all methods to ensure a fair comparison. In the main **GridWorld and Risky Bandit** experiments, all methods use seeds 1–5 and run for 2000 and 5000 updates, respectively. For visualization, we truncate excessively long plateau regions that do not affect the qualitative comparison. Hyperparameters are fixed across the Bradley–Terry and Weibull feedback conditions within each environment. For ZPG, ZBCPG, ZOPGM, and ZOPGAM, the smoothing radius is 5 in GridWorld and 2 in Risky Bandit; their learning rates, in that order, are $(0.30, 0.50, 0.15, 0.50)$ and $(0.025, 0.040, 0.180, 0.20)$. ZBCPG uses block sizes 20 and 4. We set $\beta = 0.1$ for ZOPGM and $(\beta_1, \beta_2, \varepsilon_{\text{ad}}) = (0.1, 0.001, 10^{-8})$ for ZOPGAM. Each zeroth-order update uses $N = M = 1000$ and clipping threshold $\Delta = 0.001$.

DPO and Online DPO likewise use 1000 trajectory pairs and 1000 labels per pair at each update. Both take five full-batch steps on a Bradley–Terry cross-entropy loss, with coefficient 0.1 on the trajectory log-probability ratio; empirical preference rates are not clipped. DPO samples from and references the initial policy, whereas Online DPO samples from and references the policy before each update. Their learning rates are 0.055 and 0.74 in GridWorld, and 12 and 12 in Risky Bandit.

RM+PPO fits a tabular reward model once on $500{,}000$ initial-policy trajectory pairs, with 1000 labels per pair. It uses five training epochs, minibatches of 5000, learning rate 0.1, and an ℓ_2 coefficient of 10^{-4}; the reward model is then held fixed. PPO uses 1000 fresh episodes per update, five passes per rollout, action-ratio clipping at 0.2, a current-policy KL reference with weight 0.1, and learning rate 5, selected from $\{0.2, 1, 5\}$ on seed 0. All three baselines fit a Bradley–Terry loss under both feedback links.

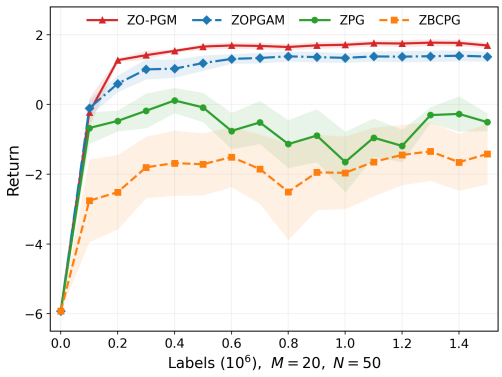
For **Acrobot**, we independently searched learning rates and smoothing radii for each method on tuning seeds 0–2. The selected (learning rate, smoothing radius) pairs are $(0.0002, 4)$ for ZPG, $(0.0002, 4)$ for ZBCPG, $(0.0005, 4)$ for ZOPGM, and $(0.3, 6)$ for ZOPGAM. ZBCPG uses block size 32; ZOPGM uses $\beta = 0.1$; and ZOPGAM uses $(\beta_1, \beta_2, \varepsilon_{\mathrm{ad}}) = (0.1, 0.001, 10^{-8})$.

For the **GridWorld budget** comparisons, we keep the smoothing radius at 5 and fix the learning rates of (ZPG, ZBCPG, ZOPGM, ZOPGAM) to $(0.30, 0.35, 0.15, 0.50)$ across all six (M, N) conditions. For the Risky Bandit query comparison, the corresponding values are $(0.025, 0.040, 0.10, 0.20)$ with smoothing radius 2. ZBCPG uses block sizes 20 and 4, respectively; the momentum parameters remain $\beta = \beta_1 = 0.1$ and $\beta_2 = 0.001$. No method is retuned separately for each (M, N) condition.

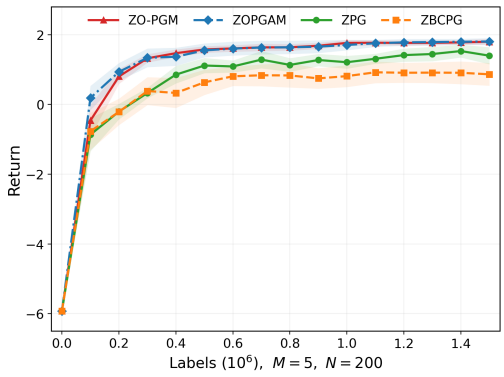


(a) $M = 1,\ N = 50$

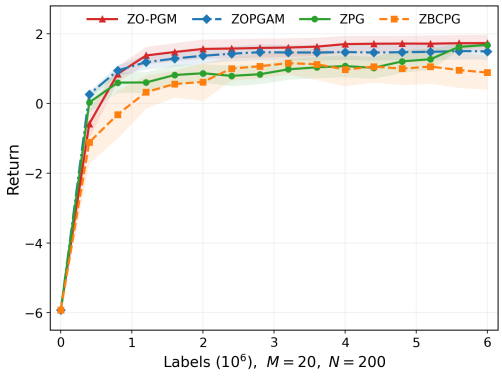
(b) $M = 5,\ N = 50$



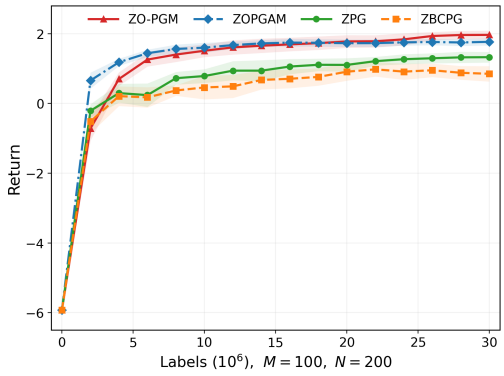
(c) $M = 20,\ N = 50$



(d) $M = 5,\ N = 200$

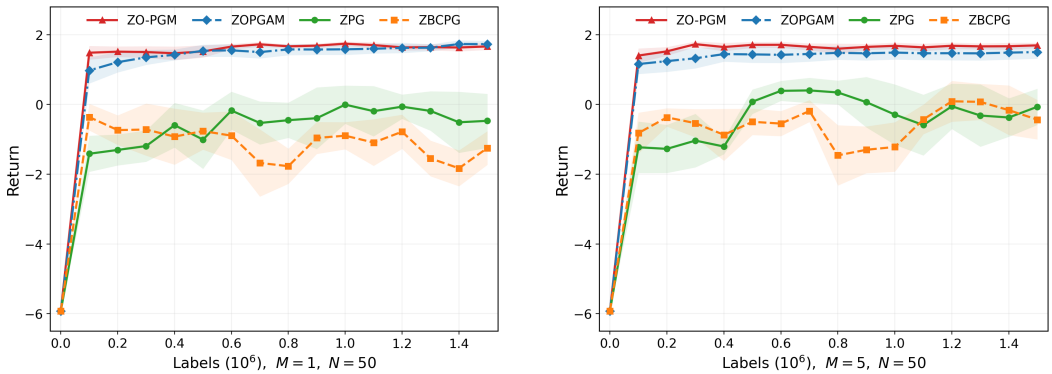


(e) $M = 20,\ N = 200$

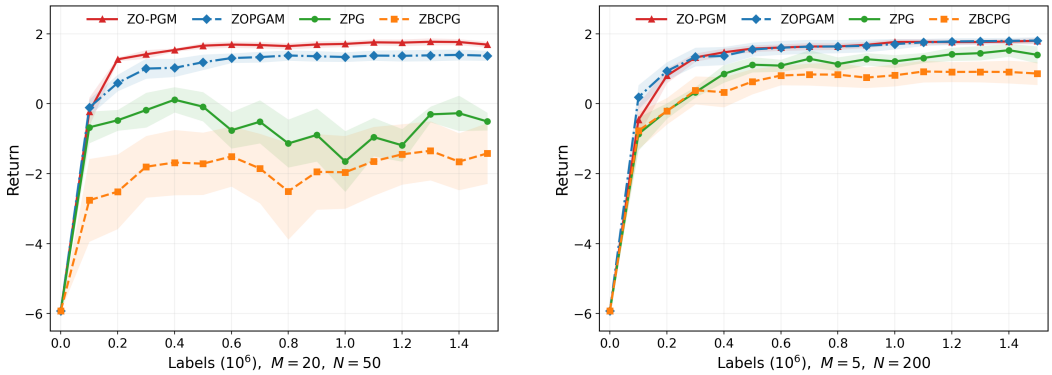


(f) $M = 100,\ N = 200$

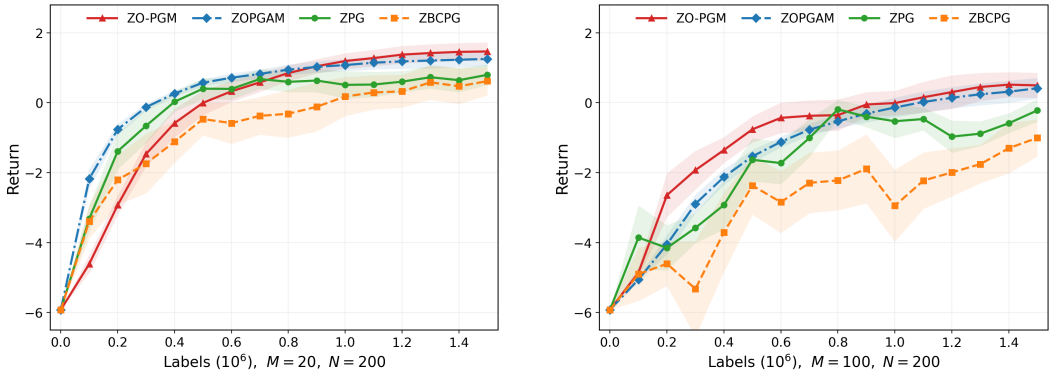
Figure 3: Query efficiency in Bradley–Terry GridWorld with a fixed number of policy updates.



(a) $M = 1,\ N = 50$ (b) $M = 5,\ N = 50$



(c) $M = 20,\ N = 50$ (d) $M = 5,\ N = 200$



(e) $M = 20,\ N = 200$ (f) $M = 100,\ N = 200$

Figure 4: Bradley–Terry GridWorld under a fixed budget of 1.5×10^6 preference labels.