Oracle-Robust Online Alignment for Large Language Models

Zimeng Li
Purdue University

Mudit Gaur
The Ohio State University

Vaneet Aggarwal
Purdue University

Abstract

We study online alignment of large language models under misspecified preference feedback, where the observed preference oracle deviates from an ideal but unknown ground-truth oracle. The online LLM alignment problem is a bi-level reinforcement problem due to the coupling between data collection and policy updates. Recently, the problem has been reduced to a tractable single-level objective in the SAIL (Self-Improving Efficient Online Alignment) framework. In this paper, we introduce a pointwise oracle uncertainty set in this problem and formulate an oracle-robust online alignment objective as a worst-case optimization problem. For log-linear policies, we show that this robust objective admits an exact closed-form decomposition into the original loss function plus an explicit sensitivity penalty. In the finite-response log-linear setting, we develop projected stochastic composite updates for the resulting weakly convex objective and prove $\widetilde{O}(\varepsilon^{-2})$ oracle complexity for reaching approximate stationarity.

1 Introduction

Reinforcement learning from human feedback (RLHF) aligns large language models by updating a policy π_θ from pairwise preferences produced by a nominal training oracle $P^\star$, which may be a pool of human annotators or a learned reward model (Christiano et al., 2017; Ouyang et al., 2022; Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022). This available oracle may be misspecified relative to an unobserved ideal oracle P^{id}: annotators disagree, labeling standards drift, and style confounders such as verbosity can systematically bias judgments (Poddar et al., 2024; Ramesh et al., 2024). In online RLHF, where the policy controls the distribution of sampled comparisons, misspecification of the nominal oracle can be amplified by the feedback loop and lead to over-optimization of oracle quirks rather than the intended preferences (Skalse et al., 2022; Gao et al., 2023).

Robustness to oracle misspecification therefore needs to be modeled through an uncertainty set and a worst-case objective, rather than only through i.i.d. label noise. Two existing lines of work each handle part of this problem. Online RLHF methods such as SAIL (Ding et al., 2024) account for policy-induced sampling through a bilevel formulation, but assume the oracle matches the model and provide no distributionally robust guarantee. Distributionally robust DPO variants (Xu et al., 2025; Wu et al., 2025) optimize a worst-case loss over an uncertainty set, but only on a fixed offline dataset, sidestepping the policy-dependent comparison distribution. The missing ingredient, which we address, is robustness against an oracle uncertainty set under on-policy sampling.

At a glance, plugging an oracle uncertainty set into the SAIL bilevel program may look like a routine application of distributionally robust optimization. It is not. The worst-case objective places a supremum over adversarial oracles $P \in U^W(P^\star, \rho)$ inside an expectation under the policy-induced law d_θ, which is generally intractable in bilevel form. Our first technical contribution is to show that for log-linear policies, this inner supremum admits an exact closed-form decomposition,

$$L_\rho^W(\theta) = L^{\mathrm{SAIL}}(\theta) + \lambda R(\theta), \qquad \lambda = \rho\beta,$$

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

where $R(\theta)$ is the expected absolute pairwise log-ratio score under d_θ. This decomposition turns a robust bilevel problem into a single-level penalized objective, but introduces a new difficulty: $R(\theta)$ is the expectation of a pointwise convex quantity under a θ-dependent sampling law, and this coupling destroys convexity. Our second contribution resolves this for log-linear policies over a finite response set, by proving that $R(\theta)$ is weakly convex with an explicit constant and by analyzing a projected stochastic composite method using the Moreau envelope as a smooth stationarity surrogate.

Our contributions are as follows.

- We formulate oracle-robust online alignment by combining SAIL with a pointwise oracle uncertainty set $U^W(P^\star, \rho)$, leading to the worst-case objective $L_\rho^W(\theta)$ under policy-induced sampling.
- Under log-linear policies, we give an exact decomposition $L_\rho^W(\theta) = L^{\mathrm{SAIL}}(\theta) + \lambda R(\theta)$ and identify $R(\theta)$ as a sensitivity penalty given by an expected absolute pairwise score.
- We prove weak convexity of the constrained robust objective, analyze projected stochastic composite updates, and establish an $\widetilde{O}(\varepsilon^{-2})$ oracle complexity for squared Moreau-envelope stationarity. Setting $\rho = 0$ gives a convergence-to-stationarity guarantee for nominal SAIL.

2 Problem Setup

Let X be the prompt (context) space and Y the response space. For each parameter vector $\theta \in \Theta \subseteq \mathbb{R}^d$, the language model induces a conditional distribution (policy) $\pi_\theta(\cdot \mid x) \in \Delta(Y)$ for each $x \in X$, where $\Delta(Y)$ denotes the probability simplex over Y. We view each $y \in Y$ as a finite token sequence.

We model preferences over response pairs by a (possibly unknown) preference oracle $\mathcal{O} := \Big\{ P \;\Big|\; P(y_1 \succ y_2 \mid x) = 1 - P(y_2 \succ y_1 \mid x) \Big\}$. For $z = (x, y_1, y_2) \in Z := X \times Y \times Y$, we write $P(1 \mid z) := P(y_1 \succ y_2 \mid x)$ and interpret the (stochastic) preference label $\mathbb{I} \in \{0, 1\}$ as $\mathbb{I} \mid z \sim \mathrm{Bernoulli}\big(P(1 \mid z)\big)$, hence $\mathbb{I} = 1 \Leftrightarrow y_1 \succ y_2$. The sampling procedure is policy-dependent: draw a prompt $x \sim \mu$ from a fixed distribution μ over X, then sample two responses independently from $\pi_\theta(\cdot \mid x)$, $y_1, y_2 \overset{\text{iid}}{\sim} \pi_\theta(\cdot \mid x)$. This induces the data-generating distribution on Z,

$$d_\theta(x, y_1, y_2) := \mu(x)\, \pi_\theta(y_1 \mid x)\, \pi_\theta(y_2 \mid x). \tag{1}$$

Let $P^\star \in \mathcal{O}$ denote the nominal training preference oracle. Fix $\rho > 0$. For each $z = (x, y_1, y_2) \in Z$, define the pointwise uncertainty set

$$U_z^W(P^\star, \rho) := \Big\{ P \in \mathcal{O} \;\Big|\; \big|P(1 \mid z) - P^\star(1 \mid z)\big| \le \rho \Big\}. \tag{2}$$

Equivalently, for fixed z, view $P(\cdot|z)$ and $P^\star(\cdot|z)$ as Bernoulli distributions over the binary preference-label space $\{0, 1\}$. Under the ground cost $c(i, i') = |i - i'|$ for labels $i, i' \in \{0, 1\}$, the 1-Wasserstein distance (Peyré and Cuturi, 2019) satisfies

$$W_1(\mathrm{Ber}(P(1|z)), \mathrm{Ber}(P^\star(1|z))) = |P(1|z) - P^\star(1|z)|.$$

Thus $U_z^W(P^\star, \rho)$ is exactly a radius-ρ Wasserstein ball around the Bernoulli preference oracle at comparison triple z.

The corresponding global uncertainty set requires the pointwise constraint to hold uniformly over all comparison triples:

$$U^W(P^\star, \rho) := \bigcap_{z \in Z} U_z^W(P^\star, \rho) = \Big\{ P \in \mathcal{O} \;\Big|\; \sup_{z \in Z} \big|P(\cdot \mid z) - P^\star(\cdot \mid z)\big| \le \rho \Big\}. \tag{3}$$

$U^W(P^\star, \rho)$ models an adversarial but uniformly bounded misspecification of the (conditional) preference probability across all prompts and response pairs that may be encountered under the policy-induced sampling in (1).

Fix a prompt (context) $x \in X$. The policy $\pi_\theta(\cdot \mid x)$ independently generates a pair of responses $y_1, y_2 \sim \pi_\theta(\cdot \mid x)$, which are then compared by a preference oracle. For the nominal training oracle

$P^\star$, we assume it satisfies the Bradley–Terry model (Bradley and Terry, 1952): for all $x \in X$ and $y_i \in Y$, there exists a latent reward function $r^\star(x, y_i)$ such that

$$P^\star(y_1 \succ y_2 \mid x) = \sigma\big(r^\star(x, y_1) - r^\star(x, y_2)\big), \tag{4}$$

where $\sigma(u) := (1 + e^{-u})^{-1}$ denotes the logistic sigmoid, $i \in \{1, 2\}$.

We impose a mild margin condition to ensure that all admissible oracles remain valid.

Assumption 1 (Nondegenerate nominal oracle and admissible radius)**.** *There exists a constant $\delta \in (0, 1/2]$ such that for all $(x, y_1, y_2) \in X \times Y \times Y$,*

$$\delta \;\le\; P^\star(y_1 \succ y_2 \mid x) \;\le\; 1 - \delta. \tag{5}$$

We restrict the oracle uncertainty radius to $\rho \in (0, \delta)$, so that every preference oracle P is nondegenerate, i.e. $\forall (x, y_1, y_2),\ \forall P \in U^W(P^\star, \rho)$

$$0 \le \delta - \rho \;\le\; P(y_1 \succ y_2 \mid x) \;\le\; 1 - (\delta - \rho) \le 1.$$

Remark 2.1. Assumption 1 bounds the nominal preference probability away from 0 and 1, so that all radius-ρ perturbations with $\rho < \delta$ remain valid Bernoulli probabilities. Under the Bradley–Terry model,

$$\delta \le \sigma(r^\star(x, y_1) - r^\star(x, y_2)) \le 1 - \delta \quad \Longleftrightarrow \quad |r^\star(x, y_1) - r^\star(x, y_2)| \le \log \frac{1 - \delta}{\delta}.$$

Thus the assumption rules out essentially deterministic comparisons over the support considered by the policy-induced sampling distribution. In particular, if $|r^\star(x, y_1) - r^\star(x, y_2)| \le M$ for all (x, y_1, y_2), then the assumption holds with $\delta = \sigma(-M)$, which gives a conservative estimate of δ. The restriction $\rho < \delta$ only ensures that every interval $[P^\star(1 \mid z) - \rho,\ P^\star(1 \mid z) + \rho]$ lies in $(0, 1)$.

Background (SAIL). SAIL is a preference-based RLHF framework for online alignment that explicitly models the coupling between (i) learning from pairwise preference feedback and (ii) updating the policy that generates the responses being compared. In the online regime, the preference data distribution is policy-dependent, and SAIL represents this dependence via a bilevel program: a reward model r is fit from Bradley–Terry comparisons on responses sampled from the KL-regularized optimal policy induced by r,

$$\begin{aligned} &\text{(upper)} \quad \min_r -\mathbb{E}_{x\sim\mu,\, y_i\sim\pi_r^\star(\cdot|x),(y_w\succ y_\ell)\sim P^\star}\Big[\log\sigma\big(r(x, y_w) - r(x, y_\ell)\big)\Big] \\ &\text{(lower)} \quad \text{s.t. } \pi_r^\star \in \arg\max_\pi \mathbb{E}_{x\sim\mu,\, y\sim\pi(\cdot|x)}\Big[r(x, y) - \beta D_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{SFT}}(\cdot \mid x)\big)\Big], \end{aligned} \tag{6}$$

Direct differentiation through the inner solution mapping $r \mapsto \pi_r^\star$ requires hypergradient computations. SAIL circumvents this by exploiting the reward–policy equivalence for KL-regularized RL: any optimizer $\pi_r^\star$ satisfies

$$r(x, y) = \beta \log \frac{\pi_r^\star(y \mid x)}{\pi_{\mathrm{SFT}}(y \mid x)} + \beta \log Z_r(x), \tag{7}$$

for a normalization $Z_r(x)$ independent of y. For a policy π_θ, define the pairwise log-ratio score

$$h_\theta(x, y_a, y_b) := \log \frac{\pi_\theta(y_a \mid x)}{\pi_{\mathrm{SFT}}(y_a \mid x)} - \log \frac{\pi_\theta(y_b \mid x)}{\pi_{\mathrm{SFT}}(y_b \mid x)}.$$

Substituting (7) into the upper-level BT likelihood reduces the bilevel program to the single-level SAIL objective

$$L^{\mathrm{SAIL}}(\theta) := -\mathbb{E}_{z\sim d_\theta,\ (y_w, y_\ell)\sim P^\star}\left[\log\sigma(\beta h_\theta(x, y_w, y_\ell))\right]. \tag{8}$$

For any oracle P, let $L_P^{\mathrm{SAIL}}(\theta)$ denote the same objective with preference outcomes generated by P instead of $P^\star$. We define the oracle-robust alignment objective as the worst-case SAIL value over adversarial oracles in the uncertainty set centered at the nominal oracle $P^\star$:

$$L_\rho^W(\theta) := \sup_{P \in U^W(P^\star, \rho)} L_P^{\mathrm{SAIL}}(\theta). \tag{9}$$

3 Proposed Approach

In this section, we first relate our oracle-robust objective L_ρ^W to the nominal SAIL objective L^{SAIL} by showing that L_ρ^W admits an *exact* decomposition into the SAIL loss and an explicit regularization term. We then study the regularity of the two components separately: we impose a smoothness condition on L^{SAIL} and analyze the policy-dependent penalty $R(\theta)$. Under mild assumptions, these results together imply that L_ρ^W is weakly convex.

3.1 Decomposition of $L_\rho^W(\theta)$

Our robust objective $L_\rho^W(\theta)$ is defined via a worst-case expectation over the global oracle uncertainty set $U^W(P^\star, \rho)$. The following assumption specifies the log-linear SAIL comparison model and enables an exact reduction of the inner worst-case problem to an explicit regularizer.

Assumption 2 (Log-linear policy class)**.** *Let $\psi : X \times Y \to \mathbb{R}^d$ be a known d-dimensional feature map. We consider the class of log-linear (softmax) policies*

$$\Pi := \left\{ \pi_\theta : \ \pi_\theta(y \mid x) = \frac{\exp\big(\theta^\top \psi(x,y)\big)}{\sum_{y' \in Y} \exp\big(\theta^\top \psi(x,y')\big)} \right\}.$$

Remark 3.1. The pointwise worst-case Bernoulli maximization does not rely on log-linearity. For any policy class with well-defined likelihood ratios, it gives

$$L_\rho^W(\theta) = L_{\mathrm{SAIL}}(\theta) + \rho\beta\, \mathbb{E}_{z \sim d_\theta} |h_\theta(z)|.$$

This assumption is used to rewrite the log-ratio score as the explicit linear pairwise score

$$h_\theta(x, y_1, y_2) = (\theta - \theta_{\mathrm{ref}})^\top (\psi(x, y_1) - \psi(x, y_2)),$$

which yields the penalty $R(\theta)$ analyzed below. The subsequent weak-convexity and oracle-complexity results rely on this linear-score structure together with the boundedness assumptions.

For notational convenience, let $\theta_{\mathrm{ref}} \in \Theta$ denote the fixed parameter of the reference (SFT) policy, i.e,

$$\pi_{\theta_{\mathrm{ref}}} \ = \ \pi_{\mathrm{SFT}}. \tag{10}$$

Theorem 3.2 (Decomposition of $L_\rho^W(\theta)$)**.** *Recall $L^{\mathrm{SAIL}}(\theta)$ denotes the non-robust SAIL objective and $L_\rho^W(\theta)$ denotes the robust objective defined by the worst-case oracle in $U^W(P^\star, \rho)$. Define the pairwise score*

$$R(\theta; x, y_1, y_2) := (\theta - \theta_{\mathrm{ref}})^\top \big(\psi(x, y_1) - \psi(x, y_2)\big),$$

and the robust penalty

$$R(\theta) := \mathbb{E}_{x \sim \mu} \mathbb{E}_{y_1, y_2 \sim \pi_\theta(\cdot|x)} \left[\big| R(\theta; x, y_1, y_2) \big| \right]. \tag{11}$$

Then under Assumptions 1, 2, the robust objective admits the exact decomposition

$$L_\rho^W(\theta) = L^{\mathrm{SAIL}}(\theta) + \lambda R(\theta), \qquad \lambda := \rho\beta. \tag{12}$$

Proof sketch of Theorem 3.2. Pointwise endpoint maximization in $p(z)$, together with the log-linear cancellation $h_\theta(x, y_1, y_2) = (\theta - \theta_{\mathrm{ref}})^\top (\psi(x, y_1) - \psi(x, y_2))$, gives the result; see Appendix B. □

Remark 3.3 (Interpretation of the decomposition). Theorem 3.2 separates the robust objective into a nominal fitting term $L^{\mathrm{SAIL}}(\theta)$ and an explicit robustness penalty $\lambda R(\theta)$ that depends on the policy-induced sampling distribution. The penalty $R(\theta)$ measures the expected magnitude of the pairwise score $R(\theta; x, y_1, y_2)$ over i.i.d. response pairs $(y_1, y_2) \sim \pi_\theta(\cdot \mid x)$, and thus quantifies sensitivity of the likelihood to adversarial perturbations of the pointwise preference probability. The uncertainty radius enters only through the linear prefactor $\lambda = \rho\beta$: increasing ρ monotonically strengthens the penalty, while $\rho \downarrow 0$ recovers the nominal objective $L^{\mathrm{SAIL}}(\theta)$.

3.2 Regularity of L^{SAIL}

Prior analyses of preference-based objectives often invoke stronger global conditions (e.g., PL-type geometries (Gaur et al., 2025)). Here we treat $L^{\mathrm{SAIL}}(\theta)$ under the following smoothness assumption.

Assumption 3 (Smooth SAIL objective)**.** *The nominal SAIL objective L^{SAIL} has L_{SAIL}-Lipschitz gradient:*

$$\|\nabla L^{\mathrm{SAIL}}(\theta) - \nabla L^{\mathrm{SAIL}}(\theta')\| \le L_{\mathrm{SAIL}}\|\theta - \theta'\| \; \forall\, \theta, \theta' \in \mathbb{R}^d.$$

This assumption is used in Theorem 3.7 and in the convergence analysis.

Remark 3.4. Assumption 3 is a standard smoothness condition in first-order optimization, and it is routinely imposed in analyses of gradient-based methods for both convex and nonconvex objectives, including stochastic settings (Nesterov, 2004; Ghadimi and Lan, 2013; Bottou et al., 2018). We adopt it here as a mild regularity requirement on L^{SAIL}.

3.3 Weak convexity of $R(\theta)$

It remains to understand the regularity of the robustness penalty $R(\theta)$. Although $|R(\theta; x, y_1, y_2)|$ is convex in θ pointwise, the expectation under the policy-dependent law d_θ destroys convexity in general; a one-dimensional counterexample is given in Appendix C.1. We therefore control the curvature of $R(\theta)$ through weak convexity: f is κ-weakly convex if $f(\cdot) + \frac{\kappa}{2}\|\cdot\|^2$ is convex. First we introduce several assumptions required for the analysis.

Assumption 4 (Finite response set and bounded features)**.** *The response space Y is finite, and the feature map satisfies $\|\psi(x, y)\| \le B_\psi < \infty$ for all $(x, y) \in X \times Y$.*

Remark 3.5. This assumption is a proof-level regularity condition used to obtain uniform score-function and curvature bounds for the policy-dependent penalty $R(\theta)$. Such assumptions are standard in analyses of log-linear and softmax models (Agarwal et al., 2021; Zhang et al., 2022).

Assumption 5 (Bounded feasible set)**.** *The feasible set Θ is nonempty, closed, convex, and bounded. In particular, with θ_{ref} as in Eq.*(10)*, define*

$$D \;:=\; \sup_{\theta \in \Theta} \|\theta - \theta_{\mathrm{ref}}\| < \infty.$$

This assumption makes projection well-defined (Bauschke and Combettes, 2011) and supplies the boundedness needed in Theorem 3.6.

Theorem 3.6 (Weak convexity of the robust penalty)**.** *Under Assumptions 2, 4, 5, the robust penalty $R : \Theta \to \mathbb{R}$ defined in Eq.* (11) *is κ_R-weakly convex on Θ, with*

$$\kappa_R \le 16B_\psi^2 + 4DB_\psi^3.$$

Proof sketch of Theorem 3.6. The proof smooths $|\cdot|$ by $\phi_\varepsilon(u) = \sqrt{u^2 + \varepsilon^2}$, applies score-function calculus to the policy-dependent expectation, and bounds the Hessian uniformly using bounded features and bounded θ. See Appendix C.3. □

Theorem 3.7 (Weak convexity of the composite objective)**.** *Under Assumptions 1, 2, 3, 4, 5, the robust objective $L_\rho^W(\theta) = L^{\mathrm{SAIL}}(\theta) + \lambda R(\theta)$ is κ-weakly convex on Θ, with*

$$\kappa := L_{\mathrm{SAIL}} + \lambda\kappa_R, \qquad \lambda = \rho\beta.$$

Proof sketch of Theorem 3.7. Sums of L-smooth and κ_R-weakly convex functions are $(L + \lambda\kappa_R)$-weakly convex; see Appendix C.4. □

3.4 Algorithm development

We optimize the oracle-robust alignment objective $L_\rho^W(\theta)$ over the bounded convex parameter set Θ. To enforce feasibility explicitly, we consider the constrained objective

$$F(\theta) \;:=\; L_\rho^W(\theta) + I_\Theta(\theta), \qquad I_\Theta(\theta) \;:=\; \begin{cases} 0, & \theta \in \Theta, \\ +\infty, & \theta \notin \Theta. \end{cases} \tag{13}$$

Because F may be nonsmooth and only weakly convex, we measure first-order stationarity using the Moreau envelope (Moreau, 1965).

Definition 3.8 (Moreau envelope and proximal point)**.** Fix $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$. The Moreau envelope of F with parameter λ_{env} is

$$F_{\lambda_{\mathrm{env}}}(\theta) \;:=\; \min_{u\in\mathbb{R}^d} \left\{ F(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u-\theta\|^2 \right\}, \tag{14}$$

and the associated proximal point mapping is

$$\mathrm{prox}_{\lambda_{\mathrm{env}}F}(\theta) \;:=\; \arg\min_{u\in\mathbb{R}^d} \left\{ F(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u-\theta\|^2 \right\}. \tag{15}$$

When convenient, we write $\hat{\theta} := \mathrm{prox}_{\lambda_{\mathrm{env}}F}(\theta)$.

Lemma 3.9 (Properties of the Moreau envelope (Davis and Drusvyatskiy, 2018))**.** *Assume F is κ-weakly convex and bounded below, and let $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$. Then:*

1. *$F_{\lambda_{\mathrm{env}}}$ is finite and continuously differentiable on $\mathbb{R}^d$.*
2. *Its gradient is $\nabla F_{\lambda_{\mathrm{env}}}(\theta) = \frac{1}{\lambda_{\mathrm{env}}}\big(\theta - \hat{\theta}\big)$, and in particular,*

$$\|\theta - \hat{\theta}\| \;=\; \lambda_{\mathrm{env}}\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|. \tag{16}$$

3. *The gradient $\nabla F_{\lambda_{\mathrm{env}}}$ is Lipschitz with constant*

$$L_{\mathrm{env}} \;:=\; \frac{1}{\lambda_{\mathrm{env}}(1-\kappa\lambda_{\mathrm{env}})}. \tag{17}$$

4. *(Approximate stationarity) For $\hat{\theta} = \mathrm{prox}_{\lambda_{\mathrm{env}}F}(\theta)$,*

$$\mathrm{dist}\big(0, \partial F(\hat{\theta})\big) \;\le\; \|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|. \tag{18}$$

Lemma 3.9 motivates $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|$ as a smooth stationarity measure: controlling $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|$ ensures (i) θ is close to its proximal point $\hat{\theta}$ via (16) and (ii) $\hat{\theta}$ is nearly stationary for the original constrained problem via (18). The envelope parameter λ_{env} governs the smoothing–bias trade-off: smaller λ_{env} reduces smoothing bias but increases L_{env} in (17).

Lemma 3.10 (Stochastic gradient/subgradient oracles)**.** *Suppose Assumptions 2, 4, and 5 hold. At iteration t, Algorithm 1 samples a mini-batch $Z_t = \{z_i\}_{i=1}^B$ with i.i.d. draws $z_i \sim d_{\theta_t}$. There exist stochastic oracles G^{SAIL} and G^R satisfying , for all $\theta, u \in \Theta$,*

$$\mathbb{E}\big[G^{\mathrm{SAIL}}(\theta; Z_t) \mid \theta\big] = \nabla L^{\mathrm{SAIL}}(\theta),$$

$$R(u) \ge R(\theta) + \langle \mathbb{E}[G^R(\theta; Z_t) \mid \theta], u-\theta\rangle - \frac{\kappa_R}{2}\|u-\theta\|^2, \tag{19}$$

Moreover, there exist constants $\sigma^2_{\mathrm{SAIL}}, \sigma^2_R < \infty$ such that,

$$\mathbb{E}\Big[\big\|G^{\mathrm{SAIL}}(\theta; Z_t) - \nabla L^{\mathrm{SAIL}}(\theta)\big\|^2 \mid \theta\Big] \le \frac{\sigma^2_{\mathrm{SAIL}}}{B},$$

$$\mathbb{E}\big[\|G^R(\theta; Z_t) - \mathbb{E}[G^R(\theta; Z_t) \mid \theta]\|^2 \mid \theta\big] \le \frac{\sigma^2_R}{B}$$

We minimize F in (13) by a projected stochastic composite gradient method (Algorithm 1). At each iterate θ_t, we draw a mini-batch under the policy-induced sampling d_{θ_t}, query preference labels for the SAIL term, and combine the two stochastic oracles with weight $\lambda = \rho\beta$:

$$G(\theta_t; Z_t) \;:=\; G^{\mathrm{SAIL}}(\theta_t; Z_t) + \lambda\, G^R(\theta_t; Z_t). \tag{20}$$

Projection onto Θ keeps the iterates feasible and the constrained objective F well-defined:

$$\theta_{t+1} \;:=\; \mathrm{proj}_\Theta(\theta_t - \eta\, G(\theta_t; Z_t)), \tag{21}$$

where $\eta > 0$ is a stepsize.

Algorithm 1 Robust Stochastic Composite Gradient Descent (R-SCGD)

Input: $\theta_0 \in \Theta$, stepsize $\eta > 0$, horizon T, batch size B, weight $\lambda = \rho\beta$.
for $t = 0, 1, \ldots, T-1$ **do**
  Sample a mini-batch $Z_t = \{(x_i, y_{1,i}, y_{2,i})\}_{i=1}^{B}$ with $x_i \sim \mu$ and $y_{1,i}, y_{2,i} \overset{\text{i.i.d.}}{\sim} \pi_{\theta_t}(\cdot \mid x_i)$.
  Query preference labels from the nominal training oracle $P^\star$ for the sampled pairs to construct $(y_{w,i}, y_{\ell,i})$ (equivalently, construct binary labels $y_i \in \{0, 1\}$).
  Compute stochastic oracles $G^{\mathrm{SAIL}}(\theta_t; Z_t)$ and $G^R(\theta_t; Z_t)$ of Lemma 3.10, and form $G(\theta_t; Z_t) = G^{\mathrm{SAIL}}(\theta_t; Z_t) + \lambda G^R(\theta_t; Z_t)$.
  Update $\theta_{t+1} \leftarrow \mathrm{proj}_\Theta\big(\theta_t - \eta\, G(\theta_t; Z_t)\big)$.
end for
Sample $R \sim \mathrm{Unif}\{0, 1, \ldots, T-1\}$.
Output: θ_R .

4 Analysis

We analyze Algorithm 1 for the constrained objective F in (13) using the Moreau-envelope stationarity measure from Definition 3.8. Throughout, we use the weak convexity constant κ from Theorem 3.7 and assume the smoothness condition on L^{SAIL} stated in Assumption 3.

Assumption 6 (Lower bounded objective)**.** *The constrained objective $F(\theta) = L_\rho^W(\theta) + I_\Theta(\theta)$ is proper, lower semicontinuous, and bounded from below on $\mathbb{R}^d$; denote*

$$F_{\inf} \;:=\; \inf_{\theta \in \mathbb{R}^d} F(\theta) \;>\; -\infty.$$

Main Result. We next state the convergence guarantee for Algorithm 1 and the resulting oracle complexity bound.

For the convergence bound, define the deterministic gradient-size constants

$$G_{\nabla \mathrm{SAIL}} := 2\beta B_\psi + 4B_\psi\big(\log 2 + 2\beta D B_\psi\big), \qquad G_{\partial R} := 2B_\psi + 8DB_\psi^2.$$

By Lemma 3.10, $G(\theta; Z) = G^{\mathrm{SAIL}}(\theta; Z) + \lambda G^R(\theta; Z)$ satisfies the uniform conditional second-moment bound

$$\mathbb{E}\big[\|G(\theta; Z)\|^2 \mid \theta\big] \le G_{\mathrm{tot}}^2, \qquad G_{\mathrm{tot}}^2 := 4\left(G_{\nabla \mathrm{SAIL}}^2 + \lambda^2 G_{\partial R}^2 + \frac{\sigma_{\mathrm{SAIL}}^2 + \lambda^2 \sigma_R^2}{B}\right),$$

for every $\theta \in \Theta$, where B is the mini-batch size and $\lambda = \rho\beta$. The proof of this bound is given in Lemma C.4.

Theorem 4.1 (Convergence rate for the Moreau envelope)**.** *Let Assumptions 1–6 hold and let κ be the weak convexity constant from Theorem 3.7. Fix any $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$ and run Algorithm 1 for T iterations with stepsize $\eta > 0$. Let $R \sim \mathrm{Unif}\{0, 1, \ldots, T-1\}$ denote the output index and θ_R the corresponding iterate. Then*

$$\mathbb{E}\Big[\big\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\big\|^2\Big] \;\le\; \frac{F_{\lambda_{\mathrm{env}}}(\theta_0) - F_{\inf}}{\eta(1 - \kappa\lambda_{\mathrm{env}})\, T} \;+\; \frac{L_{\mathrm{env}}\, \eta}{2(1 - \kappa\lambda_{\mathrm{env}})}\, G_{\mathrm{tot}}^2, \tag{22}$$

where L_{env} is defined in (17)*.*

Corollary 4.2 (Sample complexity for envelope stationarity)**.** *Under the conditions of Theorem 4.1, fix $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$ and set*

$$\eta \;:=\; \sqrt{\frac{2\lambda_{\mathrm{env}}(1 - \kappa\lambda_{\mathrm{env}})\big(F_{\lambda_{\mathrm{env}}}(\theta_0) - F_{\inf}\big)}{G_{\mathrm{tot}}^2\, T}}.$$

Then Algorithm 1 guarantees

$$\mathbb{E}\Big[\big\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\big\|^2\Big] \;\le\; \sqrt{\frac{2\, G_{\mathrm{tot}}^2\big(F_{\lambda_{\mathrm{env}}}(\theta_0) - F_{\inf}\big)}{\lambda_{\mathrm{env}}(1 - \kappa\lambda_{\mathrm{env}})^3\, T}}.$$

Consequently, to achieve $\mathbb{E}[\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|^2] \le \varepsilon$*, it suffices to take*

$$T \;\ge\; \frac{2\,G_{\mathrm{tot}}^2\big(F_{\lambda_{\mathrm{env}}}(\theta_0) - F_{\mathrm{inf}}\big)}{\lambda_{\mathrm{env}}(1-\kappa\lambda_{\mathrm{env}})^3\,\varepsilon^2} = \frac{8\left(C + \frac{\sigma_{\mathrm{SAIL}}^2 + \lambda^2\sigma_R^2}{B}\right)\big(F_{\lambda_{\mathrm{env}}}(\theta_0) - F_{\mathrm{inf}}\big)}{\lambda_{\mathrm{env}}(1-\kappa\lambda_{\mathrm{env}})^3\,\varepsilon^2}.$$

where $C = G_{\nabla\mathrm{SAIL}}^2 + \lambda^2 G_{\partial R}^2$*. If* $B = \widetilde{O}(1)$*, the total oracle complexity is* $BT = \widetilde{O}(\varepsilon^{-2})$*.*

Remark 4.3 (Role of smoothness). Theorem 4.1 does not require L_ρ^W to be smooth; R, and hence L_ρ^W, may be nonsmooth. Its proof uses the κ-weak convexity of L_ρ^W on Θ and the lower model of Lemma 3.10. The only smoothness assumption is Assumption 3, which is imposed on the nominal term L^{SAIL} rather than derived from the log-linear structure; together with Theorem 3.6, it gives the constant $\kappa = L_{\mathrm{SAIL}} + \rho\beta\kappa_R$ of Theorem 3.7. The smoothness of the Moreau envelope $F_{\lambda_{\mathrm{env}}}$ then follows from Lemma 3.9.

Remark 4.4 (Choice of the radius). Theorem 4.1 and Corollary 4.2 hold for every fixed radius $\rho \in (0, \delta)$, with $\lambda = \rho\beta$ and the corresponding κ and G_{tot}. The guarantee concerns the objective L_ρ^W defined by the chosen radius; different radii may lead to different stationary points. Since the actual level of misspecification is unknown in practice, we treat ρ as a model-selection parameter; Appendices F.5 and F.6 give validation-based and utility-constrained selection rules.

5 On the practical role of the proximal point

Our analysis measures stationarity of the *constrained* robust objective

$$F(\theta) \;:=\; L_\rho^W(\theta) + I_\Theta(\theta),$$

via the Moreau envelope $F_{\lambda_{\mathrm{env}}}$ with parameter $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$ (Definition 3.8), where κ is the weak convexity constant from Theorem 3.7. Algorithm 1 outputs a uniformly random iterate θ_R. The theory (Theorem 4.1) certifies *envelope stationarity* at θ_R by controlling $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2$, rather than directly bounding $\mathrm{dist}(0, \partial F(\theta_R))$. The interpretation is instead through the associated proximal point

$$\hat\theta_R \;:=\; \mathrm{prox}_{\lambda_{\mathrm{env}}F}(\theta_R) \;\in\; \arg\min_{u\in\mathbb{R}^d}\Big\{F(u) + \tfrac{1}{2\lambda_{\mathrm{env}}}\|u-\theta_R\|_2^2\Big\},$$

which is unique since F is κ-weakly convex and $\lambda_{\mathrm{env}} < 1/\kappa$.

Lemma 3.9 provides the key link between envelope stationarity and proximity:

$$\|\theta_R - \hat\theta_R\|_2 \;=\; \lambda_{\mathrm{env}}\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2,$$

and moreover $\mathrm{dist}(0, \partial F(\hat\theta_R)) \le \|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2$. Thus, small $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2$ simultaneously implies (i) θ_R lies in a small neighborhood of $\hat\theta_R$, and (ii) $\hat\theta_R$ is nearly first-order stationary for the original constrained objective.

A practical issue is that computing $\hat\theta_R$ exactly still requires solving the (generally nontrivial) strongly convex proximal subproblem above. Even when $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2$ is small (hence $\hat\theta_R$ is nearby), proximal operators frequently do not admit closed forms and are evaluated via iterative inner solvers, which can be expensive in realistic models (Salzo and Villa, 2012).

Lemma E.1 shows that while $\mathrm{dist}(0, \partial F(\theta)) \ge (1 - \kappa\lambda_{\mathrm{env}})\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|$ always holds, under κ-weak convexity alone there is *no* universal constant C such that $\mathrm{dist}(0, \partial F(\theta)) \le C\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|$ for all θ (even when $\kappa = 0$). This motivates phrasing guarantees in terms of the proximal point $\hat\theta_R$ rather than the raw iterate θ_R.

In practice, one can compute an *inexact* proximal refinement $\bar\theta_R \approx \hat\theta_R$ via a warm-started inner loop on $\Psi_R(\theta) := F(\theta) + \frac{1}{2\lambda_{\mathrm{env}}}\|\theta - \theta_R\|_2^2$, where proximal subproblems are solved only up to a prescribed accuracy (e.g., (Schmidt et al., 2011)). Appendix E.1 formalizes this viewpoint for our setting and provides a residual-based stopping rule. Lemma E.3 yields the practical takeaway: if $\bar\theta_R$ satisfies the proximal residual condition $\mathrm{dist}(0, \partial\Psi_R(\bar\theta_R)) \le \varepsilon_{\mathrm{prox}}$, then

$$\mathrm{dist}(0, \partial F(\bar\theta_R)) \;\le\; \|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2 \;+\; \varepsilon_{\mathrm{prox}} \;+\; \tfrac{1}{\lambda_{\mathrm{env}}}\|\bar\theta_R - \hat\theta_R\|_2.$$

Table 1: Main stress metrics at $\rho_{\mathrm{mis}} = 0.075$, averaged over 10 paired seeds. Lower is better for all columns.

Method	L_{id}	$R(\theta)$	L_{wc}	$\mathrm{AUC}_{0.5:1}$
Nominal SAIL ($q = 0$)	0.6923	0.4928	0.6960	0.6951
Robust SAIL (ours) ($q^\star = 1$)	0.6931	0.0025	0.6932	0.6932

Consequently, controlling $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2$ together with the inexactness terms $\varepsilon_{\mathrm{prox}}$ and $\|\bar{\theta}_R - \hat{\theta}_R\|_2$ yields an explicit near-stationarity certificate for the original objective F.

6 Experiments

We conduct a controlled finite-response experiment to validate the mechanism predicted by our robust decomposition: oracle-robust SAIL should reduce policy sensitivity to adversarial oracle perturbations, possibly at a small cost in ideal-oracle fitting.

Experimental setup. For each prompt x, we fix a candidate set $\mathcal{Y}_x = \{y_1, \ldots, y_K\}$ with $K = 12$ and train a $d = 7$-dimensional log-linear policy over this set. Comparisons are sampled on-policy from $d_\theta(x, y_i, y_j) = \mu(x)\pi_\theta(y_i \mid x)\pi_\theta(y_j \mid x)$. Let $p_{\mathrm{id}}(z) = P^{\mathrm{id}}(y_i \succ y_j \mid x)$ for $z = (x, y_i, y_j)$. The observed training oracle is a bounded misspecification of the ideal oracle, $p_{\mathrm{obs}}(z) = p_{\mathrm{id}}(z) + \rho_{\mathrm{mis}} b(z)$, where b is an antisymmetric nuisance-bias direction: it is induced by a style/redundancy score orthogonal to the true-reward direction, pushes preference toward the response with larger nuisance score only when that ordering conflicts with the true reward, and is normalized per prompt so that $|b(z)| \le 1$. See Appendix F for details. We sweep $\rho_{\mathrm{mis}} \in \{0.05, 0.075, 0.10\}$ over 10 paired seeds.

Nominal SAIL corresponds to $q = 0$ and minimizes $L_{\mathrm{SAIL}}(\theta; P^{\mathrm{obs}})$. Robust SAIL minimizes $L_{\mathrm{SAIL}}(\theta; P^{\mathrm{obs}}) + \rho_{\mathrm{alg}}\beta R(\theta)$, where $\beta = 0.1$, $q = \rho_{\mathrm{alg}}/\rho_{\mathrm{mis}}$, $R(\theta) = \mathbb{E}_{z\sim d_\theta}|h_\theta(z)|$, and $h_\theta(z) = \log \frac{\pi_\theta(y_i|x)}{\pi_{\mathrm{ref}}(y_i|x)} - \log \frac{\pi_\theta(y_j|x)}{\pi_{\mathrm{ref}}(y_j|x)}$. The operating point $q^\star$ is selected by seed-pooled validation high-stress AUC; validation selects $q^\star = 1$ for all three radii, so the main robust comparison uses $q^\star = 1$. See Appendix F.5 for details.

True-to-worst oracle stress metric. The policy-specific worst oracle chooses the loss-maximizing endpoint around p_{obs}: $p^\theta_{\mathrm{wc}}(z) = \mathrm{clip}_{[0,1]}\{p_{\mathrm{obs}}(z) + \rho_{\mathrm{mis}}\,\mathrm{sgn}(\ell^1_\theta(z) - \ell^0_\theta(z))\}$. The true-to-worst path is $p^\theta_\alpha(z) = (1-\alpha)p_{\mathrm{id}}(z) + \alpha p^\theta_{\mathrm{wc}}(z)$ for $\alpha \in [0,1]$, and $L_{\mathrm{stress}}(\theta;\alpha) = \mathbb{E}_{z\sim d_\theta}[p^\theta_\alpha(z)\ell^1_\theta(z) + (1 - p^\theta_\alpha(z))\ell^0_\theta(z)]$. We report $L_{\mathrm{id}} = L_{\mathrm{stress}}(\theta;0)$, $L_{\mathrm{wc}} = L_{\mathrm{stress}}(\theta;1)$, and $\mathrm{AUC}_{0.5:1} = \frac{1}{6}\sum_{\alpha\in\{0.5,0.6,0.7,0.8,0.9,1.0\}} L_{\mathrm{stress}}(\theta;\alpha)$. All loss-style metrics are lower-is-better.

Ignoring endpoint clipping when the interval is already feasible, the worst-case loss decomposes as

$$L_{\mathrm{wc}}(\theta) = L_{\mathrm{id}}(\theta) + \rho_{\mathrm{mis}}\,\mathbb{E}_{z\sim d_\theta}\big[b(z)\big(\ell^1_\theta(z) - \ell^0_\theta(z)\big)\big] + \rho_{\mathrm{mis}}\beta R(\theta). \tag{23}$$

Thus L_{wc} separates ideal-oracle performance, the fixed observed-oracle bias, and the algorithmic robustness penalty controlled by $\rho_{\mathrm{alg}}R(\theta)$.

Results. Table 1 and Fig. 1 show the main comparison at $\rho_{\mathrm{mis}} = 0.075$. Robust SAIL slightly sacrifices ideal-oracle fitting, increasing L_{id} from 0.6923 to 0.6931, but reduces the sensitivity penalty from $R(\theta) = 0.4928$ to 0.0025. As a result, its stress curve is nearly flat: the worst-case endpoint improves from $L_{\mathrm{wc}} = 0.6960$ to 0.6932, and the AUC improves from 0.6951 to 0.6932. The paired curves cross at $\alpha \approx 0.23$, meaning Robust SAIL becomes preferable once the evaluation oracle moves about 23% of the way from the ideal oracle toward the worst-case oracle. Across $\rho_{\mathrm{mis}} \in \{0.05, 0.075, 0.10\}$, validation selects $q^\star = 1$ and Robust SAIL improves L_{wc}, $\mathrm{AUC}_{0.5:1}$, and $R(\theta)$ on all $10/10$ paired seeds; this is a controlled radius-calibration result, with deployment-time radius selection discussed in Appendix F.

7 Conclusion

We studied oracle-robust online alignment of language models, where preference feedback is collected on-policy from a nominal training oracle $P^\star$ that may be misspecified. We introduced a pointwise

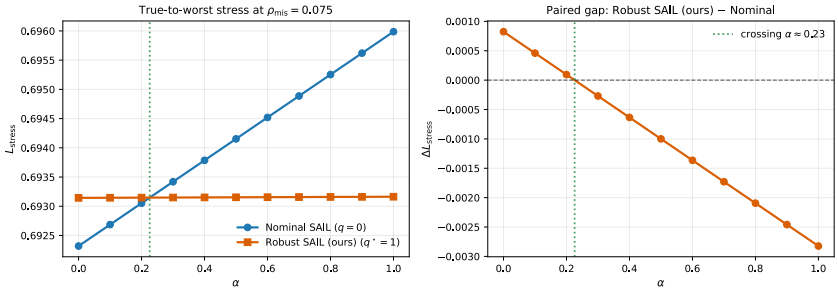

Figure 1: Left: true-to-worst oracle stress curve at $\rho_{\text{mis}} = 0.075$. Right: Robust SAIL sacrifices ideal-oracle fitting but remains nearly flat along the stress path and becomes better after $\alpha \approx 0.23$.

oracle uncertainty set $U^W(P^\star, \rho)$ and defined the robust objective $L_\rho^W(\theta)$ as the worst-case negative log-likelihood over $P \in U^W(P^\star, \rho)$. Our main novelty is an exact closed-form decomposition of this otherwise difficult minimax objective: $L_\rho^W(\theta) = L_{\text{SAIL}}(\theta) + \lambda R(\theta)$ with $\lambda = \rho\beta$. For log-linear policies over a finite response set, we establish an $\widetilde{O}(\varepsilon^{-2})$ stochastic-oracle complexity bound for reaching a squared Moreau-envelope stationarity, which implies proximity to a nearly stationary point of the original robust objective; as a special case, setting $\rho = 0$ recovers a convergence-to-stationarity guarantee for optimizing $L_{\text{SAIL}}(\theta)$. Our experiments show that, under structured oracle misspecification, Robust SAIL trades a small increase in ideal-oracle loss for a sharp reduction in oracle sensitivity, improving worst-case loss and high-stress AUC across all tested radii and paired seeds.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Science Foundation (NSF) Future Manufacturing (FM) program through Award No. 2229250. The authors declare no competing interests.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Chenjia Bai, Yang Zhang, Shuang Qiu, Qiaosheng Zhang, Kang Xu, and Xuelong Li. Online preference alignment for language models via count-based exploration. In *International Conference on Learning Representations*, volume 2025, pages 102164–102186, 2025.

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

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 1st edition, 2011.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pages 24410–24449, 2024.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

Mucong Ding, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, Amrit Bedi, and Furong Huang. SAIL: Self-improving efficient online alignment of large language models. *arXiv preprint arXiv:2406.15567*, 2024.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

Mudit Gaur, Utsav Singh, Amrit Singh Bedi, Raghu Pasupathy, and Vaneet Aggarwal. On the sample complexity bounds of bilevel reinforcement learning. *Advances in Neural Information Processing Systems*, 38:67538–67570, 2025.

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

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, New York, NY, 2004.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 52516–52544, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741, 2023.

Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. In *Advances in Neural Information Processing Systems*, volume 37, pages 37100–37137, 2024.

Saverio Salzo and Silvia Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex Analysis*, 19:1167–1192, 2012.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015. PMLR, 2023.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and RLHF. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020.

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

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *International Conference on Learning Representations*, volume 2025, pages 16123–16150, 2025.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. *Advances in Neural Information Processing Systems*, 38:27039–27076, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Matthew S Zhang, Murat A Erdogdu, and Animesh Garg. Convergence and optimality of policy gradient methods in weakly smooth settings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9066–9073, 2022.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan Awadalla, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *Transactions on Machine Learning Research*, 2025.

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

A Related Work

Bilevel optimization provides a principled abstraction for hierarchical learning problems and is increasingly relevant to online alignment, where preference learning and policy optimization are coupled through data generated by the current policy (Christiano et al., 2017; Ouyang et al., 2022; Chakraborty et al., 2024). SAIL reduces this online RLHF bilevel structure to a single-level first-order surrogate through reward-policy equivalence (Ding et al., 2024), while recent theory studies nonconvex bilevel RL and penalty-based bilevel methods (Gaur et al., 2025; Shen and Chen, 2023; Shen et al., 2025). Recent online preference-alignment methods further study exploration and active preference elicitation for on-policy feedback collection, including count-based online preference optimization and self-exploring language models (Bai et al., 2025; Zhang et al., 2025). These works model policy-dependent data collection, but they do not optimize against an explicit oracle uncertainty set.

Robust alignment work studies preference and data shift through distributionally robust objectives, including robust DPO variants with Wasserstein or KL uncertainty sets and group-robust preference optimization (Rafailov et al., 2023; Xu et al., 2025; Wu et al., 2025; Ramesh et al., 2024). These methods build on the broader DRO literature (Rahimian and Mehrotra, 2022; Mohajerin Esfahani and Kuhn, 2018; Kuhn et al., 2019; Duchi and Namkoong, 2021), but are typically formulated for fixed offline datasets. Our setting differs by placing oracle uncertainty inside the on-policy sampling loop.

B Proof of Theorem 3.2: Decomposition of $L_\rho^W(\theta)$

B.1 A pointwise maximization identity

Lemma B.1 (Pointwise worst-case Bernoulli perturbation)**.** *Fix any $p^\star \in [0,1]$, any $\rho \geq 0$ such that $[p^\star - \rho, p^\star + \rho] \subseteq [0,1]$, and any real numbers $a, b \in \mathbb{R}$. Then*

$$\sup_{p \in [p^\star - \rho, p^\star + \rho]} \Big(p\, a + (1-p)\, b \Big) = p^\star a + (1-p^\star) b + \rho\, |a - b|. \tag{24}$$

Moreover, an optimizer is $p^\star + \rho$ if $a \geq b$ and $p^\star - \rho$ if $a < b$.

Proof. The map $p \mapsto p\, a + (1-p)\, b = b + p(a-b)$ is affine in p, hence it attains its maximum over the interval $[p^\star - \rho, p^\star + \rho]$ at an endpoint. If $a - b \geq 0$ the maximizer is $p^\star + \rho$, giving value $b + (p^\star + \rho)(a-b) = p^\star a + (1-p^\star)b + \rho(a-b)$. If $a - b < 0$ the maximizer is $p^\star - \rho$, giving $b + (p^\star - \rho)(a-b) = p^\star a + (1-p^\star)b + \rho(b-a)$. Combining the two cases yields (24). □

B.2 Proof of Theorem 3.2

Proof of Theorem 3.2. Recall: (i) the policy-dependent sampling law d_θ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ (Eq. (1)), (ii) the pointwise oracle uncertainty set $U^W(P^\star, \rho)$ (Eq. (3)), (iii) the robust objective $L_\rho^W(\theta)$ (Eq. (9)), and (iv) the SAIL objective $L^{\mathrm{SAIL}}(\theta)$.

Step 1: Rewrite the robust objective pointwise in $p(z)$. For $z = (x, y_1, y_2)$, let $p(z) := P(1 \mid z) = P(y_1 \succ y_2 \mid x)$ and $p^\star(z) := P^\star(1 \mid z)$. By Eq. (3), $P \in U^W(P^\star, \rho)$ iff for all z,

$$|p(z) - p^\star(z)| \leq \rho. \tag{25}$$

Assumption 1 ensures $p^\star(z) \in [\delta, 1-\delta]$ and $\rho \in (0, \delta)$, so $[p^\star(z) - \rho, p^\star(z) + \rho] \subseteq [0,1]$ for all z.

Define the pairwise logit as

$$h_\theta(z) \equiv h_\theta(x, y_1, y_2) := \log \frac{\pi_\theta(y_1 \mid x)}{\pi_{\mathrm{SFT}}(y_1 \mid x)} - \log \frac{\pi_\theta(y_2 \mid x)}{\pi_{\mathrm{SFT}}(y_2 \mid x)}.$$

For notational convenience, set the two label-conditional losses

$$\ell_\theta^1(z) := -\log \sigma\big(\beta\, h_\theta(z)\big), \qquad \ell_\theta^0(z) := -\log \sigma\big(\beta\, h_\theta(x, y_2, y_1)\big) = -\log \sigma\big(-\beta\, h_\theta(z)\big), \tag{26}$$

where we used $h_\theta(x, y_2, y_1) = -h_\theta(x, y_1, y_2)$.

Then, for any oracle P with Bernoulli parameter $p(z)$, the conditional risk equals

$$\mathbb{E}_{y\sim P(\cdot|z)}\big[\ell_\theta(z,y)\big] = p(z)\,\ell_\theta^1(z) + (1-p(z))\,\ell_\theta^0(z) =: \ell\big(p(z); h_\theta(z)\big).$$

Therefore $L_\rho^W(\theta)$ can be written as

$$L_\rho^W(\theta) = \sup_{P\in U^W(P^\star,\rho)} \mathbb{E}_{z\sim d_\theta}\Big[p(z)\,\ell_\theta^1(z) + (1-p(z))\,\ell_\theta^0(z)\Big]. \tag{27}$$

Because the constraint (25) is pointwise in z and the objective is an integral (expectation) of a pointwise affine function of $p(z)$, the worst-case oracle can be chosen pointwise in z. Equivalently, the supremum in (27) equals the expectation of the pointwise supremum:

$$L_\rho^W(\theta) = \mathbb{E}_{z\sim d_\theta}\left[\sup_{p\in[p^\star(z)-\rho,\,p^\star(z)+\rho]} \Big(p\,\ell_\theta^1(z) + (1-p)\,\ell_\theta^0(z)\Big)\right]. \tag{28}$$

By Lemma B.1, an optimizer is always an endpoint, hence a measurable selector can be obtained by taking $p(z) = p^\star(z) + \rho$ when $\ell_\theta^1(z) \geq \ell_\theta^0(z)$ and $p(z) = p^\star(z) - \rho$ otherwise.

Apply Lemma B.1 to (28) with $a = \ell_\theta^1(z)$, $b = \ell_\theta^0(z)$ and $p^\star = p^\star(z)$. This gives

$$L_\rho^W(\theta) = \mathbb{E}_{z\sim d_\theta}\Big[p^\star(z)\,\ell_\theta^1(z) + (1-p^\star(z))\,\ell_\theta^0(z)\Big] + \rho\,\mathbb{E}_{z\sim d_\theta}\Big[|\ell_\theta^1(z) - \ell_\theta^0(z)|\Big]. \tag{29}$$

The first expectation in (29) is exactly $L^{\mathrm{SAIL}}(\theta)$.

For the second term, use the identity

$$\log\sigma(t) - \log\sigma(-t) = t \quad (\text{since } \sigma(t)/\sigma(-t) = e^t).$$

With $t = \beta h_\theta(z)$ and definitions (26), we obtain

$$\ell_\theta^1(z) - \ell_\theta^0(z) = -\log\sigma(\beta h_\theta(z)) + \log\sigma(-\beta h_\theta(z)) = -\beta h_\theta(z).$$

Hence $|\ell_\theta^1(z) - \ell_\theta^0(z)| = \beta|h_\theta(z)|$, and (29) becomes

$$L_\rho^W(\theta) = L_{\mathrm{SAIL}}(\theta) + \rho\beta\,\mathbb{E}_{z\sim d_\theta}\big[|h_\theta(z)|\big]. \tag{30}$$

Define $\lambda := \rho\beta$.

Under Assumption 2, for fixed x we have the log-linear / softmax form

$$\pi_\theta(y \mid x) = \frac{\exp(\theta^\top\psi(x,y))}{\sum_{y'\in\mathcal{Y}}\exp(\theta^\top\psi(x,y'))}.$$

By Eq. (10), $\pi_{\mathrm{SFT}} = \pi_{\theta_{\mathrm{ref}}}$ for some θ_{ref}. Then

$$\log\frac{\pi_\theta(y\mid x)}{\pi_{\mathrm{SFT}}(y\mid x)} = (\theta-\theta_{\mathrm{ref}})^\top\psi(x,y) - \log\sum_{y'} e^{\theta^\top\psi(x,y')} + \log\sum_{y'} e^{\theta_{\mathrm{ref}}^\top\psi(x,y')}.$$

Taking the difference between y_1 and y_2 cancels the log-partition terms, yielding

$$h_\theta(x,y_1,y_2) = (\theta-\theta_{\mathrm{ref}})^\top\big(\psi(x,y_1) - \psi(x,y_2)\big). \tag{31}$$

Then (31) shows $h_\theta(z) = R(\theta; x, y_1, y_2)$ where $R(\theta; x, y_1, y_2)$ is defined in Theorem 3.2, and thus the robust correction term in (30) is precisely $\lambda R(\theta)$ with

$$R(\theta) := \mathbb{E}_{z\sim d_\theta}\big[|R(\theta; x, y_1, y_2)|\big].$$

Substituting into (30) gives the claimed decomposition

$$L_\rho^W(\theta) = L_{\mathrm{SAIL}}(\theta) + \lambda R(\theta), \qquad \lambda = \rho\beta.$$

This completes the proof. □

C Weak convexity: proofs of Theorems 3.6 and 3.7

C.1 A one-dimensional nonconvexity example

Consider a single prompt $X = \{x\}$, two responses $Y = \{a, b\}$, and $d = 1$. Fix $\theta_{\mathrm{ref}} = 0$ and define features by $\psi(x, a) = 1$ and $\psi(x, b) = 0$, so that

$$R(\theta; x, y_1, y_2) = \theta\big(\psi(x, y_1) - \psi(x, y_2)\big) \in \{0, \pm\theta\}.$$

Let $\pi_\theta(a \mid x) = \sigma(\theta)$ and $\pi_\theta(b \mid x) = \sigma(-\theta)$, and sample $y_1, y_2 \overset{\text{iid}}{\sim} \pi_\theta(\cdot \mid x)$. Then

$$R(\theta) = |\theta| \Pr(y_1 \neq y_2) = 2|\theta|\sigma(\theta)\sigma(-\theta) = \frac{2|\theta|e^\theta}{(1+e^\theta)^2}.$$

Convexity would imply $R(1) \leq \frac{1}{2}(R(0) + R(2)) = \frac{1}{2}R(2)$, but direct calculation gives

$$R(1) - \tfrac{1}{2}R(2) = \frac{2e(e-1)(e^3-1)}{(1+e)^2(1+e^2)^2} > 0.$$

Thus R is not convex in general.

C.2 Auxiliary lemmas

Lemma C.1 (Policy-regularity constants for the log-linear policy class)**.** *Assume Assumptions 2 and 4. Define the policy score $g_\theta(x, y) := \nabla_\theta \log \pi_\theta(y \mid x)$. Then for all $\theta, \theta' \in \mathbb{R}^d$ and all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$\|g_\theta(x, y)\|_2 \leq G, \qquad \|g_\theta(x, y) - g_{\theta'}(x, y)\|_2 \leq M\|\theta - \theta'\|_2,$$

with explicit constants $G := 2B_\psi$ and $M := B_\psi^2$.

Moreover, for $z = (x, y_1, y_2)$ and $d_\theta(z) = \mu(x)\pi_\theta(y_1 \mid x)\pi_\theta(y_2 \mid x)$, define score (Williams, 1992)

$$S_\theta(z) := \nabla_\theta \log d_\theta(z) = g_\theta(x, y_1) + g_\theta(x, y_2).$$

Then

$$\|S_\theta(z)\|_2 \leq 2G = 4B_\psi, \qquad \|\nabla_\theta S_\theta(z)\|_{\mathrm{op}} \leq 2M = 2B_\psi^2.$$

Proof. Under Assumption 2, for fixed x the policy is

$$\pi_\theta(y \mid x) = \frac{\exp(\theta^\top \psi(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top \psi(x, y'))}.$$

Hence

$$\log \pi_\theta(y \mid x) = \theta^\top \psi(x, y) - \log \sum_{y' \in \mathcal{Y}} e^{\theta^\top \psi(x, y')},$$

so differentiating gives

$$g_\theta(x, y) = \psi(x, y) - \sum_{y' \in \mathcal{Y}} \pi_\theta(y' \mid x)\psi(x, y') = \psi(x, y) - \mu_\theta(x), \quad \mu_\theta(x) := \mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}[\psi(x, y')].$$

By Assumption 4, $\|\psi(x, y)\|_2 \leq B_\psi$ and Jensen implies $\|\mu_\theta(x)\|_2 \leq B_\psi$, hence $\|g_\theta(x, y)\|_2 \leq 2B_\psi$. This gives $G := 2B_\psi$.

Next, $\nabla_\theta g_\theta(x, y) = -\nabla_\theta \mu_\theta(x)$. A standard calculation yields that the Jacobian of $\mu_\theta(x)$ is the covariance matrix

$$\nabla_\theta \mu_\theta(x) = \mathrm{Cov}_{y' \sim \pi_\theta(\cdot|x)}\big(\psi(x, y')\big).$$

For any unit vector $v \in \mathbb{R}^d$,

$$v^\top \nabla_\theta \mu_\theta(x) v = \mathrm{Var}\big(v^\top \psi(x, y')\big) \leq \mathbb{E}\big[(v^\top \psi(x, y'))^2\big] \leq \mathbb{E}\|\psi(x, y')\|_2^2 \leq B_\psi^2.$$

Taking the supremum over $\|v\|_2 = 1$ yields $\|\nabla_\theta \mu_\theta(x)\|_{\mathrm{op}} \leq B_\psi^2$, so $\|\nabla_\theta g_\theta(x, y)\|_{\mathrm{op}} \leq B_\psi^2$ as well. By the mean value theorem,

$$\|g_\theta(x, y) - g_{\theta'}(x, y)\|_2 \leq B_\psi^2 \|\theta - \theta'\|_2,$$

so we may take $M := B_\psi^2$.

Finally, since $S_\theta(z) = g_\theta(x, y_1) + g_\theta(x, y_2)$, the bounds $\|S_\theta(z)\|_2 \leq 2G$ and $\|\nabla_\theta S_\theta(z)\|_{\mathrm{op}} \leq 2M$ follow by triangle inequality. □

Lemma C.2 (Bounded Fisher information)**.** *Define the Fisher information matrix*

$$\mathcal{F}(\theta) := \mathbb{E}_{x\sim\mu,\ y\sim\pi_\theta(\cdot|x)}\big[g_\theta(x,y)g_\theta(x,y)^\top\big].$$

Then $\|\mathcal{F}(\theta)\|_{\mathrm{op}} \le G^2$ *for all* $\theta \in \Theta$.

Proof. For any unit vector $u \in \mathbb{R}^d$,

$$u^\top \mathcal{F}(\theta) u = \mathbb{E}\big[\langle u, g_\theta(x,y)\rangle^2\big] \le \mathbb{E}\big[\|g_\theta(x,y)\|_2^2\big] \le G^2.$$

Taking the supremum over $\|u\|_2 = 1$ yields $\|\mathcal{F}(\theta)\|_{\mathrm{op}} \le G^2$. □

Remark C.3. In the weak convexity analysis of $R(\theta)$, we differentiate an expectation taken under the θ-*dependent* law d_θ. The resulting Hessian formula contains curvature terms involving the score and its Jacobian, notably $r_{\varepsilon,\theta}(z)\,\nabla_\theta S_\theta(z)$ where $S_\theta(z) = \nabla_\theta \log d_\theta(z)$ and $r_{\varepsilon,\theta}(z)$ is a smoothed absolute value. To obtain a *uniform* lower bound $\nabla^2 R_\varepsilon(\theta) \succeq -\kappa_\varepsilon I$ (hence weak convexity), we must control the worst-case quadratic form $v^\top(\nabla_\theta S_\theta(z))v$ uniformly over $\|v\|_2 = 1$. This is exactly an operator-norm control. Lemma C.2 exemplifies this type of bound for second-moment (information) matrices, and the proof of Lemma C.1 uses the same operator-norm reasoning to bound $\|\nabla_\theta S_\theta(z)\|_{\mathrm{op}}$, which is the quantity that directly enters the proof of Theorem 3.6.

C.3 Proof of Theorem 3.6 (weak convexity of the robust penalty)

Proof of Theorem 3.6. Recall the robust penalty

$$R(\theta) := \mathbb{E}_{z\sim d_\theta}\big[|R(\theta; x, y_1, y_2)|\big],\ R(\theta; x, y_1, y_2) := (\theta - \theta_{\mathrm{ref}})^\top\big(\psi(x,y_1) - \psi(x,y_2)\big),$$

where $z = (x, y_1, y_2)$. By (31), $R(\theta; x, y_1, y_2) = h_\theta(z)$, the pairwise log-ratio score.

Step 1: Smooth approximation. Fix $\varepsilon > 0$ and define $\phi_\varepsilon(u) := \sqrt{u^2 + \varepsilon^2}$. Set

$$R_\varepsilon(\theta) := \mathbb{E}_{z\sim d_\theta}\big[\phi_\varepsilon(h_\theta(z))\big].$$

Since $\mathcal{Y}$ is finite, the expectation over d_θ is a finite sum, so R_ε is twice continuously differentiable and we may differentiate under the sum.

We will show that $\nabla^2 R_\varepsilon(\theta) \succeq -\kappa_\varepsilon I$ for all $\theta \in \Theta$ with

$$\kappa_\varepsilon := 8GB_\psi + 4MDB_\psi + 2M\varepsilon, \tag{32}$$

where G, M are as in Lemma C.1. Since R_ε is C^2 and has a global Hessian lower bound, the function $R_\varepsilon(\cdot) + \frac{\kappa_\varepsilon}{2}\|\cdot\|_2^2$ has positive semidefinite Hessian and is convex; equivalently, R_ε is κ_ε-weakly convex.

Step 2: Uniform bounds. Define $\Delta\psi(z) := \psi(x,y_1) - \psi(x,y_2)$. By Assumption 4,

$$\|\Delta\psi(z)\|_2 \le \|\psi(x,y_1)\|_2 + \|\psi(x,y_2)\|_2 \le 2B_\psi.$$

By Assumption 5, $\|\theta - \theta_{\mathrm{ref}}\|_2 \le D$ for all $\theta \in \Theta$, hence

$$|h_\theta(z)| \le \|\theta - \theta_{\mathrm{ref}}\|_2\,\|\Delta\psi(z)\|_2 \le 2DB_\psi.$$

Moreover, $\phi_\varepsilon(u) \le |u| + \varepsilon$, so

$$0 \le \phi_\varepsilon(h_\theta(z)) \le 2DB_\psi + \varepsilon. \tag{33}$$

Step 3: Score-function calculus for d_θ. Write $d_\theta(z) = \mu(x)\pi_\theta(y_1 \mid x)\pi_\theta(y_2 \mid x)$ and define the score (Williams, 1992)

$$S_\theta(z) := \nabla_\theta \log d_\theta(z).$$

Then $S_\theta(z) = g_\theta(x,y_1) + g_\theta(x,y_2)$, hence Lemma C.1 gives

$$\|S_\theta(z)\|_2 \le 2G, \qquad \|\nabla_\theta S_\theta(z)\|_{\mathrm{op}} \le 2M. \tag{34}$$

For the integrand $r_{\varepsilon,\theta}(z) := \phi_\varepsilon(h_\theta(z))$, note that

$$\nabla_\theta h_\theta(z) = \Delta\psi(z), \qquad \nabla_\theta r_{\varepsilon,\theta}(z) = \phi_\varepsilon'(h_\theta(z))\,\Delta\psi(z),$$

and since $|\phi'_\varepsilon(u)| = |u|/\sqrt{u^2+\varepsilon^2} \le 1$,

$$\|\nabla_\theta r_{\varepsilon,\theta}(z)\|_2 \le \|\Delta\psi(z)\|_2 \le 2B_\psi. \tag{35}$$

Also,

$$\nabla^2_\theta r_{\varepsilon,\theta}(z) = \phi''_\varepsilon(h_\theta(z))\,\Delta\psi(z)\Delta\psi(z)^\top, \qquad \phi''_\varepsilon(u) = \frac{\varepsilon^2}{(u^2+\varepsilon^2)^{3/2}} \ge 0,$$

so $\nabla^2_\theta r_{\varepsilon,\theta}(z) \succeq 0$ for all (θ, z).

Because $\mathcal{Y}$ is finite, we can differentiate $R_\varepsilon(\theta) = \sum_z d_\theta(z)\, r_{\varepsilon,\theta}(z)$ term-by-term and use $\nabla_\theta d_\theta(z) = d_\theta(z) S_\theta(z)$ to obtain:

$$\nabla_\theta R_\varepsilon(\theta) = \mathbb{E}_{z\sim d_\theta}\big[\nabla_\theta r_{\varepsilon,\theta}(z) + r_{\varepsilon,\theta}(z) S_\theta(z)\big], \tag{36}$$

$$\begin{aligned}\nabla^2_\theta R_\varepsilon(\theta) = \mathbb{E}_{z\sim d_\theta}\Big[&\nabla^2_\theta r_{\varepsilon,\theta}(z) + S_\theta(z)\nabla_\theta r_{\varepsilon,\theta}(z)^\top + \nabla_\theta r_{\varepsilon,\theta}(z) S_\theta(z)^\top \\ &+ r_{\varepsilon,\theta}(z)\big(S_\theta(z)S_\theta(z)^\top + \nabla_\theta S_\theta(z)\big)\Big].\end{aligned} \tag{37}$$

Step 4: Hessian lower bound. Fix any unit vector $u \in \mathbb{R}^d$. Using (37) and the facts that $\nabla^2_\theta r_{\varepsilon,\theta}(z) \succeq 0$ and $S_\theta(z)S_\theta(z)^\top \succeq 0$, we have

$$u^\top \nabla^2_\theta R_\varepsilon(\theta) u \ge \mathbb{E}_{z\sim d_\theta}\Big[u^\top\big(S_\theta \nabla_\theta r_{\varepsilon,\theta}^\top + \nabla_\theta r_{\varepsilon,\theta} S_\theta^\top\big)u + r_{\varepsilon,\theta}(z)\, u^\top(\nabla_\theta S_\theta(z))u\Big].$$

For the first term,

$$u^\top\big(S_\theta \nabla_\theta r_{\varepsilon,\theta}^\top + \nabla_\theta r_{\varepsilon,\theta} S_\theta^\top\big)u = 2\langle u, S_\theta(z)\rangle\,\langle u, \nabla_\theta r_{\varepsilon,\theta}(z)\rangle \ge -2\|S_\theta(z)\|_2\,\|\nabla_\theta r_{\varepsilon,\theta}(z)\|_2,$$

and using (34) and (35) gives the pointwise bound

$$u^\top\big(S_\theta \nabla_\theta r_{\varepsilon,\theta}^\top + \nabla_\theta r_{\varepsilon,\theta} S_\theta^\top\big)u \ge -2(2G)(2B_\psi) = -8GB_\psi.$$

For the second term, by (34), $u^\top(\nabla_\theta S_\theta(z))u \ge -\|\nabla_\theta S_\theta(z)\|_{\mathrm{op}} \ge -2M$, so with (33) we have

$$r_{\varepsilon,\theta}(z)\, u^\top(\nabla_\theta S_\theta(z))u \ge -(2DB_\psi + \varepsilon)\,2M = -4MDB_\psi - 2M\varepsilon.$$

Combining the two bounds and taking expectation yields

$$u^\top \nabla^2_\theta R_\varepsilon(\theta) u \ge -(8GB_\psi + 4MDB_\psi + 2M\varepsilon) = -\kappa_\varepsilon,$$

for all unit u, hence $\nabla^2_\theta R_\varepsilon(\theta) \succeq -\kappa_\varepsilon I$.

Step 5: Pass to the nonsmooth limit $\varepsilon \downarrow 0$. For any $u \in \mathbb{R}$, $0 \le \phi_\varepsilon(u) - |u| \le \varepsilon$, hence $0 \le R_\varepsilon(\theta) - R(\theta) \le \varepsilon$ for all θ. Thus $R_\varepsilon \to R$ uniformly as $\varepsilon \downarrow 0$. Since each R_ε is κ_ε-weakly convex and $\kappa_\varepsilon \downarrow \kappa_R := 8GB_\psi + 4MDB_\psi$, taking $\varepsilon \downarrow 0$ in the defining weak convexity inequality yields that R is κ_R-weakly convex.

Finally, by Lemma C.1, we may take $G = 2B_\psi$ and $M = B_\psi^2$, giving

$$\kappa_R \le 8(2B_\psi)B_\psi + 4(B_\psi^2)DB_\psi = 16B_\psi^2 + 4DB_\psi^3,$$

which is the claimed bound in Theorem 3.6. □

C.4 Proof of Theorem 3.7

Proof of Theorem 3.7. By Theorem 3.2 we have the exact decomposition

$$L^W_\rho(\theta) = L_{\mathrm{SAIL}}(\theta) + \lambda R(\theta), \qquad \lambda = \rho\beta.$$

Under Assumption 3, L_{SAIL} has L_{SAIL}-Lipschitz gradient on $\mathbb{R}^d$. A basic fact is that any L-smooth function is L-weakly convex. This can be verified directly from the second-order characterization: if f is L-smooth, then $\nabla^2 f(\theta) \succeq -LI$. Equivalently, the function

$$\theta \;\mapsto\; f(\theta) + \frac{L}{2}\|\theta\|_2^2$$

is convex.

By Theorem 3.6, R is κ_R-weakly convex on Θ. Scaling preserves weak convexity, so λR is $(\lambda\kappa_R)$-weakly convex. Sums of weakly convex functions add their curvature parameters, hence $L^W_\rho = L_{\mathrm{SAIL}} + \lambda R$ is κ-weakly convex on Θ with

$$\kappa := L_{\mathrm{SAIL}} + \lambda\kappa_R.$$

This concludes the proof of Theorem 3.7. □

C.5 Proof of Lemma 3.10

Proof. Fix an iteration t and condition on $\theta_t = \theta \in \Theta$. The complete triples $z_i = (x_i, y_{1,i}, y_{2,i})$, $i = 1, \dots, B$, are then i.i.d. draws from d_θ. Given these triples, the preference labels are drawn independently from the nominal oracle $P^\star$; we write $(y_{w,i}, y_{\ell,i})$ for the resulting preference ordering of $(y_{1,i}, y_{2,i})$ and regard the labels as part of the batch Z_t.

For each i, define

$$\Delta\psi_i := \psi(x_i, y_{1,i}) - \psi(x_i, y_{2,i}), \qquad h_i := (\theta - \theta_{\mathrm{ref}})^\top \Delta\psi_i,$$
$$S_i := \nabla_\theta \log \pi_\theta(y_{1,i} \mid x_i) + \nabla_\theta \log \pi_\theta(y_{2,i} \mid x_i),$$

and let $\Delta\psi_i^{w\ell}$ and $h_i^{w\ell}$ be defined in the same way, with $(y_{1,i}, y_{2,i})$ replaced by $(y_{w,i}, y_{\ell,i})$. Let $s_0(0) = 0$ and $s_0(a) = \mathrm{sgn}(a)$ for $a \neq 0$, and define the sample directions

$$H_i^{\mathrm{SAIL}} := -\beta\sigma\big(-\beta h_i^{w\ell}\big)\Delta\psi_i^{w\ell} - \log\sigma\big(\beta h_i^{w\ell}\big) S_i,$$
$$H_i^{R} := s_0(h_i)\Delta\psi_i + |h_i| S_i.$$

The oracles are their mini-batch averages,

$$G^{\mathrm{SAIL}}(\theta; Z_t) := \frac{1}{B}\sum_{i=1}^{B} H_i^{\mathrm{SAIL}}, \tag{38}$$

$$G^{R}(\theta; Z_t) := \frac{1}{B}\sum_{i=1}^{B} H_i^{R}. \tag{39}$$

In both sample directions the first term is pathwise, i.e., a (sub)gradient of the per-sample loss with the sample held fixed, while the S_i-term is the score correction required because the sampling law d_θ depends on θ.

In the notation of Appendix C.3, $\Delta\psi_i = \Delta\psi(z_i)$ and $S_i = S_\theta(z_i)$, where $S_\theta(z)$ is the score of d_θ (Lemma C.1). Moreover, $h_i = h_\theta(z_i)$, where by (31) $h_\theta(z) = (\theta - \theta_{\mathrm{ref}})^\top \Delta\psi(z)$ is the pairwise log-ratio score; thus $R(\theta) = \mathbb{E}_{z\sim d_\theta}\big[|h_\theta(z)|\big]$ by (11). By Assumptions 4 and 5, the Cauchy–Schwarz inequality and Lemma C.1, for all $z \in Z$ and $\theta \in \Theta$,

$$\|\Delta\psi(z)\| \le 2B_\psi, \quad |h_\theta(z)| \le 2DB_\psi, \quad \|S_\theta(z)\| \le 4B_\psi, \quad \|\nabla_\theta S_\theta(z)\|_{\mathrm{op}} \le 2B_\psi^2. \tag{40}$$

Since both z_i and $(x_i, y_{w,i}, y_{\ell,i})$ lie in Z, these bounds apply to $\Delta\psi_i$, $\Delta\psi_i^{w\ell}$, h_i, $h_i^{w\ell}$ and S_i.

We first show that G^{SAIL} is unbiased. For a labeled comparison (x, y_w, y_ℓ), let $\Delta\psi^{w\ell}$ and $h^{w\ell}$ be defined as above and write $\ell_\theta := -\log\sigma(\beta h^{w\ell})$. By (8) and (31), $L^{\mathrm{SAIL}}(\theta)$ is the expectation of ℓ_θ when $z = (x, y_1, y_2) \sim d_\theta$ and (y_w, y_ℓ) is the ordering drawn from $P^\star$. Since $\nabla_\theta h^{w\ell} = \Delta\psi^{w\ell}$ and $(\log\sigma)'(a) = \sigma(-a)$, the pathwise derivative is

$$\nabla_\theta \ell_\theta = -\beta\sigma\big(-\beta h^{w\ell}\big)\Delta\psi^{w\ell}.$$

The prompt distribution and the conditional nominal oracle do not depend on θ, so the score of the sampling law of the labeled comparison comes entirely from the two responses and equals $S_\theta(z)$. Differentiating the policy-dependent expectation therefore gives

$$\nabla L^{\mathrm{SAIL}}(\theta) = \mathbb{E}\big[\nabla_\theta \ell_\theta + \ell_\theta S_\theta(z) \,\big|\, \theta\big] = \mathbb{E}\big[H_i^{\mathrm{SAIL}} \,\big|\, \theta\big].$$

Differentiation under the expectation is justified by dominated convergence: by (40) and $0 \le -\log\sigma(a) \le \log 2 + |a|$,

$$0 \le \ell_\theta \le \log 2 + 2\beta DB_\psi, \qquad \|\nabla_\theta \ell_\theta\| \le 2\beta B_\psi, \qquad \|S_\theta(z)\| \le 4B_\psi,$$

and the same bounds hold, with D replaced by $D + 1$, for all θ within distance 1 of Θ. Averaging over i in (38) yields

$$\mathbb{E}\big[G^{\mathrm{SAIL}}(\theta; Z_t) \mid \theta\big] = \nabla L^{\mathrm{SAIL}}(\theta).$$

Next we verify (19). Let $\bar g_R(\theta) := \mathbb{E}[H_i^R \mid \theta]$; by (39), $\mathbb{E}[G^R(\theta; Z_t) \mid \theta] = \bar g_R(\theta)$. For $\varepsilon > 0$, define

$$q_\varepsilon(a) := \sqrt{a^2 + \varepsilon^2} - \varepsilon, \qquad R_\varepsilon(\theta) := \mathbb{E}_{z\sim d_\theta}\big[q_\varepsilon(h_\theta(z))\big].$$

Then

$$0 \le q_\varepsilon(a) \le |a|, \qquad q'_\varepsilon(a) = \frac{a}{\sqrt{a^2+\varepsilon^2}}, \qquad q''_\varepsilon(a) = \frac{\varepsilon^2}{(a^2+\varepsilon^2)^{3/2}} \ge 0;$$

in particular $|q'_\varepsilon| \le 1$, and as $\varepsilon \downarrow 0$, $q_\varepsilon(a) \to |a|$ and $q'_\varepsilon(a) \to s_0(a)$ for every $a \in \mathbb{R}$, including $a = 0$.

We begin with a lower model for R_ε whose constant κ_R does not depend on ε; it follows from the calculation of Appendix C.3 with q_ε in place of ϕ_ε. Write $r_{\varepsilon,\theta}(z) := q_\varepsilon(h_\theta(z))$. Since $h_\theta(z)$ is affine in θ with gradient $\Delta\psi(z)$,

$$\nabla_\theta r_{\varepsilon,\theta}(z) = q'_\varepsilon(h_\theta(z))\Delta\psi(z), \qquad \nabla^2_\theta r_{\varepsilon,\theta}(z) = q''_\varepsilon(h_\theta(z))\Delta\psi(z)\Delta\psi(z)^\top \succeq 0.$$

The identities (36) and (37) were derived using only $\nabla_\theta d_\theta = d_\theta S_\theta$; hence they hold for this $r_{\varepsilon,\theta}$, and, as there, R_ε is twice continuously differentiable. In (37), the terms $\nabla^2_\theta r_{\varepsilon,\theta}$ and $r_{\varepsilon,\theta} S_\theta S_\theta^\top$ are positive semidefinite. Hence, for every $\theta \in \Theta$ and every unit vector $v \in \mathbb{R}^d$, using (40), $\|\nabla_\theta r_{\varepsilon,\theta}(z)\| \le \|\Delta\psi(z)\|$ and $0 \le r_{\varepsilon,\theta}(z) \le |h_\theta(z)|$,

$$\begin{aligned} v^\top \nabla^2 R_\varepsilon(\theta) v &\ge -\mathbb{E}_{z\sim d_\theta}\Big[2\|S_\theta(z)\|\,\|\nabla_\theta r_{\varepsilon,\theta}(z)\| + r_{\varepsilon,\theta}(z)\,\|\nabla_\theta S_\theta(z)\|_{\mathrm{op}}\Big] \\ &\ge -2(4B_\psi)(2B_\psi) - (2DB_\psi)(2B_\psi^2) = -\kappa_R, \end{aligned}$$

where $\kappa_R = 16B_\psi^2 + 4DB_\psi^3$ is the constant of Theorem 3.6. (The bound $q_\varepsilon(a) \le |a|$, in place of $\phi_\varepsilon(a) \le |a| + \varepsilon$, removes the term $2M\varepsilon$ in (32).) Now let $u \in \Theta$. Since Θ is convex, the segment from θ to u lies in Θ, and integrating the Hessian bound along it (Taylor's formula with integral remainder) gives the smooth lower inequality

$$R_\varepsilon(u) \ge R_\varepsilon(\theta) + \langle \nabla R_\varepsilon(\theta), u - \theta\rangle - \frac{\kappa_R}{2}\|u-\theta\|^2. \tag{41}$$

It remains to let $\varepsilon \downarrow 0$. Since $0 \le |a| - q_\varepsilon(a) \le \varepsilon$, we have $R_\varepsilon \to R$ uniformly on Θ. By (36),

$$\nabla R_\varepsilon(\theta) = \mathbb{E}_{z\sim d_\theta}\Big[q'_\varepsilon(h_\theta(z))\Delta\psi(z) + q_\varepsilon(h_\theta(z))S_\theta(z)\Big].$$

As $\varepsilon \downarrow 0$, the integrand converges pointwise to $s_0(h_\theta(z))\Delta\psi(z) + |h_\theta(z)|S_\theta(z)$, whose expectation under d_θ is $\bar g_R(\theta)$; by (40), its norm is at most $2B_\psi + (2DB_\psi)(4B_\psi)$, uniformly in ε. Dominated convergence therefore gives $\nabla R_\varepsilon(\theta) \to \bar g_R(\theta)$. Letting $\varepsilon \downarrow 0$ in (41), we obtain

$$R(u) \ge R(\theta) + \langle \bar g_R(\theta), u - \theta\rangle - \frac{\kappa_R}{2}\|u-\theta\|^2,$$

which is (19), because $\bar g_R(\theta) = \mathbb{E}[G^R(\theta; Z_t) \mid \theta]$. The argument uses only the segment from θ to u, so (19) holds for all $\theta, u \in \Theta$, including boundary points of Θ.

We now bound the conditional moments. Recall the constants $G_{\partial R}$ and $G_{\nabla\mathrm{SAIL}}$ defined before Theorem 4.1. By (40), $0 < \sigma < 1$ and $0 \le -\log\sigma(a) \le \log 2 + |a|$, we have

$$\begin{aligned} \|H_i^R\| &\le |s_0(h_i)|\,\|\Delta\psi_i\| + |h_i|\,\|S_i\| \le 2B_\psi + 8DB_\psi^2 = G_{\partial R}, \\ \|H_i^{\mathrm{SAIL}}\| &\le \beta\sigma\big(-\beta h_i^{w\ell}\big)\|\Delta\psi_i^{w\ell}\| + \big[-\log\sigma\big(\beta h_i^{w\ell}\big)\big]\|S_i\| \\ &\le 2\beta B_\psi + 4B_\psi(\log 2 + 2\beta DB_\psi) = G_{\nabla\mathrm{SAIL}}. \end{aligned}$$

Let H_i denote either H_i^{SAIL} or H_i^R, and let $m_H(\theta) := \mathbb{E}[H_i \mid \theta]$. Conditionally on θ, the vectors $H_1, \dots, H_B$ are i.i.d., because the complete triples are i.i.d. and each label depends only on its own triple. Hence the cross terms vanish, and

$$\begin{aligned} \mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{i=1}^B H_i - m_H(\theta)\Big\|^2 \;\Big|\; \theta\Big] &= \frac{1}{B^2}\sum_{i=1}^B \mathbb{E}\big[\|H_i - m_H(\theta)\|^2 \mid \theta\big] \\ &= \frac{1}{B}\Big(\mathbb{E}\big[\|H_1\|^2 \mid \theta\big] - \|m_H(\theta)\|^2\Big) \le \frac{1}{B}\,\mathbb{E}\big[\|H_1\|^2 \mid \theta\big]. \end{aligned}$$

Since $m_H(\theta) = \nabla L^{\mathrm{SAIL}}(\theta)$ for $H_i = H_i^{\mathrm{SAIL}}$ and $m_H(\theta) = \bar g_R(\theta)$ for $H_i = H_i^R$, the variance bounds of the lemma hold with

$$\sigma^2_{\mathrm{SAIL}} = G^2_{\nabla\mathrm{SAIL}}, \qquad \sigma_R^2 = G^2_{\partial R}.$$

Moreover, $\|H_i^R\| \le G_{\partial R}$ implies $\|\bar{g}_R(\theta)\| \le G_{\partial R}$ and $\|G^R(\theta; Z_t)\| \le G_{\partial R}$ almost surely; in particular, $\mathbb{E}[\|G^R(\theta; Z_t)\|^2 \mid \theta] \le G_{\partial R}^2$.

Finally, the triples and labels of iteration t are sampled afresh at θ_t, from a conditional law that depends on the past only through θ_t. Hence the mean identities, the lower model and the moment bounds above also hold conditionally on the pre-sampling history $\mathcal{F}_t$. □

C.6 Auxiliary lemma for stochastic gradient oracle

Lemma C.4 (Second-moment bound for gradient oracle)**.** *Fix any mini-batch size $B \in \mathbb{N}$, and recall the composite direction in Eq.* (20)*,*

$$G(\theta; Z) \;:=\; G_{\mathrm{SAIL}}(\theta; Z) \;+\; \lambda\, G_R(\theta; Z),$$

where Z denotes a generic mini-batch of size B drawn i.i.d. from d_θ (Lemma 3.10). Under Assumptions 2, 4, 5, the composite direction has bounded conditional second moment: for all $\theta \in \Theta$,

$$\mathbb{E}\big[\|G(\theta; Z)\|_2^2 \,\big|\, \theta\big] \;\le\; G_{\mathrm{tot}}^2,$$

with the explicit choice

$$G_{\mathrm{tot}}^2 \;:=\; 4\left(G_{\nabla\mathrm{SAIL}}^2 \;+\; \lambda^2\, G_{\partial R}^2 \;+\; \frac{\sigma_{\mathrm{SAIL}}^2 + \lambda^2 \sigma_R^2}{B}\right),$$

where

$$G_{\nabla\mathrm{SAIL}} \;:=\; 2\beta B_\psi \;+\; 4B_\psi\big(\log 2 + 2\beta D B_\psi\big), \qquad G_{\partial R} \;:=\; 2B_\psi \;+\; 8DB_\psi^2,$$

and $\sigma_{\mathrm{SAIL}}^2, \sigma_R^2$ are as in Lemma 3.10.

Proof. We use the stochastic oracles constructed in the proof of Lemma 3.10, writing $G_{\mathrm{SAIL}} = G^{\mathrm{SAIL}}$ and $G_R = G^R$, with Z denoting a generic batch of triples and preference labels. Define

$$\bar{g}_R(\theta) := \mathbb{E}[G_R(\theta; Z) \mid \theta].$$

We bound the mean components $\|\nabla L_{\mathrm{SAIL}}(\theta)\|_2$ and $\|\bar{g}_R(\theta)\|_2$, then combine these bounds with the centered-variance bounds in Lemma 3.10 and the definition (20).

Step 1: Uniform bounds on the pairwise logit and score. For $z = (x, y_1, y_2) \in \mathcal{Z}$, define $\Delta\psi(z) := \psi(x, y_1) - \psi(x, y_2)$. By Assumption 4, $\|\Delta\psi(z)\|_2 \le 2B_\psi$. Under Assumption 2 (log-linear policy class), the pairwise logit admits the cancellation

$$h_\theta(z) \;=\; h_\theta(x, y_1, y_2) \;=\; (\theta - \theta_{\mathrm{ref}})^\top \Delta\psi(z)$$

Hence, by Assumption 5,

$$|h_\theta(z)| \;\le\; \|\theta - \theta_{\mathrm{ref}}\|_2\, \|\Delta\psi(z)\|_2 \;\le\; D \cdot 2B_\psi \;=\; 2DB_\psi.$$

Moreover, recalling $d_\theta(z) = \mu(x)\pi_\theta(y_1|x)\pi_\theta(y_2|x)$ (Eq.(1)) and $S_\theta(z) := \nabla_\theta \log d_\theta(z)$, Lemma C.1 gives $\|S_\theta(z)\|_2 \le 4B_\psi$.

Step 2: Uniform bound on $\|\nabla L_{\mathrm{SAIL}}(\theta)\|_2$. Recall that $L_{\mathrm{SAIL}}(\theta) = \mathbb{E}_{z\sim d_\theta}\big[p^\star \ell^1(h) + (1 - p^\star)\ell^0(h)\big]$, where $\ell^1(h) = -\log\sigma(\beta h)$ and $\ell^0(h) = -\log\sigma(-\beta h)$, and $p^\star(z) \in [0, 1]$. Define the pointwise SAIL loss

$$\ell_\theta(z) \;:=\; p^\star(z)\, \ell^1\big(h_\theta(z)\big) \;+\; (1 - p^\star(z))\, \ell^0\big(h_\theta(z)\big).$$

Since $\ell^1(h) = \log(1+e^{-\beta h})$ and $\ell^0(h) = \log(1+e^{\beta h})$, we have for all $h \in \mathbb{R}$, $\max\{\ell^1(h), \ell^0(h)\} = \log(1 + e^{|\beta h|}) \le \log 2 + |\beta h|$. Using $|h_\theta(z)| \le 2DB_\psi$,

$$0 \;\le\; \ell_\theta(z) \;\le\; \log 2 + 2\beta D B_\psi.$$

Next, by direct differentiation, the derivatives of ℓ^1, ℓ^0 with respect to h satisfy $\big|\frac{d}{dh}\ell^1(h)\big| \le \beta$ and $\big|\frac{d}{dh}\ell^0(h)\big| \le \beta$ for all h. Together with $\nabla_\theta h_\theta(z) = \Delta\psi(z)$ and $\|\Delta\psi(z)\|_2 \le 2B_\psi$,

$$\|\nabla_\theta \ell_\theta(z)\|_2 \;\le\; \beta\, \|\Delta\psi(z)\|_2 \;\le\; 2\beta B_\psi.$$

Because $\mathcal{Y}$ is finite, we may apply the same score-function calculus as in Eq. (36) to write

$$\nabla L_{\mathrm{SAIL}}(\theta) \;=\; \mathbb{E}_{z\sim d_\theta}[\nabla_\theta \ell_\theta(z) + \ell_\theta(z)\, S_\theta(z)]\,.$$

Therefore,

$$\begin{aligned}\|\nabla L_{\mathrm{SAIL}}(\theta)\|_2 &\le \mathbb{E}_{z\sim d_\theta}[\|\nabla_\theta \ell_\theta(z)\|_2 + \ell_\theta(z)\, \|S_\theta(z)\|_2] \\ &\le 2\beta B_\psi + \big(\log 2 + 2\beta D B_\psi\big)\, 4B_\psi \;=:\; G_{\nabla\mathrm{SAIL}}.\end{aligned}$$

Step 3: Bound on $\|\bar{g}_R(\theta)\|_2$. We now bound $\|\bar{g}_R(\theta)\|_2$. By the proof of Lemma 3.10,

$$\bar{g}_R(\theta) = \mathbb{E}_{z\sim d_\theta}\big[s_0(h_\theta(z))\Delta\psi(z) + |h_\theta(z)|\, S_\theta(z)\big],$$

with s_0 as defined there. Since $|s_0| \le 1$, the uniform bounds above give

$$\|\bar{g}_R(\theta)\|_2 \;\le\; \mathbb{E}_{z\sim d_\theta}\big[\|\Delta\psi(z)\|_2 + |h_\theta(z)|\,\|S_\theta(z)\|_2\big] \;\le\; 2B_\psi + (2DB_\psi)(4B_\psi) \;=\; G_{\partial R}.$$

Fix $\theta \in \Theta$ and abbreviate $G_{\mathrm{SAIL}} := G_{\mathrm{SAIL}}(\theta; Z)$ and $G_R := G_R(\theta; Z)$. By (20) and $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$,

$$\mathbb{E}\big[\|G(\theta;Z)\|_2^2 \mid \theta\big] \;\le\; 2\,\mathbb{E}\big[\|G_{\mathrm{SAIL}}\|_2^2 \mid \theta\big] \;+\; 2\lambda^2\,\mathbb{E}\big[\|G_R\|_2^2 \mid \theta\big].$$

For the SAIL term, write $G_{\mathrm{SAIL}} = \nabla L_{\mathrm{SAIL}}(\theta) + (G_{\mathrm{SAIL}} - \nabla L_{\mathrm{SAIL}}(\theta))$ and apply $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ together with Lemma 3.10:

$$\mathbb{E}\big[\|G_{\mathrm{SAIL}}\|_2^2 \mid \theta\big] \;\le\; 2\|\nabla L_{\mathrm{SAIL}}(\theta)\|_2^2 + 2\,\mathbb{E}\big[\|G_{\mathrm{SAIL}} - \nabla L_{\mathrm{SAIL}}(\theta)\|_2^2 \mid \theta\big] \;\le\; 2G_{\nabla\mathrm{SAIL}}^2 + 2\frac{\sigma_{\mathrm{SAIL}}^2}{B}.$$

For the robust-penalty term, write $G_R = \bar{g}_R(\theta) + (G_R - \bar{g}_R(\theta))$ and apply $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ together with the bound on $\|\bar{g}_R(\theta)\|_2$ and the centered-variance bound in Lemma 3.10:

$$\mathbb{E}\big[\|G_R\|_2^2 \mid \theta\big] \;\le\; 2\|\bar{g}_R(\theta)\|_2^2 \;+\; 2\,\mathbb{E}\big[\|G_R - \bar{g}_R(\theta)\|_2^2 \mid \theta\big] \;\le\; 2G_{\partial R}^2 \;+\; 2\frac{\sigma_R^2}{B}.$$

Combining the last three displays yields, for all $\theta \in \Theta$,

$$\begin{aligned}\mathbb{E}\big[\|G(\theta;Z)\|_2^2 \mid \theta\big] &\le 2\left(2G_{\nabla\mathrm{SAIL}}^2 + 2\frac{\sigma_{\mathrm{SAIL}}^2}{B}\right) \;+\; 2\lambda^2\left(2G_{\partial R}^2 + 2\frac{\sigma_R^2}{B}\right)\\ &= 4\left(G_{\nabla\mathrm{SAIL}}^2 + \lambda^2 G_{\partial R}^2 + \frac{\sigma_{\mathrm{SAIL}}^2 + \lambda^2\sigma_R^2}{B}\right).\end{aligned}$$

This is exactly the claimed bound with G_{tot}^2 as stated. □

D Moreau envelope and convergence analysis

D.1 Proof of Lemma 3.9 (Properties of the Moreau envelope)

Proof of Lemma 3.9. Fix $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$ and recall the definition of the Moreau envelope and proximal point (Definition 3.8)

$$F_{\lambda_{\mathrm{env}}}(\theta) := \min_{u\in\mathbb{R}^d}\left\{F(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u-\theta\|_2^2\right\},$$

$$\hat{\theta} = \mathrm{prox}_{\lambda_{\mathrm{env}}F}(\theta) := \arg\min_{u\in\mathbb{R}^d}\left\{F(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u-\theta\|_2^2\right\}.$$

1–2. Finiteness, uniqueness, and the gradient formula. Since F is κ-weakly convex, the function $u \mapsto F(u) + \frac{\kappa}{2}\|u\|_2^2$ is convex. Therefore, for any fixed θ, the function

$$u \mapsto F(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u-\theta\|_2^2$$

is $(\frac{1}{\lambda_{\mathrm{env}}} - \kappa)$-strongly convex in u (because $\lambda_{\mathrm{env}} < 1/\kappa$), hence it has a unique minimizer $\hat{\theta}$ and the minimum value $F_{\lambda_{\mathrm{env}}}(\theta)$ is finite.

Moreover, by first-order optimality of $\hat{\theta}$ for the strongly convex problem,

$$0 \in \partial F(\hat{\theta}) + \frac{1}{\lambda_{\mathrm{env}}}(\hat{\theta}-\theta), \qquad \text{i.e.,} \qquad \frac{1}{\lambda_{\mathrm{env}}}(\theta-\hat{\theta}) \in \partial F(\hat{\theta}). \tag{42}$$

Standard properties of the Moreau envelope for weakly convex functions imply that $F_{\lambda_{\mathrm{env}}}$ is continuously differentiable and that

$$\nabla F_{\lambda_{\mathrm{env}}}(\theta) = \frac{1}{\lambda_{\mathrm{env}}}(\theta-\hat{\theta}).$$

Eq. (16) follows immediately: $\|\theta-\hat{\theta}\|_2 = \lambda_{\mathrm{env}}\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|_2$.

3. Lipschitz continuity of $\nabla F_{\lambda_{\text{env}}}$. A standard result for κ-weakly convex F is that $F_{\lambda_{\text{env}}}$ has Lipschitz gradient with constant

$$L_{\text{env}} = \frac{1}{\lambda_{\text{env}}(1 - \kappa\lambda_{\text{env}})},$$

which is Eq. (17).

4. Approximate stationarity of the proximal point. By (42), the vector $\frac{1}{\lambda_{\text{env}}}(\theta - \hat{\theta})$ belongs to $\partial F(\hat{\theta})$. Therefore

$$\operatorname{dist}\big(0, \partial F(\hat{\theta})\big) \leq \left\| \frac{1}{\lambda_{\text{env}}}(\theta - \hat{\theta}) \right\|_2 = \|\nabla F_{\lambda_{\text{env}}}(\theta)\|_2$$

This completes the proof. □

D.2 Auxiliary lemmas for Theorem 4.1

Lemma D.1 (Monotonicity inequality)**.** *Assume F is κ-weakly convex and let $\lambda_{\text{env}} \in (0, 1/\kappa)$. Fix any $\theta \in \mathbb{R}^d$ and let*

$$\hat{\theta} = \operatorname{prox}_{\lambda_{\text{env}} F}(\theta), \qquad \xi := \nabla F_{\lambda_{\text{env}}}(\theta) = \frac{1}{\lambda_{\text{env}}}(\theta - \hat{\theta}).$$

Then for every $v \in \partial F(\theta)$,

$$\langle \xi, v \rangle \geq (1 - \kappa\lambda_{\text{env}})\|\xi\|_2^2.$$

Proof. By κ-weak convexity, for all x, y and all $g \in \partial F(x)$,

$$F(y) \geq F(x) + \langle g, y - x \rangle - \frac{\kappa}{2}\|y - x\|_2^2. \tag{43}$$

Apply (43) twice:

(i) with $x = \theta$, $y = \hat{\theta}$, and $g = v \in \partial F(\theta)$:

$$F(\hat{\theta}) \geq F(\theta) + \langle v, \hat{\theta} - \theta \rangle - \frac{\kappa}{2}\|\hat{\theta} - \theta\|_2^2. \tag{44}$$

(ii) with $x = \hat{\theta}$, $y = \theta$, and any $\hat{v} \in \partial F(\hat{\theta})$:

$$F(\theta) \geq F(\hat{\theta}) + \langle \hat{v}, \theta - \hat{\theta} \rangle - \frac{\kappa}{2}\|\theta - \hat{\theta}\|_2^2. \tag{45}$$

By optimality of the proximal point (as in (42)), we may choose $\hat{v} = \frac{1}{\lambda_{\text{env}}}(\theta - \hat{\theta}) = \xi \in \partial F(\hat{\theta})$. Substituting this choice into (45) gives

$$F(\theta) \geq F(\hat{\theta}) + \langle \xi, \theta - \hat{\theta} \rangle - \frac{\kappa}{2}\|\theta - \hat{\theta}\|_2^2.$$

Add this inequality to (44) to eliminate $F(\theta)$ and $F(\hat{\theta})$, yielding

$$0 \geq \langle v, \hat{\theta} - \theta \rangle + \langle \xi, \theta - \hat{\theta} \rangle - \kappa\|\theta - \hat{\theta}\|_2^2.$$

Rearrange and use $\theta - \hat{\theta} = \lambda_{\text{env}}\xi$:

$$\langle \xi, \theta - \hat{\theta} \rangle \leq \langle v, \theta - \hat{\theta} \rangle + \kappa\|\theta - \hat{\theta}\|_2^2 \quad \Longrightarrow \quad \lambda_{\text{env}}\|\xi\|_2^2 \leq \lambda_{\text{env}}\langle v, \xi \rangle + \kappa\lambda_{\text{env}}^2\|\xi\|_2^2.$$

Divide by $\lambda_{\text{env}} > 0$ and rearrange:

$$\langle v, \xi \rangle \geq (1 - \kappa\lambda_{\text{env}})\|\xi\|_2^2,$$

as claimed. □

Lemma D.2 (One-step inequality)**.** *Let Assumptions 1–6 hold and let κ be the weak convexity constant from Theorem 3.7.*

Fix $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$ and define $L_{\mathrm{env}} := \frac{1}{\lambda_{\mathrm{env}}(1-\kappa\lambda_{\mathrm{env}})}$ as in Lemma 3.9.

Consider the projected update (Algorithm 1)

$$\theta_{t+1} := \mathrm{proj}_{\Theta}\big(\theta_t - \eta\, G(\theta_t; Z_t)\big),$$

where $G(\theta_t; Z_t) = G^{\mathrm{SAIL}}(\theta_t; Z_t) + \lambda G^R(\theta_t; Z_t)$ is formed from Lemma 3.10.

Then for any stepsize $\eta > 0$,

$$\mathbb{E}\big[F_{\lambda_{\mathrm{env}}}(\theta_{t+1}) \mid \theta_t\big] \le F_{\lambda_{\mathrm{env}}}(\theta_t) - \eta(1-\kappa\lambda_{\mathrm{env}})\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_t)\|_2^2 + \frac{L_{\mathrm{env}}\eta^2}{2}\,\mathbb{E}\big[\|G(\theta_t; Z_t)\|_2^2 \mid \theta_t\big].$$

Proof. Write $F = L_\rho^W + I_\Theta$, where I_Θ is the indicator of Θ (Eq.(13)). Then the envelope can be written as

$$F_{\lambda_{\mathrm{env}}}(z) = \min_{u\in\Theta}\left\{L_\rho^W(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u - z\|_2^2\right\}.$$

Let $\tilde{z} := \mathrm{proj}_\Theta(z)$. For any $u \in \Theta$, Euclidean projection optimality implies $\langle z - \tilde{z}, u - \tilde{z}\rangle \le 0$, hence

$$\|u - \tilde{z}\|_2^2 = \|u - z + (z - \tilde{z})\|_2^2 = \|u - z\|_2^2 + \|z - \tilde{z}\|_2^2 + 2\langle u - z, z - \tilde{z}\rangle \le \|u - z\|_2^2.$$

Taking the minimum over $u \in \Theta$ gives the monotonicity under projection:

$$F_{\lambda_{\mathrm{env}}}(\mathrm{proj}_\Theta(z)) \le F_{\lambda_{\mathrm{env}}}(z) \quad \text{for all } z \in \mathbb{R}^d. \tag{46}$$

Now define the unprojected point $z_t := \theta_t - \eta G(\theta_t; Z_t)$, so $\theta_{t+1} = \mathrm{proj}_\Theta(z_t)$. By (46),

$$F_{\lambda_{\mathrm{env}}}(\theta_{t+1}) \le F_{\lambda_{\mathrm{env}}}(z_t).$$

By Lemma 3.9, $F_{\lambda_{\mathrm{env}}}$ is L_{env}-smooth, hence

$$F_{\lambda_{\mathrm{env}}}(\theta_t - \eta G(\theta_t; Z_t)) \le F_{\lambda_{\mathrm{env}}}(\theta_t) - \eta\langle\nabla F_{\lambda_{\mathrm{env}}}(\theta_t), G(\theta_t; Z_t)\rangle + \frac{L_{\mathrm{env}}\eta^2}{2}\|G(\theta_t; Z_t)\|_2^2.$$

Combining with the previous display and taking conditional expectation yields

$$\mathbb{E}\big[F_{\lambda_{\mathrm{env}}}(\theta_{t+1}) \mid \theta_t\big] \le F_{\lambda_{\mathrm{env}}}(\theta_t) - \eta\,\langle\nabla F_{\lambda_{\mathrm{env}}}(\theta_t),\ \mathbb{E}[G(\theta_t; Z_t) \mid \theta_t]\rangle + \frac{L_{\mathrm{env}}\eta^2}{2}\,\mathbb{E}\big[\|G(\theta_t; Z_t)\|_2^2 \mid \theta_t\big].$$

It remains to bound the inner product. Let $m_t := \mathbb{E}[G(\theta_t; Z_t) \mid \theta_t]$; we show that

$$\big\langle\nabla F_{\lambda_{\mathrm{env}}}(\theta_t), m_t\big\rangle \ \ge\ (1 - \kappa\lambda_{\mathrm{env}})\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_t)\|_2^2. \tag{47}$$

We use the lower model of Lemma 3.10. By that lemma, $m_t = \nabla L^{\mathrm{SAIL}}(\theta_t) + \lambda\,\mathbb{E}[G^R(\theta_t; Z_t) \mid \theta_t]$, and the second mean satisfies (19) at $\theta = \theta_t$. By Assumption 3, for every $u \in \Theta$,

$$L^{\mathrm{SAIL}}(u) \ \ge\ L^{\mathrm{SAIL}}(\theta_t) + \langle\nabla L^{\mathrm{SAIL}}(\theta_t), u - \theta_t\rangle - \frac{L_{\mathrm{SAIL}}}{2}\|u - \theta_t\|_2^2.$$

Adding λ times (19) and using $L_\rho^W = L^{\mathrm{SAIL}} + \lambda R$ (Theorem 3.2), we obtain, for every $u \in \Theta$,

$$L_\rho^W(u) \ \ge\ L_\rho^W(\theta_t) + \langle m_t, u - \theta_t\rangle - \frac{\kappa}{2}\|u - \theta_t\|_2^2. \tag{48}$$

Let $\hat{\theta}_t := \mathrm{prox}_{\lambda_{\mathrm{env}}F}(\theta_t)$ and $d_t := \theta_t - \hat{\theta}_t$. Since $F = L_\rho^W + I_\Theta$, the point $\hat{\theta}_t$ lies in Θ and minimizes

$$\phi_t(u) := L_\rho^W(u) + \frac{1}{2\lambda_{\mathrm{env}}}\|u - \theta_t\|_2^2$$

over Θ; moreover, $d_t = \lambda_{\mathrm{env}}\nabla F_{\lambda_{\mathrm{env}}}(\theta_t)$ by Lemma 3.9. Because L_ρ^W is κ-weakly convex on Θ, the function ϕ_t is $(1/\lambda_{\mathrm{env}} - \kappa)$-strongly convex on Θ. For $u \in \Theta$ and $s \in (0, 1]$, the point $\hat{\theta}_t + s(u - \hat{\theta}_t)$ lies in Θ, so

$$\begin{aligned}\phi_t(\hat{\theta}_t) &\le \phi_t\big(\hat{\theta}_t + s(u - \hat{\theta}_t)\big)\\ &\le (1 - s)\,\phi_t(\hat{\theta}_t) + s\,\phi_t(u) - \frac{1/\lambda_{\mathrm{env}} - \kappa}{2}\,s(1 - s)\|u - \hat{\theta}_t\|_2^2.\end{aligned}$$

Dividing by s and letting $s \downarrow 0$ gives the quadratic growth bound $\phi_t(u) \geq \phi_t(\hat{\theta}_t) + \frac{1}{2}(1/\lambda_{\text{env}} - \kappa)\|u - \hat{\theta}_t\|_2^2$. It uses only the segment from $\hat{\theta}_t$ to u, not a first-order condition at $\hat{\theta}_t$, and therefore remains valid when $\hat{\theta}_t$ lies on the boundary of Θ. Taking $u = \theta_t$ yields

$$L_\rho^W(\theta_t) - L_\rho^W(\hat{\theta}_t) \;\geq\; \Big(\frac{1}{\lambda_{\text{env}}} - \frac{\kappa}{2}\Big)\|d_t\|_2^2.$$

On the other hand, (48) with $u = \hat{\theta}_t$ gives $L_\rho^W(\theta_t) - L_\rho^W(\hat{\theta}_t) \leq \langle m_t, d_t\rangle + \frac{\kappa}{2}\|d_t\|_2^2$. Combining the two bounds,

$$\langle m_t, d_t\rangle \;\geq\; \Big(\frac{1}{\lambda_{\text{env}}} - \kappa\Big)\|d_t\|_2^2.$$

Substituting $d_t = \lambda_{\text{env}} \nabla F_{\lambda_{\text{env}}}(\theta_t)$ and dividing by λ_{env} gives (47); inserting (47) into the previous display completes the proof. □

D.3 Proof of Theorem 4.1 (convergence rate of the Moreau envelope)

Proof of Theorem 4.1. Fix $\lambda_{\text{env}} \in (0, 1/\kappa)$ and stepsize $\eta > 0$. By Lemma D.2, for each t,

$$\mathbb{E}\big[F_{\lambda_{\text{env}}}(\theta_{t+1}) \mid \theta_t\big] \leq F_{\lambda_{\text{env}}}(\theta_t) - \eta(1 - \kappa\lambda_{\text{env}})\|\nabla F_{\lambda_{\text{env}}}(\theta_t)\|_2^2 + \frac{L_{\text{env}}\eta^2}{2}\,\mathbb{E}\big[\|G(\theta_t; Z_t)\|_2^2 \mid \theta_t\big].$$

Take total expectation and sum from $t = 0$ to $T - 1$ to obtain the telescoping inequality

$$\eta(1 - \kappa\lambda_{\text{env}}) \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F_{\lambda_{\text{env}}}(\theta_t)\|_2^2\big] \leq F_{\lambda_{\text{env}}}(\theta_0) - \mathbb{E}\big[F_{\lambda_{\text{env}}}(\theta_T)\big] + \frac{L_{\text{env}}\eta^2}{2} \sum_{t=0}^{T-1} \mathbb{E}\big[\|G(\theta_t; Z_t)\|_2^2\big]. \tag{49}$$

Since $F_{\lambda_{\text{env}}}(\theta) \geq \inf_u F(u) = F_{\text{inf}}$ for all θ, we have $\mathbb{E}[F_{\lambda_{\text{env}}}(\theta_T)] \geq F_{\text{inf}}$. By Lemma C.4, $\mathbb{E}\|G(\theta_t; Z_t)\|_2^2 \leq G_{\text{tot}}^2$ for all t. Thus (49) implies

$$\eta(1 - \kappa\lambda_{\text{env}}) \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F_{\lambda_{\text{env}}}(\theta_t)\|_2^2\big] \leq F_{\lambda_{\text{env}}}(\theta_0) - F_{\text{inf}} + \frac{L_{\text{env}}\eta^2}{2}\, T\, G_{\text{tot}}^2.$$

Divide by $\eta(1 - \kappa\lambda_{\text{env}})T$:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F_{\lambda_{\text{env}}}(\theta_t)\|_2^2\big] \leq \frac{F_{\lambda_{\text{env}}}(\theta_0) - F_{\text{inf}}}{\eta(1 - \kappa\lambda_{\text{env}})T} + \frac{L_{\text{env}}\eta}{2(1 - \kappa\lambda_{\text{env}})}\, G_{\text{tot}}^2.$$

Let R be uniformly distributed on $\{0, \ldots, T - 1\}$, independent of the algorithmic randomness. Then

$$\mathbb{E}\big[\|\nabla F_{\lambda_{\text{env}}}(\theta_R)\|_2^2\big] = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F_{\lambda_{\text{env}}}(\theta_t)\|_2^2\big],$$

and we obtain exactly Eq.(22) of Theorem 4.1. □

Proof of Corollary 4.2. Plug the stated stepsize choice

$$\eta := \frac{1}{G_{\text{tot}}}\sqrt{\frac{2\big(F_{\lambda_{\text{env}}}(\theta_0) - F_{\text{inf}}\big)}{L_{\text{env}}T}}$$

into the right-hand side of Eq.(22) and simplify to obtain

$$\mathbb{E}\big[\|\nabla F_{\lambda_{\text{env}}}(\theta_R)\|_2^2\big] \leq \frac{G_{\text{tot}}\sqrt{2L_{\text{env}}\big(F_{\lambda_{\text{env}}}(\theta_0) - F_{\text{inf}}\big)}}{(1 - \kappa\lambda_{\text{env}})\sqrt{T}}.$$

By Lemma C.4, we have

$$G_{\text{tot}}^2 \;:=\; 4\left(G_{\nabla\text{SAIL}}^2 \;+\; \lambda^2\, G_{\partial R}^2 \;+\; \frac{\sigma_{\text{SAIL}}^2 + \lambda^2\sigma_R^2}{B}\right).$$

To ensure that $\mathbb{E}\big[\|\nabla F_{\lambda_{\rm env}}(\theta_R)\|_2^2\big] \le \varepsilon$, we substitute $G_{\rm tot}^2$ into Theorem 4.1 and obtain that it suffices to take

$$\begin{aligned} T &\ge \frac{2\,G_{\rm tot}^2\big(F_{\lambda_{\rm env}}(\theta_0) - F_{\rm inf}\big)}{\lambda_{\rm env}(1-\kappa\lambda_{\rm env})^3\,\varepsilon^2} \\ &= \frac{8\left(C + \frac{\sigma_{\rm SAIL}^2 + \lambda^2\sigma_R^2}{B}\right)\big(F_{\lambda_{\rm env}}(\theta_0) - F_{\rm inf}\big)}{\lambda_{\rm env}(1-\kappa\lambda_{\rm env})^3\,\varepsilon^2}, \end{aligned}$$

where $C := G_{\nabla \rm SAIL}^2 + \lambda^2 G_{\partial R}^2$.

□

E On the practical role of the proximal point

This section provides a practical interpretation of the proximal-point and Moreau-envelope constructions used in our analysis. Throughout, recall the constrained objective

$$F(\theta) := L_\rho^W(\theta) + I_\Theta(\theta),$$

and the envelope parameter $\lambda_{\rm env} \in (0, 1/\kappa)$ from Definition 3.8, where κ is the weak convexity constant from Theorem 3.7.

The Moreau envelope

$$F_{\lambda_{\rm env}}(\theta) \;=\; \min_{u\in\mathbb{R}^d}\Big\{F(u) + \tfrac{1}{2\lambda_{\rm env}}\|u-\theta\|_2^2\Big\}$$

can be viewed as a *smoothed* surrogate of the potentially nonsmooth, nonconvex objective F. The quadratic term discourages large moves away from the current iterate θ, and the minimizer of the inner problem—the *proximal point*—acts as a locally stabilized refinement of θ. Concretely, for any iterate θ we write

$$\hat\theta := \operatorname{prox}_{\lambda_{\rm env}F}(\theta) \in \arg\min_{u\in\mathbb{R}^d}\Big\{F(u) + \tfrac{1}{2\lambda_{\rm env}}\|u-\theta\|_2^2\Big\}.$$

Because F is κ-weakly convex and $\lambda_{\rm env} < 1/\kappa$, the proximal subproblem is strongly convex in u and hence has a unique minimizer. This is the fundamental reason the envelope is differentiable and why $\|\nabla F_{\lambda_{\rm env}}(\theta)\|_2$ is a meaningful stationarity proxy (Lemma 3.9).

Algorithm 1 does *not* require solving the proximal subproblem at each iteration. The algorithm updates θ_t using stochastic subgradient information for $L_\rho^W = L_{\rm SAIL} + \lambda R$ followed by projection onto Θ, and then outputs a random iterate θ_R. The proximal point

$$\hat\theta_R := \operatorname{prox}_{\lambda_{\rm env}F}(\theta_R)$$

is introduced only as a *theoretical device* to translate envelope stationarity at θ_R into near-stationarity of the original constrained objective at $\hat\theta_R$ (see Lemma 3.9 and Theorem 4.1). In particular, Lemma 3.9 implies that θ_R and $\hat\theta_R$ are close whenever the envelope gradient is small:

$$\|\theta_R - \hat\theta_R\|_2 \;=\; \lambda_{\rm env}\,\|\nabla F_{\lambda_{\rm env}}(\theta_R)\|_2,$$

and also that $\hat\theta_R$ is nearly stationary for F in the sense that

$$\operatorname{dist}\big(0, \partial F(\hat\theta_R)\big) \;\le\; \|\nabla F_{\lambda_{\rm env}}(\theta_R)\|_2.$$

Thus, driving $\|\nabla F_{\lambda_{\rm env}}(\theta_R)\|_2$ small (as guaranteed in expectation by Theorem 4.1, with sample complexity summarized in Corollary 4.2) simultaneously guarantees proximity to, and near-stationarity of, a point for the original constrained robust objective.

Lemma E.1 (What if we do not compute the proximal point?)**.** *Assume $F : \mathbb{R}^d \to (-\infty, +\infty]$ is κ-weakly convex and bounded below, and fix $\lambda_{\rm env} \in (0, 1/\kappa)$ as in Definition 3.8. For any $\theta \in \mathbb{R}^d$, let*

$$\hat\theta := \operatorname{prox}_{\lambda_{\rm env}F}(\theta), \qquad \xi := \nabla F_{\lambda_{\rm env}}(\theta) = \frac{1}{\lambda_{\rm env}}(\theta - \hat\theta)$$

(Definition 3.8 and Lemma 3.9). Then:

$$\|\theta - \hat{\theta}\| = \lambda_{\mathrm{env}}\|\xi\|, \tag{50}$$

$$\operatorname{dist}\big(0, \partial F(\hat{\theta})\big) \le \|\xi\|, \tag{51}$$

$$\operatorname{dist}\big(0, \partial F(\theta)\big) \ge (1 - \kappa\lambda_{\mathrm{env}})\|\xi\|. \tag{52}$$

On the other hand, under κ-weak convexity alone, there is no *universal constant C such that*

$$\operatorname{dist}\big(0, \partial F(\theta)\big) \;\le\; C\,\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\| \quad \textit{for all } \theta \textit{ and all } \kappa\textit{-weakly convex } F$$

(even in the convex case $\kappa = 0$).

Proof. We first prove (50)–(51). By Lemma 3.9 (see also the first-order optimality condition for the proximal subproblem),

$$0 \in \partial F(\hat{\theta}) + \frac{1}{\lambda_{\mathrm{env}}}(\hat{\theta} - \theta), \qquad \text{equivalently} \qquad \frac{1}{\lambda_{\mathrm{env}}}(\theta - \hat{\theta}) \in \partial F(\hat{\theta}).$$

With $\xi := \frac{1}{\lambda_{\mathrm{env}}}(\theta - \hat{\theta}) = \nabla F_{\lambda_{\mathrm{env}}}(\theta)$, taking norms yields $\|\theta - \hat{\theta}\| = \lambda_{\mathrm{env}}\|\xi\|$, which is (50). Moreover, since $\xi \in \partial F(\hat{\theta})$, we have $\operatorname{dist}\big(0, \partial F(\hat{\theta})\big) \le \|\xi\|$, which is (51).

Next we prove (52). By Lemma D.1 (Monotonicity inequality), for every $v \in \partial F(\theta)$,

$$\langle \xi, v\rangle \ge (1 - \kappa\lambda_{\mathrm{env}})\|\xi\|^2.$$

By Cauchy–Schwarz, $\langle \xi, v\rangle \le \|\xi\|\,\|v\|$, hence for every $v \in \partial F(\theta)$,

$$\|v\| \ge (1 - \kappa\lambda_{\mathrm{env}})\|\xi\|.$$

Taking the infimum over $v \in \partial F(\theta)$ gives (52).

For the stated consequences with the algorithmic output, apply (50)–(51) with $\theta = \theta_R$ and $\hat{\theta} = \hat{\theta}_R$. Then

$$\|\theta_R - \hat{\theta}_R\|^2 = \lambda_{\mathrm{env}}^2\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|^2, \qquad \operatorname{dist}^2\big(0, \partial F(\hat{\theta}_R)\big) \le \|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|^2.$$

Taking expectations, using Jensen's inequality for $\mathbb{E}\|\theta_R - \hat{\theta}_R\|$, and plugging $\mathbb{E}\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|^2 \le \varepsilon$ yields the two bounds.

Finally, we show that no universal constant C can upper bound $\operatorname{dist}(0, \partial F(\theta))$ by $\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|$ under weak convexity alone. Consider $d = 1$ and $\Theta = \mathbb{R}$, and define $F(\theta) = |\theta|$, which is convex (thus $\kappa = 0$). Fix $\lambda_{\mathrm{env}} = 1$ and any $\theta \in (0, 1)$. The proximal point is

$$\hat{\theta} = \arg\min_{u \in \mathbb{R}} \Big\{ |u| + \tfrac{1}{2}(u - \theta)^2 \Big\} = 0,$$

so $\nabla F_{\lambda_{\mathrm{env}}}(\theta) = \theta - \hat{\theta} = \theta$. However, for $\theta > 0$, $\partial F(\theta) = \{1\}$ and therefore $\operatorname{dist}(0, \partial F(\theta)) = 1$. Hence

$$\frac{\operatorname{dist}(0, \partial F(\theta))}{\|\nabla F_{\lambda_{\mathrm{env}}}(\theta)\|} = \frac{1}{\theta} \to \infty \quad \text{as } \theta \downarrow 0,$$

which rules out any finite universal C. □

Remark E.2 (Interpretation for practice). Even if Algorithm 1 outputs only θ_R (without solving the proximal subproblem), Theorem 4.1 still certifies that θ_R lies within $O(\lambda_{\mathrm{env}}\sqrt{\varepsilon})$ of a point $\hat{\theta}_R$ that is $O(\sqrt{\varepsilon})$-stationary for the original constrained objective F. In contrast, without additional regularity beyond κ-weak convexity, one cannot generally convert envelope stationarity into a bound on $\operatorname{dist}(0, \partial F(\theta_R))$ itself; this is why the analysis (and the stationarity notion) is phrased in terms of the proximal point.

E.1 Practical computation of the proximal point

In the analysis we introduce the proximal point

$$\hat{\theta}_R \;:=\; \operatorname{prox}_{\lambda_{\mathrm{env}} F}(\theta_R) \;=\; \arg\min_{\theta \in \Theta} \Big\{ L_\rho^W(\theta) + \frac{1}{2\lambda_{\mathrm{env}}}\|\theta - \theta_R\|_2^2 \Big\}, \tag{53}$$

where $F(\theta) := L_\rho^W(\theta) + I_\Theta(\theta)$ and $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$. By weak convexity of F and $\lambda_{\mathrm{env}} < 1/\kappa$, the objective in (53) is strongly convex and hence admits a unique minimizer $\hat{\theta}_R$. Although Algorithm 1 outputs θ_R, one can compute an approximation $\bar{\theta}_R \approx \hat{\theta}_R$ as an optional post-processing step, warm-started at θ_R. A simple inner-loop solver is provided below.

A simple inner-loop solver. Let $\theta^{(0)} := \theta_R$. For $k = 0, 1, \ldots, K-1$, sample $Z^{(k)} \sim d_{\theta^{(k)}}$ and compute the stochastic direction

$$G^{\mathrm{prox}}\big(\theta^{(k)}; Z^{(k)}\big) \;:=\; G_{\mathrm{SAIL}}\big(\theta^{(k)}; Z^{(k)}\big) \;+\; \lambda\, G_R\big(\theta^{(k)}; Z^{(k)}\big) \;+\; \frac{1}{\lambda_{\mathrm{env}}}\big(\theta^{(k)} - \theta_R\big),$$

which is the usual oracle from Lemma 3.10 plus the deterministic gradient of the proximal quadratic. Then take a projected step

$$\theta^{(k+1)} \;:=\; \Pi_\Theta\Big(\theta^{(k)} - \alpha_k\, G^{\mathrm{prox}}\big(\theta^{(k)}; Z^{(k)}\big)\Big), \qquad \bar{\theta}_R := \theta^{(K)}. \tag{54}$$

Since the proximal quadratic improves conditioning, this inner loop is typically stable in practice. Moreover, the theory only requires $\bar{\theta}_R$ to be a sufficiently accurate approximation; exact solves of (53) are not necessary. This is consistent with standard practice in inexact proximal methods, where the proximal subproblem is solved only approximately while maintaining meaningful convergence guarantees (e.g., Schmidt et al. 2011).

Lemma E.3 (Stationarity of an inexact proximal point)**.** *Fix $\lambda_{\mathrm{env}} \in (0, 1/\kappa)$ and an index R. Let $\hat{\theta}_R = \mathrm{prox}_{\lambda_{\mathrm{env}}} F(\theta_R)$ and let $\bar{\theta}_R \in \Theta$ be any point. Define the proximal objective*

$$\Psi_R(\theta) \;:=\; F(\theta) + \frac{1}{2\lambda_{\mathrm{env}}}\|\theta - \theta_R\|_2^2.$$

If $\bar{\theta}_R$ satisfies the (first-order) proximal residual bound

$$\mathrm{dist}\big(0, \partial\Psi_R(\bar{\theta}_R)\big) \;\le\; \varepsilon_{\mathrm{prox}}, \tag{55}$$

then

$$\mathrm{dist}\big(0, \partial F(\bar{\theta}_R)\big) \;\le\; \|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2 \;+\; \varepsilon_{\mathrm{prox}} \;+\; \frac{1}{\lambda_{\mathrm{env}}}\|\bar{\theta}_R - \hat{\theta}_R\|_2. \tag{56}$$

In particular, if one ensures $\|\bar{\theta}_R - \hat{\theta}_R\|_2 \le \lambda_{\mathrm{env}}\varepsilon_{\mathrm{geom}}$, then $\mathrm{dist}\big(0, \partial F(\bar{\theta}_R)\big) \le \|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2 + \varepsilon_{\mathrm{prox}} + \varepsilon_{\mathrm{geom}}$.

Proof. By subdifferential calculus, $\partial\Psi_R(\theta) = \partial F(\theta) + \frac{1}{\lambda_{\mathrm{env}}}(\theta - \theta_R)$. The residual condition (55) implies there exists $g \in \partial F(\bar{\theta}_R)$ such that

$$\left\| g + \frac{1}{\lambda_{\mathrm{env}}}(\bar{\theta}_R - \theta_R) \right\|_2 \;\le\; \varepsilon_{\mathrm{prox}}.$$

Therefore,

$$\mathrm{dist}\big(0, \partial F(\bar{\theta}_R)\big) \;\le\; \|g\|_2 \;\le\; \varepsilon_{\mathrm{prox}} + \frac{1}{\lambda_{\mathrm{env}}}\|\bar{\theta}_R - \theta_R\|_2.$$

Next, by the triangle inequality, $\|\bar{\theta}_R - \theta_R\|_2 \le \|\bar{\theta}_R - \hat{\theta}_R\|_2 + \|\hat{\theta}_R - \theta_R\|_2$. By Lemma 3.9 (property $\nabla F_{\lambda_{\mathrm{env}}}(\theta) = \frac{1}{\lambda_{\mathrm{env}}}(\theta - \mathrm{prox}_{\lambda_{\mathrm{env}}} F(\theta))$), we have $\|\hat{\theta}_R - \theta_R\|_2 = \lambda_{\mathrm{env}}\|\nabla F_{\lambda_{\mathrm{env}}}(\theta_R)\|_2$. Substituting yields (56). □

Remark E.4 (Practical stopping criteria). Lemma E.3 suggests two inexpensive targets for the inner loop: (i) keep the proximal residual $\mathrm{dist}(0, \partial\Psi_R(\theta^{(k)}))$ small (as estimated by a minibatch), and (ii) exploit warm-starting so that $\|\bar{\theta}_R - \hat{\theta}_R\|_2$ remains modest.

F Finite-response experiment details

This appendix gives the finite-response experimental details supporting Section 6. The goal is to document the controlled oracle-stress construction behind the main-text claim: Robust SAIL (ours) sacrifices ideal-oracle fitting in order to reduce oracle sensitivity and improve worst-case oracle stress loss. The code for these experiments is available at `https://github.com/anonymousneurips2026/oracle-robust-online-alignment`.

F.1 Controlled finite-response setup and nuisance-bias construction

For each prompt x, we fix a candidate response set

$$\mathcal{Y}_x = \{y_1, \dots, y_K\}, \qquad K = 12.$$

The learned policy is a log-linear distribution over this fixed candidate set with feature dimension $d = 7$:

$$\pi_\theta(y_i \mid x) = \frac{\exp(\theta^\top \psi(x, y_i))}{\sum_{j=1}^{K} \exp(\theta^\top \psi(x, y_j))}.$$

The reference policy π_{ref} is fixed and induced by frozen reference logits. The policy-induced comparison distribution is

$$d_\theta(x, y_i, y_j) = \mu(x)\pi_\theta(y_i \mid x)\pi_\theta(y_j \mid x).$$

The ideal oracle is denoted by P^{id}. For $z = (x, y_i, y_j)$, write

$$p_{\mathrm{id}}(z) = P^{\mathrm{id}}(y_i \succ y_j \mid x).$$

The ideal oracle follows a Bradley–Terry model induced by a synthetic true reward $r^\star$:

$$p_{\mathrm{id}}(x, i, j) = \sigma\left(\frac{r^\star(x, y_i) - r^\star(x, y_j)}{\tau}\right).$$

The observed training oracle is a bounded perturbation of this ideal oracle:

$$p_{\mathrm{obs}}(z) = p_{\mathrm{id}}(z) + \rho_{\mathrm{mis}} b(z), \qquad \rho_{\mathrm{mis}} \in \{0.05, 0.075, 0.10\}.$$

Here ρ_{mis} is the environment misspecification radius in this controlled benchmark; it controls the distance between the observed training oracle and the ideal oracle. The algorithmic robust radius ρ_{alg} is a training hyperparameter. We use $q = \rho_{\mathrm{alg}}/\rho_{\mathrm{mis}}$ only because this controlled experiment knows ρ_{mis}.

The perturbation direction b is a structured nuisance bias. Let $s(x, y_i)$ be a redundancy-style scalar attribute given by a fixed projection of $\psi(x, y_i)$ that is orthogonal to the true-reward direction. For a pair (y_i, y_j), define

$$\Delta s_{ij}(x) = s(x, y_i) - s(x, y_j), \qquad \Delta r_{ij}(x) = r^\star(x, y_i) - r^\star(x, y_j),$$

and the per-prompt normalization

$$M_x = \max_{u,v} |s(x, y_u) - s(x, y_v)| + \varepsilon_{\mathrm{norm}}.$$

The nuisance-bias direction is

$$b_{ij}(x) = \mathrm{sign}(\Delta s_{ij}(x)) \frac{|\Delta s_{ij}(x)|}{M_x} \mathbf{1}\{\Delta s_{ij}(x) \Delta r_{ij}(x) \le 0\}.$$

By construction,

$$|b_{ij}(x)| \le 1, \qquad b_{ji}(x) = -b_{ij}(x).$$

Thus the observed oracle is biased toward the nuisance attribute only when that attribute conflicts with the true reward. This creates structured oracle misspecification rather than random label noise.

All experiments use 10 paired seeds for every ρ_{mis}. Evaluation uses exact finite-response expectations, so the reported losses have no Monte Carlo evaluation noise. The candidate pool is fixed before optimization. The frozen checkpoint `Qwen/Qwen2.5-0.5B-Instruct` (Yang et al., 2025) is used only to instantiate language-like finite candidate sets and reference logits; it is not fine-tuned or evaluated as the object of study.

F.2 Training objectives and hyperparameter grid

The pairwise log-ratio score is

$$h_\theta(x, y_i, y_j) = \log \frac{\pi_\theta(y_i \mid x)}{\pi_{\mathrm{ref}}(y_i \mid x)} - \log \frac{\pi_\theta(y_j \mid x)}{\pi_{\mathrm{ref}}(y_j \mid x)}.$$

The sensitivity penalty is

$$R(\theta) = \mathbb{E}_{z\sim d_\theta}\big[|h_\theta(z)|\big].$$

Nominal SAIL is trained under the observed oracle:

$$\theta_{\text{Nominal}} \in \arg\min_\theta L_{\text{SAIL}}(\theta; P^{\text{obs}}).$$

For $q > 0$, Robust SAIL trains

$$\theta_q \in \arg\min_\theta \left[L_{\text{SAIL}}(\theta; P^{\text{obs}}) + \rho_{\text{alg}}\beta R(\theta)\right], \qquad \rho_{\text{alg}} = q\rho_{\text{mis}},$$

with $\beta = 0.1$. The baseline $q = 0$ is Nominal SAIL, while $q > 0$ are Robust SAIL candidates.

For $\rho_{\text{mis}} \in \{0.05, 0.10\}$, the trained grid is

$$q \in \{0, 0.025, 0.05, 0.10, 0.25, 0.50, 1.00\}.$$

For the central radius $\rho_{\text{mis}} = 0.075$, we also train smaller radii:

$$q \in \{0, 0.005, 0.01, 0.025, 0.05, 0.10, 0.25, 0.50, 1.00\}.$$

The robust operating point $q^\star$ is selected globally for each ρ_{mis} using the validation split rather than the test split, with seed-pooled validation high-stress AUC rather than per-seed selection.

F.3 True-to-worst oracle stress metric

For each trained policy θ, define the two label-conditional losses

$$\ell^1_\theta(z) = -\log\sigma(\beta h_\theta(z)), \qquad \ell^0_\theta(z) = -\log\sigma(-\beta h_\theta(z)).$$

The stress evaluation uncertainty interval is centered at the observed oracle and uses the environment evaluation radius ρ_{mis}:

$$p_{\text{low}}(z) = \max\{0, p_{\text{obs}}(z) - \rho_{\text{mis}}\}, \qquad p_{\text{high}}(z) = \min\{1, p_{\text{obs}}(z) + \rho_{\text{mis}}\}.$$

This radius is not ρ_{alg}: ρ_{alg} controls training, while ρ_{mis} controls the stress evaluation.

The policy-specific worst-case oracle is

$$p^\theta_{\text{wc}}(z) = \begin{cases} p_{\text{high}}(z), & \ell^1_\theta(z) \geq \ell^0_\theta(z), \\ p_{\text{low}}(z), & \ell^1_\theta(z) < \ell^0_\theta(z). \end{cases}$$

The true-to-worst oracle path is

$$p^\theta_\alpha(z) = (1-\alpha)p_{\text{id}}(z) + \alpha p^\theta_{\text{wc}}(z), \qquad \alpha \in [0, 1].$$

Thus $\alpha = 0$ is the ideal oracle, $\alpha = 1$ is the policy-specific worst-case oracle, and larger α means stronger adversarial oracle stress. The stress loss is

$$L_{\text{stress}}(\theta; \alpha) = \mathbb{E}_{z\sim d_\theta}\left[p^\theta_\alpha(z)\ell^1_\theta(z) + \big(1 - p^\theta_\alpha(z)\big)\ell^0_\theta(z)\right].$$

We report

$$L_{\text{id}}(\theta) = L_{\text{stress}}(\theta; 0), \qquad L_{\text{wc}}(\theta) = L_{\text{stress}}(\theta; 1),$$

and

$$\text{AUC}_{0.5:1}(\theta) = \frac{1}{6} \sum_{\alpha\in\{0.5,0.6,0.7,0.8,0.9,1.0\}} L_{\text{stress}}(\theta; \alpha).$$

All loss-style metrics are lower-is-better.

For fixed θ, the stress loss is affine in α:

$$L_{\text{stress}}(\theta; \alpha) = (1-\alpha)L_{\text{id}}(\theta) + \alpha L_{\text{wc}}(\theta).$$

Since $p_{\text{obs}}(z) = p_{\text{id}}(z) + \rho_{\text{mis}}b(z)$ and the worst-case endpoint around p_{obs} uses radius ρ_{mis}, the unclipped endpoint decomposes as

$$L_{\text{wc}}(\theta) = L_{\text{id}}(\theta) + \rho_{\text{mis}}\mathbb{E}_{z\sim d_\theta}\left[b(z)\big(\ell^1_\theta(z) - \ell^0_\theta(z)\big)\right] + \rho_{\text{mis}}\beta R(\theta).$$

This separates ideal-oracle fitting, the fixed observed-oracle bias direction, and the worst-case sensitivity term. In all reported runs the clipped mass is zero, so this unclipped decomposition is exact. The mean interval widths are $2\rho_{\text{mis}} = 0.10, 0.15, 0.20$ for $\rho_{\text{mis}} = 0.05, 0.075, 0.10$, respectively.

Table F.1: Paired deltas, Robust SAIL (ours, $q^\star$) minus Nominal SAIL, on the test split. Losses and $R(\theta)$ are lower-is-better, so negative deltas indicate improvement for Robust SAIL.

ρ_{mis}	$q^\star$	ΔL_{id}	ΔR	ΔL_{wc}	$\Delta\text{AUC}_{0.5:1}$	crossing α	wins
0.05	1.0	$+0.0011$	-0.4931	-0.0016	-0.0009	0.417	10/10
0.075	1.0	$+0.0008$	-0.4904	-0.0028	-0.0019	0.226	10/10
0.10	1.0	$+0.0004$	-0.5144	-0.0042	-0.0030	0.092	10/10

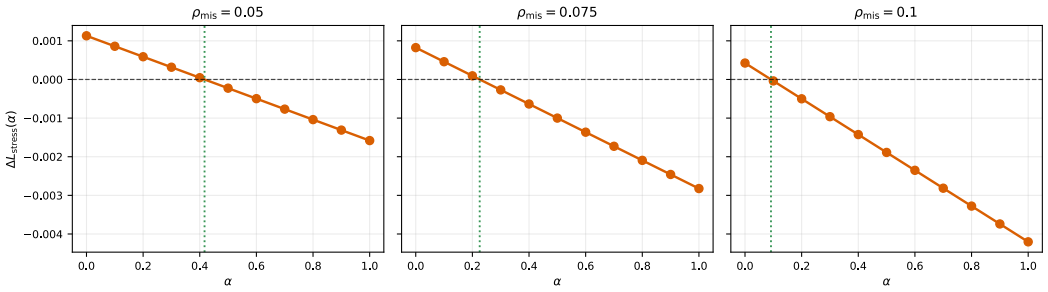

Figure F.1: Paired true-to-worst stress gaps across misspecification radii. The plotted quantity is $\Delta L_{\text{stress}}(\alpha) = L_{\text{stress}}(\theta_{\text{Robust}}; \alpha) - L_{\text{stress}}(\theta_{\text{Nominal}}; \alpha)$. Negative values indicate lower stress loss for Robust SAIL (ours). Robust SAIL is worse near the ideal oracle but becomes better as the evaluation oracle moves toward the worst-case endpoint.

F.4 Main all-radius stress results

Table F.1 reports paired test deltas between Robust SAIL (ours) at $q^\star$ and Nominal SAIL. Robust SAIL (ours) has higher L_{id}, which is the intended ideal-oracle fitting sacrifice. It simultaneously reduces $R(\theta)$ by roughly two orders of magnitude and improves both L_{wc} and $\text{AUC}_{0.5:1}$ on all $10/10$ paired seeds for every misspecification radius.

The crossing point moves earlier as ρ_{mis} increases: under larger oracle stress, the reduction in sensitivity dominates the ideal-oracle fitting cost sooner along the true-to-worst path.

F.5 Radius selection and calibration

For each ρ_{mis}, the robust radius is selected by seed-pooled validation high-stress AUC:

$$q^\star(\rho_{\text{mis}}) = \arg\min_{q>0} \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \text{AUC}^{\text{val}}_{0.5:1}(\theta_{q,s}).$$

The nominal baseline $q = 0$ is excluded from the robust candidate pool. Selection is global per ρ_{mis}, not per seed, and the test split is used only after $q^\star$ is fixed. The same selected value is obtained if validation L_{wc} is used instead of validation AUC.

The selected value has a precise controlled-benchmark interpretation:

$$q^\star = 1 \quad \Longleftrightarrow \quad \rho_{\text{alg}} = \rho_{\text{mis}}.$$

This is a radius-calibration result, not a deployment-time rule. In real applications, ρ_{mis} is unknown. One should tune ρ_{alg} directly, or equivalently $\lambda = \rho_{\text{alg}}\beta$, using one of the following criteria: robust validation performance, uncertainty calibration from annotator disagreement or a domain-specific risk tolerance.

F.6 Utility-preserving operating point selection

The main experiment selects $q^\star$ using a robust-first validation criterion. If a reader instead wants to preserve ideal reward, the same grid can be used to choose a smaller ρ_{alg} under a utility constraint. The quantity $E[r^\star]$ is available only in this controlled benchmark and is used here only as a diagnostic.

Table F.2: Validation selection of $q^\star$. Both validation high-stress AUC and validation L_{wc} select the full-radius robust objective for all three radii.

ρ_{mis}	$q^\star$ by AUC	$q^\star$ by L_{wc}	val $\mathrm{AUC}_{0.5:1}$	val L_{wc}	val R
0.05	1.0	1.0	0.6931	0.6932	0.0038
0.075	1.0	1.0	0.6932	0.6932	0.0025
0.10	1.0	1.0	0.6932	0.6932	0.0026

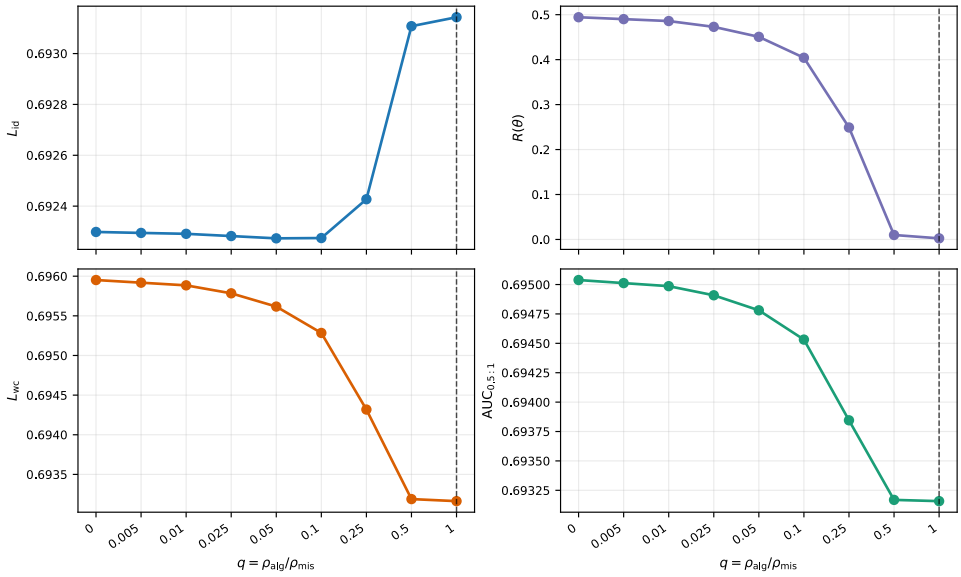

Figure F.2: Radius-selection sweep at $\rho_{\mathrm{mis}} = 0.075$. Increasing $q = \rho_{\mathrm{alg}}/\rho_{\mathrm{mis}}$ reduces the sensitivity penalty and the high-stress oracle loss, while increasing the ideal-oracle loss. Validation high-stress AUC selects the full-radius robust objective $q^\star = 1$.

For a utility tolerance ϵ_{util}, define

$$q_{\mathrm{util}} = \max\left\{q : E[r^\star(\theta_q)] \geq E[r^\star(\theta_0)] - \epsilon_{\mathrm{util}}\right\}.$$

At the central radius $\rho_{\mathrm{mis}} = 0.075$, the utility values are:

q	0	0.005	0.01	0.025	0.05	0.10	0.25	0.50	1.00
$E[r^\star]$	0.2190	0.2188	0.2186	0.2181	0.2172	0.2154	0.2098	0.2011	0.2008

For $\epsilon_{\mathrm{util}} = 10^{-3}$, this rule selects $q_{\mathrm{util}} = 0.025$. Thus $q^\star = 1$ is best under the robust-first high-stress AUC criterion, while $q_{\mathrm{util}} = 0.025$ is a much lighter robust setting that preserves true utility within the specified tolerance. Radius selection is therefore a risk-tolerance problem, not something the robust objective solves automatically.

F.7 KL-to-reference smoothing baseline

We also evaluate KL-to-reference smoothing:

$$\min_\theta L_{\mathrm{SAIL}}(\theta; P^{\mathrm{obs}}) + \eta_{\mathrm{KL}} \mathbb{E}_x D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x)\right).$$

The hyperparameter η_{KL} is a generic smoothing weight. It is not an oracle uncertainty radius and does not correspond to ρ_{mis}.

KL-to-reference smoothing is selected by validation high-stress AUC over its own grid. At $\rho_{\mathrm{mis}} = 0.075$, the selected value is $\eta_{\mathrm{KL}} = 0.1$. Table F.3 shows that tuned KL smoothing can produce a stress profile close to Robust SAIL (ours) in this finite-response benchmark. Robust SAIL differs in that its penalty is derived from the worst-case oracle objective and directly controls the explicit sensitivity term $R(\theta)$.

Table F.3: KL-to-reference smoothing baseline at $\rho_{\text{mis}} = 0.075$. Lower is better for all columns.

Method	L_{id}	$R(\theta)$	L_{wc}	$\text{AUC}_{0.5:1}$
Nominal SAIL	0.6923	0.4928	0.6960	0.6951
Robust SAIL (ours) ($q^\star = 1$)	0.6931	0.0025	0.6932	0.6932
KL-to-reference ($\eta_{\text{KL}} = 0.1$)	0.6930	0.0385	0.6933	0.6932

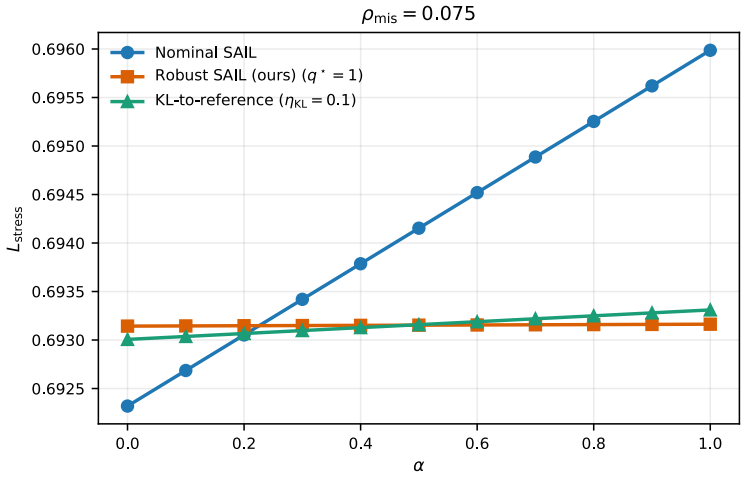

Figure F.3: KL-to-reference smoothing baseline at $\rho_{\text{mis}} = 0.075$. A tuned KL baseline also flattens the stress curve, showing that strong reference smoothing can approximate the robust operating point in this finite-response benchmark. Robust SAIL (ours) differs in that its penalty is derived from the worst-case oracle objective and directly controls the sensitivity term $R(\theta)$.

F.8 Shared-adversary common-worst diagnostic

The main stress metric uses a policy-specific worst-case oracle, which is the appropriate robust-risk evaluation. As a sanity check, we also construct a shared common adversary using the Nominal policy:

$$p_{\text{wc}}^{\theta_{\text{Nominal}}}(z).$$

Both methods are then evaluated along the same interpolation path

$$p_\alpha^{\text{common}}(z) = (1-\alpha)p_{\text{id}}(z) + \alpha p_{\text{wc}}^{\theta_{\text{Nominal}}}(z).$$

The common-adversary stress loss is

$$L_{\text{stress}}^{\text{common}}(\theta;\alpha) = \mathbb{E}_{z\sim d_\theta}\left[p_\alpha^{\text{common}}(z)\ell_\theta^1(z) + \left(1 - p_\alpha^{\text{common}}(z)\right)\ell_\theta^0(z)\right].$$

We plot the paired gap

$$\Delta L_{\text{stress}}^{\text{common}}(\alpha) = L_{\text{stress}}^{\text{common}}(\theta_{\text{Robust}};\alpha) - L_{\text{stress}}^{\text{common}}(\theta_{\text{Nominal}};\alpha).$$

F.9 Statistical reporting and reproducibility

All reported averages are over 10 paired seeds. Error bars, where shown, are standard errors over paired seeds unless otherwise stated. Pairwise comparisons use seed-level paired differences. For all main stress results, clipped mass is zero. All reported losses use exact finite-response sums. Validation selection is performed before test reporting.

All experiments are finite-response experiments over fixed candidate sets. Each method-radius-seed run is short; in our implementation each cell trains in less than one minute on a single GPU. The full sweep was run on an internal workstation with four NVIDIA GeForce RTX 2080 Ti GPUs with 11 GB memory per GPU. The controlled true reward $E[r^\star]$ is reported only as an appendix diagnostic and is never used for selecting the robust-first $q^\star$.

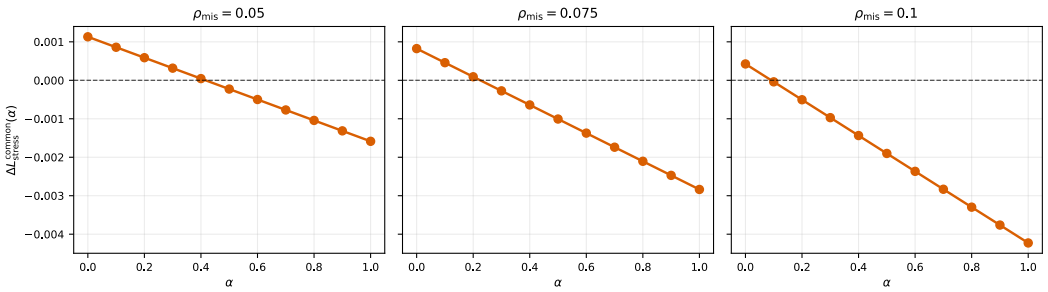

Figure F.4: Shared-adversary common-worst diagnostic. Negative values mean Robust SAIL (ours) has lower stress loss. Robust SAIL still becomes better under high stress when both methods are evaluated against the same Nominal-policy adversary, confirming that the main result is not an artifact of using different endpoint oracles for different policies.

G Limitations and future work

Our results apply to a finite-response, log-linear abstraction of online alignment. Assumptions 2, 4, and 5 require a softmax policy over a finite candidate set, bounded features, and a bounded convex parameter set. These assumptions enable explicit weak-convexity and stochastic-oracle bounds, but they do not directly cover transformer-scale autoregressive policies over all token sequences. The decomposition of Theorem 3.2 itself holds for any policy with well-defined likelihood ratios with respect to the reference policy; what does not carry over is the subsequent analysis. For an autoregressive neural policy, h_θ is nonlinear in θ and the score involves full-sequence gradients, so the bounded-feature arguments no longer yield uniform curvature or moment constants. Extending the analysis beyond finite log-linear policies, for instance under local curvature and sequence-gradient moment conditions, is an important direction for future work.

Our convergence guarantee is stated through the Moreau proximal point of the returned iterate rather than the iterate itself. This is a limitation, but not a purely surrogate one: standard envelope theory ensures the proximal point is close to the returned iterate and nearly stationary for the original constrained objective.

H Broader impacts

This work may have positive societal impact by improving the robustness of online preference-based alignment when the available preference oracle is misspecified. In particular, explicitly modeling oracle uncertainty can help reduce over-optimization of annotator, reward-model, or feedback-loop quirks, and can make the assumptions behind robustness guarantees more transparent. At the same time, robust alignment objectives do not by themselves guarantee safe deployment: if the uncertainty set is too narrow, miscentered, or based on biased feedback, the resulting policy may still preserve or amplify undesirable preferences. Improved alignment optimization tools could also be misused to optimize systems toward manipulative, biased, or otherwise harmful preference oracles. Our empirical study is limited to controlled finite-response experiments and does not release a deployable language model or dataset. We therefore emphasize explicit assumptions, validation-based radius selection, and transparency about the scope of the robustness guarantee as partial mitigations.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: These claims are supported by Theorems 3.2, 3.7, 4.1 and Corollary 4.2, and the experiments in Section 6 and Appendix F.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please see Appendix G.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The assumptions are stated in Assumptions 1–6 and Lemma 3.10, and complete proofs are given in Appendices B–D.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 6 and Appendix. F specify the experimental setups.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The code for these experiments is available at `https://github.com/anonymousneurips2026/oracle-robust-online-alignment`.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 6 and Appendix F specify the finite-response setups.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: See Section 6

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See appendix F.1

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is theoretical and uses controlled finite-response experiments with no human subjects, no personally identifiable data, and no release of high-risk models or datasets.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: See appendix H

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release pretrained language models, image generators, scraped datasets, or other assets with high misuse risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Appendix F discloses that the experiments use the frozen checkpoint `Qwen/Qwen2.5-0.5B-Instruct` and its Apache-2.0 license is respected.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: No new assets are introduced in the paper.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, human annotation, or human-subject data collection.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or equivalent institutional review is not applicable.

    Guidelines:

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

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: Appendix F describes the limited LLM usage in the experiments: `Qwen/Qwen2.5-0.5B-Instruct` is used as a frozen reference-logit provider for constructing finite response sets.

    Guidelines:

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