SHARPNESS-AWARE MINIMIZATION IN LOGIT SPACE EFFICIENTLY ENHANCES DIRECT PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) has emerged as a popular algorithm for aligning pretrained large language models with human preferences, owing to its simplicity and training stability. However, DPO suffers from the recently identified *squeezing effect* (also known as *likelihood displacement*), where the probability of preferred responses decreases unintentionally during training. To understand and mitigate this phenomenon, we develop a theoretical framework that models the coordinate-wise dynamics in the logit space. Our analysis reveals that gradient descent with a negative learning rate causes residuals to expand rapidly along high-curvature directions, which underlies the squeezing effect, whereas Sharpness-Aware Minimization (SAM) can suppress this behavior through its curvature-regularization effect. Building on this insight, we investigate *logits-SAM*, a computationally efficient variant that perturbs only the output layer with negligible overhead. Extensive experiments on Pythia-2.8B and Mistral-7B across multiple datasets demonstrate that logits-SAM consistently improves the effectiveness of DPO.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022) is a crucial technique for aligning pretrained large language models (LLMs) with human preferences to ensure helpfulness, harmlessness and safety (Bai et al., 2022; Dai et al., 2023). Its pipeline typically comprises three stages: supervised fine-tuning (SFT), reward modeling, and policy optimization. Classical policy optimization methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), while widely used for their effectiveness, depend heavily on the quality of the learned reward model, rendering training complex and often unstable. Direct Preference Optimization (DPO) (Rafailov et al., 2024b) is a recently proposed and promising offline alternative that, by reparameterizing the implicit reward and optimizing a closed-form objective on preference data, trains the policy directly without explicitly fitting a reward model. DPO has gained traction due to its algorithmic simplicity and training stability.

Despite DPO and its many variants demonstrating state-of-the-art performance across a range of tasks, several potential issues remain. A particularly important one is the recently identified *squeezing effect* (Ren & Sutherland, 2024) (also known as *likelihood displacement* (Razin et al., 2024)), which describes an unintended decrease in the generation probability of preferred responses during DPO training, contrary to the intended goal of increasing it embodied in the DPO objective. This phenomenon can lead to performance degradation, reduced safety, and even alignment failure (Pal et al., 2024; Yuan et al., 2024; Rafailov et al., 2024a; Tajwar et al., 2024; Pang et al., 2024).

To understand the mechanism behind the squeezing effect and to identify an effective remedy, we develop a theoretical framework that elucidates the learning dynamics in both the parameter space and the logit space. Our analysis shows that gradient descent (GD) with a negative learning rate causes the residual vector to expand rapidly along high-curvature directions, i.e., along the eigenvectors associated with large eigenvalues of the Hessian, which is the source of the squeezing effect. This raises a natural question: *can curvature-aware training mitigate this unintended drift?*

We investigate *Sharpness-Aware Minimization* (SAM) (Foret et al., 2021), a bilevel optimization method widely used in supervised learning, and establish its dynamics in both the parameter and logit spaces. Our theory demonstrates that SAM effectively alleviates the squeezing effect through its intrinsic curvature regularization. Guided by these insights, we advocate using *logits-SAM* for DPO training, a computationally efficient variant of SAM that perturbs only the output-layer parameters. Although logits-SAM has been mentioned merely as a byproduct in prior work (Baek et al., 2024; Singh et al., 2025) and often overlooked, our study turns this neglected variant into a practically useful and effective technique by integrating it into DPO, where it efficiently mitigates the squeezing effect and consistently improves performance. To the best of our knowledge, this is the first work to analyze and apply SAM in the context of DPO.

Contributions. Our contributions are summarized as follows:

- We develop a theoretical framework that connects the parameter space and the logit space through geometric properties, enabling a unified analysis of learning dynamics in both domains. This framework yields unified dynamical equations for GD and SAM that precisely track coordinate-wise evolution with controlled error terms.
- Our analysis identifies the root cause of the squeezing effect: under a negative learning rate, residuals expand rapidly along high-curvature directions. We rigorously show that SAM, through its intrinsic curvature regularization, effectively alleviates this phenomenon.
- Bridging theory and practice, we implement an efficient variant, *logits-SAM*, which perturbs only the output-layer parameters. Unlike vanilla SAM, it incurs virtually no additional overhead. Experiments on Pythia-2.8B and Mistral-7B across multiple datasets and benchmarks validate its effectiveness, demonstrating consistent performance gains for DPO and its variants.

2 PRELIMINARIES

2.1 PREFERENCE OPTIMIZATION

SFT–RLHF pipeline. Classical RLHF alignment proceeds in three phases: (i) *supervised fine-tuning* of a base policy on instruction-following data; (ii) *reward modeling* by fitting a scalar reward function on pairwise human preferences; and (iii) *policy optimization* to maximize the learned reward under a KL regularizer toward a reference policy.

DPO reparameterization. DPO (Rafailov et al., 2024b) bypasses training an explicit reward model by expressing an *implicit* reward for a policy $\pi_{\boldsymbol{\theta}}$ as a log-likelihood ratio to a fixed reference policy π_{ref} (typically the SFT model):

$$r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) \;=\; \beta \,\log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y} \,|\, \boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y} \,|\, \boldsymbol{x})} \;+\; \beta \log Z(\boldsymbol{x}), \tag{1}$$

where $\beta > 0$ is a temperature and $Z(\boldsymbol{x})$ is a partition term independent of $\boldsymbol{\theta}$. Combining equation 1 with the Bradley–Terry preference model (Bradley & Terry, 1952) $p(\boldsymbol{y}^+ \succ \boldsymbol{y}^- \mid \boldsymbol{x}) = \sigma\big(r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^+) - r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^-)\big)$ yields the standard DPO objective, optimized over a dataset $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y}^+, \boldsymbol{y}^-)\}$ of preferred/dispreferred pairs:

$$\mathcal{L}_{\mathrm{DPO}}(\pi_{\boldsymbol{\theta}}; \pi_{\mathrm{ref}}) = -\,\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}^+, \boldsymbol{y}^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+ \mid \boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y}^+ \mid \boldsymbol{x})} - \beta \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^- \mid \boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y}^- \mid \boldsymbol{x})} \right) \right], \tag{2}$$

where $\sigma(\cdot)$ is the logistic function.

2.2 SHARPNESS-AWARE MINIMIZATION

SAM regularizes training by explicitly penalizing *parameter-space sharpness*: it chooses parameters that minimize the worst-case loss within an ℓ_2 ball of radius ρ around $\boldsymbol{\theta}$. Concretely, for supervised learning with examples $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}$ and per-example loss $f(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$, the SAM objective is

$$\min_{\boldsymbol{\theta}} \;\; \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \left[\max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \; f\big(\boldsymbol{\theta} + \boldsymbol{\epsilon}\,;\, \boldsymbol{x}, \boldsymbol{y}\big) \right]. \tag{3}$$

This formulation can be interpreted as a form of *curvature regularization*: by seeking minimizers whose neighborhoods exhibit consistently low loss, SAM favors flatter minima that often correlate with improved generalization. In practice, the inner maximization is approximated to first order by the perturbation $\boldsymbol{\epsilon}^*(\boldsymbol{\theta}) = \rho \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) / \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})\|$, and one takes a descent step using the gradient at the perturbed point, $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta} + \boldsymbol{\epsilon}^*; \boldsymbol{x}, \boldsymbol{y})$.

3 Learning dynamics in logit space

3.1 Setting

We adopt the same theoretical setting as in Ren & Sutherland (2024), namely multiclass logistic classification, where the features of the samples are fixed (also referred to as the kernel regime), and the learning rate can be either positive or negative, corresponding respectively to the objectives of $\boldsymbol{y}^+$ and $\boldsymbol{y}^-$ in DPO. Let $\boldsymbol{x}$ be a training example with one-hot label $\boldsymbol{y} \in \{0,1\}^V,\ \mathbf{1}^\top \boldsymbol{y} = 1$. In the fixed-feature (kernel) regime, $\phi(\boldsymbol{x}) \in \mathbb{R}^d$ are fixed and

$$\boldsymbol{z}^t = \boldsymbol{W}^t \phi(\boldsymbol{x}) \in \mathbb{R}^V, \qquad \boldsymbol{p}^t = \mathrm{softmax}(\boldsymbol{z}^t), \qquad f(\boldsymbol{z}^t, \boldsymbol{y}) = -\sum_{k=1}^{V} \boldsymbol{y}_k \log \boldsymbol{p}_k^t,$$

where $\boldsymbol{W}^t \in \mathbb{R}^{V \times d}$ are trainable parameters, $\boldsymbol{z}^t$ are the logits. For notational convenience, we write $\phi(\boldsymbol{x})$ as ϕ. We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the Frobenius norm for matrices. We use $\otimes$ to denote the Kronecker product.

We denote the parameter Hessian by $\boldsymbol{H}_{\boldsymbol{W}}^t \coloneqq \nabla_{\boldsymbol{W}}^2 f(\boldsymbol{z}^t, \boldsymbol{y}) \in \mathbb{R}^{Vd \times Vd}$, and $\mu \coloneqq \|\phi\|^2$. In logit space, we denote the logit gradient by $\boldsymbol{g}^t \coloneqq \nabla_{\boldsymbol{z}} f(\boldsymbol{z}^t, \boldsymbol{y}) = \boldsymbol{p}^t - \boldsymbol{y} \in \mathbb{R}^V$, and denote the logit Hessian by $\boldsymbol{H}_{\boldsymbol{z}}^t \coloneqq \nabla_{\boldsymbol{z}}^2 f(\boldsymbol{z}^t, \boldsymbol{y}) \in \mathbb{R}^{V \times V}$.

3.2 Theory

The theoretical results of Ren & Sutherland (2024) demonstrate that the *squeezing effect* arises from the objective with a negative learning rate. Specifically, they prove that the probability of the ground-truth label necessarily decreases, while the probability of the model's most confident incorrect class necessarily increases. In this work, we provide a finer-grained analysis of the learning dynamics under this setting. We establish a unified modeling framework for the residuals of all classes and derive the linear convergence rate up to higher-order remainders. Furthermore, we apply our framework to prior analyses and further establish a rigorous conclusion that SAM can effectively mitigate the squeeze effect.

For GD, first-order derivatives are sufficient to characterize its dynamics. However, the intrinsic curvature regularization effect of SAM motivates us to further investigate the geometric structure of the parameter space through the Hessian matrix. To this end, we develop a theoretical framework that connects the geometry of the parameter space and the logit space, via the link between the parameter Hessian and the logit Hessian.

Proposition 3.1 (Geometry of the logit space; simplified version of Proposition A.1)**.** *In coordinates,* $\boldsymbol{H}_{\boldsymbol{W}} = \boldsymbol{H}_{\boldsymbol{z}} \otimes \left(\phi\phi^\top\right)$. *Thus, if* $\phi \neq \mathbf{0}$, *then* $\mathrm{rank}(\boldsymbol{H}_{\boldsymbol{W}}) = \mathrm{rank}(\boldsymbol{H}_{\boldsymbol{z}})$. *Moreover, the second-order effect of any parameter perturbation depends only on the induced logits perturbation* $T_\phi(\Delta \boldsymbol{W}) \coloneqq \Delta \boldsymbol{W}\, \phi$.

This proposition establishes that all second-order effects in the parameter space, whose Hessian $\boldsymbol{H}_{\boldsymbol{W}}$ lies in $\mathbb{R}^{Vd \times Vd}$, can be equivalently studied through the logit Hessian $\boldsymbol{H}_{\boldsymbol{z}}$ in $\mathbb{R}^{V \times V}$, thereby greatly simplifying the analysis of second-order dynamics. Next, we establish a unified framework to track the SAM dynamics in both the parameter space and the logit space, thanks to their favorable geometric structures. Unlike prior work, our framework can simultaneously trace the evolution of all coordinates of the parameters, logits, and residuals, while providing precise control over the error terms.

Theorem 3.2 (SAM dynamics in parameter and logit space; informal version of Theorem A.2)**.** *Assume that we conduct the SAM update for* $\boldsymbol{W}$. *Under mild assumptions, there exists a constant*

$C > 0$ *such that the following expansions hold with* $O(\eta^2)$ *remainders:*

$$\textbf{(parameters)} \quad \boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \eta\Big(\boldsymbol{g}^t\,\phi^\top + \underbrace{\tilde{\rho}^t\,\boldsymbol{H}_z^t \boldsymbol{g}^t\,\phi^\top}_{\textit{SAM's correction}}\Big) + \boldsymbol{R}_{\boldsymbol{W}}^t, \qquad \|\boldsymbol{R}_{\boldsymbol{W}}^t\| \le C\,\eta^2,$$

$$\textbf{(logits)} \quad \boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \eta\,\mu\Big(\boldsymbol{g}^t + \underbrace{\tilde{\rho}^t\,\boldsymbol{H}_z^t \boldsymbol{g}^t}_{\textit{SAM's correction}}\Big) + \boldsymbol{r}_{\boldsymbol{z}}^t, \qquad \|\boldsymbol{r}_{\boldsymbol{z}}^t\| \le C\,\eta^2,$$

$$\textbf{(residuals)} \quad \boldsymbol{g}^{t+1} = \boldsymbol{p}^{t+1} - \boldsymbol{y} = \Big(\boldsymbol{I} - \eta\,\mu\,\boldsymbol{H}_z^t - \underbrace{\eta\,\mu\,\tilde{\rho}^t\,(\boldsymbol{H}_z^t)^2}_{\textit{SAM's correction}}\Big)(\boldsymbol{p}^t - \boldsymbol{y}) + \boldsymbol{r}_{\boldsymbol{g}}^t, \quad \|\boldsymbol{r}_{\boldsymbol{g}}^t\| \le C\,\eta^2,$$

where $\tilde{\rho}^t \coloneqq \rho\,\sqrt{\mu}/\|\boldsymbol{g}^t\|$ *is the equivalent perturbation coefficient.*

It is worth noting that when $\rho = 0$, the dynamics reduce to the GD dynamics. This theorem, viewed through the lens of the logit Hessian, provides a precise theory for characterizing GD and SAM dynamics across spaces. In both parameter and logit space, GD amounts to scaling by the logit gradient, whereas SAM introduces an additional $\boldsymbol{H}_{\boldsymbol{z}}$ correction term that can be regarded as a preconditioning matrix. Moreover, the updates of the residual vector under GD and SAM are both preconditioned by $\boldsymbol{H}_{\boldsymbol{z}}$ (and, for SAM, by $(\boldsymbol{H}_{\boldsymbol{z}})^2$). This implies that if we choose the eigenvectors of the logit Hessian as a basis, the curvature coupling effects of both the first-order and second-order terms can be unified. To formalize this intuition, we show that $\boldsymbol{g}$ lies precisely in the column space of $\boldsymbol{H}_{\boldsymbol{z}}$, thus we can select the nonzero eigenvectors of $\boldsymbol{H}_{\boldsymbol{z}}$ as a basis to obtain the coordinate representation of $\boldsymbol{g}$.

Proposition 3.3. $\boldsymbol{H}_{\boldsymbol{z}}$ *is symmetric positive semidefinite with* $\ker(\boldsymbol{H}_{\boldsymbol{z}}) = \operatorname{span}\{\mathbf{1}\}$ *and* $\operatorname{rank}(\boldsymbol{H}_{\boldsymbol{z}}) = V - 1$. *Moreover, for the residual* $\boldsymbol{g}$ *we have* $\mathbf{1}^\top \boldsymbol{g} = 0$, *hence* $\boldsymbol{g} \in \mathbf{1}^\perp = \operatorname{range}(\boldsymbol{H}_{\boldsymbol{z}})$; *in particular, given any eigenbasis of* $\boldsymbol{H}_{\boldsymbol{z}}$ *restricted to* $\mathbf{1}^\perp$, $\boldsymbol{g}$ *admits a unique coordinate representation in that basis.*

Corollary 3.4 (Modal dynamics in the eigenbasis of $\boldsymbol{H}_z^t$)**.** *Under the same assumptions as Theorem 3.2. For each* t, *let the spectral decomposition of the symmetric positive–semidefinite matrix* $\boldsymbol{H}_z^t$ *be*

$$\boldsymbol{H}_z^t \;=\; \sum_{k=1}^{V-1} \lambda_k^t\,\boldsymbol{v}_k^t (\boldsymbol{v}_k^t)^\top,$$

where $\lambda_k^t > 0, \;\; (\boldsymbol{v}_k^t)^\top \boldsymbol{v}_\ell^t = \delta_{k\ell}$ *are the non-zero eigenvalues and eigenvectors. Define the* modal coefficients *of the residual* $\boldsymbol{g}^t = \boldsymbol{p}^t - \boldsymbol{y}$ *by*

$$e_k^t \coloneqq (\boldsymbol{v}_k^t)^\top \boldsymbol{g}^t, \quad e_k^{t+1} \;\coloneqq (\boldsymbol{v}_k^t)^\top \boldsymbol{g}^{t+1}, \qquad k = 1, \ldots, V-1. \tag{4}$$

Then there exists a constant $C > 0$ *such that for all nonzero modes* $k \ge 1$,

$$e_k^{t+1} = \Big(1 \;-\; \eta\,\mu\,\big[\lambda_k^t + \underbrace{\tilde{\rho}^t (\lambda_k^t)^2}_{\textit{SAM's correction}}\big]\Big)\, e_k^t \;+\; r_k^t, \qquad |r_k^t| \;\le\; C\,\eta^2. \tag{5}$$

Proofs are deferred to Appendix A. The corollary diagonalizes the vector dynamics into coordinate-wise scalars in the eigenbasis of $\boldsymbol{H}_{\boldsymbol{z}}$, making SAM's effect transparent. We now characterize the additional SAM correction in two regimes.

Case 1: Positive η, corresponding to the $\boldsymbol{y}^+$ objective in DPO. In this case, GD induces a stronger contraction of the residual $\boldsymbol{g}$ along the high-curvature directions, i.e., those associated with large eigenvalues of $\boldsymbol{H}_{\boldsymbol{z}}$. The additional correction term introduced by SAM has the same sign as that of GD, thereby amplifying this effect. **Case 2: Negative** η, corresponding to the $\boldsymbol{y}^-$ objective in DPO. Here, GD causes the residual $\boldsymbol{g}$ to expand more rapidly along high-curvature directions. Furthermore, standard SAM with positive ρ exacerbates this phenomenon, causing the residual to expand even faster along high-curvature directions compared to GD. By contrast, choosing a negative ρ counteracts this expansion.

Next, we extend our theoretical framework to the result of Ren & Sutherland (2024), which introduced the squeezing effect. For consistency with their notation, we let y denote the ground–truth class index (with one–hot label $\boldsymbol{y} = \boldsymbol{e}_y$).

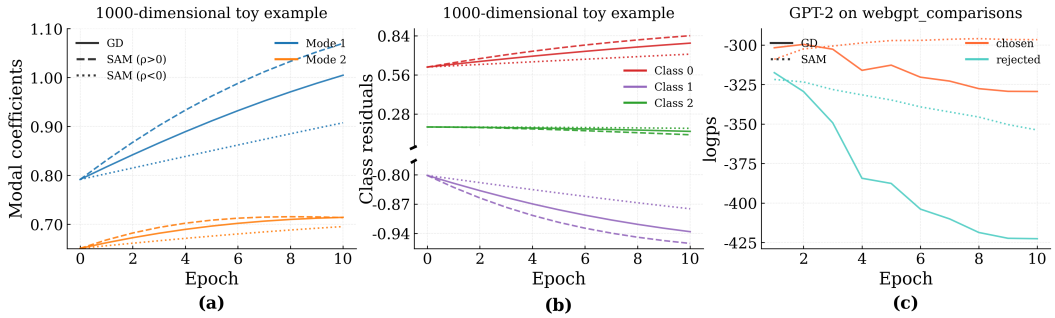

Figure 1: Training dynamics under different settings. (a–b) 1000-dimensional toy example with three classes, trained with a negative learning rate under GD, SAM ($\rho > 0$), and SAM ($\rho < 0$). Panel (a) shows the modal coefficients, and panel (b) shows the class residuals. (c) Real-data experiment on `webgpt_comparisons` with GPT-2, comparing GD and SAM: the panel reports the log-probabilities of the chosen and rejected responses.

Lemma 3.5 (One–step confidence ratios under GD, Lemma 1 of Ren & Sutherland (2024))**.** *For each class $i \in [V]$, define the one–step confidence ratio $\alpha_i := p_i^{t+1}/p_i^t$. Consider the objective with a negative learning rate $\eta < 0$, and denote its ground–truth label by y^-. Let $y^* = \arg\max_{j \neq y^-} p_j^t$ be the most confident incorrect class. Then*

$$\alpha_{y^*}^{\mathrm{GD}} > 1, \qquad \alpha_{y^-}^{\mathrm{GD}} < 1.$$

Lemma 3.5 formalizes the squeezing effect under GD with a negative learning rate: the probability of the most confident incorrect class increases, while that of the ground–truth class decreases. Within our framework, we next analyze the ratio of these two probabilities after a one–step SAM update.

Corollary 3.6 (One–step confidence ratios under SAM, informal version of Corollary A.5)**.** *Under the same assumptions as Theorem 3.2, assume that $\eta\rho > 0$. Then, for sufficiently small step size $|\eta|$, the following inequalities hold:*

$$\alpha_{y^*}^{\mathrm{SAM}} \leq \alpha_{y^*}^{\mathrm{GD}}, \qquad \alpha_{y}^{\mathrm{SAM}} \geq \alpha_{y}^{\mathrm{GD}}. \tag{6}$$

Here $y \in \{y^+, y^-\}$ denotes the ground–truth label corresponding to the positive or negative learning rate, respectively. Moreover, the inequalities in equation 6 are strict whenever $p_{y^*}^t \in (0, 1)$ and $p_y^t \leq \frac{1}{2}$.

The proof is deferred to Appendix A. Corollary 3.6 and Lemma 3.5 together imply that, when $\eta < 0$, using SAM with a negative $\rho < 0$ moderates the growth of the most confident incorrect class and slows the decay of the ground-truth class, thereby preventing excessive expansion and premature collapse. Our analysis thus reveals a key, albeit somewhat counterintuitive, fact: for negative η, one should choose a *negative* ρ (interpreted as a perturbation along the gradient descent direction), which effectively alleviates the squeezing effect.

We empirically validate our theoretical findings using a 1000-dimensional toy example with three classes. Specifically, we first train for 10 epochs using class 0 as the label for initialization, mimicking the SFT process, and then switch to class 1 as the label while continuing training with a negative learning rate. As shown in Figure 1, this setup faithfully reproduces the squeezing effect observed in prior work (Ren & Sutherland, 2024): both modal coefficients expand rapidly, the probabilities of class 1 and class 2 decrease, and only the probability of class 0, the model's most confident incorrect prediction, increases. Moreover, SAM with positive ρ exacerbates this effect, whereas SAM with negative ρ hinders this trend, exactly as predicted by our theory.

Additionally, Corollary 3.6 shows that for $\eta > 0$, SAM with $\rho > 0$ likewise mitigates the effect: the contraction of y^* is accelerated, while the growth of the ground-truth y^+ is enhanced. Taken together, these results establish a simple rule: during training, choosing ρ with the *same sign* as the learning rate alleviates the squeezing effect—specifically, it restrains the growth of y^* and promotes (or reduces the suppression of) y^+ and y^-. To validate this idea, we train a GPT-2 (Radford et al., 2019) on a subset of the WebGPT comparisons dataset (Nakano et al., 2022) using both GD and SAM. The probability dynamics of the chosen and rejected responses are shown in Figure 1.

We observe that SAM increases the probability of the chosen responses, whereas GD decreases it due to the squeezing effect; meanwhile, SAM slows down the decrease of the rejected responses' probability. These results are consistent with our theoretical predictions.

3.3 From theory to practice

An important challenge in applying SAM to DPO is that it requires an additional forward and backward pass, thereby nearly doubling the computational cost. However, our dynamical analysis shows that curvature regularization can still be achieved even when the perturbation is applied solely in the logit space (with an appropriate choice of the sign of ρ), which also alleviates the squeezing effect. Motivated by this observation, we suggest to use a computationally efficient SAM variant that perturbs only in the last layer, called *logits-SAM*, to improve the effectiveness and robustness of DPO. Its objective can be formulated as follows:

$$\mathcal{L}_{\text{DPO}}^{\text{logits-SAM}}(\boldsymbol{W},\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-) = \mathcal{L}_{\text{DPO}}\left(\boldsymbol{W}+\rho\frac{\nabla_{\boldsymbol{W}}\mathcal{L}_{\text{DPO}}(\boldsymbol{W},\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-)}{\left\|\nabla_{\boldsymbol{W}}\mathcal{L}_{\text{DPO}}(\boldsymbol{W},\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-)\right\|},\boldsymbol{\theta};\boldsymbol{x},\boldsymbol{y}^+,\boldsymbol{y}^-\right).$$

where $\boldsymbol{W}$ denotes the parameters in the output layer, and $\boldsymbol{\theta}$ denote the parameters except $\boldsymbol{W}$.

Implementation. Unlike our theoretical setting, common DPO implementations[1][2] typically encode the $\boldsymbol{y}^-$ objective as negative while using a single positive learning rate, rather than assigning positive and negative rates to $\boldsymbol{y}^+$ and $\boldsymbol{y}^-$, respectively. Accordingly, we adopt this convention in our SAM implementation. Our dynamical analysis further indicates that ρ should share the sign of the learning rate; hence we consistently use a positive ρ. We summarize the differences between the theoretical and practical settings in Table 4 of Appendix B.

Remark. This choice does not render our analysis of the negative learning rate redundant. For first-order methods such as GD, using a negative objective with a positive learning rate is equivalent (in dynamics) to using a positive objective with a negative learning rate. Therefore, our analysis applies fully to the case of negative objectives.

The implementation pseudocode can be found in Algorithm 1 of Appendix B. We compute the perturbation manually using the hidden states from the penultimate layer and the parameters of the final layer, requiring only a single full forward–backward pass instead of the two full passes required in standard SAM. Since the parameters of the final layer typically constitute only a small fraction of all trainable parameters (e.g., 4.64% in Pythia-2.8B and 1.81% in Mistral-7B), the additional training overhead introduced by logits-SAM is negligible. A detailed comparison of wall-clock time and peak memory usage is provided in Section 4.3.

4 Experiments

4.1 Experimental setup

Datasets. We conduct DPO training on three widely used datasets to evaluate our algorithm: Anthropic-HH (Bai et al., 2022), the Reddit TL;DR summarization dataset (Stiennon et al., 2020), and the UltraFeedback Binarized dataset (Cui et al., 2023).

Models. Following common practice, we adopt SFT models as our base models. We use Pythia-2.8B (Biderman et al., 2023) for experiments on Anthropic-HH and Reddit TL;DR, and Mistral-7B-v0.1 (Jiang et al., 2023) for UltraFeedback. For Pythia-2.8B, we initialize from the Hugging Face open-source checkpoint[3], which was SFT for one epoch on Anthropic-HH. For the TL;DR experiments, we use the checkpoint[4], which was SFT for one epoch on Reddit TL;DR. For Mistral-7B-v0.1, we use the Alignment Handbook (Tunstall et al., 2023a) checkpoint Zephyr-7b[5] (Tunstall et al., 2023b), which was SFT for one epoch on UltraChat-200k.

[1] https://github.com/eric-mitchell/direct-preference-optimization

[2] https://github.com/huggingface/trl

[3] https://huggingface.co/lomahony/eleuther-pythia2.8b-hh-sft

[4] https://huggingface.co/trl-lib/pythia-2.8b-deduped-tldr-sft

[5] https://huggingface.co/alignment-handbook/zephyr-7b-sft-full

Table 1: Evaluation results (WR %) on HH and TL;DR datasets using Pythia-2.8B. The judge is GPT-5-mini. The highest value within each method group (baseline vs. logits-SAM) is **bolded**.

Method	HH		TL;DR	
	vs SFT	vs chosen	vs SFT	vs chosen
DPO	70.52	56.35	84.21	34.78
DPO+logits-SAM	**72.28**	**60.51**	**89.58**	**36.57**
SLiC-HF	65.27	54.72	91.88	31.36
SLiC-HF+logits-SAM	**71.87**	**62.21**	**94.40**	**32.80**
CPO	66.60	58.19	90.99	39.38
CPO+logits-SAM	**70.24**	**59.90**	**93.29**	**45.41**

Evaluation. For Pythia-2.8B, we evaluate model performance on Anthropic-HH and Reddit TL;DR by measuring win rates (WR) against both the SFT baseline and the human-preferred responses, using GPT-5-mini (version 2025-08-07) as the automatic judge. Following the DPO paper, we set the decoding temperature to 0 for HH and 1 for TL;DR. For Mistral-7B-v0.1, we conduct evaluation on three popular open-ended instruction-following benchmarks: AlpacaEval 2 (Dubois et al., 2024), Arena-Hard v0.1 (Li et al., 2024), and MT-Bench (Zheng et al., 2023). Details of each benchmark can be found in Appendix C. We adopt the default generation parameters provided by each benchmark. Specifically, we report both length-controlled win rates (LC) and raw WR for AlpacaEval 2, model WR for Arena-Hard v0.1, and averaged judge scores (1–10) for MT-Bench, all following the standard evaluation protocols, with default decoding configurations.

Baselines. We apply logits-SAM to DPO and two SOTA variants, SLiC-HF (Zhao et al., 2023) and CPO (Xu et al., 2024). We use AdamW optimizer (Loshchilov & Hutter, 2019) in all experiments. For Pythia-2.8B, we set batch size 64 and learning rate 1×10^{-6}, following the DPO paper; for Mistral-7B, we use batch size 128 and learning rate 5×10^{-7}, following the Alignment Handbook's recommended settings.

Hyperparameters. For DPO, we adopt the recommended β values from the DPO paper and the Alignment Handbook, which are widely used and well tuned. For SLiC-HF and CPO, we select hyperparameters following the tuning protocol from Meng et al. (2024b). For logits-SAM, we keep all hyperparameters identical to each corresponding baseline to ensure fairness; the only additional hyperparameter is ρ, which we tune over $\{1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}\}$. Full hyperparameter settings are provided in Table 5 and Table 6 of Appendix C.

4.2 Experimental results

Performance of summarization and dialogue generation tasks. We present the results in Table 1. We find that logits-SAM consistently improves performance across both HH and TL;DR datasets. All three baselines (DPO, SLiC-HF, and CPO) achieve higher win rates against both SFT and chosen responses when augmented with logits-SAM. Notably, SLiC-HF shows the largest gains on HH (+6.60 pp vs SFT, +7.49 pp vs chosen), while CPO achieves strong improvements on TL;DR (+2.30 pp vs SFT, +6.03 pp vs chosen), demonstrating that logits-SAM provides stable and generalizable benefits across different optimization methods.

Performance on open-ended instruction-following benchmarks. We present the results in Table 2. The results demonstrate that combining logits-SAM with different DPO variants consistently yields performance gains across all benchmarks. On open-ended instruction-following evaluations, logits-SAM improves both length-controlled and original win rates on AlpacaEval 2 (e.g., with CPO: +4.35 pp LC, +3.65 pp WR), increases head-to-head win rate on Arena-Hard v0.1 (e.g., with DPO: +4.1 pp WR), and provides steady gains on MT-Bench (e.g., DPO: +0.30, SLiC-HF: +0.17, CPO: +0.27). These findings indicate that logits-SAM is a generally effective and robust enhancement across diverse evaluation settings.

Table 2: Evaluation results on AlpacaEval 2 (LC and WR), Arena-Hard v0.1 (WR), and MT-Bench using Mistral-7B-v0.1. Judges are GPT-4 Turbo for AlpacaEval 2, and GPT-4.1 for Arena-Hard v0.1 and MT-Bench. The highest value within each method group (baseline vs. logits-SAM) is **bolded**.

Method	AlpacaEval 2		Arena-Hard v0.1	MT-Bench
	LC (%)	WR (%)	WR (%)	(score)
DPO	13.08	10.96	19.0	5.49
DPO+logits-SAM	**13.90**	**11.62**	**23.1**	**5.79**
SLiC-HF	8.92	8.97	19.1	5.05
SLiC-HF+logits-SAM	**10.63**	**9.23**	**21.1**	**5.22**
CPO	8.97	8.13	19.2	5.22
CPO+logits-SAM	**13.32**	**11.78**	**21.4**	**5.49**

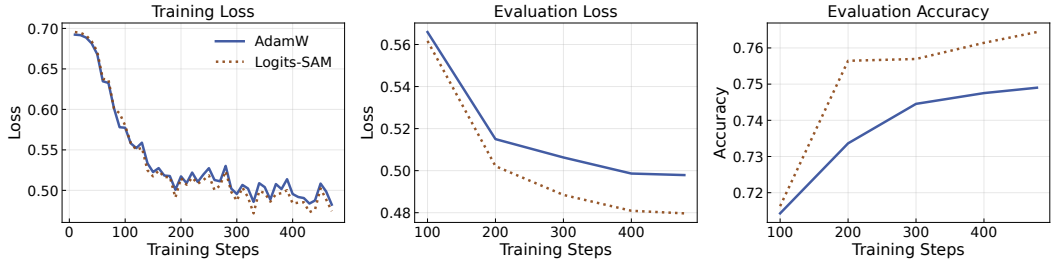

Figure 2: Learning dynamics of Mistral-7B on UltraFeedback. We compare AdamW and logits-SAM in terms of training loss, evaluation loss, and evaluation accuracy.

4.3 Additional Analysis

Learning dynamics. In Figure 2, we present a comparison of the learning dynamics between AdamW and SAM when training Mistral-7B on the UltraFeedback dataset. The figure reports training loss, evaluation loss, and evaluation accuracy across training steps. We observe that both optimizers achieve similar reductions in training loss, but SAM yields consistently lower evaluation loss and higher evaluation accuracy throughout training. These results suggest that SAM provides better generalization ability compared to AdamW.

Sharpness. To further probe the reasons underlying the generalization gains of logits-SAM, we measure the traces of the parameter Hessian and the logit Hessian at the final checkpoint of Mistral-7B. For AdamW, the traces are 1.337×10^4 / 2.732×10^2 (parameter / logit Hessian), while for logits-SAM they are reduced to 1.186×10^4 / 2.586×10^2. This reduction indicates that logits-SAM converges to a flatter solution, which is widely believed to be beneficial for generalization.

Computational overhead. Compared to vanilla SAM, logits-SAM minimizes additional computational overhead. We report wall-clock training time and peak memory on Pythia-2.8B trained on the Reddit TL;DR dataset (Figure 3), using data-parallel training (DDP) across two NVIDIA A100 GPUs with a per-device batch size of 4. The results show that logits-SAM adds only $\sim$ 2–3% extra time, with negligible peak-memory overhead. By contrast, vanilla SAM is practically infeasible for Pythia-2.8B on A100s with DDP: it nearly doubles the step time (due to an extra full forward–backward pass) and requires a perturbation buffer comparable to the model size (for billion-parameter models, this entails more than 10 GB of additional GPU memory), which leads to out-of-memory even with batch size 1. These observations highlight the clear compute-cost advantage of logits-SAM.

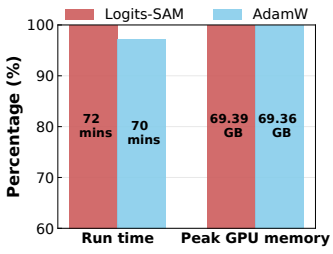

Figure 3: Efficiency comparison.

Sensitivity analysis. We present a sensitivity analysis of the additional hyperparameter ρ for logits-SAM in Table 3. The results indicate that, within a reasonable range of ρ, performance is typically improved consistently, whereas further enlarging ρ leads to a marked degradation. Notably, unlike original SAM, logits-SAM perturbs only the output layer, so the appropriate scale of ρ is much smaller than the range (0.01–0.5) recommended in the SAM paper. We recommend starting the search for logits-SAM's ρ at 10^{-5} or 10^{-4} and, if resources permit, performing a finer sweep in this neighborhood.

Table 3: Performance on HH and TL;DR datasets under different ρ values. Each entry reports win rate vs SFT (left) and vs chosen (right).

Dataset	$\rho = 0$ (AdamW)	$\rho = 10^{-5}$	$\rho = 10^{-4}$	$\rho = 10^{-3}$	$\rho = 10^{-2}$
HH	70.52 / 56.35	69.47 / 58.27	72.28 / 60.51	68.49 / 59.52	65.49 / 56.31
TL;DR	84.21 / 34.78	87.79 / 33.97	89.58 / 36.57	84.25 / 29.93	81.56 / 29.31

5 Related work

Reinforcement learning from human feedback. RLHF has emerged as the de facto post-training recipe for aligning large language models (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022), typically combining supervised fine-tuning (Zhou et al., 2023; Taori et al., 2023; Conover et al., 2023; Wang et al., 2023b), reward modeling (Gao et al., 2023; Luo et al., 2023; Lambert et al., 2024), and policy optimization (Schulman et al., 2017; Anthony et al., 2017). To reduce the complexity and instability of online preference optimization, offline methods such as SLiC-HF (Zhao et al., 2023) and RRHF (Yuan et al., 2023) learn policies from comparisons using closed-form objectives. DPO (Rafailov et al., 2024b) is a central example that maximizes the log-probability margin between preferred and rejected responses relative to a reference policy. Thanks to its simplicity and training stability, DPO has rapidly gained popularity, spurring a line of variants aimed at improving performance. For example, Azar et al. (2024) propose IPO, a more theoretically grounded variant; CPO (Xu et al., 2024) approximates the reference policy as uniform to eliminate the reference term; f-DPO (Wang et al., 2023a) generalizes DPO via a family of f-divergences; SimPO (Meng et al., 2024a) uses length-normalized scores that better reflect generation-time preferences; and Cal-DPO (Xiao et al., 2024) aligns the implicit reward scale with likelihoods.

Sharpness-aware minimization. A widely held belief in the deep learning community is that flatter solutions typically generalize better (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Dinh et al., 2017; Jiang et al., 2019; Xie et al., 2020; Liu et al., 2023). Motivated by this view, SAM (Foret et al., 2021) is a bilevel optimization method that explicitly seeks flatter minima, and it has gained popularity for delivering consistent improvements across a wide range of supervised learning tasks (Foret et al., 2021; Kwon et al., 2021; Kaddour et al., 2022; Liu et al., 2022; Kim et al., 2022; Li & Giannakis, 2023). Most relevant to our work are its recent applications in LLMs. Singh et al. (2025) propose Functional-SAM for LLM pretraining and demonstrate strong performance, while Lee & Yoon (2025) apply SAM to Proximal Policy Optimization to improve robustness in both the reward and action spaces. Logits-SAM is a byproduct mentioned in recent studies, yet it is often overlooked. Baek et al. (2024) analyze the effect of label noise on SAM in linear regression and argue that Jacobian-SAM, the counterpart of logits-SAM, plays the dominant role. Similarly, Singh et al. (2025) identify Jacobian-SAM, also referred to as Functional-SAM, as more important and show that it can effectively improve the generalization performance of LLM pretraining.

6 Conclusion

We analyzed the squeezing effect in DPO via coordinate-wise dynamics in parameter and logit spaces. Our framework shows that GD with negative η drives residuals to expand along high-curvature directions, and that SAM suppresses this behavior via curvature regularization; in particular, negative η calls for negative ρ. Motivated by this, we adopt *logits-SAM*, which perturbs only the output layer and adds negligible overhead, and demonstrate consistent gains in effectiveness and robustness across models and datasets. We expect these insights to inform curvature-aware preference optimization going forward.

REPRODUCIBILITY STATEMENT

All theoretical results presented in this paper are accompanied by complete proofs, which can be found in Appendix A. To further facilitate reproducibility, we will release the source code upon publication, allowing the community to verify and build upon our results.

REFERENCES

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, 2017.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Christina Baek, Zico Kolter, and Aditi Raghunathan. Why is sam robust to label noise? *arXiv preprint arXiv:2405.03676*, 2024.

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.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

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.

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

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instructiontuned llm. 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

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

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Hyun Kyu Lee and Sung Whan Yoon. Flat reward in policy parameter space implies robust reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bingcong Li and Georgios B. Giannakis. Enhancing sharpness-aware optimization through variance suppression, 2023. URL https://arxiv.org/abs/2309.15639.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024. URL https://arxiv.org/abs/2406.11939.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pp. 22188–22214. PMLR, 2023.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024a.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024b. URL https://arxiv.org/abs/2405.14734.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

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

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024a.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024b. URL https://arxiv.org/abs/2305.18290.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*, 2024.

Yi Ren and Danica J Sutherland. Learning dynamics of llm finetuning. *arXiv preprint arXiv:2407.10490*, 2024.

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

Sidak Pal Singh, Hossein Mobahi, Atish Agarwala, and Yann Dauphin. Avoiding spurious sharpness minimization broadens applicability of sam. *arXiv preprint arXiv:2502.02407*, 2025.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *NeurIPS*, 2020.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alvaro Bartolome, Carlos M. Patiño, Alexander M. Rush, and Thomas Wolf. The Alignment Handbook, 2023a. URL https://github.com/huggingface/alignment-handbook.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023b. URL https://arxiv.org/abs/2310.16944.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints, 2023a. URL https://arxiv.org/abs/2309.16240.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023b.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. Cal-dpo: Calibrated direct preference optimization for language model alignment. *Advances in Neural Information Processing Systems*, 37:114289–114320, 2024.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

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

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slic-hf: Sequence likelihood calibration with human feedback, 2023. URL https://arxiv.org/abs/2305.10425.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

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 Formal theorems and proofs

Proposition A.1 (Geometry of the logit space and the parameter–logit correspondence)**.** *Let* $\ell : \mathbb{R}^V \to \mathbb{R}$ *be* C^2*. Fix an input* $\boldsymbol{x}$ *and a feature map* $\phi(\boldsymbol{x}) \in \mathbb{R}^d$*. For* $\boldsymbol{W} \in \mathbb{R}^{V\times d}$ *set*

$$\boldsymbol{z} = \boldsymbol{W}\,\phi \in \mathbb{R}^V, \qquad F(\boldsymbol{W}) = \ell(\boldsymbol{z}).$$

Denote $\boldsymbol{H}_{\boldsymbol{z}} \coloneqq \nabla^2_{\boldsymbol{z}}\ell(\boldsymbol{z}) \in \mathbb{R}^{V\times V}$ *and* $\boldsymbol{H}_{\boldsymbol{W}} \coloneqq \nabla^2_{\boldsymbol{W}}F(\boldsymbol{W}) \in \mathbb{R}^{Vd\times Vd}$*.*

Equip $\mathbb{R}^{V\times d}$ *with the Frobenius inner product* $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_F = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ *and* $\mathbb{R}^V$ *with the Euclidean inner product. Let*

$$T_\phi : \mathbb{R}^{V\times d} \to \mathbb{R}^V, \qquad T_\phi(\Delta \boldsymbol{W}) = \Delta \boldsymbol{W}\,\phi$$

be the differential of the map $\boldsymbol{W} \mapsto \boldsymbol{W}\phi$, and let $T_\phi^* : \mathbb{R}^V \to \mathbb{R}^{V\times d}$ be its adjoint with respect to these inner products, i.e., $\langle T_\phi(\Delta \boldsymbol{W}),\ \boldsymbol{v}\rangle = \langle \Delta \boldsymbol{W},\ T_\phi^*(\boldsymbol{v})\rangle_F$ for all $\Delta \boldsymbol{W}, \boldsymbol{v}$. Then $T_\phi^*(\boldsymbol{v}) = \boldsymbol{v}\,\phi^\top$. The following statements hold.

(1) ***Pullback identity (operator form).***

$$\boldsymbol{H}_{\boldsymbol{W}} = T_\phi^*\, \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi$$

as linear operators on $\mathbb{R}^{V\times d}$. Equivalently, in coordinates,

$$\nabla_{\boldsymbol{W}} F(\boldsymbol{W}) = \big(\nabla_{\boldsymbol{z}} \ell(\boldsymbol{z})\big)\,\phi^\top, \qquad \boldsymbol{H}_{\boldsymbol{W}} = \boldsymbol{H}_{\boldsymbol{z}} \otimes \big(\phi\phi^\top\big).$$

Consequently, if $\phi \neq \mathbf{0}$, then

$$\operatorname{rank}(\boldsymbol{H}_{\boldsymbol{W}}) = \operatorname{rank}(\boldsymbol{H}_{\boldsymbol{z}}).$$

(2) ***Pullback of the bilinear form.*** *For every $\Delta \boldsymbol{W}, \Delta \boldsymbol{W}' \in \mathbb{R}^{V\times d}$,*

$$\langle \Delta \boldsymbol{W},\ \boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}']\rangle_F = \big\langle T_\phi(\Delta \boldsymbol{W}),\ \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W}')\big\rangle$$

and,

$$\boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}] = T_\phi^*(\boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W})) = \boldsymbol{H}_{\boldsymbol{z}}\, \Delta \boldsymbol{W}\, \big(\phi\phi^\top\big).$$

Thus the second–order effect of any parameter perturbation depends only on the induced logits perturbation $T_\phi(\Delta \boldsymbol{W}) = \Delta \boldsymbol{W}\,\phi$.

(3) ***Surjectivity, kernel, and quotient-space view.*** *If $\phi \neq \mathbf{0}$, then T_ϕ is surjective. For any $\Delta \boldsymbol{z} \in \mathbb{R}^V$, a minimum-Frobenius-norm preimage is*

$$\Delta \boldsymbol{W}_\star = \frac{\Delta \boldsymbol{z}\,\phi^\top}{\|\phi\|^2} \quad \textit{with} \quad T_\phi(\Delta \boldsymbol{W}_\star) = \Delta \boldsymbol{z}.$$

The kernel is

$$\ker(T_\phi) = \{\ \Delta \boldsymbol{W} \in \mathbb{R}^{V\times d} :\ \Delta \boldsymbol{W}\,\phi = \mathbf{0}\ \},$$

of dimension $V(d-1)$. Consequently, $\boldsymbol{H}_{\boldsymbol{W}}$ descends to the quotient $\mathbb{R}^{V\times d}/\ker(T_\phi) \cong \mathbb{R}^V$.

Proof. A direct computation gives

$$\langle T_\phi(\Delta \boldsymbol{W}), \boldsymbol{v}\rangle = \operatorname{tr}\big((\Delta \boldsymbol{W}\phi)^\top \boldsymbol{v}\big) = \operatorname{tr}\big(\Delta \boldsymbol{W}^\top\, \boldsymbol{v}\,\phi^\top\big) = \langle \Delta \boldsymbol{W},\ \boldsymbol{v}\,\phi^\top\rangle_F,$$

hence

$$T_\phi^*(\boldsymbol{v}) = \boldsymbol{v}\,\phi^\top\ .$$

(1) Pullback identity and coordinate forms. Let $F(\boldsymbol{W}) = \ell(\boldsymbol{W}\phi)$. The first differential of F is

$$\mathrm{d}F[\Delta \boldsymbol{W}] = \big\langle \nabla_{\boldsymbol{z}} \ell(\boldsymbol{z}),\ T_\phi(\Delta \boldsymbol{W})\big\rangle = \big\langle T_\phi^*\big(\nabla_{\boldsymbol{z}} \ell(\boldsymbol{z})\big),\ \Delta \boldsymbol{W}\big\rangle_F,$$

so

$$\nabla_{\boldsymbol{W}} F(\boldsymbol{W}) = T_\phi^*\big(\nabla_{\boldsymbol{z}} \ell(\boldsymbol{z})\big) = \big(\nabla_{\boldsymbol{z}} \ell(\boldsymbol{z})\big)\,\phi^\top.$$

Differentiating once more and using $\mathrm{d}(\nabla_{\boldsymbol{z}} \ell)(\boldsymbol{z})[\Delta \boldsymbol{z}] = \boldsymbol{H}_{\boldsymbol{z}}\,\Delta \boldsymbol{z}$ with $\Delta \boldsymbol{z} = T_\phi(\Delta \boldsymbol{W})$ yields, for all $\Delta \boldsymbol{W}, \Delta \boldsymbol{W}'$,

$$\mathrm{d}^2 F[\Delta \boldsymbol{W}, \Delta \boldsymbol{W}'] = \big\langle T_\phi(\Delta \boldsymbol{W}),\ \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W}')\big\rangle.$$

By the Riesz representation on $(\mathbb{R}^{V\times d}, \langle\cdot,\cdot\rangle_F)$, this means

$$\boldsymbol{H}_{\boldsymbol{W}}\ =\ T_\phi^*\, \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi\ .$$

Using $T_\phi^*(\boldsymbol{v}) = \boldsymbol{v}\phi^\top$ and $T_\phi(\Delta \boldsymbol{W}) = \Delta \boldsymbol{W}\phi$,

$$\boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}] = T_\phi^*(\boldsymbol{H}_{\boldsymbol{z}}\,(\Delta \boldsymbol{W}\phi)) = \big(\boldsymbol{H}_{\boldsymbol{z}}\,(\Delta \boldsymbol{W}\phi)\big)\,\phi^\top = \boldsymbol{H}_{\boldsymbol{z}}\,\Delta \boldsymbol{W}\,(\phi\phi^\top),$$

which is the coordinate (Kronecker) form used in the main text.

For the rank statement, assume $\phi \neq \mathbf{0}$. Then T_ϕ is surjective and T_ϕ^* is injective. Hence

$$\operatorname{rank}(\boldsymbol{H}_{\boldsymbol{W}}) = \operatorname{rank}\big(T_\phi^*\, \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi\big) = \operatorname{rank}\big(\boldsymbol{H}_{\boldsymbol{z}}\, T_\phi\big) = \operatorname{rank}(\boldsymbol{H}_{\boldsymbol{z}}),$$

because $\operatorname{range}(T_\phi) = \mathbb{R}^V$.

(2) Pullback of the bilinear form. By the operator identity above,

$$\langle \Delta \boldsymbol{W},\ \boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}'] \rangle_F = \big\langle \Delta \boldsymbol{W},\ T_\phi^*\, \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W}') \big\rangle_F = \big\langle T_\phi(\Delta \boldsymbol{W}),\ \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W}') \big\rangle.$$

Equivalently, $\boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}] = T_\phi^*(\boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W})) = \boldsymbol{H}_{\boldsymbol{z}}\, \Delta \boldsymbol{W}\, (\phi\phi^\top)$. Thus the bilinear form on parameter space is the pullback of the bilinear form induced by $\boldsymbol{H}_{\boldsymbol{z}}$ on logit space.

(3) Surjectivity, kernel and quotient view. If $\phi \neq \mathbf{0}$, then for any $\Delta \boldsymbol{z} \in \mathbb{R}^V$

$$\Delta \boldsymbol{W}_\star \ =\ \frac{\Delta \boldsymbol{z}\, \phi^\top}{\|\phi\|^2} \quad \text{satisfies} \quad T_\phi(\Delta \boldsymbol{W}_\star) = \Delta \boldsymbol{z},$$

so T_ϕ is surjective. The same choice minimizes the Frobenius norm among all preimages (row-wise Cauchy–Schwarz). The kernel is $\ker(T_\phi) = \{\Delta \boldsymbol{W} :\ \Delta \boldsymbol{W}\, \phi = \mathbf{0}\}$, and rank–nullity gives $\dim \ker(T_\phi) = V(d-1)$. Finally, if $\Delta \boldsymbol{W}_1 - \Delta \boldsymbol{W}_2 \in \ker(T_\phi)$, then $T_\phi(\Delta \boldsymbol{W}_1) = T_\phi(\Delta \boldsymbol{W}_2)$ and

$$\begin{aligned}\langle \Delta \boldsymbol{W}_1,\ \boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}_1] \rangle_F &= \big\langle T_\phi(\Delta \boldsymbol{W}_1),\ \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W}_1) \big\rangle = \big\langle T_\phi(\Delta \boldsymbol{W}_2),\ \boldsymbol{H}_{\boldsymbol{z}}\, T_\phi(\Delta \boldsymbol{W}_2) \big\rangle \\ &= \langle \Delta \boldsymbol{W}_2,\ \boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}_2] \rangle_F,\end{aligned}$$

so the bilinear form descends to the quotient $\mathbb{R}^{V\times d}/\ker(T_\phi) \cong \mathbb{R}^V$.

If $\phi = \mathbf{0}$ then $T_\phi \equiv 0$ and $\boldsymbol{H}_{\boldsymbol{W}} \equiv \mathbf{0}$, the degenerate case. □

Theorem A.2 (Dynamics of SAM)**.** *Fix a SAMple $\boldsymbol{x}$ and set $\mu = \|\phi\|^2 < \infty$. Assume:*

(1) $f(\boldsymbol{z}, \boldsymbol{y})$ is C^3 in $\boldsymbol{z}$ and there exists $L < \infty$ such that $\sup_{\boldsymbol{z}} \|\nabla^3_{\boldsymbol{z}} f(\boldsymbol{z}, \boldsymbol{y})\| \le L$.

(2) The step size $|\eta| \in (0, 1]$ and the SAM radius satisfies $|\rho| \le \kappa\sqrt{|\eta|}$ with a constant $\kappa \ge 0$.

(3) If $\|\boldsymbol{g}^t\| = 0$, set the inner perturbation to 0 and define $\tilde\rho^{\,t} = 0$; otherwise $\tilde\rho^{\,t} := \rho\,\sqrt{\mu}/\|\boldsymbol{g}^t\|$.

Consider standard SAM*:*

$$\Delta \boldsymbol{W}^t_{\mathrm{adv}} = \rho\, \frac{\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^t)}{\|\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^t)\|_F}, \qquad \widetilde{\boldsymbol{W}}^{\,t} = \boldsymbol{W}^t + \Delta \boldsymbol{W}^t_{\mathrm{adv}}, \qquad \boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \eta\, \nabla_{\boldsymbol{W}} f(\widetilde{\boldsymbol{W}}^{\,t}).$$

Then, there exists a constant $C > 0$ (depending only on L, μ, κ) such that the following expansions hold with $O(\eta^2)$ remainders:

$$\textbf{\textit{(parameters)}} \quad \boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \eta\Big(\boldsymbol{g}^t\, \phi^\top + \tilde\rho^{\,t}\, \boldsymbol{H}^t_z \boldsymbol{g}^t\, \phi^\top\Big) + \boldsymbol{R}^t_{\boldsymbol{W}}, \qquad \|\boldsymbol{R}^t_{\boldsymbol{W}}\|_F \le C\,\eta^2,$$

$$\textbf{\textit{(logits)}} \quad \boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \eta\,\mu\Big(\boldsymbol{g}^t + \tilde\rho^{\,t}\, \boldsymbol{H}^t_z \boldsymbol{g}^t\Big) + \boldsymbol{r}^t_{\boldsymbol{z}}, \qquad \|\boldsymbol{r}^t_{\boldsymbol{z}}\| \le C\,\eta^2,$$

$$\textbf{\textit{(logit gradient)}} \quad \boldsymbol{g}^{t+1} = \Big(\boldsymbol{I} - \eta\,\mu\, \boldsymbol{H}^t_z - \eta\,\mu\,\tilde\rho^{\,t}\, (\boldsymbol{H}^t_z)^2\Big)\boldsymbol{g}^t + \boldsymbol{r}^t_{\boldsymbol{g}}, \qquad \|\boldsymbol{r}^t_{\boldsymbol{g}}\| \le C\,\eta^2.$$

In particular, for softmax cross-entropy where $\boldsymbol{g}^t = \boldsymbol{p}^t - \boldsymbol{y}$ and

$$\textbf{\textit{(residual)}} \quad \boldsymbol{p}^{t+1} - \boldsymbol{y} = \Big(\boldsymbol{I} - \eta\,\mu\, \boldsymbol{H}^t_z - \eta\,\mu\,\tilde\rho^{\,t}\, (\boldsymbol{H}^t_z)^2\Big)(\boldsymbol{p}^t - \boldsymbol{y}) + \boldsymbol{r}^t_{\boldsymbol{g}}, \quad \|\boldsymbol{r}^t_{\boldsymbol{g}}\| \le C\,\eta^2.$$

Proof. Write $F(\boldsymbol{W}) := f(\boldsymbol{W}\phi, \boldsymbol{y})$ and $\boldsymbol{z} = \boldsymbol{W}\phi$. By Proposition A.1 (Pullback/Kronecker and operator forms),

$$\nabla_{\boldsymbol{W}} F(\boldsymbol{W}) = \boldsymbol{g}\, \phi^\top, \qquad \boldsymbol{H}_{\boldsymbol{W}}[\Delta \boldsymbol{W}] = \boldsymbol{H}_z\, \Delta \boldsymbol{W}\, \big(\phi\phi^\top\big),$$

and $T_\phi(\Delta \boldsymbol{W}) = \Delta \boldsymbol{W}\, \phi$ with $\|T_\phi\| \le \|\phi\| = \sqrt{\mu}$. Moreover, by the multilinear chain rule applied to $F(\boldsymbol{W}) = f(\boldsymbol{W}\phi, \boldsymbol{y})$,

$$\nabla^3_{\boldsymbol{W}} F(\boldsymbol{W})[\Delta_1, \Delta_2, \Delta_3] = \nabla^3_{\boldsymbol{z}} f(\boldsymbol{z}, \boldsymbol{y})\big[T_\phi(\Delta_1),\, T_\phi(\Delta_2),\, T_\phi(\Delta_3)\big], \tag{7}$$

hence the operator norm satisfies

$$\sup_{\boldsymbol{W}} \big\|\nabla^3_{\boldsymbol{W}} F(\boldsymbol{W})\big\| \ \le\ \Big(\sup_{\boldsymbol{z}} \big\|\nabla^3_{\boldsymbol{z}} f(\boldsymbol{z}, \boldsymbol{y})\big\|\Big)\, \|T_\phi\|^3 \ \le\ L\,\mu^{3/2}. \tag{8}$$

(i) Parameter update. Let

$$\Delta \boldsymbol{W}_{\text{adv}}^{\,t} = \rho \, \frac{\nabla_{\boldsymbol{W}} F(\boldsymbol{W}^t)}{\|\nabla_{\boldsymbol{W}} F(\boldsymbol{W}^t)\|_F} \quad \text{and} \quad \widetilde{\boldsymbol{W}}^{\,t} = \boldsymbol{W}^t + \Delta \boldsymbol{W}_{\text{adv}}^{\,t}.$$

If $\|\boldsymbol{g}^t\| > 0$, then $\nabla_{\boldsymbol{W}} F(\boldsymbol{W}^t) = \boldsymbol{g}^t \phi^\top$ and $\|\boldsymbol{g}^t \phi^\top\|_F = \|\boldsymbol{g}^t\| \, \|\phi\| = \|\boldsymbol{g}^t\| \sqrt{\mu}$, so

$$\Delta \boldsymbol{W}_{\text{adv}}^{\,t} = \rho \, \frac{\boldsymbol{g}^t \phi^\top}{\|\boldsymbol{g}^t\| \sqrt{\mu}}, \qquad \big\|\Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big\|_F = |\rho| \le \kappa \sqrt{|\eta|}.$$

(If $\|\boldsymbol{g}^t\| = 0$, our convention sets $\Delta \boldsymbol{W}_{\text{adv}}^{\,t} = \mathbf{0}$.) A second-order Taylor expansion of $\nabla_{\boldsymbol{W}} F$ at $\boldsymbol{W}^t$ gives, for some $\theta \in (0,1)$,

$$\nabla_{\boldsymbol{W}} F(\widetilde{\boldsymbol{W}}^{\,t}) = \nabla_{\boldsymbol{W}} F(\boldsymbol{W}^t) + \boldsymbol{H}_{\boldsymbol{W}}^t\big[\Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big] + \tfrac{1}{2} \, \nabla_{\boldsymbol{W}}^3 F\big(\boldsymbol{W}^t + \theta \Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big)\big[\Delta \boldsymbol{W}_{\text{adv}}^{\,t}, \Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big].$$

By equation 8 and $\|\Delta \boldsymbol{W}_{\text{adv}}^{\,t}\|_F \le \kappa \sqrt{|\eta|}$,

$$\Big\| \tfrac{1}{2} \, \nabla_{\boldsymbol{W}}^3 F(\,\cdot\,)\big[\Delta \boldsymbol{W}_{\text{adv}}^{\,t}, \Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big] \Big\| \;\le\; \tfrac{1}{2} \, L \, \mu^{3/2} \, \|\Delta \boldsymbol{W}_{\text{adv}}^{\,t}\|_F^2 \;\le\; C_0 \, |\eta|,$$

for a constant $C_0 = C_0(L, \mu, \kappa)$. Using the operator identity from Proposition A.1,

$$\boldsymbol{H}_{\boldsymbol{W}}^t\big[\Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big] = \boldsymbol{H}_z^t \, \Delta \boldsymbol{W}_{\text{adv}}^{\,t}\big(\phi \phi^\top\big) = \frac{\rho \sqrt{\mu}}{\|\boldsymbol{g}^t\|} \, \boldsymbol{H}_z^t \boldsymbol{g}^t \, \phi^\top = \tilde{\rho}^{\,t} \, \boldsymbol{H}_z^t \boldsymbol{g}^t \, \phi^\top.$$

Therefore

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^t - \eta \Big(\boldsymbol{g}^t \phi^\top + \tilde{\rho}^{\,t} \, \boldsymbol{H}_z^t \boldsymbol{g}^t \phi^\top \Big) - \eta \, \boldsymbol{R}_\nabla^t,$$

where $\|\boldsymbol{R}_\nabla^t\|_F \le C_0 \, |\eta|$. Setting $\boldsymbol{R}_{\boldsymbol{W}}^t \coloneqq -\eta \, \boldsymbol{R}_\nabla^t$ yields $\|\boldsymbol{R}_{\boldsymbol{W}}^t\|_F \le C \, \eta^2$ with $C = C(L, \mu, \kappa)$, proving the parameter expansion.

(ii) Logit update. Right-multiplying by ϕ and using $\mu = \|\phi\|^2$,

$$\boldsymbol{z}^{t+1} - \boldsymbol{z}^t = (\boldsymbol{W}^{t+1} - \boldsymbol{W}^t)\phi = -\eta \, \mu \Big(\boldsymbol{g}^t + \tilde{\rho}^{\,t} \, \boldsymbol{H}_z^t \boldsymbol{g}^t \Big) + \boldsymbol{r}_{\boldsymbol{z}}^t,$$

with $\|\boldsymbol{r}_{\boldsymbol{z}}^t\| \le \|\boldsymbol{R}_{\boldsymbol{W}}^t\|_F \, \|\phi\| \le C \, \eta^2$ (absorbing $\sqrt{\mu}$ into C). This proves the logits expansion.

(iii) logit gradient update. Since $\boldsymbol{g} = \nabla_{\boldsymbol{z}} f(\boldsymbol{z}, \boldsymbol{y})$, a first-order Taylor expansion at $\boldsymbol{z}^t$ gives

$$\boldsymbol{g}^{t+1} = \boldsymbol{g}^t + \boldsymbol{H}_z^t(\boldsymbol{z}^{t+1} - \boldsymbol{z}^t) + \tfrac{1}{2} \, \nabla_{\boldsymbol{z}}^3 f(\boldsymbol{z}^t + \xi^t, \boldsymbol{y})\big[\Delta \boldsymbol{z}^t, \Delta \boldsymbol{z}^t\big], \quad \Delta \boldsymbol{z}^t = \boldsymbol{z}^{t+1} - \boldsymbol{z}^t.$$

By assumption $\|\nabla_{\boldsymbol{z}}^3 f\| \le L$ and $\|\Delta \boldsymbol{z}^t\| = O(\eta)$ from the previous step, hence the remainder has norm $\le C_1 \, \eta^2$. Substituting the logits expansion from step (ii) yields

$$\boldsymbol{g}^{t+1} = \Big(\boldsymbol{I} - \eta \, \mu \, \boldsymbol{H}_z^t - \eta \, \mu \, \tilde{\rho}^{\,t} \, (\boldsymbol{H}_z^t)^2 \Big) \boldsymbol{g}^t + \boldsymbol{r}_{\boldsymbol{g}}^t, \qquad \|\boldsymbol{r}_{\boldsymbol{g}}^t\| \le C \, \eta^2,$$

after absorbing constants into C. This proves the logit gradient statement.

Combining (i)–(iii) completes the proof, with a constant C depending only on (L, μ, κ), and the bounds holding for all $|\eta| \in (0, 1]$ and $|\rho| \le \kappa \sqrt{|\eta|}$.

For softmax cross-entropy,

$$\nabla_{\boldsymbol{z}} f(\boldsymbol{z}, \boldsymbol{y}) = \boldsymbol{p}(\boldsymbol{z}) - \boldsymbol{y}, \qquad \boldsymbol{H}_z(\boldsymbol{z}) = \nabla_{\boldsymbol{z}}^2 f(\boldsymbol{z}, \boldsymbol{y}) = \mathrm{Diag}(\boldsymbol{p}(\boldsymbol{z})) - \boldsymbol{p}(\boldsymbol{z}) \boldsymbol{p}(\boldsymbol{z})^\top.$$

Since $\boldsymbol{p}(\boldsymbol{z}) \in \Delta^{V-1} \subset [0,1]^V$ for all $\boldsymbol{z}$, every entry of the third derivative tensor $\nabla_{\boldsymbol{z}}^3 f(\boldsymbol{z}, \boldsymbol{y})$ is a bounded polynomial in $\boldsymbol{p}(\boldsymbol{z})$ (hence in $[0,1]$). Therefore there exists a finite constant $L_{\text{sm}}(V)$ depending only on V such that

$$\sup_{\boldsymbol{z}} \big\| \nabla_{\boldsymbol{z}}^3 f(\boldsymbol{z}, \boldsymbol{y}) \big\| \le L_{\text{sm}}(V).$$

In particular, f is C^∞ and Assumption (1) of the theorem holds with $L = L_{\text{sm}}(V)$. □

Proposition A.3. *$\boldsymbol{H_z}$ is symmetric positive semidefinite with $\ker(\boldsymbol{H_z}) = \operatorname{span}\{\mathbf{1}\}$ and $\operatorname{rank}(\boldsymbol{H_z}) = V - 1$. Moreover, for the residual $\boldsymbol{g}$ we have $\mathbf{1}^\top \boldsymbol{g} = 0$, hence $\boldsymbol{g} \in \mathbf{1}^\perp = \operatorname{range}(\boldsymbol{H_z})$; in particular, given any eigenbasis of $\boldsymbol{H_z}$ restricted to $\mathbf{1}^\perp$, $\boldsymbol{g}$ admits a unique coordinate representation in that basis.*

Proof. Let $\boldsymbol{p} = \operatorname{softmax}(\boldsymbol{z}) \in (0,1)^V$ so that $\mathbf{1}^\top \boldsymbol{p} = 1$, and recall

$$\boldsymbol{H_z} = \operatorname{Diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top.$$

For any $\boldsymbol{v} \in \mathbb{R}^V$,

$$\boldsymbol{v}^\top \boldsymbol{H_z}\, \boldsymbol{v} = \sum_{i=1}^{V} p_i v_i^2 - \Big(\sum_{i=1}^{V} p_i v_i\Big)^2 = \operatorname{Var}_{\boldsymbol{p}}(v) \;\geq 0,$$

hence $\boldsymbol{H_z}$ is symmetric positive semidefinite. Moreover, $\boldsymbol{v}^\top \boldsymbol{H_z}\, \boldsymbol{v} = 0$ iff $\operatorname{Var}_{\boldsymbol{p}}(v) = 0$, i.e., v_i is constant across i. Since $p_i > 0$ for all i, this means $\boldsymbol{v} = c\,\mathbf{1}$, thus

$$\ker(\boldsymbol{H_z}) = \operatorname{span}\{\mathbf{1}\} \quad \Rightarrow \quad \operatorname{rank}(\boldsymbol{H_z}) = V - \dim\ker(\boldsymbol{H_z}) = V - 1.$$

Then $\mathbf{1}^\top \boldsymbol{g} = \mathbf{1}^\top \boldsymbol{p} - \mathbf{1}^\top \boldsymbol{y} = 0$, so $\boldsymbol{g} \in \mathbf{1}^\perp$. For any symmetric matrix, $\operatorname{range}(\boldsymbol{H_z}) = (\ker(\boldsymbol{H_z}))^\perp$; using $\ker(\boldsymbol{H_z}) = \operatorname{span}\{\mathbf{1}\}$ yields $\mathbf{1}^\perp = \operatorname{range}(\boldsymbol{H_z})$, hence $\boldsymbol{g} \in \operatorname{range}(\boldsymbol{H_z})$.

Restrict $\boldsymbol{H_z}$ to the invariant subspace $\mathbf{1}^\perp$. Being symmetric, $\boldsymbol{H_z}|_{\mathbf{1}^\perp}$ admits an orthonormal eigenbasis $\{\boldsymbol{v}_k\}_{k=1}^{V-1}$ associated with its positive eigenvalues. Since $\boldsymbol{g} \in \mathbf{1}^\perp$, it has the unique expansion $\boldsymbol{g} = \sum_{k=1}^{V-1} e_k\, \boldsymbol{v}_k$, with $e_k = (\boldsymbol{v}_k)^\top \boldsymbol{g}$. □

Corollary A.4 (Modal dynamics in the eigenbasis of $\boldsymbol{H_z^t}$)**.** *Under the same assumptions as Theorem A.2. For each t, let the spectral decomposition of the symmetric positive–semidefinite matrix $\boldsymbol{H}_z^t$ be*

$$\boldsymbol{H}_z^t \;=\; \sum_{k=1}^{V-1} \lambda_k^t\, \boldsymbol{v}_k^t (\boldsymbol{v}_k^t)^\top,$$

where $\lambda_k^t > 0$, $(\boldsymbol{v}_k^t)^\top \boldsymbol{v}_\ell^t = \delta_{k\ell}$ are the non-zero eigenvalues and eigenvectors. Define the modal coefficients *of the residual $\boldsymbol{g}^t = \boldsymbol{p}^t - \boldsymbol{y}$ by*

$$e_k^t \;\coloneqq\; (\boldsymbol{v}_k^t)^\top \boldsymbol{g}^t, \qquad k = 1, \ldots, V-1.$$

Then there exists a constant $C > 0$ such that for all nonzero modes $k \geq 1$,

$$(\boldsymbol{v}_k^t)^\top \boldsymbol{g}^{t+1} = \Big(1 \;-\; \eta\,\mu\,\big[\lambda_k^t + \tilde{\rho}^{\,t}(\lambda_k^t)^2\big]\Big)\, e_k^t \;+\; r_k^t, \qquad |r_k^t| \;\leq\; C\,\eta^2. \tag{9}$$

Proof. By Theorem A.2 (residual expansion), we have

$$\boldsymbol{g}^{t+1} = \Big(\boldsymbol{I} - \eta\,\mu\,\boldsymbol{H}_z^t - \eta\,\mu\,\tilde{\rho}^{\,t}\,(\boldsymbol{H}_z^t)^2\Big)\boldsymbol{g}^t + \boldsymbol{r}_{\boldsymbol{g}}^t, \qquad \|\boldsymbol{r}_{\boldsymbol{g}}^t\| \leq C\,\eta^2. \tag{10}$$

Fix t and let the eigendecomposition of $\boldsymbol{H}_z^t$ be $\boldsymbol{H}_z^t = \sum_{k=1}^{V-1} \lambda_k^t\, \boldsymbol{v}_k^t (\boldsymbol{v}_k^t)^\top$ with $\lambda_k^t > 0$ and $\{\boldsymbol{v}_k^t\}_{k=1}^{V-1}$ orthonormal. (The zero mode corresponding to $\lambda = 0$ is orthogonal to $\boldsymbol{g}^t$ in the softmax–CE case and is therefore omitted.)

Project equation 10 onto the eigenvector $\boldsymbol{v}_k^t$:

$$(\boldsymbol{v}_k^t)^\top \boldsymbol{g}^{t+1} = (\boldsymbol{v}_k^t)^\top \Big(\boldsymbol{I} - \eta\,\mu\,\boldsymbol{H}_z^t - \eta\,\mu\,\tilde{\rho}^{\,t}\,(\boldsymbol{H}_z^t)^2\Big)\boldsymbol{g}^t \;+\; (\boldsymbol{v}_k^t)^\top \boldsymbol{r}_{\boldsymbol{g}}^t.$$

Using the eigen-relations $\boldsymbol{H}_z^t \boldsymbol{v}_k^t = \lambda_k^t \boldsymbol{v}_k^t$ and $(\boldsymbol{H}_z^t)^2 \boldsymbol{v}_k^t = (\lambda_k^t)^2 \boldsymbol{v}_k^t$ and the definition $e_k^t = (\boldsymbol{v}_k^t)^\top \boldsymbol{g}^t$, we obtain

$$(\boldsymbol{v}_k^t)^\top \boldsymbol{g}^{t+1} = \Big(1 - \eta\,\mu\,\lambda_k^t - \eta\,\mu\,\tilde{\rho}^{\,t}(\lambda_k^t)^2\Big) e_k^t \;+\; r_k^t, \qquad r_k^t \coloneqq (\boldsymbol{v}_k^t)^\top \boldsymbol{r}_{\boldsymbol{g}}^t.$$

Finally, since $\|\boldsymbol{v}_k^t\| = 1$ we have $|r_k^t| \leq \|\boldsymbol{r}_{\boldsymbol{g}}^t\| \leq C\,\eta^2$, which is exactly equation 9. This completes the proof. □

Corollary A.5 (One–step confidence ratios under SAM)**.** *Under the assumptions of Theorem A.2. Fix an iteration t and write $\boldsymbol{p}^t = \mathrm{softmax}(\boldsymbol{z}^t)$, $\boldsymbol{g}^t = \boldsymbol{p}^t - \boldsymbol{e}_y$, and $\boldsymbol{H}_z^t = \mathrm{diag}(\boldsymbol{p}^t) - \boldsymbol{p}^t(\boldsymbol{p}^t)^\top$. For each class $i \in [V]$, define the one–step confidence ratio*

$$\alpha_i^\bullet \;:=\; \frac{p_i^{t+1}(\bullet)}{p_i^t}, \qquad \bullet \in \{\mathrm{GD}, \mathrm{SAM}\}.$$

Then $\alpha_i^\bullet$ admits the representation

$$\alpha_i^\bullet = \frac{\sum_{j=1}^V e^{z_j^t}}{\sum_{j=1}^V \beta_j^\bullet\, e^{z_j^t}}, \qquad \beta_j^{\mathrm{GD}} \;=\; \exp\!\big\{-\eta'\big[(p_j^t - y_j) - (p_i^t - y_i)\big]\big\},$$

and the SAM correction appears multiplicatively as

$$\beta_j^{\mathrm{SAM}} \;=\; \beta_j^{\mathrm{GD}}\, \exp\!\Big\{ -\eta'\tilde\rho^{\,t}\big[(\boldsymbol{H}_z^t\boldsymbol{g}^t)_j - (\boldsymbol{H}_z^t\boldsymbol{g}^t)_i\big]\Big\},$$

where $\eta' = \eta\,\mu$ and, when $\|\boldsymbol{g}^t\| > 0$, $\tilde\rho^{\,t} = \rho\,\sqrt{\mu}/\|\boldsymbol{g}^t\|$ (otherwise $\tilde\rho^{\,t} = 0$ by convention).

Let y be the ground–truth label and $y^* = \arg\max_{j\neq y} p_j^t$ the most confident incorrect class. Assume the sign condition $\eta'\tilde\rho^{\,t} > 0$ and the radius scaling $|\rho| \le \kappa\sqrt{|\eta|}$. Then there exists $\eta_0 = \eta_0(\boldsymbol{p}^t, \boldsymbol{H}_z^t, \|\boldsymbol{g}^t\|, \mu, \kappa, L) > 0$ such that, for all step sizes $0 < |\eta| \le \eta_0$, the following one–step inequalities hold without remainder terms*:*

$$\alpha_{y^*}^{\mathrm{SAM}} \;\le\; \alpha_{y^*}^{\mathrm{GD}}, \qquad \alpha_y^{\mathrm{SAM}} \;\ge\; \alpha_y^{\mathrm{GD}}.$$

Here $y \in \{y^+, y^-\}$ denotes the ground–truth label corresponding to the positive or negative learning rate, respectively. Moreover, the inequalities are strict whenever $p_{y^*}^t \in (0,1)$ and $p_y^t \le \frac{1}{2}$. In particular, when $\tilde\rho^{\,t} = 0$ (no SAM), the two equalities hold.

Proof. Fix t and a class $i \in [V]$. Set $\eta' = \eta\,\mu$. By Theorem A.2 (logits line),

$$\Delta\boldsymbol{z} \;:=\; \boldsymbol{z}^{t+1} - \boldsymbol{z}^t = -\,\eta'\big(\boldsymbol{g}^t + \tilde\rho^{\,t}\boldsymbol{H}_z^t\boldsymbol{g}^t\big) + \boldsymbol{r}_{\boldsymbol{z}}^t, \qquad \|\boldsymbol{r}_{\boldsymbol{z}}^t\|_\infty \le C_1\,\eta^2,$$

where C_1 depends only on (L, μ, κ) and the hypothesis $|\rho| \le \kappa\sqrt{|\eta|}$ is in force.

For any increment $\Delta\boldsymbol{z}$,

$$\alpha_i = \frac{p_i(\boldsymbol{z}^t + \Delta\boldsymbol{z})}{p_i(\boldsymbol{z}^t)} = \frac{\sum_j e^{z_j^t}}{\sum_j \exp\{\Delta z_j - \Delta z_i\} e^{z_j^t}} = \frac{\sum_j e^{z_j^t}}{\sum_j \beta_j e^{z_j^t}}.$$

With the above $\Delta\boldsymbol{z}$,

$$\beta_j^{\mathrm{SAM}} = \underbrace{\exp\{-\eta'(g_j^t - g_i^t)\}}_{\beta_j^{\mathrm{GD}}}\ \underbrace{\exp\big\{-\eta'\tilde\rho^{\,t}\Delta_{j,i}^t\big\}}_{\text{curvature factor}}\ \underbrace{\exp\{r_j^t - r_i^t\}}_{\text{remainder factor}}, \quad \Delta_{j,i}^t := (\boldsymbol{H}_z^t\boldsymbol{g}^t)_j - (\boldsymbol{H}_z^t\boldsymbol{g}^t)_i.$$

From $\|\boldsymbol{r}_{\boldsymbol{z}}^t\|_\infty \le C_1\eta^2$, we have $e^{-2C_1\eta^2} \le \exp\{r_j^t - r_i^t\} \le e^{2C_1\eta^2}$ for all i, j.

With $\boldsymbol{H}_z^t = \mathrm{diag}(\boldsymbol{p}^t) - \boldsymbol{p}^t(\boldsymbol{p}^t)^\top$ and $\boldsymbol{g}^t = \boldsymbol{p}^t - \boldsymbol{e}_y$,

$$(\boldsymbol{H}_z^t\boldsymbol{g}^t)_i = p_i^t\,(p_i^t - y_i - C^t), \qquad C^t := \sum_k (p_k^t)^2 - p_y^t.$$

Let $y^* = \arg\max_{j\neq y} p_j^t$. Then $C^t \le p_{y^*}^t$ and one checks: (i) for $i = y^*$, $\Delta_{j,y^*}^t \le 0$ for all j, and $\Delta_{j,y^*}^t < 0$ for some j whenever $p_{y^*}^t \in (0,1)$; (ii) for $i = y$, $\Delta_{j,y}^t \ge 0$ for all j whenever $p_y^t \le \frac{1}{2}$, and $\Delta_{j,y}^t > 0$ for some j if $p_y^t \in (0, \frac{1}{2}]$.

Define

$$D_i^{\mathrm{GD}} := \sum_j \beta_j^{\mathrm{GD}} e^{z_j^t}, \qquad \widetilde{D}_i := \sum_j \beta_j^{\mathrm{GD}} e^{z_j^t}\exp\{-\eta'\tilde\rho^{\,t}\Delta_{j,i}^t\}, \qquad D_i^{\mathrm{SAM}} := \sum_j \beta_j^{\mathrm{SAM}} e^{z_j^t}.$$

By the remainder bounds, $e^{-2C_1\eta^2}\widetilde{D}_i \le D_i^{\mathrm{SAM}} \le e^{2C_1\eta^2}\widetilde{D}_i$. Next, by $e^x \ge 1+x$ and the sign structure of $\Delta_{j,i}^t$,

$$\frac{\widetilde{D}_{y^*}}{D_{y^*}^{\mathrm{GD}}} = \sum_j w_j^{(y^*)} e^{-\eta'\tilde{\rho}^t\Delta_{j,y^*}^t} \ge 1+\eta'\tilde{\rho}^t\sum_j w_j^{(y^*)}(-\Delta_{j,y^*}^t) \ge 1+c_{y^*}\eta'\tilde{\rho}^t,$$

for some $c_{y^*} > 0$ whenever $p_{y^*}^t \in (0,1)$; here $w_j^{(i)} := \beta_j^{\mathrm{GD}} e^{z_j^t}/D_i^{\mathrm{GD}}$ are positive weights. Similarly, for $p_y^t \le \frac{1}{2}$,

$$\frac{\widetilde{D}_y}{D_y^{\mathrm{GD}}} = \sum_j w_j^{(y)} e^{-\eta'\tilde{\rho}^t\Delta_{j,y}^t} \le 1-\eta'\tilde{\rho}^t\sum_j w_j^{(y)}\Delta_{j,y}^t \le 1-c_y\eta'\tilde{\rho}^t$$

for some $c_y > 0$ (strict in the stated nondegenerate case).

Now use the scaling $|\rho| \le \kappa\sqrt{|\eta|}$: then $\eta'\tilde{\rho}^t = \Theta(\eta^{3/2})$, whereas $e^{\pm 2C_1\eta^2} = 1 \pm O(\eta^2)$. Hence there exists $\eta_0 > 0$ (depending only on $(\boldsymbol{p}^t, \boldsymbol{H}_z^t, \|\boldsymbol{g}^t\|, \mu, \kappa, L)$) such that for $0 < |\eta| \le \eta_0$,

$$D_{y^*}^{\mathrm{SAM}} \ge e^{-2C_1\eta^2}\widetilde{D}_{y^*} \ge D_{y^*}^{\mathrm{GD}}\big(1+\tfrac{1}{2}c_{y^*}\eta'\tilde{\rho}^t\big), \qquad D_y^{\mathrm{SAM}} \le e^{2C_1\eta^2}\widetilde{D}_y \le D_y^{\mathrm{GD}}\big(1-\tfrac{1}{2}c_y\eta'\tilde{\rho}^t\big).$$

Since $\alpha_i = \big(\sum_j e^{z_j^t}\big)/D_i$, we obtain for $0 < |\eta| \le \eta_0$:

$$\alpha_{y^*}^{\mathrm{SAM}} \le \alpha_{y^*}^{\mathrm{GD}}, \qquad \alpha_y^{\mathrm{SAM}} \ge \alpha_y^{\mathrm{GD}},$$

with strict inequalities under the stated nondegeneracy conditions (because then $c_{y^*}, c_y > 0$). If $\|\boldsymbol{g}^t\| = 0$ (so $\tilde{\rho}^t = 0$ by convention), both become equalities. This completes the proof. □

B Implementation

Algorithm 1 Logits-SAM pseudocode

Require: model, batch, ρ

1: Let $W \leftarrow$ `lm_head.weight`
2: Run forward to get `loss_pre` and hidden states H
3: $g \leftarrow \mathrm{grad}(\texttt{loss\_pre},\ W)$
4: $e \leftarrow \rho\, g/\|g\|_2$
5: `logits_perturbed` $\leftarrow \mathrm{linear}(H,\ W+e)$
6: Compute `loss_post` with `logits_perturbed`
7: Backward `loss_post`

Table 4: Comparison between theoretical and practical settings of DPO with SAM. Although the signs differ for y^-, the resulting dynamics are equivalent. For y^+, the settings coincide.

Class	Objective	Learning rate	ρ	Setting
y^+	Positive objective $f = -\log p$	Positive ($\eta > 0$)	Positive	Theory = Practice
y^- (Theory)	Positive objective $f = -\log p$	Negative ($\eta < 0$)	Negative	Theory
y^- (Practice)	Negative objective $f = \log p$	Positive ($\eta > 0$)	Positive	Practice

C Additional experimental details

Benchmark details. **AlpacaEval 2** (Dubois et al., 2024) is a large-scale preference benchmark for open-ended instruction following that uses LLM-as-a-judge calibrated to human preferences; its evaluation set contains 805 single-turn instructions, and models are typically compared in pairwise settings against a baseline. **Arena-Hard v0.1** (Li et al., 2024) is a challenging subset of difficult user instructions mined from Chatbot Arena; it enables fine-grained, head-to-head comparisons between models via pairwise judging and comprises 500 hard prompts. **MT-Bench** (Zheng et al., 2023) is a multi-turn dialogue benchmark that tests a model's ability to handle diverse conversational tasks across several categories; the standard evaluation set consists of 80 multi-turn questions.

Method	Objective	Hyperparameter
SLiC-HF	$\max\big(0,\ \delta - \log \pi_\theta(y_w \mid x) + \log \pi_\theta(y_l \mid x)\big) - \lambda \log \pi_\theta(y_w \mid x)$	$\lambda \in \{0.1, 0.5, 1.0, 10.0\}$; $\delta \in \{0.5, 1.0, 2.0, 10.0\}$
CPO	$-\log \sigma\big(\beta \log \pi_\theta(y_w \mid x) - \beta \log \pi_\theta(y_l \mid x)\big) - \lambda \log \pi_\theta(y_w \mid x)$	$\lambda = 1.0$; $\beta \in \{0.01, 0.05, 0.1\}$

Table 5: Objectives and hyperparameters for SLiC-HF and CPO.

Method	Pythia-2.8B	Mistral-7B
DPO	1×10^{-4}	1×10^{-5}
SLiC-HF	1×10^{-3}	1×10^{-4}
CPO	1×10^{-4}	1×10^{-5}

Table 6: Choice of ρ for logits-SAM.

Training details. For experiments on Pythia-2.8B, we use two NVIDIA A100 GPUs with data-parallel training under DDP; for Mistral-7B, we use four NVIDIA A100 GPUs with parallel training via DeepSpeed ZeRO-3 (Rasley et al., 2020).

D LLM USAGE STATEMENT

In preparing this manuscript, we employed a large language model (LLM) as an auxiliary tool. Specifically, the LLM was used to assist with proofreading, formatting, and grammar checking of the text.