GSAFEPO: GEOMETRIC OPTIMIZERS FOR REWARD-SAFE PREFERENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) and related preference-learning methods optimize relative margins between chosen and rejected responses, but provide limited control over their absolute reward dynamics. This can lead to *likelihood displacement*, where chosen responses become less likely despite improved preference margins, and to *reward over-optimization*, where excessive absolute reward shifts drive unnecessary policy drift that can degrade generation quality even as the preference objective continues to improve. Recent methods address these effects largely through loss-level interventions. We introduce GSAFEPO (Geometrically Safe Preference Optimization), an optimizer framework that instead directly constrains the geometry of the realized parameter update. GSAFEPO projects the effective optimizer step onto a reward-safe set so that rewards for chosen responses do not decrease and rewards for rejected responses do not increase to first order, while deviating minimally from the unconstrained update. We characterize the low-dimensional geometry of such updates, derive conditions under which they remain aligned with descent on the original preference objective, and establish population-level guarantees for stochastic minibatch training. We instantiate the framework as two adaptive optimizers, GSAFERMSPROP and GSAFEADAM, extending the construction to adaptive preconditioned optimizers. Across multiple preference objectives and model configurations, GSafePO consistently converts undesirable common-mode reward motion into chosen-up/rejected-down dynamics while preserving competitive preference performance. Generation-quality evaluations further show that these corrected dynamics translate into significantly improved behavior in settings exhibiting likelihood displacement. These results establish update-geometry control as a modular mechanism for adding explicit reward safeguards to existing preference-learning objectives.

1 INTRODUCTION

Direct Preference Optimization (DPO) (Rafailov et al., 2023) and related methods (Meng et al., 2024; Hong et al., 2024; Xu et al., 2024; Tang et al., 2024) have become a central paradigm for preference-based post-training. Given a preference $\mathbf{y}_w \succ \mathbf{y}_l \mid \mathbf{x}$, DPO optimizes an implicit reward $r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) := \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}$ through the margin-based objective

$$\mathcal{L}_{\text{DPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}} \left[\log \sigma(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l))\right]. \tag{1}$$

Because this objective depends only on the reward difference, it does not control the absolute motion of chosen and rejected rewards. Thus preference improvement can be achieved while both rewards decrease, both increase, or otherwise drift in ways not determined by the pairwise label. Prior work has documented such behavior (Pal et al., 2024; Liu et al., 2024; Pang et al., 2024; Yuan et al., 2025b; Rafailov et al., 2024b; Tajwar et al., 2024; Ren & Sutherland, 2025; Yuan et al., 2025a; Razin et al., 2025). In particular, Razin et al. (2025) identify *likelihood displacement*, where the chosen response becomes less likely even as the preference margin improves, while broader *reward over-optimization* can induce unnecessary policy drift and eventually degrade generation quality (Rafailov et al., 2024a).

This exposes a gap between optimizing relative preference and controlling absolute reward dynamics. Ideally, preference optimization should improve the original preference objective while also avoiding decreases in expected chosen reward and increases in expected rejected reward. We refer to this

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.

secondary requirement as *reward safety*[1]: an update is *reward-safe* if it does not decrease the expected chosen rewards and does not increase expected rejected rewards. Existing approaches attempt to primarily achieve these dynamics through loss-level interventions. We instead ask whether the same preference objective can be retained while controlling the geometry of the update actually taken.

We introduce GSAFEPO, an optimizer-level framework that projects the effective update onto a reward-safe set, enforcing chosen-up/rejected-down first-order dynamics while deviating minimally from the underlying optimization direction. Our contributions are:

- We formalize *reward-safe* preference optimization and characterize the corresponding first-order safe update directions through a compact low-dimensional geometry (Section 2).
- We derive GSAFEPO, a projected update rule that enforces the desired reward dynamics without modifying the underlying preference loss (Section 3).
- We analyze stochastic minibatch-defined projections, giving conditions under which they preserve preference-loss descent and transfer minibatch safety to population-level reward guarantees (Section 4).
- We instantiate the framework as GSAFERMSPROP and GSAFEADAM. Across preference objectives and model configurations, these optimizers induce consistent chosen-up/rejected-down dynamics while preserving preference performance, and yield substantial generation gains when likelihood displacement is present (Sections 5–6).

Related Work: Recent work has established that preference-margin improvement alone is insufficient, and has identified increasing chosen rewards while decreasing rejected rewards as a key optimization desideratum (Xiao et al., 2024; Yuan et al., 2025a; Chen et al., 2026). Gradient-entanglement analyses (Yuan et al., 2025a; Chen et al., 2026) aim to characterize, for an individual preference pair, when a pairwise gradient update yields the desired chosen-up/rejected-down dynamics (disentangled dynamics) rather than synchronized motion of the two likelihoods (entangled dynamics). However, pairwise disentanglement does not in general imply the desired change in expected rewards nor do they guarantee per-example disentanglement after a realized update: gradients from different pairs are aggregated within a minibatch, and the resulting direction may be further transformed by the optimizer, so pairwise disentanglement conditions need not hold for the parameter update actually taken or have the desired effect on expected rewards (see Appendix F). This leaves two distinct questions open at the level of the realized optimization step: (1) *Does the aggregate update produce the desired change in expected rewards?* and (2) *Can such updates be constructed while preserving progress on the preference objective?*. Our notion of reward safety addresses exactly these questions.

Existing reward dynamics interventions act primarily at the loss level. DPOP adds a chosen-likelihood penalty (Pal et al., 2024), Cal-DPO calibrates chosen and rejected rewards toward target values (Xiao et al., 2024), and Chen et al. (2026) wrap the rewards for chosen/rejected responses in a pairwise surrogate loss whose backward derivatives are implicitly rescaled to steer individual preference pairs toward a pairwise disentanglement band which seeks to preserve chosen responses and supress rejected responses. These mechanisms shape the gradient signal upstream of the optimizer, but do not directly constrain the realized minibatch update or account for the geometry introduced by adaptive optimization. In contrast, GSAFEPO leaves the preference objective unchanged and imposes chosen-up/rejected-down constraints directly on the realized update, providing an optimizer-level mechanism for controlling aggregate reward dynamics.

2 REWARD-SAFE PREFERENCE OPTIMIZATION

Reward-Safe Optimization Trajectories. Consider a generalized training objective over pairwise rewards,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\big[f\big(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l,r_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}_w),r_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y}_l)\big)\big], \tag{2}$$

where f is differentiable in the score arguments and $r_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{y})$ is any differentiable scalar implicit reward induced by the policy. This general form includes several DPO-variants as special cases (see Appendix D). Rather than changing the objective, our goal is to find an optimization algorithm

[1]The term *reward safety* refers only to this reward-dynamics property and does not denote broader notions of model safety or harmlessness.

for the same objective that guards against likelihood displacement by engineering updates that do not reduce expected reward for a chosen response and also do not increase expected reward for a rejected response. We therefore seek a sequence $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}_t$ that remains stationarity-seeking for the original loss, while also satisfying reward safety constraints at any iterate. This results in the following reward-safe trajectory problem (RSTP) framing:

$$\text{find } \{\boldsymbol{\theta}_t\} \text{ s.t. } \lim_{t\to\infty} \|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\| = 0, r_w(\boldsymbol{\theta}_{t+1}) \geq r_w(\boldsymbol{\theta}_t),\ r_l(\boldsymbol{\theta}_{t+1}) \leq r_l(\boldsymbol{\theta}_t),\ \forall t \geq 0 \quad \text{(RSTP)}$$

where $r_w(\boldsymbol{\theta}_t) := \mathbb{E}[r_{\boldsymbol{\theta}_t}(\mathbf{x}, \mathbf{y}_w)]$, $r_l(\boldsymbol{\theta}_t) := \mathbb{E}[r_{\boldsymbol{\theta}_t}(\mathbf{x}, \mathbf{y}_l)]$. We assume rewards are measured relative to the reference initialization, so that $r_w(\boldsymbol{\theta}_0) = r_l(\boldsymbol{\theta}_0) = 0$. RSTP attempts to protect expected rewards, imposing a non-decreasing constraint on reward assigned to chosen responses and a non-increasing constraint on reward assigned to rejected responses throughout training. Thus, reward safety is a constraint on the reward dynamics of an optimizer that still targets stationary points of the original preference loss.

First-order conditions for reward-safe trajectories. We pass to a local first-order analysis of a single update step and derive necessary conditions for instantaneous chosen-up/rejected-down behavior. Consider the update $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\mathbf{v}$ along direction $\mathbf{v} \in \mathbb{R}^N$ at the current iterate $\boldsymbol{\theta}$. Let $\boldsymbol{\mu}_w := \nabla_{\boldsymbol{\theta}} r_w(\boldsymbol{\theta})$, $\boldsymbol{\mu}_l := \nabla_{\boldsymbol{\theta}} r_l(\boldsymbol{\theta})$. A first-order expansion gives

$$r_w(\boldsymbol{\theta}^+) - r_w(\boldsymbol{\theta}) = -\eta\langle\boldsymbol{\mu}_w, \mathbf{v}\rangle + o(\eta), \qquad r_l(\boldsymbol{\theta}^+) - r_l(\boldsymbol{\theta}) = -\eta\langle\boldsymbol{\mu}_l, \mathbf{v}\rangle + o(\eta) \tag{3}$$

Thus, the desired dynamics are equivalent, to first-order, to the directional conditions:

$$-\langle\boldsymbol{\mu}_w, \mathbf{v}\rangle \geq 0, \qquad \langle\boldsymbol{\mu}_l, \mathbf{v}\rangle \geq 0 \tag{4}$$

We use the geometry of these reward-safe directions to construct stochastic minibatch updates that are locally consistent with the RSTP objective and admit population-level loss and reward guarantees. We call this update-level approach Geometrically Safe Preference Optimization (GSAFEPO).

2.1 The Geometry of Reward-Safe Updates

We consider an alternate formulation of the set of reward-safe update directions satisfying Equation 4 and show that it is possible to determine safety of an update direction $\mathbf{v}$, entirely from its orthogonal projection $\mathbf{v}_{\text{span}}$ onto the subspace spanned by the vectors $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$. This allows us to equivalently characterize the *safe* region in terms of the coefficients of the subspace basis in $\mathbb{R}^2$.

While it may seem natural to consider the representation of $\mathbf{v}_{\text{span}}$ w.r.t. the obvious $(\boldsymbol{\mu}_w, \boldsymbol{\mu}_l)$ basis so that $\mathbf{v}_{\text{span}} = a_w\boldsymbol{\mu}_w + a_l\boldsymbol{\mu}_l$ with $\mathbf{a} = (a_w, a_l)^T \in \mathbb{R}^2$, in such a *primal coefficient-space* formulation the corresponding safe regions in coefficient space do not have a simple characterization (see Section B for a development of this primal approach).

We consider a more elegant *dual coefficient space* with basis $(\tilde{\boldsymbol{\mu}}_w, \tilde{\boldsymbol{\mu}}_l)$, in which $\mathbf{v}_{\text{span}} = -\alpha_w\tilde{\boldsymbol{\mu}}_w + \alpha_l\tilde{\boldsymbol{\mu}}_l$, with coefficients $\boldsymbol{\beta} = (-\alpha_w, \alpha_l)^T \in \mathbb{R}^2$. The vectors $(\tilde{\boldsymbol{\mu}}_w, \tilde{\boldsymbol{\mu}}_l)$ are defined as follows:

$$\tilde{\boldsymbol{\mu}}_w := \frac{\|\boldsymbol{\mu}_l\|^2}{\|\boldsymbol{\mu}_w\|^2\|\boldsymbol{\mu}_l\|^2 - \langle\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\rangle^2}\left(\boldsymbol{\mu}_w - \frac{\langle\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\rangle}{\|\boldsymbol{\mu}_l\|^2}\boldsymbol{\mu}_l\right) \tag{5}$$

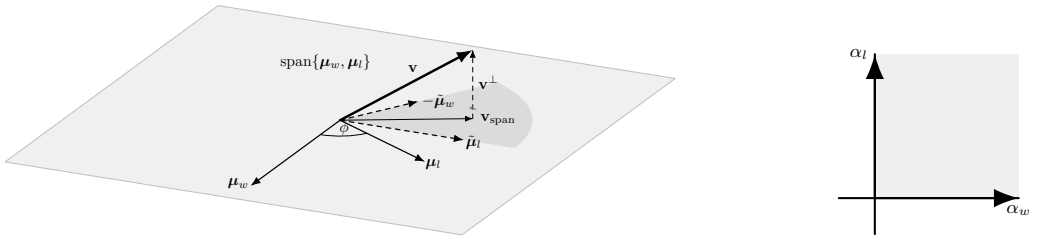
$$\tilde{\boldsymbol{\mu}}_l := \frac{\|\boldsymbol{\mu}_w\|^2}{\|\boldsymbol{\mu}_w\|^2\|\boldsymbol{\mu}_l\|^2 - \langle\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\rangle^2}\left(\boldsymbol{\mu}_l - \frac{\langle\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\rangle}{\|\boldsymbol{\mu}_w\|^2}\boldsymbol{\mu}_w\right) \tag{6}$$

Remark (Link between primal and dual bases). $(\tilde{\boldsymbol{\mu}}_w, \tilde{\boldsymbol{\mu}}_l)$ form a dual basis (a.k.a. a bi-orthogonal basis) for $\text{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$ with the following properties:

$$\langle\tilde{\boldsymbol{\mu}}_w, \boldsymbol{\mu}_w\rangle = 1,\ \langle\tilde{\boldsymbol{\mu}}_w, \boldsymbol{\mu}_l\rangle = 0,\ \langle\tilde{\boldsymbol{\mu}}_l, \boldsymbol{\mu}_w\rangle = 0,\ \langle\tilde{\boldsymbol{\mu}}_l, \boldsymbol{\mu}_l\rangle = 1 \tag{7}$$

Theorem 2.1 (Safe update geometry)**.** *Any safe update direction* $\mathbf{v}$ *has an orthogonal projection onto* $\text{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$ *given by* $\mathbf{v}_{\text{span}} = -\alpha_w\tilde{\boldsymbol{\mu}}_w + \alpha_l\tilde{\boldsymbol{\mu}}_l$*, with* $\alpha_w, \alpha_l \geq 0$*. The safe set is therefore the first quadrant in* (α_w, α_l)*-space.*

From the form of $\mathbf{v}_{\text{span}}$ for safe updates, we can see that the vectors $-\tilde{\boldsymbol{\mu}}_w$ and $\tilde{\boldsymbol{\mu}}_l$ are boundary vectors for the feasible set in $\text{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$ (Figure 1(a)). Note that, unlike the safe update geometry in (a_w, a_l)-space (Figure 3, Appendix B), the safe set in (α_w, α_l)-space (Figure 1(b)) does not change



(a) Safe geometry illustrated in $\mathbb{R}^3$; only part of the safe set is shaded.

(b) (α_w, α_l)-space

Figure 1: Illustration of the safe update geometry described by Theorem 2.1.

with the gradient magnitudes of $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$ or with the angle between them, making it an elegant characterization of DPO update geometry. The transformation between the coefficient spaces is straightforward:

$$\underbrace{\begin{bmatrix} -\alpha_w \\ \alpha_l \end{bmatrix}}_{\boldsymbol{\beta}} = \underbrace{\begin{bmatrix} \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_w \rangle & \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \\ \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle & \langle \boldsymbol{\mu}_l, \boldsymbol{\mu}_l \rangle \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} a_w \\ a_l \end{bmatrix}}_{\mathbf{a}} \tag{8}$$

In the following sections, we assume $\mathbf{M} \succ 0$, excluding the degenerate case in which $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$ are collinear. Let ϕ denote the angle between them, so that degeneracy corresponds to $\sin^2 \phi = 0$. We treat this unlikely case in Appendix C.

3 Construction of Geometrically-Safe Updates

Having characterized the local reward-safe directions, we construct updates by projecting the loss gradient onto this safe set. However, we additionally impose the constraint that the update move winner rewards up by the same amount that loser rewards are moved down, up to first order. There are two reasons for this. First, because DPO and variants operate on paired preferences with no explicit rating scale, a useful inductive bias is to require balanced reward updates. Second, and more importantly, balancing reward updates reduces bias that can cause winner and loser rewards to move in the same direction. To see this, consider the following for an update $\Delta\boldsymbol{\theta}$:

$$\Delta r_w \approx \langle \boldsymbol{\mu}_w, \Delta\boldsymbol{\theta} \rangle = \frac{1}{2}\langle \boldsymbol{\mu}_w + \boldsymbol{\mu}_l, \Delta\boldsymbol{\theta} \rangle + \frac{1}{2}\langle \boldsymbol{\mu}_w - \boldsymbol{\mu}_l, \Delta\boldsymbol{\theta} \rangle \tag{9}$$

$$\Delta r_l \approx \langle \boldsymbol{\mu}_l, \Delta\boldsymbol{\theta} \rangle = \frac{1}{2}\langle \boldsymbol{\mu}_w + \boldsymbol{\mu}_l, \Delta\boldsymbol{\theta} \rangle - \frac{1}{2}\langle \boldsymbol{\mu}_w - \boldsymbol{\mu}_l, \Delta\boldsymbol{\theta} \rangle \tag{10}$$

It is easy to see that the first term is a shared bias that pushes winner and loser rewards in the same direction. Requiring balanced updates is equivalent to enforcing, up to first order, $\Delta r_w + \Delta r_l = 0$ at each update, which seeks to eliminate this shared bias term, thereby increasing the fraction of the step devoted to true preference separation.

Given the loss gradient $\mathbf{g} := \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$, finding the optimal safe projected gradient $\hat{\mathbf{g}}$ with safety margin $\delta \geq 0$ requires solving:

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{v}} \|\mathbf{g} - \mathbf{v}\|^2, \text{ s.t. } \langle \boldsymbol{\mu}_w + \boldsymbol{\mu}_l, \mathbf{v} \rangle = 0, \ -\langle \boldsymbol{\mu}_w, \mathbf{v} \rangle \geq \delta, \ \langle \boldsymbol{\mu}_l, \mathbf{v} \rangle \geq \delta. \tag{11}$$

Working in dual space, this reduces to computing a single scalar:

Proposition 3.1 (Optimal projected gradient)**.** *Let*

$$\mathbf{g} = \mathbf{g}^\perp + \mathbf{g}_{\text{span}}, \qquad \mathbf{g}_{\text{span}} = -\alpha_w \tilde{\boldsymbol{\mu}}_w + \alpha_l \tilde{\boldsymbol{\mu}}_l, \tag{12}$$

where $\mathbf{g}^\perp$ *is orthogonal to* $\text{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$*. The problem in Equation (11) is equivalent to the one-dimensional problem*

$$\hat{\alpha} = \arg\min_{\alpha \geq \delta} (\boldsymbol{\beta} - \alpha \mathbf{d})^T \mathbf{M}^{-1} (\boldsymbol{\beta} - \alpha \mathbf{d}), \tag{13}$$

where $\boldsymbol{\beta} = (-\alpha_w, \alpha_l)^T$*,* $\mathbf{d} = (-1, 1)^T$*, and* $\mathbf{M}$ *is the Gram matrix of the primal basis. The optimal safe projected gradient is*

$$\hat{\mathbf{g}} = \mathbf{g}^\perp - \hat{\alpha}\, \tilde{\boldsymbol{\mu}}_w + \hat{\alpha}\, \tilde{\boldsymbol{\mu}}_l, \qquad \hat{\alpha} = \max\left(\delta, \frac{\boldsymbol{\beta}^T \mathbf{M}^{-1} \mathbf{d}}{\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}}\right). \tag{14}$$

We now prove some key properties of the projected gradient.

Corollary 3.1 (Uniform bound on projected-gradient norm)**.** *Let $\mathbf{M} \succ 0$ and $\lambda_i(\mathbf{M}) > 0$ be the ith eigenvalue of $\mathbf{M}$. Let $\delta = \kappa\|\boldsymbol{g}\|$ for some constant $\kappa \geq 0$. Suppose there exist constants $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ such that at every iterate: $\underline{\lambda} \leq \lambda_i(\mathbf{M}) \leq \overline{\lambda}$, $i = 1, 2$. Then,*

$$\|\hat{\boldsymbol{g}}\| \leq \sqrt{1 + \frac{2}{\underline{\lambda}} \max\left\{\kappa, \sqrt{\frac{\overline{\lambda}}{2}}\right\}^2} \|\boldsymbol{g}\| \tag{15}$$

Thus, uniformly well-conditioned reward geometry implies that the projection cannot inflate the update norm by more than a fixed multiplicative factor.

Corollary 3.2 (Pointwise and uniform descent alignment)**.** *Let $\mathbf{d} = (-1, 1)^T$, $p := \|\mathbf{g}^{\perp}\|^2$, and $s := \mathbf{d}^T\mathbf{M}^{-1}\boldsymbol{\beta}$ with $\mathbf{M} \succ 0$. Let $\delta = \kappa\|\mathbf{g}\|$ for some constant $\kappa \geq 0$. Then,*

$$\langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq 0 \iff s\delta \geq -p \tag{16}$$

Moreover, suppose there exist constants $\kappa_s > 0$ and $\kappa_p \in (0, 1]$ such that, at every iterate, $s \geq \kappa_s\|\mathbf{g}\|$ and $p \geq \kappa_p\|\mathbf{g}\|^2$, then, at every iterate,

$$\langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq (\kappa_p + \kappa\kappa_s)\|\mathbf{g}\|^2 \tag{17}$$

Corollary 3.2 shows that failure to obtain a descent-aligned update requires a specific adverse geometry. The quantity $s := \mathbf{d}^T\mathbf{M}^{-1}\boldsymbol{\beta}$ measures the orientation of the unprojected gradient within the constraint span along the direction that increases chosen reward relative to rejected reward. For $\delta = 0$, GSAFEPO is *always non-negatively aligned with the original loss gradient*. For $\delta > 0$, loss of descent alignment can occur only when the in-span orientation is sufficiently adverse that $s\delta < -p$, i.e., its negative contribution overwhelms the positive alignment $p = \|g^{\perp}\|^2$ supplied by the orthogonal component. Thus, except for such strongly incompatible configurations, GSAFEPO remains descent-aligned with the original loss gradient. Further, under the stronger uniform conditions, $\langle \mathbf{g}, \hat{\mathbf{g}} \rangle$ is bounded below by a positive multiple of $\|\mathbf{g}\|^2$ at every iterate.

4 Analysis of Stochastic GSafePO Updates

We consider the stochastic setting in which both the gradient and projection geometry are minibatch-dependent. Unlike standard projected SGD with a fixed feasible set, here the same minibatch determines both the stochastic gradient and the constraint geometry, coupling projection error to minibatch noise. Our analysis therefore characterizes when such minibatch-defined projections preserve the intended population-level loss descent and reward dynamics.

The following theorem establishes how stochastic minibatch updates realize the population-level RSTP objective. Under unbiased minibatch updates, sufficient alignment of the edited update with the population descent direction, and chosen-up/rejected-down progress in conditional expectation at each step, the population loss approaches stationarity in expectation over stochastic training trajectories, with reward-safe updates.

Theorem 4.1 (Population guarantees for reward-safe stochastic updates)**.** *Let $\mathcal{L}(\boldsymbol{\theta})$ denote an L-smooth population loss that is bounded below and parameterized by $\boldsymbol{\theta} \in \mathbb{R}^N$. Consider an algorithm that produces updates $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}_t$, with $\Delta\boldsymbol{\theta}_t = -\eta_t\hat{\mathbf{g}}_t^{(b)}$ measurable with respect to the filtration $\mathcal{F}_t \vee \sigma(B_t)$, where $\mathcal{F}_t := \sigma(B_0, B_1, \cdots, B_{t-1})$ and B_t represents the random minibatch drawn i.i.d. at step t. Let $r_w(\boldsymbol{\theta}_t; B_t)$ and $r_l(\boldsymbol{\theta}_t; B_t)$ denote the winner and loser rewards averaged over minibatch B_t. We define the following minibatch quantities:*

$$r_{w,t}^{(b)}(\boldsymbol{\theta}_t) := r_w(\boldsymbol{\theta}_t; B_t)\ ,\ r_{l,t}^{(b)}(\boldsymbol{\theta}_t) := r_l(\boldsymbol{\theta}_t; B_t)\ ,\ \boldsymbol{\mu}_{w,t}^{(b)} := \nabla r_{w,t}^{(b)}(\boldsymbol{\theta}_t)\ ,\ \boldsymbol{\mu}_{l,t}^{(b)} := \nabla r_{l,t}^{(b)}(\boldsymbol{\theta}_t)$$

Since the minibatch reward is a sample-mean estimator, under unbiased minibatch sampling:

$$\mathbb{E}[r_{w,t}^{(b)}(\boldsymbol{\theta}_t)] = r_w(\boldsymbol{\theta}_t),\ \mathbb{E}[r_{l,t}^{(b)}(\boldsymbol{\theta}_t)] = r_l(\boldsymbol{\theta}),\ \mathbb{E}[\boldsymbol{\mu}_{w,t}^{(b)} \mid \mathcal{F}_t] = \boldsymbol{\mu}_{w,t},\ \mathbb{E}[\boldsymbol{\mu}_{l,t}^{(b)} \mid \mathcal{F}_t] = \boldsymbol{\mu}_{l,t}$$

We take $r_w(\boldsymbol{\theta}_0) = r_l(\boldsymbol{\theta}_0) = 0$, corresponding to initialization with the reference policy. We assume that step sizes are chosen to satisfy $\eta_t > 0$, $\sum_{t=1}^{\infty}\eta_t = \infty$, $\sum_{t=1}^{\infty}\eta_t^2 < \infty$. Additionally, we assume:

1. Stochastic gradient with bounded second moment: *For a minibatch gradient* $\mathbf{g}_t^{(b)}$ *and population loss gradient* $\mathbf{g}_t := \nabla\mathcal{L}(\boldsymbol{\theta}_t)$*:*

$$\mathbb{E}\left[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\right] < G^2, \;\; \sup_t \mathbb{E}\|\mathbf{g}_t\|^2 < \infty \tag{18}$$

2. Sufficient, stable population descent for the edited gradient: *There exists a sequence* $c_t \geq c_{\min} > 0$ *and a constant* $C < \infty$ *such that:*

$$\mathbb{E}\left[\langle \mathbf{g}_t, \hat{\mathbf{g}}_t^{(b)}\rangle \mid \mathcal{F}_t\right] \geq c_t\|\mathbf{g}_t\|^2, \;\; \mathbb{E}\left[\|\hat{\mathbf{g}}_t^{(b)}\|^2 \mid \mathcal{F}_t\right] \leq C^2\mathbb{E}\left[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\right] \tag{19}$$

3. Population safety constraints on the update in conditional expectation: *There exists a nonnegative* $\mathcal{F}_t$*-measurable sequence* $\bar{\gamma}_t \geq 0$ *such that*

$$\mathbb{E}[\langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \geq \bar{\gamma}_t, \qquad \mathbb{E}[\langle \boldsymbol{\mu}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \leq -\bar{\gamma}_t. \tag{20}$$

4. First-order dominance over second-order effects: *The population reward functions* r_w *and* r_l *are differentiable with Lipschitz gradients, with constants* L_w *and* L_l*, respectively. Assume moreover that, for every* $t \geq 0$*,*

$$\mathbb{E}[\bar{\gamma}_t] \geq \frac{\max\{L_w, L_l\}}{2}\mathbb{E}\|\Delta\boldsymbol{\theta}_t\|^2 \tag{21}$$

Then, we have the following consequences:

a. Loss and gradient convergence:

$$\mathcal{L}(\boldsymbol{\theta}_t) \xrightarrow{a.s.} L_\infty < \infty, \quad \liminf_{t\to\infty} \mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right] = 0 \tag{22}$$

b. Positive reward for the winner and negative reward for the loser: *For every* $t \geq 0$*,*

$$\mathbb{E}\left[r_w(\boldsymbol{\theta}_{t+1})\right] \geq \mathbb{E}\left[r_w(\boldsymbol{\theta}_t)\right], \qquad \mathbb{E}\left[r_l(\boldsymbol{\theta}_{t+1})\right] \leq \mathbb{E}\left[r_l(\boldsymbol{\theta}_t)\right] \tag{23}$$

The theorem is deliberately stated at the population level, and therefore does not by itself resolve the central difficulty of minibatch-defined edits: the descent and safety conditions available during training are typically verified only on the sampled minibatch, while the theorem requires corresponding population conditions in conditional expectation. The propositions below address this gap. They provide sufficient conditions under which minibatch alignment and minibatch safety constraints transfer to the population-level descent and safety assumptions of the theorem, with explicit dependence on the noise-to-signal ratios and update size. In this way, they connect the implementable minibatch rules used by the algorithm to the abstract population guarantees established above.

Proposition 4.1 (Sufficient condition for population descent (Assumption 2))**.** *Let* $\mathbf{g}_t^{(b)} = \mathbf{g}_t + \boldsymbol{\zeta}_t$*, where* $\boldsymbol{\zeta}_t$ *represents minibatch gradient noise with* $\mathbb{E}[\boldsymbol{\zeta}_t \mid \mathcal{F}_t] = 0$*. Assume that we can bound noise variance relative to signal as follows:*

$$\exists\ \rho_t \geq 0 \quad \textit{such that} \quad \mathbb{E}\left[\|\boldsymbol{\zeta}_t\|^2 \mid \mathcal{F}_t\right] \leq \rho_t^2\|\mathbf{g}_t\|^2 \tag{24}$$

Let $\hat{\mathbf{g}}_t^{(b)}$ *be the edited minibatch gradient, and suppose, for* $C \geq c > 0$ *and* $C < \infty$*:*

$$\langle \mathbf{g}_t^{(b)}, \hat{\mathbf{g}}_t^{(b)}\rangle \geq c\|\mathbf{g}_t^{(b)}\|^2, \qquad \|\hat{\mathbf{g}}_t^{(b)}\| \leq C\|\mathbf{g}_t^{(b)}\|, \;\; a.s. \tag{25}$$

Then, if $\rho_t \leq \rho_{\max}$*, Assumption 2 of Theorem 4.1 is satisfied when* $\frac{1}{\rho_{\max}^2} > \left(\frac{C}{c}\right)^2 - 1$*.*

Since $\frac{1}{\rho_t^2}$ is a lower bound on the signal-to-noise ratio (SNR), Proposition 4.1 says that noisier gradients can be tolerated if the edit is both strongly aligned (large c) and tightly norm-controlled (small C); if alignment is weak or norm inflation is large, higher SNR is needed for population descent to survive. When $C = c$, there is no gap between alignment quality and norm inflation, and the edit is maximally efficient. For the practical projection rule of the previous section, the upper bound on projected norm required in Proposition 4.1 is ensured by Corollary 3.1. Likewise, the alignment condition in Proposition 4.1 is certified by Corollary 3.2.

Proposition 4.2 (Sufficient condition for population safety (Assumption 3))**.** *Let* $\boldsymbol{\mu}_{w,t}^{(b)} = \boldsymbol{\mu}_{w,t} + \boldsymbol{\xi}_{w,t}$*,* $\boldsymbol{\mu}_{l,t}^{(b)} = \boldsymbol{\mu}_{l,t} + \boldsymbol{\xi}_{l,t}$ *where* $\boldsymbol{\xi}_{w,t}, \boldsymbol{\xi}_{l,t}$ *are minibatch noise terms satisfying* $\mathbb{E}[\boldsymbol{\xi}_{w,t} \mid \mathcal{F}_t] = \mathbb{E}[\boldsymbol{\xi}_{l,t} \mid \mathcal{F}_t] = 0$*. Let* $\Delta\boldsymbol{\theta}_t$ *be the update produced from the minibatch, and suppose, for some* $\mathcal{F}_t$*-measurable* $\gamma_t \geq 0$*,*

$$\langle \boldsymbol{\mu}_{w,t}^{(b)} + \boldsymbol{\mu}_{l,t}^{(b)}, \Delta\boldsymbol{\theta}_t\rangle = 0, \quad \langle \boldsymbol{\mu}_{w,t}^{(b)}, \Delta\boldsymbol{\theta}_t\rangle \geq \gamma_t, \qquad \langle \boldsymbol{\mu}_{l,t}^{(b)}, \Delta\boldsymbol{\theta}_t\rangle \leq -\gamma_t \qquad a.s. \tag{26}$$

Assume moreover that there exist $\rho_{w,t}, \rho_{l,t} \geq 0$ *and* $S_t \geq 0$ *such that*

$$\mathbb{E}[\|\boldsymbol{\xi}_{w,t}\|^2 \mid \mathcal{F}_t] \leq \rho_{w,t}^2, \quad \mathbb{E}[\|\boldsymbol{\xi}_{l,t}\|^2 \mid \mathcal{F}_t] \leq \rho_{l,t}^2, \text{ and } \mathbb{E}[\|\Delta\boldsymbol{\theta}_t\|^2 \mid \mathcal{F}_t] \leq S_t^2 \tag{27}$$

Then Assumption 3 of Theorem 4.1 is satisfied when $\gamma_t \geq \bar{\gamma}_t + \max\{\rho_{w,t}, \rho_{l,t}\}S_t$.

Proposition 4.2 gives sufficient noise-control requirements for safety. The required minibatch safety margin grows not only with the noise scale in the reward-gradient estimates, but also with the update size, because the perturbation to the first-order reward change enters through inner products of the form $\langle \boldsymbol{\xi}, \Delta\boldsymbol{\theta} \rangle$. Thus, larger or noisier updates require a stronger margin for minibatch-safe behavior to transfer to population-safe behavior. This creates a no-free-lunch tradeoff in the safety margin: stronger margins improve robustness to stochastic geometric error, but also push the update farther from the unconstrained direction, increasing optimization bias, norm inflation, and higher-order effects. Under the practical parameterization $\delta = \kappa\|\mathbf{g}\|$ introduced in Corollaries 3.1–3.2, this tradeoff is controlled by κ. Beyond population-level guarantees on expected reward dynamics, an additional variance bound yields a high-probability first-order safety guarantee for a randomly sampled preference example; see Proposition A.1 in the Appendix.

5 GSafePO Stochastic Optimizers

Practical preference optimization typically uses adaptive methods such as RMSProp or Adam, whose preconditioning changes the geometry of the realized parameter update. Following the standard preconditioned-gradient view (Staib et al., 2019), we write $\Delta\boldsymbol{\theta}_t = -\eta_t \mathbf{P}_t \mathbf{v}_t, \ \mathbf{P}_t \succ 0$, where $\mathbf{v}_t$ is the direction before preconditioning and $\mathbf{P}_t$ is the effective preconditioner. For RMSProp, $\mathbf{P}_t = \mathrm{diag}((\sqrt{s_t} + \epsilon)^{-1})$, where s_t is the exponential moving average of elementwise squared gradients; this is the practical diagonal approximation to an inverse-square-root second-moment preconditioner. To recover the Euclidean geometry of Sections 2–3, we introduce the following *bar-space* transformation:

$$\bar{\mathbf{v}}_t = \mathbf{P}_t^{1/2}\mathbf{v}_t, \qquad \bar{\boldsymbol{\mu}}_{w,t} = \mathbf{P}_t^{1/2}\boldsymbol{\mu}_{w,t}, \qquad \bar{\boldsymbol{\mu}}_{l,t} = \mathbf{P}_t^{1/2}\boldsymbol{\mu}_{l,t}, \qquad \bar{\mathbf{g}}_t = \mathbf{P}_t^{1/2}\mathbf{g}_t. \tag{28}$$

Since $\langle \mathbf{a}, \mathbf{P}_t\mathbf{b} \rangle = \langle \mathbf{P}_t^{1/2}\mathbf{a}, \mathbf{P}_t^{1/2}\mathbf{b} \rangle$, the balance, safety, and descent constraints retain exactly the same form in *bar-space*, specifically,

$$\langle \bar{\boldsymbol{\mu}}_{w,t} + \bar{\boldsymbol{\mu}}_{l,t}, \bar{\mathbf{v}}_t \rangle = 0, \quad -\langle \bar{\boldsymbol{\mu}}_{w,t}, \bar{\mathbf{v}}_t \rangle \geq 0, \quad \langle \bar{\boldsymbol{\mu}}_{l,t}, \bar{\mathbf{v}}_t \rangle \geq 0, \quad \langle \bar{\mathbf{g}}_t, \bar{\mathbf{v}}_t \rangle \geq 0. \tag{29}$$

After constructing a safe edited direction $\hat{\bar{\mathbf{v}}}_\mathbf{t}$ in *bar-space*, we map back via

$$\hat{\mathbf{v}}_t = \mathbf{P}_t^{-1/2}\hat{\bar{\mathbf{v}}}_\mathbf{t}, \qquad \boldsymbol{\Delta\theta_t} = -\eta_\mathbf{t}\mathbf{P_t}\hat{\mathbf{v}}_\mathbf{t} = -\eta_\mathbf{t}\mathbf{P}_\mathbf{t}^{\mathbf{1/2}}\hat{\bar{\mathbf{v}}}_\mathbf{t}. \tag{30}$$

Thus, the same geometric construction applies to adaptive optimizers after a change of metric. Appendix C gives pseudocode for GSafeRMSProp and GSafeAdam, together with the conditions under which the results of Section 4 extend to the adaptive setting.

6 Experiments

We evaluate GSafePO across three model configurations (0.5B, 1B, and 8B-LoRA) and three standard preference datasets. For compactness, the main paper reports the full method comparison on an SFT checkpoint of `Llama-3.2-1B`; results for the remaining model configurations, together with training, evaluation, and implementation details, as well as additional experiments and ablations are provided in Appendix E.

Table 1 compares GSafeRMSProp against standard RMSProp across several preference-learning objectives. Base methods include DPO (Rafailov et al., 2023) and reference-free SimPO (Meng et al., 2024). Among methods designed to combat displacement, we compare against DPO-Positive (DPOP) (Pal et al., 2024), RC-DPO (Chen et al., 2026) and Cal-DPO (Xiao et al., 2024). For each objective, the "-S" variant replaces the underlying optimizer with GSafeRMSProp while leaving the preference loss unchanged. We report the change in unscaled average chosen and rejected rewards relative to initialization, denoted by r^+ and r^-, together with preference win rate (WR), under both in-distribution (ID) and, where applicable, out-of-distribution (OOD) evaluation. This comparison

Method	Train	UltraFeedback $r^+ \uparrow$	UltraFeedback $r^- \downarrow$	UltraFeedback WR $\uparrow$	Gen. H2H WR $\uparrow$	RewardBench-v2 $r^+ \uparrow$	RewardBench-v2 $r^- \downarrow$	RewardBench-v2 WR $\uparrow$	MMLU-Pro $r^+ \uparrow$	MMLU-Pro $r^- \downarrow$	MMLU-Pro WR $\uparrow$
DPO	ID	-12.278	-14.634	0.667	0.330	-3.648	-6.555	0.771	-0.547	-1.382	0.699
DPO-S	ID	0.698	-2.780	**0.717**	**0.670**	0.354	-1.183	**0.772**	0.083	-0.182	**0.701**
DPO	OOD			–		-0.506	-0.586	0.527	-0.416	-0.426	0.520
DPO-S	OOD			–		0.272	-0.346	**0.532**	0.001	-0.002	**0.528**
SimPO	ID	-39.25	-49.36	0.745	0.205	-34.212	-41.481	0.734	-8.261	-11.195	0.711
SimPO-S	ID	0.59	-3.149	**0.746**	**0.795**	0.35	-1.14	**0.739**	1.02	-1.65	**0.729**
SimPO	OOD			–		-34.749	-35.663	0.564	-30.253	-30.564	0.541
SimPO-S	OOD			–		0.033	-0.065	**0.569**	-5.653	-5.670	**0.544**
RC DPO	ID	-0.061	-0.280	0.642	0.300	-1.998	-3.121	0.768	-0.503	-0.902	0.604
RC DPO-S	ID	0.017	-0.018	**0.710**	**0.700**	1.143	-0.483	**0.773**	0.204	-0.183	**0.699**
RC DPO	OOD			–		-0.101	-0.125	0.527	-0.247	-0.249	0.525
RC DPO-S	OOD			–		0.056	-0.060	**0.549**	0.044	-0.025	**0.561**
DPOP	ID	8.200	6.090	0.650	0.497	9.115	4.508	0.710	4.983	4.177	0.676
DPOP-S	ID	4.637	-1.527	**0.673**	**0.503**	1.080	-0.205	**0.758**	1.273	-3.129	**0.687**
DPOP	OOD			–		1.997	1.930	0.524	0.192	0.191	0.502
DPOP-S	OOD			–		0.020	-0.018	**0.527**	0.015	-0.002	**0.528**
Cal-DPO	ID	0.069	-0.067	0.707	0.462	0.079	-0.079	0.761	0.031	-0.034	**0.688**
Cal-DPO-S	ID	0.400	-0.809	**0.724**	**0.538**	1.813	-3.632	**0.776**	0.219	-0.301	0.687
Cal-DPO	OOD			–		0.148	0.167	0.511	-0.132	-0.136	0.521
Cal-DPO-S	OOD			–		0.015	-0.005	**0.537**	0.001	-0.015	**0.534**

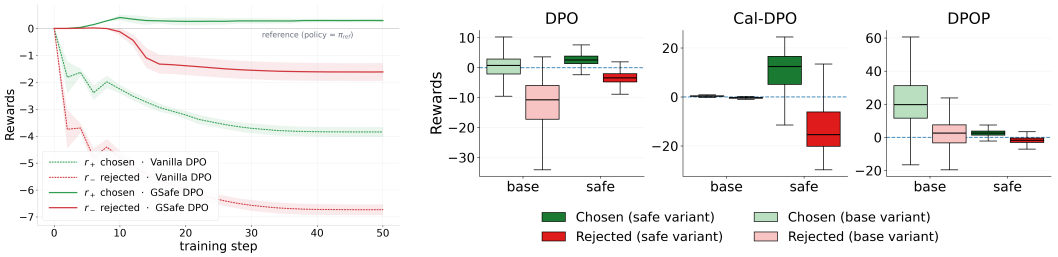
Table 1: Results on `Llama-3.2-1B`. ID rows use the dataset's own eval split; OOD rows use UltraFeedback training and are evaluated on RewardBench-v2 and MMLU-Pro. Gen. H2H WR is evaluated on UltraFeedback eval split. Red entries indicate violations of the desired chosen-up/rejected-down behavior.

therefore tests whether the proposed optimizer can impose the desired reward dynamics independently of how the underlying preference objective itself is constructed.

We also directly evaluate the quality of responses generated by the final policies. For each ID baseline/GSAFERMSPROP pair, we generate responses to the same `UltraFeedback` eval split prompts and compare them head-to-head using LLAMA-3.1-70B-INSTRUCT as a judge. The *Generation H2H WR* column reports the complementary win rates of the two models in each comparison; thus the baseline and its "-S" variant sum to one. This evaluation is deliberately pairwise: it tests whether changing the optimizer improves or degrades the generations produced by an otherwise identical preference objective, rather than relying solely on reward-space diagnostics.

We see that across objectives, high preference win rates can be achieved *without* inducing the desired chosen-up/rejected-down reward dynamics. DPO and SimPO may improve the preference margin while decreasing both chosen and rejected rewards, whereas DPOP can increase both, protecting the chosen response without preventing rejected-response inflation. Reward-calibrated methods such as RC-DPO and Cal-DPO substantially reduce these effects, but act through the loss and do not explicitly constrain the geometry of the realized optimizer update. In contrast, the corresponding GSAFERMSPROP variants consistently steer reward dynamics toward chosen-up/rejected-down behavior while preserving competitive preference WR, including under distribution shift. Figure 2(a) shows that this effect persists throughout training rather than arising only at the final checkpoint, while the final reward distributions (Figure 2(b)) show a corresponding shift toward positive chosen and negative rejected rewards. Projection diagnostics provide empirical support for the conditions underlying Corollaries 3.1 and 3.2: despite nontrivial edits, $\langle g, \hat{g}\rangle/\|g\|^2$ and $\|\hat{g}\|/\|g\|$ remain close to one over training, while $\sin^2\phi$ indicates that the winner-loser gradient geometry is rarely near degeneracy (Appendix E.11).

Improvement is also visible at the level of individual preference pairs. As shown in Table 2(a), GSAFERMSPROP increases the fraction of reward-safe $(+,-)$ examples from 5.6% to 66.2% for DPO, 45.3% to 100% for Cal-DPO, and 13.7% to 65.6% for DPOP. The dominant baseline



(a) Reward trajectories on the eval set.

(b) Distribution of final rewards.

Figure 2: Reward dynamics and reward distribution on `RewardBench-v2`.

(a) Reward-sign percentage

Method	$(+,-)$	$(+,+)$	$(-,-)$	$(-,+)$
DPO	5.6	0.0	94.4	0.0
DPO-S	**66.2**	15.6	14.8	3.5
Cal-DPO	45.3	29.2	17.2	8.3
Cal-DPO-S	**100.0**	0.0	0.0	0.0
DPOP	13.7	83.4	1.1	1.9
DPOP-S	**65.6**	16.9	14.3	3.2
RC DPO	7.7	1.9	89.3	1.1
RC DPO-S	**46.5**	19.1	28.6	5.8

(b) OOD Gen. H2H WR
Cal-DPO-S vs. Cal-DPO

Model	**RB-v2**	**MMLU-Pro**
Qwen-0.5B	**54.5**	**57.6**
Llama-1B	**60.0**	**59.5**
Llama-8B-LoRA	**59.0**	**60.2**

Table 2: Reward-safety and OOD generation diagnostics. (a) Percentage of held-out REWARDBENCH-V2 examples in each reward-sign category for `Llama-3.2-1B`; $(+,-)$ is reward-safe. (b) OOD generation H2H win rate of Cal-DPO-S against Cal-DPO; values above 50% favor Cal-DPO-S. Generations are judged by LLAMA-3.1-70B-INSTRUCT with randomized presentation order.

failure modes differ: DPO and RC DPO are overwhelmingly characterized by $(-,-)$ likelihood displacement, whereas DPOP is dominated by $(+,+)$ rejected-reward inflation; GSAFERMSPROP sharply reduces both.

The effect on generation quality depends on the underlying reward dynamics. For DPO, RC-DPO, and SimPO, which exhibit chosen-response likelihood displacement, GSAFERMSPROP yields substantial ID head-to-head generation gains. For DPOP and Cal-DPO, ID generation quality is approximately preserved while their undesirable reward dynamics are corrected. Importantly, the benefit of combining reward calibration with update-level safety becomes more apparent under distribution shift: Table 2(b) shows that Cal-DPO-S outperforms Cal-DPO in OOD generation across all three model configurations and both evaluation datasets, with H2H win rates of 54.5–60.2%.

A long-horizon experiment in Appendix E.10 probes reward over-optimization by tracking win rate against policy divergence. Both DPO and DPO-S eventually turn over (Rafailov et al., 2024a), but DPO-S reaches higher quality at lower divergence and degrades more gradually, indicating a better quality–drift tradeoff.

7 CONCLUSION

We introduced GSAFEPO, an optimizer framework that controls first-order chosen/rejected reward dynamics through the geometry of the parameter update while leaving the underlying preference objective unchanged. Across preference objectives and model configurations, GSAFEPO consistently induces chosen-up/rejected-down dynamics while preserving preference performance, and these improved reward dynamics translate into improved generation quality, with substantial head-to-head gains for several objectives. Our guarantees are first-order and primarily population or expectation-based, and therefore do not rule out finite-sample or higher-order violations. GSAFEPO also incurs additional computation and memory from estimating reward-gradient geometry; detailed overhead measurements are given in Appendix E.4. Proofs of all theoretical results are provided in Appendix A.

AI USE STATEMENT

In this work, we used generative AI tools to assist with developing and/or refining theoretical formulations, formulating and/or verifying mathematical arguments, checking and improving proof arguments, refining hypotheses, designing and critiquing experimental methodology, interpreting experimental results, and reviewing aspects of the implementation. We did not use generative AI to generate synthetic datasets, translate research material, clean or reformat datasets, or perform qualitative or thematic data analysis; the remaining required-disclosure tasks were not applicable. We additionally used generative AI tools for literature search and synthesis, brainstorming, identifying related work, suggesting experimental parameters and diagnostics, reviewing and editing code, drafting and editing portions of the manuscript, improving readability and organization, formatting references, and proposing titles and keywords.

All AI-assisted theoretical arguments, mathematical derivations, experimental conclusions, code changes, and literature claims were independently reviewed by the authors; mathematical results were checked against their stated assumptions, code was tested independently certified by at least 3 authors, and cited literature was verified against the original sources. The authors take responsibility for the final content of the work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We do not identify any specific ethical concerns arising from this work beyond those generally associated with research on large language models and preference optimization. The work involves no human-subject studies or collection of new personal data.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce both the theoretical and empirical results. The main paper specifies the GSAFEPO formulation and optimizer construction, while the appendix contains complete proofs, supporting derivations, and details of the stochastic minibatch analysis. The experimental appendix documents datasets, model configurations, preprocessing, training and evaluation protocols, hyperparameters, implementation choices, computational overheads, ablations, and additional results, including reward-dynamics, generation-quality, OOD, and long-horizon reward-overoptimization evaluations. Source code will be released upon acceptance.

REFERENCES

Gheshlaghi Azar, Mohammad, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455, 02–04 May 2024.

Debangshu Banerjee and Aditya Gopalan. Towards Reliable Alignment: Uncertainty-aware RLHF, 2024. URL https://arxiv.org/abs/2410.23726.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

Wei Chen, Yubing Wu, Junmei Yang, Delu zeng, Qibin Zhao, John Paisley, Min Chen, and Zhou Wang. Towards disentangled preference optimization dynamics: Suppress the loser, preserve the winner. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=TaNH4XiQ6P.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting Language Models with Scaled AI Feedback, 2024. URL https://arxiv.org/abs/2310.01377.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=oEsYs3WRc3.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF Workflow: From Reward Modeling to Online RLHF, 2024. URL https://arxiv.org/abs/2405.07863.

Aditya Gopalan, Sayak Ray Chowdhury, and Debangshu Banerjee. Why DPO is a misspecified estimator and how to fix it. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=btEiAfnLsX.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189. Association for Computational Linguistics, nov 2024. URL https://aclanthology.org/2024.emnlp-main.626/.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer, 2024. URL https://arxiv.org/abs/2405.16436.

Haocheng Luo, Zehang Deng, Thanh-Toan Do, Mehrtash Harandi, Dinh Phung, and Trung Le. Sharpness-aware minimization in logit space efficiently enhances direct preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4mE2FlL66E.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3Tzcot1LKb.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024. URL https://arxiv.org/abs/2402.13228.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4XIKfvNYvx.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=pf4OuJyn4Q.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to Q^*: Your language model is secretly a Q-function. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=kEVcNxtqXk.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uaMSBJDnRv.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tPNHOoZFl9.

Matthew Staib, Sashank Reddi, Satyen Kale, Sanjiv Kumar, and Suvrit Sra. Escaping saddle points with adaptive gradient methods. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/staib19a.html.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data, 2024. URL https://arxiv.org/abs/2404.14367.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In *Forty-first International Conference on Machine Learning*, 2024.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *The 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024b.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. Cal-dpo: Calibrated direct preference optimization for language model alignment. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 114289–114320. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cf8b2205e39f81726a8d828ecbe00ad0-Paper-Conference.pdf.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *ICML*, 2024.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024. URL https://arxiv.org/abs/2401.08417.

Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A common pitfall of margin-based language model alignment: Gradient entanglement. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=YaBiGjuDiC.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing LLM reasoning generalists with preference trees. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=2ea5TNVR0c.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0qSOodKmJaN.

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 Proofs of Theorems

A.1 Proof of Theorem 2.1

Proof. Consider the decomposition of $\mathbf{v} = \mathbf{v}^{\perp} + \mathbf{v}_{\text{span}}$ where $\mathbf{v}_{\text{span}} = -\alpha_w \tilde{\boldsymbol{\mu}}_w + \alpha_l \tilde{\boldsymbol{\mu}}_l$ and $\mathbf{v}^{\perp} \in \text{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}^{\perp}$. Taking inner products with $-\boldsymbol{\mu}_w, \boldsymbol{\mu}_l$ and using bi-orthogonality of the dual basis, we have:

$$\langle \boldsymbol{\mu}_w, \mathbf{v} \rangle = -\alpha_w, \ \langle \boldsymbol{\mu}_l, \mathbf{v} \rangle = \alpha_l \tag{31}$$

Therefore, by the first-order safety conditions in equation (4), if the update v is safe, then $\alpha_w, \alpha_l \geq 0$.

□

A.2 Proof of Proposition 3.1

Proof. Since coefficients in the dual basis are given by Equation (31), requiring $\langle \boldsymbol{\mu}_w + \boldsymbol{\mu}_l, \mathbf{v} \rangle = \mathbf{0}$ is equivalent to requiring $\alpha_w = \alpha_l := \alpha$. The safety conditions require $\alpha \geq \delta$ in dual space. Expanding $\mathbf{g} = \mathbf{g}^{\perp} + \mathbf{g}_{span}$ and $\mathbf{v} = \mathbf{v}^{\perp} + \mathbf{v}_{span}$, we see that

$$\|\mathbf{g} - \mathbf{v}\|^2 = \|\mathbf{g}^{\perp} - \mathbf{v}^{\perp}\|^2 + \|\mathbf{g}_{span} - \mathbf{v}_{span}\|^2 \tag{32}$$

Since all constraints are on the span components, we can always choose $\mathbf{v}^{\perp} = \mathbf{g}^{\perp}$. Therefore we may restrict attention to vectors $\mathbf{v}$ expressed as $\mathbf{v} = \mathbf{g}^{\perp} - \alpha \tilde{\boldsymbol{\mu}}_w + \alpha \tilde{\boldsymbol{\mu}}_l$. Also, this transforms the problem to minimizing $\|\mathbf{g}_{span} - \mathbf{v}_{span}\|^2$. The dual-basis coefficients of $\mathbf{g}_{span}$ and $\mathbf{v}_{span}$ are $\boldsymbol{\beta}$ and $\alpha\ \mathbf{d}$ respectively. Noting that the Gram matrix of the dual basis is $\mathbf{M}^{-1}$, we may express $\|\mathbf{g}_{span} - \mathbf{v}_{span}\|^2 = (\boldsymbol{\beta} - \alpha\ \mathbf{d})^T \mathbf{M}^{-1} (\boldsymbol{\beta} - \alpha\ \mathbf{d})$. Solving for optimal $\hat{\alpha}$ is a matter of setting derivative w.r.t α equal to zero and solving. If the obtained $\alpha < \delta$, the constraint is active and therefore $\hat{\alpha} = \delta$. Therefore the overall optimal projection in the non-degenerate case is:

$$\hat{\mathbf{g}} = \mathbf{g}^{\perp} - \hat{\alpha}\, \tilde{\boldsymbol{\mu}}_w + \hat{\alpha}\, \tilde{\boldsymbol{\mu}}_l, \qquad \hat{\alpha} = \max\left(\delta, \frac{\boldsymbol{\beta}^T \mathbf{M}^{-1} \mathbf{d}}{\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}}\right). \tag{33}$$

□

A.3 Proof of Corollary 3.1

Proof. Let $p := \|\mathbf{g}^{\perp}\|^2$, $r := \mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}$, and $q := \boldsymbol{\beta}^T \mathbf{M}^{-1} \boldsymbol{\beta}$ with $\boldsymbol{\beta} = (-\alpha_w, \alpha_l)^T$. Using the corresponding bi-orthogonal basis representations

$$\hat{\mathbf{g}} = \mathbf{g}^{\perp} - \hat{\alpha} \tilde{\boldsymbol{\mu}}_w + \hat{\alpha} \tilde{\boldsymbol{\mu}}_l = \mathbf{g}^{\perp} + \hat{a}_w \boldsymbol{\mu}_w + \hat{a}_l \boldsymbol{\mu}_l \tag{34}$$

$$\mathbf{g} = \mathbf{g}^{\perp} - \alpha_w \tilde{\boldsymbol{\mu}}_w + \alpha_l \tilde{\boldsymbol{\mu}}_l = \mathbf{g}^{\perp} + a_w \boldsymbol{\mu}_w + a_l \boldsymbol{\mu}_l \tag{35}$$

and using change of basis relations $\mathbf{a} = \mathbf{M^{-1}} \boldsymbol{\beta}$ and $\hat{\mathbf{a}} = \hat{\alpha} \mathbf{M}^{-1} \mathbf{d}$, we can write

$$\|\mathbf{g}\|^2 = \langle \mathbf{g}^{\perp} + a_w \boldsymbol{\mu}_w + a_l \boldsymbol{\mu}_l, \mathbf{g}^{\perp} - \alpha_w \tilde{\boldsymbol{\mu}}_w + \alpha_l \tilde{\boldsymbol{\mu}}_l \rangle = \|\mathbf{g}^{\perp}\|^2 + -\alpha_w a_w + \alpha_l a_l \tag{36}$$

$$= \|\mathbf{g}^{\perp}\|^2 + \boldsymbol{\beta}^T a = \|\mathbf{g}^{\perp}\|^2 + \boldsymbol{\beta}^T \mathbf{M}^{-1} \boldsymbol{\beta} \tag{37}$$

$$= p + q \tag{38}$$

Similarly,

$$\|\hat{\mathbf{g}}\|^2 = \langle \mathbf{g}^{\perp} + \hat{a}_w \boldsymbol{\mu}_w + \hat{a}_l \boldsymbol{\mu}_l, \mathbf{g}^{\perp} - \hat{\alpha} \tilde{\boldsymbol{\mu}}_w + \hat{\alpha} \tilde{\boldsymbol{\mu}}_l \rangle = \|\mathbf{g}^{\perp}\|^2 + -\hat{\alpha} \hat{a}_w + \hat{\alpha} \hat{a}_l \tag{39}$$

$$= \|\mathbf{g}^{\perp}\|^2 + \hat{\alpha} \mathbf{d^T} \hat{\mathbf{a}} = \|\mathbf{g}^{\perp}\|^{\mathbf{2}} + \hat{\alpha}^{\mathbf{2}} \mathbf{d^T M^{-1} d} \tag{40}$$

$$= p + \hat{\alpha}^2 r \tag{41}$$

From the QP solution we have $\hat{\alpha} = \max\left(\delta, \frac{s}{r}\right)$. Applying the Cauchy–Schwarz inequality in the inner product induced by $\mathbf{M}^{-1}$,

$$s^2 = (\boldsymbol{\beta}^T \mathbf{M}^{-1} \mathbf{d})^2 \leq (\boldsymbol{\beta}^T \mathbf{M}^{-1} \boldsymbol{\beta})(\mathbf{d}^T \mathbf{M}^{-1} \mathbf{d}) = qr \tag{42}$$

Therefore we have $\hat{\alpha} \leq \max\left(\delta, \frac{\sqrt{q}}{\sqrt{r}}\right)$.

Further, since $\sqrt{q} \leq \|\mathbf{g}\| = \sqrt{p+q}$, and $\delta \leq \kappa\|\mathbf{g}\|$, we have:

$$\hat{\alpha} \leq \max\left(\delta, \frac{\|\mathbf{g}\|}{\sqrt{r}}\right) \leq \max\left(\kappa, \frac{1}{\sqrt{r}}\right)\|\mathbf{g}\| \tag{43}$$

Since $\mathbf{M} \succ 0$ and symmetric, and using Rayleigh quotient bounds for $\mathbf{M}^{-1}$

$$\frac{2}{\lambda_{\max}(\mathbf{M})} = \lambda_{\min}(\mathbf{M}^{-1})\|\mathbf{d}\|^2 \leq r = \mathbf{d}^T\mathbf{M}^{-1}\mathbf{d} \leq \lambda_{\max}(\mathbf{M}^{-1})\|\mathbf{d}\|^2 = \frac{2}{\lambda_{\min}(\mathbf{M})}, \tag{44}$$

because $\|\mathbf{d}\|^2 = 2$. Therefore,

$$\frac{2}{\overline{\lambda}} \leq \frac{2}{\lambda_{\max}(\mathbf{M})} \leq r \leq \frac{2}{\lambda_{\min}(\mathbf{M})} \leq \frac{2}{\underline{\lambda}} \tag{45}$$

We can now substitute for $\hat{\alpha}$:

$$\|\hat{\mathbf{g}}\|^2 = p + \hat{\alpha}^2 r \tag{46}$$

$$\leq \left[\|g\|^2 + \frac{2}{\underline{\lambda}}\left\{\max\left(\kappa, \sqrt{\frac{\overline{\lambda}}{2}}\right)\|\mathbf{g}\|\right\}^2\right] \tag{47}$$

which yields:

$$\|\hat{\mathbf{g}}\| \leq \sqrt{1 + \frac{2}{\underline{\lambda}}\max\left\{\kappa, \sqrt{\frac{\overline{\lambda}}{2}}\right\}^2}\|\mathbf{g}\| \tag{48}$$

□

A.4 Proof of Corollary 3.2

Proof. Let $p := \|\mathbf{g}^{\perp}\|^2$, $s := \boldsymbol{\beta}^T\mathbf{M}^{-1}\mathbf{d}$ with $\boldsymbol{\beta} = (-\alpha_w, \alpha_l)^T$, $\mathbf{d} = (-1, 1)^T$ and $\mathbf{a} = (a_w, a_l)^T$. Using the corresponding bi-orthogonal basis representations

$$\hat{\mathbf{g}} = \mathbf{g}^{\perp} - \hat{\alpha}\tilde{\boldsymbol{\mu}}_w + \hat{\alpha}\tilde{\boldsymbol{\mu}}_l = \mathbf{g}^{\perp} + \hat{a}_w\boldsymbol{\mu}_w + \hat{a}_l\boldsymbol{\mu}_l \tag{49}$$

$$\mathbf{g} = \mathbf{g}^{\perp} - \alpha_w\tilde{\boldsymbol{\mu}}_w + \alpha_l\tilde{\boldsymbol{\mu}}_l = \mathbf{g}^{\perp} + a_w\boldsymbol{\mu}_w + a_l\boldsymbol{\mu}_l \tag{50}$$

and using change of basis relations $\mathbf{a} = \mathbf{M}^{-1}\boldsymbol{\beta}$ and $\hat{\mathbf{a}} = \hat{\alpha}\mathbf{M}^{-1}\mathbf{d}$, we can write

$$\langle \mathbf{g}, \hat{\mathbf{g}} \rangle = \langle \mathbf{g}^{\perp} + a_w\boldsymbol{\mu}_w + a_l\boldsymbol{\mu}_l, \mathbf{g}^{\perp} - \hat{\alpha}\tilde{\boldsymbol{\mu}}_w + \hat{\alpha}\tilde{\boldsymbol{\mu}}_l \rangle = \|\mathbf{g}^{\perp}\|^2 + -\hat{\alpha}a_w + \hat{\alpha}a_l \tag{51}$$

$$= \|\mathbf{g}^{\perp}\|^2 + \hat{\alpha}\boldsymbol{\beta}^T\mathbf{M}^{-1}\mathbf{d} \tag{52}$$

$$= p + \hat{\alpha}s \tag{53}$$

This implies that

$$\langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq 0 \iff \hat{\alpha}s \geq -p \tag{54}$$

Now, we will prove $\langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq 0 \iff \delta s \geq -p$.

$\langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq 0 \implies \delta s \geq -p$ follows from assuming $\delta s < -p$, which implies $s < 0$, so:

$$\hat{\alpha} = \max\left(\delta, \frac{s}{r}\right) = \delta \implies \hat{\alpha}s < -p \implies \langle \mathbf{g}, \hat{\mathbf{g}} \rangle < 0 \tag{55}$$

$\delta s \geq -p \implies \langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq 0$ follows from

$$\delta s \geq -p \implies \hat{\alpha}s \geq -p \implies \langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq 0 \tag{56}$$

For the uniform bound, since $\langle \mathbf{g}, \hat{\mathbf{g}} \rangle = p + \hat{\alpha}s$, the assumptions imply

$$\langle \mathbf{g}, \hat{\mathbf{g}} \rangle \geq \kappa_p\|\mathbf{g}\|^2 + (\kappa\|\mathbf{g}\|)(\kappa_s\|\mathbf{g}\|) = (\kappa_p + \kappa\kappa_s)\|\mathbf{g}\|^2$$

□

A.5 Proof of Theorem 4.1

Proof. Let $\mathbf{g}_t := \nabla \mathcal{L}(\boldsymbol{\theta}_t)$ and recall that $\Delta\boldsymbol{\theta}_t = -\eta_t \hat{\mathbf{g}}_t^{(b)}$

We first prove part (a). Since $\mathcal{L}$ is L-smooth,

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) + \langle \mathbf{g}_t, \Delta\boldsymbol{\theta}_t \rangle + \frac{L}{2}\|\Delta\boldsymbol{\theta}_t\|^2 \tag{57}$$

Substituting $\Delta\boldsymbol{\theta}_t = -\eta_t \hat{\mathbf{g}}_t^{(b)}$ gives

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \left\langle \mathbf{g}_t, \hat{\mathbf{g}}_t^{(b)} \right\rangle + \frac{L}{2}\eta_t^2 \|\hat{\mathbf{g}}_t^{(b)}\|^2 \tag{58}$$

Taking conditional expectation with respect to $\mathcal{F}_t$ and using Assumption 2 and the fact that $\boldsymbol{\theta}_t$ and $\mathbf{g}_t$ are $\mathcal{F}_t$-measurable, we have:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}) \mid \mathcal{F}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t c_t \|\mathbf{g}_t\|^2 + \frac{L}{2}\eta_t^2 \mathbb{E}\Big[\|\hat{\mathbf{g}}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] \tag{59}$$

Again by Assumptions 2 and 1,

$$\mathbb{E}\Big[\|\hat{\mathbf{g}}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] \leq C^2 \mathbb{E}\Big[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] \leq C^2 G^2 \tag{60}$$

Therefore taking expectation over the trajectory:

$$\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{t+1})] \leq \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_t)] - c_{\min}\eta_t \,\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] + \frac{L}{2}C^2G^2\,\eta_t^2 \tag{61}$$

Rearranging, we get

$$c_{\min}\eta_t \,\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_t)] - \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{t+1})] + \frac{L}{2}C^2G^2\,\eta_t^2 \tag{62}$$

Summing from $t = 0$ to $T-1$ and invoking the lower bound on $\mathcal{L}$, $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_t)] \geq \inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) := \underline{L}$ yields

$$c_{\min}\sum_{t=0}^{T-1} \eta_t \,\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_0)] - \underline{L} + \frac{L}{2}C^2G^2\sum_{t=0}^{T-1}\eta_t^2 \tag{63}$$

Since $\sum_t \eta_t^2 < \infty$, we obtain

$$\sum_{t=0}^{\infty} \eta_t \,\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] < \infty \tag{64}$$

Because $\sum_t \eta_t = \infty$, it follows that

$$\liminf_{t\to\infty} \mathbb{E}\big[\|\mathbf{g}_t\|^2\big] = 0 \tag{65}$$

Moreover, since the descent inequality is in the standard Robbins-Siegmund setting, $\mathcal{L}(\boldsymbol{\theta}_t) \xrightarrow{a.s.} L_\infty < \infty$. This proves part (a).

We now prove part (b). Since r_w has L_w-Lipschitz gradient, the descent lemma gives

$$r_w(\boldsymbol{\theta}_{t+1}) \geq r_w(\boldsymbol{\theta}_t) + \langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t \rangle - \frac{L_w}{2}\|\Delta\boldsymbol{\theta}_t\|^2 \tag{66}$$

Taking conditional expectation and using Assumption 3,

$$\mathbb{E}[r_w(\boldsymbol{\theta}_{t+1}) \mid \mathcal{F}_t] \geq r_w(\boldsymbol{\theta}_t) + \bar{\gamma}_t - \frac{L_w}{2}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_t\|^2 \mid \mathcal{F}_t\big] \tag{67}$$

Using the law of total expectation,

$$\mathbb{E}[r_w(\boldsymbol{\theta}_{t+1})] \geq \mathbb{E}[r_w(\boldsymbol{\theta}_t)] + \mathbb{E}[\bar{\gamma}_t] - \frac{L_w}{2}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_t\|^2\big] \tag{68}$$

Re-arranging and using assumption 4,

$$\mathbb{E}[r_w(\boldsymbol{\theta}_{t+1})] - \mathbb{E}[r_w(\boldsymbol{\theta}_t)] \geq \mathbb{E}[\bar{\gamma}_t] - \frac{L_w}{2}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_t\|^2\big] \geq 0 \tag{69}$$

Similarly, since r_l has L_l-Lipschitz gradient,

$$r_l(\boldsymbol{\theta}_{t+1}) \leq r_l(\boldsymbol{\theta}_t) + \langle \boldsymbol{\mu}_{l,t}, \Delta\boldsymbol{\theta}_t \rangle + \frac{L_l}{2}\|\Delta\boldsymbol{\theta}_t\|^2 \tag{70}$$

Taking conditional expectation and using Assumption 3,

$$\mathbb{E}[r_l(\boldsymbol{\theta}_{t+1}) \mid \mathcal{F}_t] \leq r_l(\boldsymbol{\theta}_t) - \bar{\gamma}_t + \frac{L_l}{2}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_t\|^2 \mid \mathcal{F}_t\big] \tag{71}$$

Using the law of total expectation and using Assumption 4,

$$\mathbb{E}[r_l(\boldsymbol{\theta}_{t+1})] - \mathbb{E}[r_l(\boldsymbol{\theta}_t)] \leq -\mathbb{E}[\bar{\gamma}_t] + \frac{L_l}{2}\mathbb{E}\big[\|\Delta\boldsymbol{\theta}_t\|^2\big] \leq 0 \tag{72}$$

This proves part (b). □

A.6 Proof of Proposition 4.1

Proof. Write

$$\mathbf{g}_t^{(b)} = \mathbf{g}_t + \boldsymbol{\zeta}_t, \qquad \mathbb{E}[\boldsymbol{\zeta}_t \mid \mathcal{F}_t] = 0 \tag{73}$$

Since $\mathbf{g}_t$ is $\mathcal{F}_t$-measurable,

$$\mathbb{E}[\mathbf{g}_t^{(b)} \mid \mathcal{F}_t] = \mathbf{g}_t, \qquad \mathbb{E}\Big[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] = \|\mathbf{g}_t\|^2 + \mathbb{E}\big[\|\boldsymbol{\zeta}_t\|^2 \mid \mathcal{F}_t\big] \tag{74}$$

We can write $\mathbb{E}[\|\boldsymbol{\zeta}_t\|^2 \mid \mathcal{F}_t] = v_t^2\|\mathbf{g}_t\|^2$, where $v_t \leq \rho_t$. Therefore,

$$\mathbb{E}\Big[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] = (1 + v_t^2)\|\mathbf{g}_t\|^2. \tag{75}$$

We first prove the descent part of Assumption 2. Using

$$\langle \mathbf{g}_t, \hat{\mathbf{g}}_t^{(b)} \rangle = \langle \mathbf{g}_t^{(b)}, \hat{\mathbf{g}}_t^{(b)} \rangle - \langle \boldsymbol{\zeta}_t, \hat{\mathbf{g}}_t^{(b)} \rangle \tag{76}$$

Applying the pathwise norm lower bound of equation (25) and taking conditional expectation,

$$\mathbb{E}\Big[\langle \mathbf{g}_t, \hat{\mathbf{g}}_t^{(b)} \rangle \mid \mathcal{F}_t\Big] \geq c\,\mathbb{E}\Big[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] - \mathbb{E}\Big[\langle \boldsymbol{\zeta}_t, \hat{\mathbf{g}}_t^{(b)} \rangle \mid \mathcal{F}_t\Big]. \tag{77}$$

By Cauchy–Schwarz and the pathwise norm upper bound in equation (25),

$$\big|\langle \boldsymbol{\zeta}_t, \hat{\mathbf{g}}_t^{(b)} \rangle\big| \leq \|\boldsymbol{\zeta}_t\|\,\|\hat{\mathbf{g}}_t^{(b)}\| \leq C\|\boldsymbol{\zeta}_t\|\,\|\mathbf{g}_t^{(b)}\| \tag{78}$$

Therefore,

$$\mathbb{E}\Big[\langle \mathbf{g}_t, \hat{\mathbf{g}}_t^{(b)} \rangle \mid \mathcal{F}_t\Big] \geq c\,\mathbb{E}\Big[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] - C\,\mathbb{E}\Big[\|\boldsymbol{\zeta}_t\|\,\|\mathbf{g}_t^{(b)}\| \mid \mathcal{F}_t\Big] \tag{79}$$

Applying Cauchy-Schwarz for inner-product $\langle X, Y \rangle := \mathbb{E}[XY|\mathcal{F}_t]$,

$$\mathbb{E}\Big[\|\boldsymbol{\zeta}_t\|\,\|\mathbf{g}_t^{(b)}\| \mid \mathcal{F}_t\Big] \leq \Big(\mathbb{E}[\|\boldsymbol{\zeta}_t\|^2 \mid \mathcal{F}_t]\Big)^{1/2}\Big(\mathbb{E}[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t]\Big)^{1/2} \tag{80}$$

Now use

$$\mathbb{E}[\|\boldsymbol{\zeta}_t\|^2 \mid \mathcal{F}_t] = v_t^2\|\mathbf{g}_t\|^2 \tag{81}$$

and equation (75) to get

$$\mathbb{E}\Big[\|\boldsymbol{\zeta}_t\|\,\|\mathbf{g}_t^{(b)}\| \mid \mathcal{F}_t\Big] \leq v_t\sqrt{1 + v_t^2}\,\|\mathbf{g}_t\|^2 \tag{82}$$

Hence, from equation (79),

$$\mathbb{E}\Big[\langle \mathbf{g}_t, \hat{\mathbf{g}}_t^{(b)} \rangle \mid \mathcal{F}_t\Big] \geq \underbrace{\Big(c(1 + v_t^2) - C\rho_t\sqrt{1 + v_t^2}\Big)}_{:=h(v_t^2)}\|\mathbf{g}_t\|^2 \tag{83}$$

Since $h(v^2)$ is decreasing in v^2 whenever $C \geq c$, $h(v_t^2) \geq h(\rho_{\max}^2)$. Thus Assumption 2 holds with

$$c_{\min} := c(1+\rho_{\max}^2) - C\rho_t\sqrt{1+\rho_{\max}^2} \tag{84}$$

provided $c_{\min} > 0$. This is equivalent to $c(1+\rho_{\max}^2) \geq C\rho_{\max}\sqrt{1+\rho_{\max}^2}$. Squaring both sides yields

$$c^2(1+\rho_{\max}^2) \geq C^2\rho_{\max}^2 \quad \Longleftrightarrow \quad c^2 \geq (C^2-c^2)\rho_{\max}^2 \tag{85}$$

Equivalently,

$$\frac{1}{\rho_{\max}^2} \geq \frac{C^2-c^2}{c^2} = \left(\frac{C}{c}\right)^2 - 1 \tag{86}$$

Next, for the norm-bound part of Assumption 2, using equation (25),

$$\|\hat{\mathbf{g}}_t^{(b)}\|^2 \leq C^2\|\mathbf{g}_t^{(b)}\|^2 \qquad a.s. \tag{87}$$

Taking conditional expectation gives

$$\mathbb{E}\Big[\|\hat{\mathbf{g}}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] \leq C^2\mathbb{E}\Big[\|\mathbf{g}_t^{(b)}\|^2 \mid \mathcal{F}_t\Big] \tag{88}$$

which is exactly the second inequality in Assumption 2.

□

A.7 Proof of Proposition 4.2

Proof. Using

$$\boldsymbol{\mu}_{w,t}^{(b)} = \boldsymbol{\mu}_{w,t} + \boldsymbol{\xi}_{w,t}, \qquad \boldsymbol{\mu}_{l,t}^{(b)} = \boldsymbol{\mu}_{l,t} + \boldsymbol{\xi}_{l,t} \tag{89}$$

we can write

$$\langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle = \Big\langle \boldsymbol{\mu}_{w,t}^{(b)}, \Delta\boldsymbol{\theta}_t\Big\rangle - \langle \boldsymbol{\xi}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \tag{90}$$

Using equation (26) and taking conditional expectation given $\mathcal{F}_t$:

$$\mathbb{E}[\langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \geq \gamma_t - \mathbb{E}[\langle \boldsymbol{\xi}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \tag{91}$$

By conditional Cauchy-Schwarz,

$$|\mathbb{E}[\langle \boldsymbol{\xi}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t]| \leq \Big(\mathbb{E}[\|\boldsymbol{\xi}_{w,t}\|^2 \mid \mathcal{F}_t]\Big)^{1/2}\Big(\mathbb{E}[\|\Delta\boldsymbol{\theta}_t\|^2 \mid \mathcal{F}_t]\Big)^{1/2} \leq \rho_{w,t}S_t \tag{92}$$

Hence

$$\mathbb{E}[\langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \geq \gamma_t - \rho_{w,t}S_t \tag{93}$$

Similarly,

$$\langle \boldsymbol{\mu}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle = \Big\langle \boldsymbol{\mu}_{l,t}^{(b)}, \Delta\boldsymbol{\theta}_t\Big\rangle - \langle \boldsymbol{\xi}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \tag{94}$$

so by equation (26),

$$\mathbb{E}[\langle \boldsymbol{\mu}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \leq -\gamma_t + |\mathbb{E}[\langle \boldsymbol{\xi}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t]| \tag{95}$$

Again by conditional Cauchy-Schwarz,

$$|\mathbb{E}[\langle \boldsymbol{\xi}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t]| \leq \Big(\mathbb{E}[\|\boldsymbol{\xi}_{l,t}\|^2 \mid \mathcal{F}_t]\Big)^{1/2}\Big(\mathbb{E}[\|\Delta\boldsymbol{\theta}_t\|^2 \mid \mathcal{F}_t]\Big)^{1/2} \leq \rho_{l,t}S_t \tag{96}$$

and therefore

$$\mathbb{E}[\langle \boldsymbol{\mu}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \leq -\gamma_t + \rho_{l,t}S_t \tag{97}$$

Now define:

$$\bar{\gamma}_t := \gamma_t - \max\{\rho_{w,t}, \rho_{l,t}\}S_t \tag{98}$$

Then

$$\mathbb{E}[\langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \geq \bar{\gamma}_t, \qquad \mathbb{E}[\langle \boldsymbol{\mu}_{l,t}, \Delta\boldsymbol{\theta}_t\rangle \,|\, \mathcal{F}_t] \leq -\bar{\gamma}_t \tag{99}$$

Thus, Assumption 3 of Theorem 4.1 holds provided that $\bar{\gamma}_t \geq 0$, equivalently

$$\gamma_t \geq \bar{\gamma}_t + \max\{\rho_{w,t}, \rho_{l,t}\}S_t \tag{100}$$

□

A.8 RANDOM-SAMPLE WINNER SAFETY

Proposition A.1 (Random-sample winner safety)**.** *Let B_t represent a minibatch at step t, with elements sampled i.i.d. from $\mathcal{D}$. Let $\tilde{b} \sim \mathcal{D}$ be an independent fresh preference example. Define $X_{w,t}(\tilde{b}) := \langle \boldsymbol{\mu}_{w,t}(\tilde{b}), \Delta\boldsymbol{\theta}_t \rangle$, where $\boldsymbol{\mu}_{w,t}(\tilde{b}) := \nabla_{\boldsymbol{\theta}} r_w(\boldsymbol{\theta}_t; \tilde{b})$ and $\boldsymbol{\mu}_{w,t} := r_w(\boldsymbol{\theta}_t)$. Suppose that $\mathrm{Var}(X_{w,t}(\tilde{b}) \mid \mathcal{F}_t \vee \sigma(B_t)) \le \sigma_{w,t}^2$. Then, for any $\gamma_t \ge 0$ and $\epsilon \in (0,1)$,*

$$\langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t \rangle \ge \gamma_t + \sigma_{w,t}\sqrt{\frac{1-\epsilon}{\epsilon}} \implies \Pr\Big(X_{w,t}(\tilde{b}) \ge \gamma_t \mid \mathcal{F}_t \vee \sigma(B_t)\Big) \ge 1-\epsilon \tag{101}$$

Proposition A.1 shows that population-level winner safety of a minibatch-defined update can transfer to fresh samples from the population. In particular, if the population-average first-order winner improvement induced by the realized minibatch update exceeds the desired per-sample margin by a variance-dependent buffer, then a fresh population example is winner-safe with probability at least $1-\epsilon$. The bound is distribution-free and therefore conservative; its purpose is to establish that sufficiently strong population margins imply random-sample safety without assuming Gaussian reward-gradient fluctuations.

Proof. Conditional on $\mathcal{F}_t \vee \sigma(B_t)$, the update $\Delta\boldsymbol{\theta}_t$ is fixed. Since $\tilde{b}$ is independent of $\mathcal{F}_t \vee \sigma(B_t)$,

$$\mathbb{E}\Big[X_{w,t}(\tilde{b}) \mid \mathcal{F}_t \vee \sigma(B_t)\Big] = \langle \boldsymbol{\mu}_{w,t}, \Delta\boldsymbol{\theta}_t \rangle =: m_{w,t} \tag{102}$$

Let $v_{w,t} := \mathrm{Var}(X_{w,t}(\tilde{b}) \mid \mathcal{F}_t \vee \sigma(B_t))$. For $m_{w,t} > \gamma_t$, and $a > 0$, Cantelli's inequality gives

$$\Pr\Big(X_{w,t}(\tilde{b}) - m_{w,t} \le -a \mid \mathcal{F}_t \vee \sigma(B_t)\Big) \le \frac{v_{w,t}}{v_{w,t} + a^2} \tag{103}$$

Taking $a = m_{w,t} - \gamma_t > 0$, and since $v_{w,t} < \sigma_{w,t}^2$ and since $\frac{v}{v+a^2}$ is increasing in v,

$$\Pr\Big(X_{w,t}(\tilde{b}) < \gamma_t \mid \mathcal{F}_t \vee \sigma(B_t)\Big) \le \frac{\sigma_{w,t}^2}{\sigma_{w,t}^2 + (m_{w,t} - \gamma_t)^2} \tag{104}$$

Since,

$$m_{w,t} - \gamma_t \ge \sigma_{w,t}\sqrt{\frac{1-\epsilon}{\epsilon}}, \implies \frac{\sigma_{w,t}^2}{\sigma_{w,t}^2 + (m_{w,t} - \gamma_t)^2} \le \epsilon \tag{105}$$

Therefore,

$$\Pr\Big(X_{w,t}(\tilde{b}) \ge \gamma_t \mid \mathcal{F}_t \vee \sigma(B_t)\Big) \ge 1-\epsilon \tag{106}$$

□

B APPENDIX B: SAFETY IN THE PRIMAL COEFFICIENT SPACE

In the primal coefficient space, the safe regions are cones in $\mathbb{R}^2$.

Proposition B.1 (Conditions for a safe update direction)**.** *Let $\mathbf{v}$ be an update direction. Assume r_w and r_l are Fréchet differentiable at $\boldsymbol{\theta}$. Let $\boldsymbol{\mu}_w := \nabla_{\boldsymbol{\theta}} r_w(\boldsymbol{\theta})$ and $\boldsymbol{\mu}_l := \nabla_{\boldsymbol{\theta}} r_l(\boldsymbol{\theta})$. Let $\mathbf{v}_{span} = a_w\boldsymbol{\mu}_w + a_l\boldsymbol{\mu}_l$ be the projection of $\mathbf{v}$ onto span$\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$. Then, a necessary condition for $\mathbf{v}$ to be safe is:*

$$\mathbf{e}_w^T\mathbf{M}\mathbf{a} \le 0 \text{ and } \mathbf{e}_l^T\mathbf{M}\mathbf{a} \ge 0 \tag{107}$$

where:

$$\mathbf{M} := \begin{bmatrix} \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_w \rangle & \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \\ \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle & \langle \boldsymbol{\mu}_l, \boldsymbol{\mu}_l \rangle \end{bmatrix}, \ \mathbf{a} := \begin{bmatrix} a_w \\ a_l \end{bmatrix}, \ \mathbf{e}_w := \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \mathbf{e}_l := \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Proof. We start by decomposing $\mathbf{v}$ uniquely into a component $\mathbf{v}_{\mathrm{span}}$ in $\mathrm{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$ and a component $\mathbf{v}^\perp$ orthogonal to $\mathrm{span}\{\boldsymbol{\mu}_w, \boldsymbol{\mu}_l\}$. Thus,

$$\mathbf{v} = \mathbf{v}^\perp + \mathbf{v}_{\mathrm{span}} = \mathbf{v}^\perp + a_w\boldsymbol{\mu}_w + a_l\boldsymbol{\mu}_l \tag{108}$$

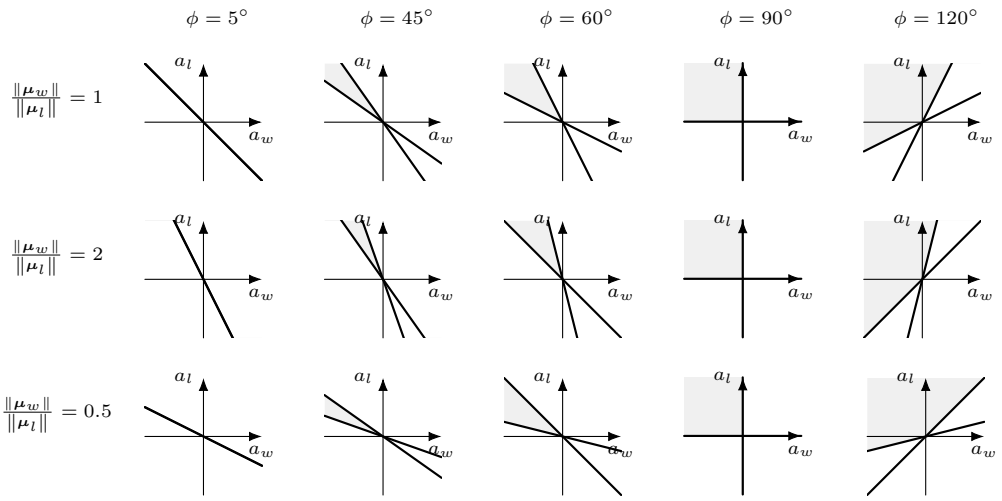


Figure 3: Safe sets in (a_w, a_l)-space for representative angles $\phi \in [0, \pi]$ and different norm ratios.

for some $a_w, a_l \in \mathbb{R}$. Therefore $\mathbf{v}^\perp = \mathbf{v} - a_w\boldsymbol{\mu}_w - a_l\boldsymbol{\mu}_l$. Since $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$ are orthogonal to $\mathbf{v}^\perp$, we have

$$\langle \boldsymbol{\mu}_w, \mathbf{v} - a_w\boldsymbol{\mu}_w - a_l\boldsymbol{\mu}_l \rangle = 0 \tag{109}$$
$$\langle \boldsymbol{\mu}_l, \mathbf{v} - a_w\boldsymbol{\mu}_w - a_l\boldsymbol{\mu}_l \rangle = 0 \tag{110}$$

which gives the 2×2 system of equations

$$\underbrace{\begin{bmatrix} \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_w \rangle & \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \\ \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle & \langle \boldsymbol{\mu}_l, \boldsymbol{\mu}_l \rangle \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} a_w \\ a_l \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} \langle \boldsymbol{\mu}_w, \mathbf{v} \rangle \\ \langle \boldsymbol{\mu}_l, \mathbf{v} \rangle \end{bmatrix}}_{\mathbf{b}} \tag{111}$$

from which we can solve for $\mathbf{a} = \mathbf{M}^{-1}\mathbf{b}$ whenever the geometry is non-degenerate, i.e., the Gram matrix is nonsingular.

The necessary condition for safety from equation (4) requires both $-\langle \boldsymbol{\mu}_w, \mathbf{v} \rangle \geq 0$ and $\langle \boldsymbol{\mu}_l, \mathbf{v} \rangle \geq 0$. The conditions for these to hold are:

$$-\langle \boldsymbol{\mu}_w, \mathbf{v} \rangle = -a_w\|\boldsymbol{\mu}_w\|^2 - a_l\langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \geq 0 \implies a_w\|\boldsymbol{\mu}_w\|^2 + a_l\langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \leq 0 \tag{112}$$
$$\langle \boldsymbol{\mu}_l, \mathbf{v} \rangle = a_w\langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle + a_l\|\boldsymbol{\mu}_l\|^2 \geq 0 \tag{113}$$

Writing these half-plane conditions in Matrix form we have:

$$\mathbf{a}^T\mathbf{M}\mathbf{e}_w \leq 0 \text{ and } \mathbf{a}^T\mathbf{M}\mathbf{e}_l \geq 0, \text{ where } \mathbf{e}_w = (1, 0)^T, \ \mathbf{e}_l = (0, 1)^T \tag{114}$$

Thus $e_w^\top Ma \leq 0$ and $e_l^\top Ma \geq 0$ are exactly the first-order safety half-plane conditions. Strict inequalities give strict first-order safety. □

Proposition B.2 (Safe region depends on angle and relative magnitude between basis vectors)**.** *The condition of Proposition B.1 may be expressed equivalently as conditions on the angle $\phi \in [0, \pi]$ between $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$, whenever the indicated divisions are well defined, via the cosine inequalities*

$$\textit{if } a_l > 0: \ \cos(\phi) \leq -\frac{\|\boldsymbol{\mu}_w\|a_w}{\|\boldsymbol{\mu}_l\|a_l} \quad , \quad \textit{if } a_l < 0: \ \cos(\phi) \geq -\frac{\|\boldsymbol{\mu}_w\|a_w}{\|\boldsymbol{\mu}_l\|a_l}, \tag{115}$$
$$\textit{if } a_w > 0: \ \cos(\phi) \geq -\frac{\|\boldsymbol{\mu}_l\|a_l}{\|\boldsymbol{\mu}_w\|a_w} \quad , \quad \textit{if } a_w < 0: \ \cos(\phi) \leq -\frac{\|\boldsymbol{\mu}_l\|a_l}{\|\boldsymbol{\mu}_w\|a_w}. \tag{116}$$

The edge cases are:

1. *If $a_l = 0$: $a_w < 0$ and $\cos(\phi) \leq 0$, or $a_w = 0$.*

2. *If* $a_w = 0$*:* $a_l > 0$ *and* $\cos(\phi) \leq 0$*, or* $a_l = 0$*.*

3. *If* $\phi = 0$*:* $a_w \|\boldsymbol{\mu}_w\| = -a_l \|\boldsymbol{\mu}_l\|$*.*

4. *If* $\phi = \pi$*:* $a_l \|\boldsymbol{\mu}_l\| \geq a_w \|\boldsymbol{\mu}_w\|$*.*

Proof. The necessary condition for safety from equation (4) requires both $-\langle \boldsymbol{\mu}_w, \mathbf{v} \rangle \geq 0$ and $\langle \boldsymbol{\mu}_l, \mathbf{v} \rangle \geq 0$. The conditions for these to hold are:

$$-\langle \boldsymbol{\mu}_w, \mathbf{v} \rangle = -a_w \|\boldsymbol{\mu}_w\|^2 - a_l \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \geq 0 \implies a_w \|\boldsymbol{\mu}_w\|^2 + a_l \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle \leq 0 \quad (117)$$

$$\langle \boldsymbol{\mu}_l, \mathbf{v} \rangle = a_w \langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle + a_l \|\boldsymbol{\mu}_l\|^2 \geq 0 \quad (118)$$

Since $\langle \boldsymbol{\mu}_w, \boldsymbol{\mu}_l \rangle = \|\boldsymbol{\mu}_w\| \|\boldsymbol{\mu}_l\| \cos(\phi)$ we have the conditions:

$$a_w \|\boldsymbol{\mu}_w\| + a_l \|\boldsymbol{\mu}_l\| \cos(\phi) \leq 0 \ , \ a_w \|\boldsymbol{\mu}_w\| \cos(\phi) + a_l \|\boldsymbol{\mu}_l\| \geq 0 \quad (119)$$

When $a_w > 0$ and $a_l > 0$ these can be rewritten as:

$$\cos(\phi) \leq -\frac{a_w \|\boldsymbol{\mu}_w\|}{a_l \|\boldsymbol{\mu}_l\|} \ , \ \cos(\phi) \geq -\frac{a_l \|\boldsymbol{\mu}_l\|}{a_w \|\boldsymbol{\mu}_w\|} \quad (120)$$

When $a_w < 0$ and $a_l < 0$ the inequalities flip direction, yielding:

$$\cos(\phi) \geq -\frac{a_w \|\boldsymbol{\mu}_w\|}{a_l \|\boldsymbol{\mu}_l\|} \ , \ \cos(\phi) \leq -\frac{a_l \|\boldsymbol{\mu}_l\|}{a_w \|\boldsymbol{\mu}_w\|} \quad (121)$$

The edge cases follow directly from Equation (119).

1. $a_l = 0$ yields:

$$a_w \|\boldsymbol{\mu}_w\| \leq 0 \ , \ a_w \|\boldsymbol{\mu}_w\| \cos(\phi) \geq 0 \implies a_w < 0 \text{ and } \cos(\phi) \leq 0, \text{ or } a_w = 0. \quad (122)$$

2. $a_w = 0$ yields:

$$a_l \|\boldsymbol{\mu}_l\| \cos(\phi) \leq 0 \ , \ a_l \|\boldsymbol{\mu}_l\| \geq 0 \implies a_l > 0 \text{ and } \cos(\phi) \leq 0, \text{ or } a_l = 0. \quad (123)$$

3. $\phi = 0$ yields:

$$a_w \|\boldsymbol{\mu}_w\| + a_l \|\boldsymbol{\mu}_l\| \leq 0 \ , \ a_w \|\boldsymbol{\mu}_w\| + a_l \|\boldsymbol{\mu}_l\| \geq 0 \implies a_w \|\boldsymbol{\mu}_w\| = -a_l \|\boldsymbol{\mu}_l\| \quad (124)$$

4. $\phi = \pi$ yields:

$$a_w \|\boldsymbol{\mu}_w\| - a_l \|\boldsymbol{\mu}_l\| \leq 0 \ , \ -a_w \|\boldsymbol{\mu}_w\| + a_l \|\boldsymbol{\mu}_l\| \geq 0 \implies a_w \|\boldsymbol{\mu}_w\| \leq a_l \|\boldsymbol{\mu}_l\| \quad (125)$$

This proves the proposition. □

Note that in the primal coefficient space, the safe region depends on the angle and relative magnitude of the basis vectors. These quantities can change at each training step because they are determined by the local update geometry. This is why we prefer the dual coefficient space for projection calculations, where the safe region is exactly the first quadrant and is invariant to the angle and relative magnitude of the basis vectors, assuming non-degenerate geometry. Figure 3 shows the safe sets in (a_w, a_l)-space under various combinations of relative gradient magnitudes and angles between $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$. Note how the safe region shrinks toward degeneracy as the angle between the basis vectors goes to zero.

C PSEUDOCODE FOR GSAFEPO OPTIMIZERS

Algorithm 1 GSAFERMSPROP update for $\boldsymbol{\theta} \in \mathbb{R}^N$ with minibatch of size K

Definitions:

$r_{w,k} := r_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{y}_{w,k}), \quad r_{l,k} := r_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{y}_{l,k}), \quad \Delta_k := r_{w,k} - r_{l,k} \quad k = 1, \dots, K$

Inputs:

$\boldsymbol{\mu}_w^{(b)} := \nabla_{\boldsymbol{\theta}} \Big[\frac{1}{K} \sum_{k=1}^{K} r_{w,k} \Big]$, $\boldsymbol{\mu}_l^{(b)} := \nabla_{\boldsymbol{\theta}} \Big[\frac{1}{K} \sum_{k=1}^{K} r_{l,k} \Big]$, $\mathbf{g}^{(b)} := \nabla_{\boldsymbol{\theta}} \Big[-\frac{1}{K} \sum_{k=1}^{K} \log \sigma(\Delta_k) \Big]$

RMSProp state $\mathbf{s} \in \mathbb{R}^N$, decay λ, learning rate η, relative safety margin $\kappa \geq 0$, $\epsilon > 0$, and $\tau > 0$

$\mathbf{s} \leftarrow \lambda \mathbf{s} + (1 - \lambda)\big(\mathbf{g}^{(b)} \odot \mathbf{g}^{(b)}\big)$ ▷ update second moment

$\mathbf{P} \leftarrow \mathrm{diag}(\sqrt{\mathbf{s}} + \epsilon)^{-1}$ ▷ RMSProp preconditioner

$\bar{\boldsymbol{\mu}}_w \leftarrow \mathbf{P}^{1/2} \boldsymbol{\mu}_w^{(b)}$, $\bar{\boldsymbol{\mu}}_l \leftarrow \mathbf{P}^{1/2} \boldsymbol{\mu}_l^{(b)}$, $\bar{\mathbf{g}} \leftarrow \mathbf{P}^{1/2} \mathbf{g}^{(b)}$ ▷ convert to bar space

$A \leftarrow \langle \bar{\boldsymbol{\mu}}_w, \bar{\boldsymbol{\mu}}_w \rangle / N$, $B \leftarrow \langle \bar{\boldsymbol{\mu}}_w, \bar{\boldsymbol{\mu}}_l \rangle / N$, $C \leftarrow \langle \bar{\boldsymbol{\mu}}_l, \bar{\boldsymbol{\mu}}_l \rangle / N$ ▷ Normalized Gram terms

$D \leftarrow AC - B^2$, $G \leftarrow \|\bar{\mathbf{g}}\| / \sqrt{N}$ ▷ $D = N^{-2} |\mathbf{M}|$

$\alpha_w \leftarrow -\langle \bar{\boldsymbol{\mu}}_w, \bar{\mathbf{g}} \rangle / N$, $\alpha_l \leftarrow \langle \bar{\boldsymbol{\mu}}_l, \bar{\mathbf{g}} \rangle / N$ ▷ dual space coefficients

if $D \leq \tau AC$ **then** ▷ guard for degenerate span

  $a_l \leftarrow \alpha_l / (C + \epsilon)$ ▷ line coefficient

  $\hat{\bar{\mathbf{v}}} \leftarrow \bar{\mathbf{g}}^{\perp} \leftarrow \bar{\mathbf{g}} - a_l \bar{\boldsymbol{\mu}}_l$ ▷ safe fallback to orthogonal residual

else

  $a_w \leftarrow \frac{-C\alpha_w - B\alpha_l}{D}$, $a_l \leftarrow \frac{B\alpha_w + A\alpha_l}{D}$ ▷ corresponding primal coefficients

  $\hat{\alpha} \leftarrow \max\Big(\kappa G, \frac{(B+C)\alpha_w + (A+B)\alpha_l}{A+2B+C}\Big)$ ▷ safe dual coefficient from Proposition 3.1

  $\hat{a}_w \leftarrow \frac{-(B+C)\hat{\alpha}}{D}$, $\hat{a}_l \leftarrow \frac{(A+B)\hat{\alpha}}{D}$ ▷ corresponding safe primal coefficients

  $\bar{\mathbf{g}}^{\perp} \leftarrow \bar{\mathbf{g}} - a_w \bar{\boldsymbol{\mu}}_w - a_l \bar{\boldsymbol{\mu}}_l$ ▷ orthogonal component

  $\hat{\bar{\mathbf{v}}} \leftarrow \bar{\mathbf{g}}^{\perp} + \hat{a}_w \bar{\boldsymbol{\mu}}_w + \hat{a}_l \bar{\boldsymbol{\mu}}_l$ ▷ safe hyperplane projection

end if

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{P}^{\frac{1}{2}} \hat{\bar{\mathbf{v}}}$ ▷ merged map back and apply update

Algorithm 2 GSAFEADAM update for $\boldsymbol{\theta} \in \mathbb{R}^N$ with batch size K

Definitions:

$r_{w,k} := r_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{y}_{w,k}), \quad r_{l,k} := r_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{y}_{l,k}), \quad \Delta_k := r_{w,k} - r_{l,k} \quad k = 1, \dots, K$

Inputs:

$\boldsymbol{\mu}_w^{(b)} := \nabla_{\boldsymbol{\theta}} \Big[\frac{1}{K} \sum_{k=1}^{K} r_{w,k} \Big]$, $\boldsymbol{\mu}_l^{(b)} := \nabla_{\boldsymbol{\theta}} \Big[\frac{1}{K} \sum_{k=1}^{K} r_{l,k} \Big]$, $\mathbf{g}^{(b)} := \nabla_{\boldsymbol{\theta}} \Big[-\frac{1}{K} \sum_{k=1}^{K} \log \sigma(\Delta_k) \Big]$

Adam states $\mathbf{m}, \mathbf{s} \in \mathbb{R}^N$, decay λ_1, λ_2, step t, learning rate η, safety margin $\kappa > 0$, $\epsilon > 0$, $\tau > 0$

$\mathbf{m} \leftarrow \lambda_1 \mathbf{m} + (1 - \lambda_1) \mathbf{g}^{(b)}$ ▷ update first moment

$\mathbf{s} \leftarrow \lambda_2 \mathbf{s} + (1 - \lambda_2)\big(\mathbf{g}^{(b)} \odot \mathbf{g}^{(b)}\big)$ ▷ update second moment

$\mathbf{m}' \leftarrow \mathbf{m} / (1 - \lambda_1^t)$ ▷ bias-corrected first moment

$\mathbf{s}' \leftarrow \mathbf{s} / (1 - \lambda_2^t)$ ▷ bias-corrected second moment

$\mathbf{P} \leftarrow \mathrm{diag}(\sqrt{\mathbf{s}'} + \epsilon)^{-1}$ ▷ Adam preconditioner

$\bar{\boldsymbol{\mu}}_w \leftarrow \mathbf{P}^{1/2} \boldsymbol{\mu}_w^{(b)}$, $\bar{\boldsymbol{\mu}}_l \leftarrow \mathbf{P}^{1/2} \boldsymbol{\mu}_l^{(b)}$, $\bar{\mathbf{g}} \leftarrow \mathbf{P}^{1/2} \mathbf{m}'$ ▷ convert to bar space

$A \leftarrow \langle \bar{\boldsymbol{\mu}}_w, \bar{\boldsymbol{\mu}}_w \rangle / N$, $B \leftarrow \langle \bar{\boldsymbol{\mu}}_w, \bar{\boldsymbol{\mu}}_l \rangle / N$, $C \leftarrow \langle \bar{\boldsymbol{\mu}}_l, \bar{\boldsymbol{\mu}}_l \rangle / N$ ▷ Normalized Gram terms

$D \leftarrow AC - B^2$, $G \leftarrow \|\bar{\mathbf{g}}\| / \sqrt{N}$ ▷ $D = N^{-2} |\mathbf{M}|$

$\alpha_w \leftarrow -\langle \bar{\boldsymbol{\mu}}_w, \bar{\mathbf{g}} \rangle / N$, $\alpha_l \leftarrow \langle \bar{\boldsymbol{\mu}}_l, \bar{\mathbf{g}} \rangle / N$ ▷ dual space coefficients

if $D \leq \tau AC$ **then** ▷ guard for degenerate span

  $a_l \leftarrow \alpha_l / (C + \epsilon)$ ▷ line coefficient

  $\hat{\bar{\mathbf{v}}} \leftarrow \bar{\mathbf{g}}^{\perp} \leftarrow \bar{\mathbf{g}} - a_l \bar{\boldsymbol{\mu}}_l$ ▷ safe fallback to orthogonal residual

else

  $a_w \leftarrow \frac{-C\alpha_w - B\alpha_l}{D}$, $a_l \leftarrow \frac{B\alpha_w + A\alpha_l}{D}$ ▷ corresponding primal coefficients

  $\hat{\alpha} \leftarrow \max\Big(\kappa G, \frac{(B+C)\alpha_w + (A+B)\alpha_l}{A+2B+C}\Big)$ ▷ safe dual coefficient from Proposition 3.1

  $\hat{a}_w \leftarrow \frac{-(B+C)\hat{\alpha}}{D}$, $\hat{a}_l \leftarrow \frac{(A+B)\hat{\alpha}}{D}$ ▷ corresponding safe primal coefficients

  $\bar{\mathbf{g}}^{\perp} \leftarrow \bar{\mathbf{g}} - a_w \bar{\boldsymbol{\mu}}_w - a_l \bar{\boldsymbol{\mu}}_l$ ▷ orthogonal component

  $\hat{\bar{\mathbf{v}}} \leftarrow \bar{\mathbf{g}}^{\perp} + \hat{a}_w \bar{\boldsymbol{\mu}}_w + \hat{a}_l \bar{\boldsymbol{\mu}}_l$ ▷ safe hyperplane projection

end if

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{P}^{\frac{1}{2}} \hat{\bar{\mathbf{v}}}$ ▷ merged map back and apply update

Practical Tweaks to GSAFERMSPROP In our experiments we use two tweaks to Algorithm 1, which improved results with SimPO, while also showing comparable performance on the other losses.

Modified floor in line 12. We replace κG in line 12 with $\kappa h G_{span}$, where

$$h = \sqrt{\frac{D}{A + 2B + C}}, \qquad G_{span} = \sqrt{a_l \alpha_l - a_w \alpha_w} = \frac{\|\bar{\mathbf{g}}_{span}\|}{\sqrt{N}},$$

and $\bar{\mathbf{g}}_{span}$ is the projection of $\bar{\mathbf{g}}$ onto $\mathrm{span}(\bar{\boldsymbol{\mu}}_w, \bar{\boldsymbol{\mu}}_l)$. When the floor of the $\max$ is active, the corrected in-plane component then has norm exactly $\kappa\|\bar{\mathbf{g}}_{span}\|$. The floor is thus only applied to the in-plane component of the gradient.

Using moments from the corrected update. After computing $\hat{\tilde{\mathbf{v}}}$, we replace the raw gradient's contribution to the second moment with that of the corrected update $\tilde{\mathbf{g}} = \mathbf{P}^{-1/2}\hat{\tilde{\mathbf{v}}}$:

$$\mathbf{s} \leftarrow \mathbf{s} + (1 - \lambda)\big(\tilde{\mathbf{g}} \odot \tilde{\mathbf{g}} - \mathbf{g}^{(b)} \odot \mathbf{g}^{(b)}\big).$$

Therefore, the preconditioner at each step is built from the corrected updates of previous steps and the raw gradient of the current step, which retains RMSProp's implicit damping of gradient spikes.

In addition to this, and only for the very first step, one can compute $\hat{\tilde{\mathbf{v}}}$ only to perform this replacement, and apply the uncorrected update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\,\mathbf{P}\mathbf{g}^{(b)}$. Empirically, this prevents early-step blowups.

Remark on degeneracy handling. For a small relative threshold $\tau > 0$. Writing ϕ for the angle between $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$, we have

$$B^2 = AC\cos^2\phi \qquad \Longrightarrow \qquad D = AC(1 - \cos^2\phi) = AC\sin^2\phi,$$

so the guard is equivalently triggered when $\sin^2\phi \le \tau$. Thus, τ has a simple geometric interpretation: it sets the minimum angular separation required to trust the two-dimensional winner/loser contrast geometry. In the degenerate branch (line 7), since $\bar{\mu}_w$ and $\bar{\mu}_l$ are nearly collinear, we approximate the collapsed span by the line generated by $\bar{\mu}_l$ and retain the residual orthogonal to this line. In this regime, the ambiguous contrast component is discarded, and the update retains only the part of the step orthogonal to the degenerate winner/loser line, reducing sensitivity to noisy or low-information preference geometry.

Remark on extending theoretical results of Section 4 to adaptive optimizers. Although the change to *bar-space* in Section 5 makes the deterministic projection geometry identical to the vanilla case, the stochastic results of Section 4 require additional assumptions because the adaptive preconditioner and, for Adam, the momentum direction depend on the training trajectory. We additionally require that the adaptive preconditioner remain uniformly well conditioned: its eigenvalues are bounded above and away from zero at every iterate, i.e., $0 < p_{\min} \le \lambda_i(\mathbf{P}_t) \le p_{\max}, \forall i$; this makes the *bar-space* and parameter-space norms uniformly equivalent, so stationarity in bar space implies stationarity of the original population loss. For stochastic minibatch-to-population arguments involving unbiased stochastic gradients, it is sufficient that $\mathbf{P}_t$ be $\mathcal{F}_t$-measurable, which can be realized for RMSProp using a one-step-lagged second-moment preconditioner. Under this condition, conditional unbiasedness and relative noise control are preserved under the bar-space transformation. For momentum-based methods such as Adam, the effective direction is generally biased relative to the current population gradient; the population theorem therefore applies when its descent, norm-control, and reward-safety conditions hold for the effective Adam update, while the sufficient unbiased-gradient condition of Proposition 4.1 is not invoked.

D SCORE CONSTRUCTIONS FOR COMMON DPO VARIANTS

The minimal requirement for applying GSAFEPO is that the objective admit a pairwise score form:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\big[f\big(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l, r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w), r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l)\big)\big] \tag{126}$$

where f is differentiable in the score arguments and $r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$ is a differentiable scalar score induced by the current policy. In that case

$$\boldsymbol{\mu}_w = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{D}}[r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w)], \qquad \boldsymbol{\mu}_l = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathcal{D}}[r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l)] \tag{127}$$

define the local first-order geometry used by GSAFEPO.

Below we list common objectives together with a compatible score choice.

DPO (Rafailov et al., 2023).

$$\mathcal{L}_{\mathrm{DPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathcal{D}}[\log \sigma(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l))]\,, \tag{128}$$

A compatible score is

$$r_{\boldsymbol{\theta}}^{\mathrm{DPO}}(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})}. \tag{129}$$

IPO (Azar et al., 2024).

$$\mathcal{L}_{\mathrm{IPO}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\left[\left(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l) - \frac{1}{2\tau}\right)^2\right], \tag{130}$$

A compatible score is

$$r_{\boldsymbol{\theta}}^{\mathrm{IPO}}(\mathbf{x}, \mathbf{y}) = \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})}. \tag{131}$$

SLiC (Zhao et al., 2023).

$$\mathcal{L}_{\mathrm{SLiC}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}[\max\{0,\, \delta - (r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l))\}]\,, \tag{132}$$

A compatible score is

$$r_{\boldsymbol{\theta}}^{\mathrm{SLiC}}(\mathbf{x}, \mathbf{y}) = \log \pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}). \tag{133}$$

DPO-Positive / DPOP (Pal et al., 2024).

$$\mathcal{L}_{\mathrm{DPOP}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathcal{D}}[\log \sigma(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l) - \lambda \max\{0,\, -r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w)\})] \tag{134}$$

A compatible score is

$$r_{\boldsymbol{\theta}}^{\mathrm{DPOP}}(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})}, \tag{135}$$

Cal-DPO (Xiao et al., 2024).

$$\begin{aligned}\mathcal{L}_{\text{Cal-DPO}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\Bigg[&- \log \sigma\left(\frac{1}{\beta}\big(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l)\big)\right) \\ &+ \left(\frac{1}{\beta} r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - \frac{1}{2\beta}\right)^2 + \left(\frac{1}{\beta} r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l) + \frac{1}{2\beta}\right)^2\Bigg].\end{aligned} \tag{136}$$

A compatible score is

$$r_{\boldsymbol{\theta}}^{\text{Cal-DPO}}(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})}, \tag{137}$$

GPO (Tang et al., 2024).

$$\mathcal{L}_{\mathrm{GPO}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}[g(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l))]\,, \tag{138}$$

where $g : \mathbb{R} \to \mathbb{R}$ is convex and decreasing, with score

$$r_{\boldsymbol{\theta}}^{\mathrm{GPO}}(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x})}. \tag{139}$$

SimPO (Meng et al., 2024).

$$\mathcal{L}_{\mathrm{SimPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathcal{D}}[\log \sigma(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l) - \gamma)]\,, \tag{140}$$

A compatible score is

$$r_{\boldsymbol{\theta}}^{\mathrm{SimPO}}(\mathbf{x}, \mathbf{y}) = \frac{\beta}{|\mathbf{y}|} \log \pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}). \tag{141}$$

ORPO (Hong et al., 2024). ORPO includes an auxiliary chosen-only SFT term together with a pairwise odds-ratio term:

$$\mathcal{L}_{\mathrm{ORPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathcal{D}}[\log P_{\boldsymbol{\theta}}(\mathbf{y}_w \mid \mathbf{x}) + \lambda \log \sigma(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l))], \tag{142}$$

where $P_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) := \exp\big(\frac{1}{|\mathbf{y}|} \log \pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})\big)$. The pairwise score component is

$$r_{\boldsymbol{\theta}}^{\mathrm{ORPO}}(\mathbf{x}, \mathbf{y}) = \log \frac{P_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{1 - P_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}. \tag{143}$$

CPO (Xu et al., 2024). CPO also includes an auxiliary chosen-only regularizer:

$$\mathcal{L}_{\mathrm{CPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathcal{D}}[\log \sigma(r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_w) - r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}_l)) - \log \pi_{\boldsymbol{\theta}}(\mathbf{y}_w \mid \mathbf{x})], \tag{144}$$

with pairwise score component

$$r_{\boldsymbol{\theta}}^{\mathrm{CPO}}(\mathbf{x}, \mathbf{y}) = \beta \log \pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}). \tag{145}$$

Remark. All of the variants listed above fit into the general score form in Equation (126), provided one chooses a scalar score $r_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$ sufficient to express the objective. In some methods, such as ORPO and CPO, the objective contains an additional chosen-only regularization term; this is still covered by Equation (126).

E Experiments and Ablations

Datasets We conduct evaluations on three benchmark datasets: RewardBench v2, MMLU-Pro, and UltraFeedback. RewardBench v2 (Malik et al., 2025) is a multi-skill reward modeling benchmark designed to bring new, challenging data for accuracy-based reward model evaluation. The dataset contains around $1.87K$ prompts across categories such as factuality, precise instruction following, and focus, each containing a chosen and a rejected response. MMLU-Pro (Wang et al., 2024b) is a robust, challenging, and well-suited massive multi-task understanding dataset for rigorously benchmarking large language models' capabilities. It contains around $12K$ complex questions across fields such as math, law, chemistry, business, history, and psychology. For each question, there is a list of 10 answer options together with the correct answer. We use UltraFeedback (Cui et al., 2024) as our final dataset. Specifically, we use the pre-processed and binarized version of UltraFeedback as presented by Dong et al. (2024), which has been shown to generate higher quality reward models (Dong et al., 2024; Wang et al., 2024a; Xiong et al., 2024; Banerjee & Gopalan, 2024). RewardBench v2 is a preference dataset that provides chosen and rejected responses. MMLU-Pro is not a preference dataset by default, so we convert it into one by treating the correct answer as the chosen response and any incorrect responses as rejected responses, following Gopalan et al. (2026).

Methodology Each base model is first instruction-tuned via supervised fine-tuning (SFT) on UltraChat-200k (Ding et al., 2023)[2] (`train_sft`/`test_sft` splits); this SFT checkpoint is the *reference policy* π_{ref} against which all reward shifts are measured, and the starting point for every alignment method compared in this work.

For the in-distribution (ID) setup, each preference dataset is split into 80/20 training and evaluation subsets. Models are trained on the training split and ID results are reported on its held-out evaluation split. For the out-of-distribution (OOD) setup, the reference policy policy is trained on UltraFeedback and evaluated OOD on the entire RewardBench-v2 and MMLU-Pro datasets. This measures whether alignment gains transfer across preference distributions rather than reflecting dataset-specific overfitting.

Results averaged over 5 random seeds are reported; error bars in the figures denote ± 2 standard deviation, and dashed lines indicate the per-panel base policy (π_{ref}). For base methods, we adopt hyperparameters from the original works or their official GitHub repositories (see Table 3).

We measure (i) the average change in chosen and rejected rewards relative to the reference policy, and (ii) win rates (WR), across models and datasets, both ID and OOD, including their safe variants. Experiments are conducted on Qwen-2.5-0.5B (Team, 2024), Llama-3.2-1B, and Llama-3.1-8B-LoRA, demonstrating consistent behavior across model configurations.

[2] `https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k`

Integration details We implement GSAFERMSPROP and GSAFEADAM as extensions of `torch`'s RMSprop and Adam optimizers (see Section C for detailed pseudocode). Additionally, we introduce a custom trainer, SAFETRAINER, extending `TRL`'s DPOTRAINER, to compute and store $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_l$ required by the optimizer step. Computing these quantities requires two backward passes with `retain_graph=True` in `torch.autograd`. The loss gradient $\mathbf{g}^{(b)}$ is automatically computed and made available to the optimizer step when `loss.backward` is executed at each step.

E.1 EXPERIMENTAL RESULTS: `QWEN-2.5-0.5B`

		UltraFeedback				RewardBench-v2			MMLU-Pro		
Method	**Train**	$r^+ \uparrow$	$r^- \downarrow$	WR $\uparrow$	**Gen. H2H WR** $\uparrow$	$r^+ \uparrow$	$r^- \downarrow$	WR $\uparrow$	$r^+ \uparrow$	$r^- \downarrow$	WR $\uparrow$
DPO	ID	-0.811	-1.997	**0.719**	0.395	-1.131	-6.759	**0.787**	-0.629	-1.608	**0.703**
DPO-S	ID	0.019	-0.253	0.712	**0.605**	0.570	-1.791	**0.787**	0.123	-0.734	0.691
DPO	OOD			–		-0.084	-0.116	**0.521**	-2.182	-2.190	**0.523**
DPO-S	OOD			–		0.012	-0.213	0.515	0.001	-0.002	0.520
SimPO	ID	-16.318	-20.135	**0.749**	0.325	-22.456	-31.013	**0.787**	-6.989	-9.710	**0.706**
SimPO-S	ID	0.01	-0.163	0.738	**0.675**	1.02	-3.536	0.784	0.011	-0.091	0.678
SimPO	OOD			–		-0.004	-0.153	**0.545**	-25.904	-26.127	0.523
SimPO-S	OOD			–		0.001	-0.015	**0.545**	-0.150	-0.189	**0.548**
RC DPO	ID	-0.137	-0.311	**709**	0.480	-1.118	-1.809	0.739	-0.279	-0.445	0.666
RC DPO-S	ID	0.003	-0.002	0.703	**0.520**	0.007	-0.164	**0.761**	0.116	-0.217	**0.672**
RC DPO	OOD			–		-0.098	-0.100	0.519	-0.409	-0.411	0.528
RC DPO-S	OOD			–		0.173	-0.125	**0.523**	0.017	-0.022	**0.536**
DPOP	ID	0.970	0.125	**0.721**	0.460	1.787	0.548	0.761	2.978	2.075	0.697
DPOP-S	ID	0.016	-0.194	0.711	**0.540**	0.092	-2.216	**0.782**	0.083	-0.005	**0.699**
DPOP	OOD			–		0.985	0.983	0.521	2.119	2.088	0.525
DPOP-S	OOD			–		0.015	-0.003	**0.552**	0.008	-0.006	**0.534**
Cal-DPO	ID	0.069	-0.067	0.706	0.490	0.108	-0.068	0.774	0.061	-0.053	0.688
Cal-DPO-S	ID	0.023	-0.094	**0.724**	**0.510**	1.763	-1.540	**0.779**	0.645	-1.005	**0.695**
Cal-DPO	OOD			–		0.023	-0.012	0.551	-0.081	-0.082	0.517
Cal-DPO-S	OOD			–		0.122	-0.015	**0.562**	0.003	-0.005	**0.524**

E.2 EXPERIMENTAL RESULTS: `LLAMA-3.1-8B-LORA`

		UltraFeedback				RewardBench-v2			MMLU-Pro		
Method	**Train**	$r^+ \uparrow$	$r^- \downarrow$	WR $\uparrow$	**Gen. H2H WR** $\uparrow$	$r^+ \uparrow$	$r^- \downarrow$	WR $\uparrow$	$r^+ \uparrow$	$r^- \downarrow$	WR $\uparrow$
DPO	ID	0.141	0.006	0.697	0.347	0.001	-0.570	0.761	-0.006	-0.096	0.719
DPO-S	ID	0.157	-0.449	**0.729**	**0.653**	0.709	-0.538	**0.793**	0.019	-0.046	**0.721**
DPO	OOD			–		0.232	0.217	0.564	-0.176	-0.185	0.570
DPO-S	OOD			–		0.016	-0.010	**0.585**	0.008	-0.004	**0.573**
SimPO	ID	-3.388	-5.623	**0.741**	0.365	-7.402	-12.080	0.771	-1.101	-3.071	0.697
SimPO-S	ID	0.232	-2.662	0.737	**0.635**	0.044	-1.212	**0.772**	1.959	-1.558	**0.709**
SimPO	OOD			–		-0.741	-1.348	**0.614**	-3.066	-3.569	**0.585**
SimPO-S	OOD			–		0.004	-0.005	0.609	0.064	-0.466	**0.585**
DPOP	ID	1.367	0.401	**0.755**	0.444	1.390	0.560	0.745	0.669	0.336	0.725
DPOP-S	ID	0.038	-0.053	0.726	**0.556**	0.678	-0.507	**0.771**	0.063	-0.185	**0.735**
DPOP	OOD			–		0.269	0.251	**0.585**	0.175	0.158	0.560
DPOP-S	OOD			–		0.054	-0.014	0.574	0.020	-0.004	**0.561**
Cal-DPO	ID	0.084	-0.069	**0.745**	0.482	0.116	-0.100	**0.779**	0.033	-0.033	**0.739**
Cal-DPO-S	ID	0.032	-0.117	0.743	**0.518**	0.116	-0.083	0.771	0.066	-0.067	0.7350
Cal-DPO	OOD			–		0.156	0.122	0.577	-0.018	-0.019	**0.574**
Cal-DPO-S	OOD			–		0.105	-0.054	**0.598**	0.006	-0.007	**0.574**

E.3 HYPERPARAMETERS

Model	Method	LR	β	γ	λ
QWEN-2.5-0.5B	DPO	1×10^{-5}	0.5	–	–
	Cal-DPO	3×10^{-6}	0.1	–	–
	DPOP	3×10^{-6}	0.5	–	0.1
	SimPO	3×10^{-5}	2.0	2.0	–
LLAMA-3.2-1B	DPO	1×10^{-5}	0.1	–	–
	Cal-DPO	3×10^{-6}	0.1	–	–
	DPOP	3×10^{-6}	0.5	–	0.1
	SimPO	3×10^{-5}	2.5	2.0	–
LLAMA-3.1-8B-LORA	DPO	1×10^{-5}	0.5	–	–
	Cal-DPO	1×10^{-5}	0.1	–	–
	DPOP	1×10^{-5}	0.5	–	0.1
	SimPO	3×10^{-5}	2.0	2.0	–

Table 3: Best hyperparameters selected using the grid search described below to maximize preference win-rate (WR) on Ultrafeedback dataset. γ denotes the margin parameter (SimPO), and λ is the DPOP regularization parameter.

Optimizer	α	β_1	β_2	ϵ	κ	τ
RMSprop	0.9	–	–	10^{-8}	–	–
GSAFERMSPROP	0.9	–	–	10^{-8}	0.4	10^{-3}

Table 4: Optimizer hyperparameters. Safe variants use the same base hyperparameters as their base counterparts.

Parameter	Value
Rank r	16
α	32
Dropout	0.05
Target modules	`q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`
Bias	none
Gradient checkpointing	enabled

Table 5: LoRA configuration used for LLAMA-3.1-8B.

Grid search. For DPO and Cal-DPO we swept learning rate $\in \{1\times10^{-6}, 3\times10^{-6}, 1\times10^{-5}\}$ and $\beta \in \{0.1, 0.5\}$ (6 points). For DPOP we additionally swept $\lambda \in \{0.1, 1.0, 10.0\}$ (18 points). For SimPO we swept learning rate $\in \{1\times10^{-6}, 3\times10^{-6}, 3\times10^{-5}\}$, $\beta \in \{2.0, 2.5\}$, and $\gamma \in \{2.0, 4.0, 6.0\}$ (18 points). The same grid was used for all three base models; Llama-3.1-8B was trained via LoRA rather than full fine-tuning for computational efficiency, with all other hyperparameters unchanged.

For the safe variants of each algorithm, we use the same base hyperparameters as their vanilla counterparts (Table 3). The parameter κ was swept over $\{0.3, 0.4, 0.5, 0.6\}$ (SimPO-S also included 0.2). We report $\kappa = 0.4$ throughout as a representative value rather than a tuned one. We present the optimizer hyperparameters in Table 4.

For ID experiments, hyper-parameter searches were performed to maximize preference win-rates on a validation subset of the train-split of each dataset, before model was retrained on the full train-split

with the best hyper-parameters. For OOD experiments, a hyper-parameter search was performed to maximize preference win-rate on UltraFeedback dataset. The best hyper-parameters (listed in Table 3) were then frozen and model was evaluated on the test-splits of RewardBench-V2 and MMLU-Pro.

LoRA configuration for Llama-3.1-8B. Llama-3.1-8B is trained via LoRA (Hu et al., 2022) rather than full fine-tuning; all other models are fully fine-tuned. Table 5 lists the LoRA configuration, applied identically across all methods and their safe variants. Gradient checkpointing is enabled for this model only.

E.4 Hardware, Memory and Compute

Experiments are run on a single node with 8 AMD Instinct MI300X GPUs (192 GB each). Training uses bfloat16 mixed precision, gradient checkpointing, maximum prompt length 512, and effective batch size 512 (per-device batch 32, gradient accumulation 16, one GPU per run). `Llama-3.2-1B` is full fine-tuned, while `Llama-3.1-8B` uses LoRA.

Table 6 reports GSafePO overhead relative to vanilla DPO. For `Llama-3.2-1B`, DPO-S incurs roughly $2.0\times$ per-step time and $1.31\times$ peak memory. The compute overhead comes from additional `torch.autograd` backward passes used to estimate reward geometry.

Under LoRA on `Llama-3.1-8B`, the time overhead remains about $2.0\times$, while memory overhead nearly disappears ($1.004\times$; 106.38 vs. 106.85 GB), since only rank-16 adapters are trainable and the frozen base model dominates memory use. All runs fit within a single MI300X GPU. Restricting reward control to a subset of layers could further reduce overhead, but we do not explore this optimization.

Model	Method	Steps/s	Mem. (GB)	Overhead
`Llama-3.2-1B` full FT	Vanilla DPO	0.051	58.86	$1.0\times$
	DPO-S (Ours)	0.026	84.43	$1.96\times$ time, $1.31\times$ mem.
`Llama-3.1-8B` LoRA r=16	Vanilla DPO	0.012	106.38	$1.0\times$
	DPO-S (Ours)	0.006	106.85	$2.00\times$ time, $1.004\times$ mem.

Table 6: Compute and memory overhead of GSafePO. Memory is per GPU. LoRA nearly eliminates DPO-S memory overhead, while its per-step time overhead persists because the reward-geometry backward passes scale with trainable parameter count rather than base model size.

E.5 Dataset Description

MMLU-Pro. MMLU-Pro (Wang et al., 2024b) is a strengthened variant of the Massive Multitask Language Understanding benchmark, designed to stress-test reasoning in large language models. It contains approximately 12,000 multiple-choice questions across 14 subjects (e.g., mathematics, law, chemistry, business, history, psychology). Each question offers 10 candidate answers with a single correct label—expanding the option set from four to ten—to curb guessing and sharpen separation among models. Compared with the original MMLU, prior work reports substantially lower accuracies and reduced sensitivity to prompt style, making MMLU-Pro a robust, reasoning-centric evaluation suite.

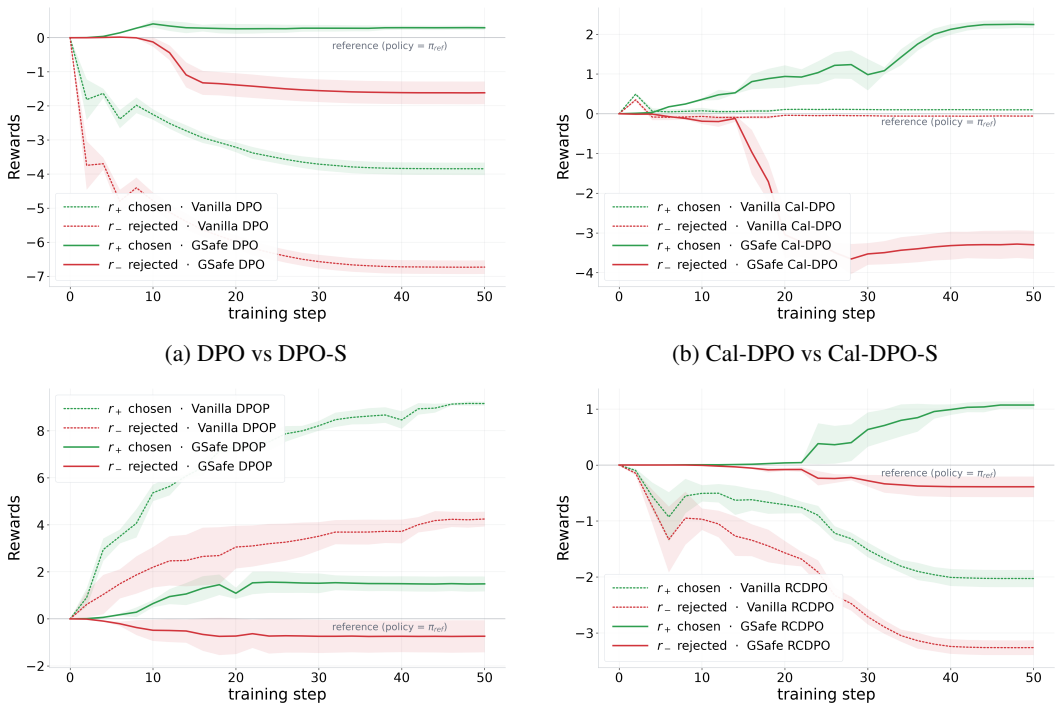
RewardBench v2. RewardBench v2 (Malik et al., 2025) is a second-generation, multi-skill benchmark for accuracy-based evaluation of reward models on unseen human data. It comprises $\sim 1{,}870$ prompts spanning six subsets—*Factuality*, *Precise Instruction Following*, *Math*, *Safety*, *Focus*, and *Ties*. Each prompt includes a preferred ("chosen") response and multiple rejected responses. Accuracy is measured by whether the reward model assigns a higher score to the chosen response than to all rejected alternatives. Compared to v1, the v2 release emphasizes harder, out-of-distribution prompts and reports per-subset counts to facilitate reproducible evaluation.[3]

[3] See the dataset card and paper for construction details, category counts, and scoring.

UltraFeedback. ULTRAFEEDBACK (Cui et al., 2024) is a large human-preference corpus widely used for preference optimization. We adopt the standardized pairwise format (chosen vs. rejected); the public *preference-standard* split contains approximately 340,000 rows, supporting stable optimization and serving as our training source for out-of-distribution experiments. Following common practice for DPO-style training, we use the preprocessed, binarized release of ULTRAFEEDBACK curated by Dong et al. (2024), which has been shown to yield higher-quality reward models (Dong et al., 2024; Wang et al., 2024a; Xiong et al., 2024; Banerjee & Gopalan, 2024).

E.6 REWARD DYNAMICS VS OTHER METHODS

Figure 2(a) in the main text shows chosen and rejected reward trajectories for DPO and its safe variant. Figure 4 shows the full comparison panel, with an identical protocol, for including the three remaining objectives: Cal-DPO, DPOP, and RC-DPO on `Llama-3.2-1B`/REWARDBENCH-V2. The vanilla objectives exhibit distinct reward dynamics. For Cal-DPO, chosen and rejected rewards remain close to the reference and move only modestly in opposite directions. For DPOP, both rewards rise together: the chosen reward increases strongly, but the rejected reward also rises above its initial value. For RC-DPO, both rewards decrease, with the chosen reward falling below the reference despite continued preference separation, exhibiting likelihood displacement. In all three cases, GSAFERMSPROP redirects the trajectories toward the desired chosen-up/rejected-down dynamics, increasing the chosen reward while decreasing the rejected reward relative to the reference.



(a) DPO vs DPO-S

(b) Cal-DPO vs Cal-DPO-S

(c) DPOP vs DPOP-S

(d) RC-DPO vs RC-DPO-S

Figure 4: Chosen and rejected reward trajectories over training for the vanilla objective (dotted) and its GSAFERMSPROP variant (solid), on `Llama-3.2-1B`/REWARDBENCH-V2. Rewards are measured relative to the frozen SFT reference policy, marked by the horizontal line at zero; shaded bands show the spread across seeds.

E.7 GSafeAdam experiment: `Llama-3.2-1B`

Category	Method	Dataset `RewardBench-v2` $r^+ \uparrow$	$r^- \downarrow$	WR
Base w/ ref	DPO	-1.648	-3.207	0.728
	DPO-S	0.460	-0.138	**0.742**

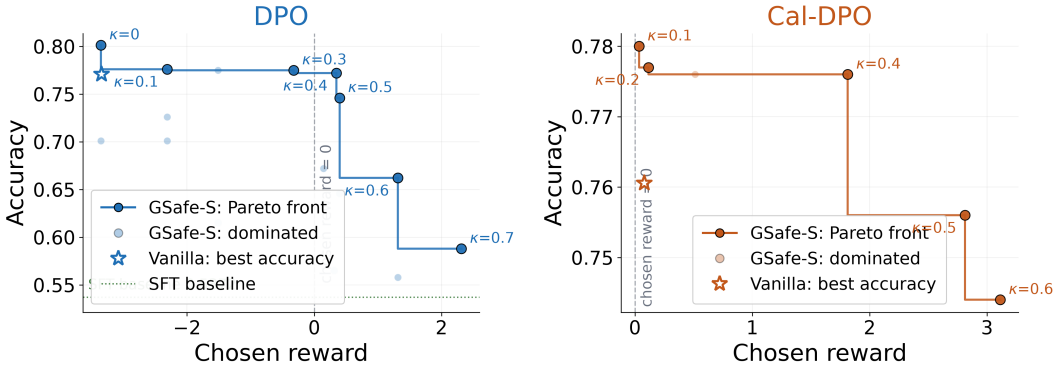
Method	β	κ	lr
Adam (DPO)	{0.05, 0.1, 0.3, 0.5*}	–	{1e−7, 3e−7, 1e−6*, 3e−6, 5e−6}
GSafeAdam (DPO-S)	{0.05*, 0.1, 0.3, 0.5}	{0.3, 0.4, 0.5, 0.6*}	{1e−7, 3e−7, 1e−6, 3e−6, 5e−6*}

Table 7: Hyperparameter grid for GSafeAdam experiment. Asterisks denote the selected values.

E.8 Finding the Optimal κ

κ	DPO-S r^+	r^-	WR	Cal-DPO-S r^+	r^-	WR
0.0	-3.354	-4.103	0.8013	–	–	–
0.1	-2.313	-3.337	0.776	0.035	-5.001	0.780
0.2	-1.513	-2.245	0.775	0.115	-4.955	0.777
0.3	-0.323	-1.395	0.775	0.512	-4.011	0.776
0.4	**0.347**	-1.192	**0.772**	**1.813**	-3.632	**0.776**
0.5	0.398	-1.142	0.746	2.813	-2.116	0.756
0.6	1.313	-0.876	0.662	3.113	-1.117	0.744
0.7	2.313	-0.039	0.588	–	–	–

Table 8: `Llama-3.2-1B` on `RewardBench`: Ablation to find optimal κ on the validation set.



(a) DPO-S (b) Cal-DPO-S

Figure 5: Accuracy versus chosen reward on `RewardBench` for `Llama-3.2-1B` as κ is swept. Solid markers on the connected line are Pareto-optimal κ values, faded markers are dominated, and the star marks the best-accuracy vanilla run. The dashed vertical line is zero chosen reward, i.e. no change relative to initialization.

Table 8 illustrates the expected no-free-lunch tradeoff in the safety margin. For small κ, the safety correction is too weak: loser rewards remain positive for DPO-S and Cal-DPO-S, indicating that the update is still not strong enough to reliably induce the desired chosen-up/rejected-down behavior.

As κ increases, r^- decreases monotonically across both methods, while win rate initially improves, showing that a larger margin indeed provides a stronger safety buffer and better directional control of the reward dynamics. However, beyond an intermediate range, further increasing κ begins to hurt r^+ and eventually win rate as well. This is consistent with the theory: a larger margin gives more robustness to minibatch noise and geometric uncertainty, but it also pushes the edited step farther from the minimally corrected update, increasing bias and making optimization less faithful to the original descent direction. In this experiment, $\kappa = 0.4$ gives the best overall tradeoff across both methods: it is the smallest margin that makes the chosen reward positive while retaining accuracy within 0.03 of the best value attained at small κ.

Figure 5 shows the same tradeoff as a Pareto front in the (chosen reward, accuracy) plane. Sweeping κ traces a frontier along which accuracy is exchanged for chosen reward: small κ retains the accuracy of the vanilla run but leaves the chosen reward at or below its initialization, while large κ drives the chosen reward well above zero at a growing cost in accuracy. Intermediate values dominate the vanilla baseline (star), attaining a positive chosen reward at comparable accuracy.

E.9 Ablation of projection variants from Section 3

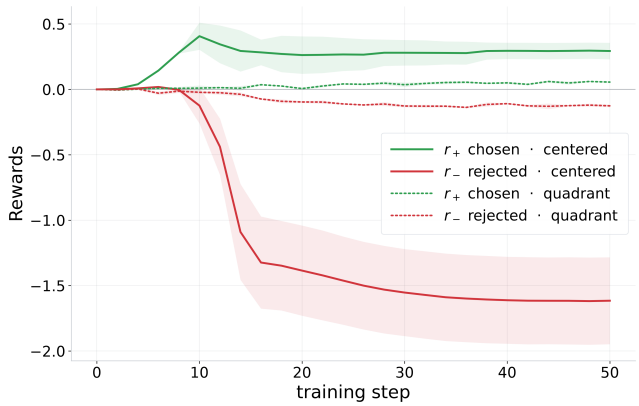


Figure 6: Ablation of projection variants for DPO on RewardBench-v2 with `Llama-3.2-1B`. Both projections increase reward separation, but the centered projection induces substantially stronger reward dynamics, producing both a larger increase in chosen reward and a larger decrease in rejected reward than the less restrictive quadrant projection.

We compare the centered projection used in GSafePO with the less restrictive quadrant projection, which directly enforces only the inequalities corresponding to winner-up and loser-down reward dynamics. Unlike the centered construction, which additionally imposes the balance constraint $\langle \boldsymbol{\mu}_w + \boldsymbol{\mu}_l, \mathbf{v} \rangle = 0$ and thereby removes the local shared reward-shift direction, the quadrant projection permits any update satisfying the two directional constraints. Figure 6 shows the resulting evaluation-set reward trajectories on RewardBench-v2 with `Llama-3.2-1B` and the DPO base loss. Both variants induce the desired qualitative reward separation, but the centered projection produces substantially stronger movement in both directions, increasing the chosen reward while driving the rejected reward markedly lower. In contrast, the quadrant projection induces only modest reward changes, consistent with its weaker constraints requiring little modification whenever the base update already lies near the feasible quadrant. These results suggest that explicitly removing the shared reward-shift component provides a stronger inductive bias toward reward-separating updates than enforcing sign constraints alone.

The theoretical conditions (e.g., Proposition 4.2) govern the expected first-order population reward dynamics under minibatch-estimated projection geometry. Since the projected direction is itself estimated from a noisy minibatch, and the realized reward changes also include higher-order effects, evaluation trajectories need not be exactly monotone or exhibit equal-magnitude chosen/rejected changes.

E.10 REWARD OVEROPTIMIZATION

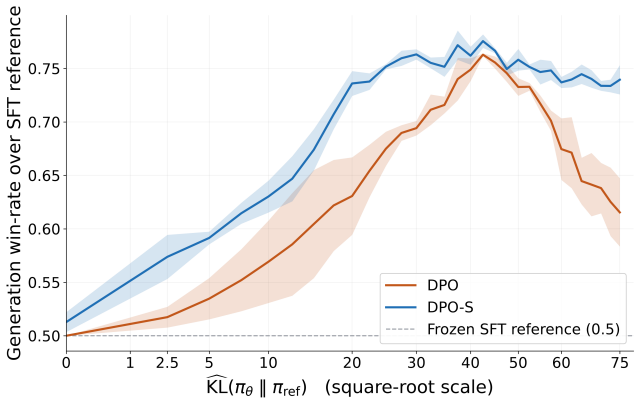


Figure 7: Generation win rate against the frozen SFT reference as a function of the estimated divergence $\widehat{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$ (square-root scale) for DPO and DPO-S on `Llama-3.2-1B` / REWARDBENCH. Curves are means over five seeds and shaded bands show the spread across seeds; the dashed line marks the reference win rate of 0.5.

To test for reward overoptimization–the regime in which continued optimization against a proxy objective increases policy divergence while independent response quality degrades (Rafailov et al., 2024a)—we track a divergence proxy together with an external LLM-judge quality signal throughout training for DPO and DPO-S on `Llama-3.2-1B`/REWARDBENCH.

Training protocol. We use the best hyperparameters established for this model/dataset pair: $\beta = 0.1$, $\mathrm{lr} = 1\mathrm{e}{-5}$ for DPO, and additionally $\kappa = 0.4$ for DPO-S. Each method is trained for 35 epochs over five seeds, yielding 27 checkpoints per run and 270 checkpoints in total.

Generation win rate over SFT reference. At each checkpoint, we greedily generate responses on the held-out REWARDBENCH split and compare them against a fixed set of responses from the frozen SFT reference model. Pairs are judged by LLAMA-3.1-70B-INSTRUCT with randomized left/right presentation order.

Divergence proxy. For each generated response y to prompt x, we compute

$$\widehat{D}(x) = \log \pi_\theta(y \mid x) - \log \pi_{\mathrm{ref}}(y \mid x), \tag{146}$$

using teacher-forced evaluation under the policy and frozen reference model, and average this quantity over the examples. Because the same greedy completion $y = \arg\max \pi_\theta(\cdot \mid x)$ is used rather than a sample from π_θ, $\widehat{D}$ is not an unbiased estimator of $\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}})$; it should be interpreted as a proxy for policy divergence along the model's generated trajectory.

Curve construction. For each run, we plot win rate over SFT reference against the checkpoint-level divergence proxy across training, yielding five trajectories per method. The x-axis is shown on a square-root scale to make the low-divergence regime easier to resolve.

Results. Figure 7 shows the mean trajectories across seeds. Both methods initially improve as policy divergence increases, consistent with reward overoptimization theory (Rafailov et al., 2024a). DPO-S is more sample-efficient in divergence space: it reaches a given generation win rate relative to reference at lower divergence during the ascent, attains a slightly higher peak, and degrades more gradually beyond that peak. In contrast, DPO loses substantially more generation perfomance at the largest observed divergences.

E.11 Projection Diagnostics

Figure 8 connects the observed optimization dynamics to Corollaries 3.1 and 3.2. Although the projection often makes a nontrivial correction, $\|\hat{g}\|/\|g\|$ remains close to one, consistent with the controlled norm inflation characterized by Corollary 3.1. Likewise, $\langle g, \hat{g}\rangle/\|g\|^2$ remains positive and close to one, indicating that the edits preserve strong descent alignment as characterized by Corollary 3.2. Finally, the observed $\sin^2\phi$ values show that the reward-gradient geometry is typically far from the degenerate regime.

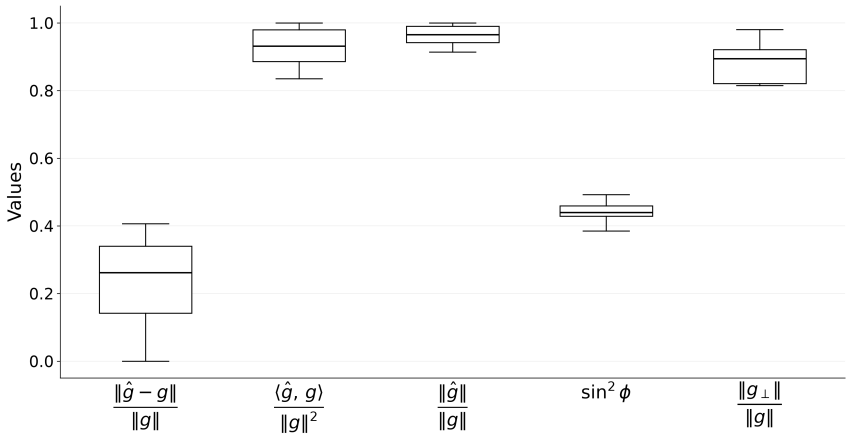


Figure 8: Distribution of five geometric diagnostics over training minibatches: the relative edit size $\|\hat{g}-g\|/\|g\|$, the normalized alignment $\langle \hat{g}, g\rangle/\|g\|^2$, the norm ratio $\|\hat{g}\|/\|g\|$, the degeneracy measure $\sin^2\phi$, where ϕ is the angle between the winner and loser score gradients, and $\frac{\|g^\perp\|}{\|g\|}$. Results using `Llama-3.2-1B` on `RewardBench`.

F Why Pairwise Gradient Disentanglement Does Not Imply Reward Safety

We clarify the distinction between pairwise disentanglement conditions (Yuan et al., 2025a; Chen et al., 2026) and the reward-safety condition imposed by GSafePO. The key difference is that pairwise conditions characterize how an individual preference pair responds to the update induced by that same pair, whereas reward safety concerns the effect of the aggregate parameter update actually taken on average chosen and rejected rewards. Consider a minibatch of B preference pairs. For pair i, define the chosen- and rejected-reward gradients

$$\boldsymbol{\mu}_{w,i} := \nabla_{\boldsymbol{\theta}} r_{w,i}(\boldsymbol{\theta}), \qquad \boldsymbol{\mu}_{l,i} := \nabla_{\boldsymbol{\theta}} r_{l,i}(\boldsymbol{\theta}), \tag{147}$$

and let $\Delta\boldsymbol{\theta}_i$ denote the parameter update that would be induced by pair i in isolation. A pairwise disentanglement condition requires

$$\langle \boldsymbol{\mu}_{w,i}, \Delta\boldsymbol{\theta}_i\rangle \geq 0, \qquad \langle \boldsymbol{\mu}_{l,i}, \Delta\boldsymbol{\theta}_i\rangle \leq 0, \qquad \forall i \tag{148}$$

Thus, to first order, each pair's own update increases its chosen reward and decreases its rejected reward. The realized minibatch update, however, aggregates contributions from all pairs:

$$\Delta\boldsymbol{\theta} = \frac{1}{B}\sum_{j=1}^{B} \Delta\boldsymbol{\theta}_j \tag{149}$$

Likewise, the minibatch-average chosen-reward gradient is

$$\boldsymbol{\mu}_w = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{\mu}_{w,i} \tag{150}$$

The first-order change in average chosen reward under the realized update is therefore

$$\Delta r_w \approx \langle \boldsymbol{\mu}_w, \Delta\boldsymbol{\theta}\rangle$$

$$= \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \langle \boldsymbol{\mu}_{w,i}, \Delta\boldsymbol{\theta}_j \rangle$$

$$= \frac{1}{B^2} \left[\underbrace{\sum_{i=1}^{B} \langle \boldsymbol{\mu}_{w,i}, \Delta\boldsymbol{\theta}_i \rangle}_{\text{controlled by pairwise disentanglement}} + \underbrace{\sum_{i \neq j} \langle \boldsymbol{\mu}_{w,i}, \Delta\boldsymbol{\theta}_j \rangle}_{\text{cross-pair terms}} \right] \tag{151}$$

Equation (148) constrains only the diagonal terms in equation (151). The cross-pair terms are unconstrained and may be negative enough to dominate the positive diagonal contribution. Consequently,

$$\forall i : \langle \boldsymbol{\mu}_{w,i}, \Delta\boldsymbol{\theta}_i \rangle \geq 0 \quad \not\Longrightarrow \quad \langle \boldsymbol{\mu}_w, \Delta\boldsymbol{\theta} \rangle \geq 0 \tag{152}$$

An identical argument gives

$$\forall i : \langle \boldsymbol{\mu}_{l,i}, \Delta\boldsymbol{\theta}_i \rangle \leq 0 \quad \not\Longrightarrow \quad \langle \boldsymbol{\mu}_l, \Delta\boldsymbol{\theta} \rangle \leq 0 \tag{153}$$

A simple two-pair scalar example makes this explicit. Let

$$\mu_{w,1} = 1, \qquad \Delta\theta_1 = 10, \qquad \mu_{w,2} = -2, \qquad \Delta\theta_2 = -1 \tag{154}$$

Each pair individually satisfies the chosen-up condition,

$$\mu_{w,1}\Delta\theta_1 = 10 > 0, \qquad \mu_{w,2}\Delta\theta_2 = 2 > 0 \tag{155}$$

However,

$$\mu_w = \frac{1-2}{2} = -\frac{1}{2}, \qquad \Delta\theta = \frac{10-1}{2} = \frac{9}{2}, \tag{156}$$

and hence

$$\mu_w \, \Delta\theta = -\frac{9}{4} < 0, \qquad \mu_{w,2} \, \Delta\theta = -9 < 0 \tag{157}$$

Thus, every pair can satisfy the pairwise chosen-up condition while the aggregate update decreases average chosen reward. Note also, that pair 2's chosen reward decreases under the actual update.

Hence, pairwise disentanglement does not imply either per-example safety under the aggregate update or safety of the average reward. This distinction becomes even more important with adaptive optimizers. If the aggregate direction is subsequently transformed by an effective preconditioner $\boldsymbol{P}_t$, the realized step takes the form

$$\Delta\boldsymbol{\theta}_t = -\eta_t \boldsymbol{P}_t \boldsymbol{g}_t^{(b)}, \tag{158}$$

so sign conditions established for pairwise gradient contributions need not be preserved after mini-batch aggregation and optimizer-induced geometric transformation.

Pairwise disentanglement therefore answers a local question: whether an individual preference pair's own update moves its chosen and rejected rewards in the desired directions. Reward safety asks a different question: whether the aggregate parameter update actually taken moves the expected chosen and rejected rewards in the desired directions. GSAFEPO is designed to constrain the latter directly.

G ILLUSTRATIVE GENERATION WINS

The examples below illustrate cases from the head-to-head generation evaluations judged by by LLAMA-3.1-70B-INSTRUCT on ULTRAFEEDBACK evaluation split which the judge preferred GSAFE over the corresponding vanilla method. These examples are intended only to illustrate the qualitative differences underlying some of those wins. Among winning comparisons, we retain English-only prompts, exclude degenerate or near-duplicate completions, and select examples short enough to present compactly. Policy identities are randomized in the evaluation.

Example 1: `Llama-3.1-8B` DPO-S

Prompt: Where will you find an automobile that has broken down?
Options:
- fast lane
- for sale
- racetrack
- side of road
- freeway
Please answer and provide answer explanation.

Completion 1 [VANILLA] Thank you.
Answer:
The correct answer is: side of road
The side of the road is where you will find a broken down automobile.

Completion 2 [GSAFE] Answer: side of road
Explanation: The side of the road is where you will find an automobile that has broken down.

Judge verdict: Comparison: Completion 2 is a better response because it directly answers the question without including unnecessary words like "Thank you" and provides a clear explanation in a concise manner, whereas Completion 1 includes an extra phrase that is not relevant to the question.
Preferred: 2
Preferred: GSafe

Example 2: `Llama-3.1-8B` SimPO-S

Prompt: What is a shorter version of this:
the top russian official responsible for human rights in chechnya said thursday that the living conditions in chechen refugee camps were unacceptable , inflicting " colossal damage " on the prospects for stability in the war-ravaged region .
Summary:

Completion 1 [GSAFE] The head of Russia's human rights commission said the living conditions in Chechen refugee camps were appalling, causing "colossal damage" to the prospects for stability in the war-torn region.

Completion 2 [VANILLA] The Chechen refugee camps' substandard living conditions, according to the Russian human rights chief, severely impair the chances of peace in Chechnya, causing immense harm.
Chechnya's Rights Chief's Critique on Inadequate Refugee Camps: Impeding Chechnya's Stability.

Judge verdict: Comparison: Completion 1 is a better response as it is more precise and concise, directly stating the official's title and the impact of the living conditions, whereas Completion 2 uses more general terms and adds the phrase "Impeding Chechnya's Stability" which is not necessary for the summary.
Preferred: 1
Preferred: GSafe

Example 3: `Llama-3.2-1B` DPOP-S

Prompt: Please write a Python program that takes a user inputted number and finds the sum of all the factors of the number. Your program should display the result to the user. It should also handle invalid inputs such as non-numeric inputs or negative numbers. Aim for efficient code and incorporate comments to explain your code logic.

Completion 1 [VANILLA]

```
# Prompt user for input
num = input("Enter a number: ")
```

```
# Convert input to integer
try:
   num = int(num)
except ValueError:
   print("Invalid input. Please enter a number.")
   exit()

# Check if input is negative
if num < 0:
   print("Invalid input. Number must be positive.")
   exit()

# Find factors of number
factor_list = []
for I in range(1, num + 1):
   if num % I == 0:
      factor_list.append(i)

# Calculate sum of factors
sum_of_factors = sum(factor_list)

# Display result
print("The sum of factors of", num, "is", sum_of_factors)
```

Completion 2 [**GSAFE**]

```
# Prompt user for input
num = input("Enter a number: ")

# Check if input is numeric
try:
   num = int(num)
except ValueError:
   print("Invalid input. Please enter a numeric value.")
   exit()

# Check if input is negative
if num < 0:
   print("Invalid input. Number must be positive.")
   exit()

# Calculate sum of factors
sum_factors = 0
for I in range(1, num + 1):
   if num % I == 0:
      sum_factors += I

# Display result
print("The sum of factors of", num, "is", sum_factors)
```

Judge verdict: Comparison: Completion 2 is preferred because it achieves the same result as Completion 1 but with more efficiency and conciseness by directly calculating the sum of factors in the loop, eliminating the need for an extra list and the sum function. Note Completion 1 has a genuine bug: the loop variable is capital "I", but the append line uses lowercase "i", an undefined name that raises a `NameError` at runtime. Completion 2 loop uses "I" consistently and runs correctly. This failure is independent of the judge's preference; it can be verified by running both snippets.
Preferred: 2
Preferred: GSafe

Example 4: `Llama-3.1-8B` RC-DPO-S

Prompt: Given the sentence "Mother and her two daughter plying swing in a park." can we conclude that "Mother and her two daughters playing tag at a funeral."?
Options:
- yes
- it is not possible to tell
- no Stream of thoughts:

Completion 1 [GSAFE] The sentence given is a simple sentence with a subject and a verb. The subject is "Mother and her two daughters" and the verb is "playing". The sentence is in present tense. The sentence is about a mother and her two daughters playing in a park. The sentence does not mention anything about a funeral. So, we cannot conclude that the mother and her two daughters are playing tag at a funeral.

Completion 2 [VANILLA] Mother-daughter relation is given in both sentences. So, it can be concluded that daughters are playing swing with mother. So, no. of daughters $= 2$. So, no. of daughters playing tag $= 2$. So, it can be concluded that daughters playing tag at funeral. Yes

Judge verdict: Comparison: Completion 1 is a better response because it directly addresses the question by explaining that the original sentence does not provide any information about a funeral, making it impossible to conclude the second sentence, whereas Completion 2 makes an incorrect assumption by trying to establish a connection between the two sentences based on the mother-daughter relationship.
Preferred: 1
Preferred: GSafe

Remark

This is a natural-language-inference question with an objectively correct answer ("no": the second sentence is unrelated to the first). The vanilla reasoning is confused and arrives at the wrong final answer ("Yes"); the GSAFE completion reasons correctly and answers "no".

Example 5: `Llama-3.1-8B` CAL-DPO-S

Prompt: In a hypothetical scenario, if we were to express a desire for someone to be more considerate, how would we rephrase the following sentence using the subjunctive mood?

Completion 1 [GSAFE] "I wish he would be more considerate of my feelings."
The subjunctive mood is used to express a wish or a hypothetical situation that is contrary to fact. In this case, we are expressing a desire for someone to be more considerate, which is contrary to the current reality. Therefore, we would rephrase the sentence using the subjunctive mood as follows: "I wish he were more considerate of my feelings." The use of the subjunctive mood here emphasizes the hypothetical nature of the statement and conveys a sense of longing or desire for a different outcome.

Completion 2 [VANILLA] "I wish he would be more considerate of my feelings."
The subjunctive mood is used to express a wish or a hypothetical situation. In this case, we are expressing a wish for someone to be more considerate. The sentence "I wish he would be more considerate of my feelings" is already in the subjunctive mood. Therefore, we would not need to rephrase it. However, if we were to rephrase it, we could say "I wish that he were more considerate of my feelings." This sentence uses the subjunctive mood to express a wish for a hypothetical situation.

Judge verdict: Comparison: Completion 1 is a better response because it accurately identifies the need to rephrase the original sentence to use the subjunctive mood, whereas Completion 2 incorrectly states that the original sentence is already in the subjunctive mood.
Preferred: 1
Preferred: GSafe

Remark

A grammar-correction task. The vanilla completion asserts a factual error, that the original sentence is "already in the subjunctive mood" (it is not; "would be" is not subjunctive), before offering a fix anyway. The GSAFE completion correctly identifies the need to rephrase and does so directly.