SAME OBJECTIVE, DIFFERENT OPTIMIZER: PREFERENCE-FLIP NOISE VERSUS LABEL SMOOTHING IN DPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Flipping preference labels with probability p and smoothing them with $\varepsilon = p$ (conservative DPO, cDPO) give the same expected DPO loss but not the same optimizer. We quantify the difference of one update from a given state and show that its order in the step size depends on the update rule. The flip noise of example i is exactly $(Z_i - p)J_i$, with J_i the margin Jacobian, whether or not the margin is saturated. For updates affine in the gradient, such as SGD, momentum and Adam with a frozen second moment, the flip-induced change of a smooth objective is second order, $\frac{\eta^2 p(1-p)}{2m^2} \sum_i (CJ_i)^\top \nabla^2 F\,(CJ_i)$; a live adaptive second moment or an active global-norm clip can bias the expected update, giving a first-order objective difference when $\nabla F^\top \beta_B \neq 0$. Exact same-state measurements on Qwen3-0.6B, enumerating all flip patterns, confirm both orders: measured-to-predicted ratios are 0.99–1.06 for SGD, and the first-order error is 0.6% for live Adam at $\eta = 2.5 \times 10^{-6}$. A crossover step size $\eta^* = |a_1|/|A_2|$, computed before any measurement, predicts whether the first-order live or the second-order frozen-denominator excess is larger at twelve unseen states of Qwen3-0.6B and Qwen3-1.7B in 599 of 607 and 582 of 598 classified cases (95% intervals over state–minibatch groups $[97.1, 99.8]\%$ and $[95.6, 98.8]\%$), and at twelve further states built from an independent data panel in 582 of 617 and 590 of 604. An unbiased stratified estimate of the first-order coefficient replaces the enumeration, with a nominal 95% interval whose coverage over 1,000 repetitions of 360 cases is 95.1% and 94.8%. In paired multi-epoch LoRA runs of Qwen3.5-2B and 4B, the flip-to-cDPO gradient-norm ratio starts near the value 1.25 that the covariance formula gives at zero margin for orthogonal Jacobians, with median 1.23, and rises to 4.7 by the third epoch. A loss audit finds that the smoothing option leaves TRL's native sigmoid loss unchanged in the newest release it runs; a code scan finds the smoothing mixture absent from that branch from one release on.

1 INTRODUCTION

Preference optimization methods often differ by changes that look small at the objective level but matter in the training run. Direct Preference Optimization (DPO) replaces reward-model fitting with a supervised preference loss over chosen and rejected responses (Rafailov et al., 2023). Two common ways to account for noisy preferences are to soften the label, as in conservative DPO (Mitchell, 2023) and robust variants (Chowdhury et al., 2024), or to inject label flips during training. With a fixed flip probability p and smoothing $\varepsilon = p$, the expected losses coincide, and the two constructions are frequently used interchangeably in ablations, curricula and robustness studies.

What remains different once the expected objective is matched? The identity is the flip expectation of the cDPO mixture (Mitchell, 2023), and label smoothing equals symmetric label noise in expectation (Lukasik et al., 2020). DisturbLabel reports that per-iteration random labels and their soft mean train differently despite equal expected gradients (Xie et al., 2016). Label-noise SGD has a noise covariance along the output gradients with a $p(1-p)$ factor for the logistic loss, which yields a long-horizon flat-minimum bias for SGD and momentum near zero-loss solutions (Damian et al., 2021), and a different sharpness bias for Adam when $1 - \beta_2 = \Theta(\eta^2)$ (Li et al., 2025); clipped

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

stochastic gradients are biased (Koloskova et al., 2023). These results do not give the size of the flip-versus-smoothing difference in one update from a given state of a preference model, its dependence on the update rule used to fine-tune language models, or on whether flips are redrawn or reused; these are the quantities that an objective-matched ablation changes.

Our contributions are as follows.

- **Covariance, Lemma 2.** Switching the label of example i changes its gradient by exactly $J_i = \nabla_\theta s_\theta(x_i)$ at every θ, so the flip noise is exactly $(Z_i - p)J_i$ at every state; we use this covariance for one-step expansions from any state, whereas label-noise SGD theory (Damian et al., 2021) uses it for the long-horizon bias near zero-loss solutions. Its coefficient does not shrink as margins saturate: at a common state $\mathbb{E}\|g_{\mathrm{flip}}\|^2 = \|g_{\mathrm{cDPO}}\|^2 + \mathrm{tr}\Sigma_B$, where $\mathrm{tr}\Sigma_B$ is free of the margins, so at fixed Jacobians the relative excess grows as $\|g_{\mathrm{cDPO}}\|$ decreases; the gradient norms of every paired run are consistent with this.
- **Affine updates: second order, Theorem 3.** For SGD, momentum and Adam with a frozen denominator, the one-step excess is $O(\eta^2)$ from any state, with a leading η^2 coefficient that is exactly proportional to $p(1-p)$ at fixed F and computable from local quantities.
- **Adaptive denominators and clipping: expected-update bias, Proposition 4 and Corollary 5.** Because the current gradient enters Adam's second moment, flips bias the expected step itself, by $\frac{u_k''}{2}p(1-p)m^{-2}\sum_i J_{ik}^2$ per coordinate to leading order, in closed form, as the flip perturbation has known variance, whereas the variance adaptation of Balles & Hennig (2018) uses the estimated gradient variance. An active global-norm clip shortens the expected flip step along the cDPO direction by a term proportional to the flip covariance over $\|g_{\mathrm{cDPO}}\|^2$. Either bias gives a first-order excess $-\eta\nabla F^\top \beta_B$ when $\nabla F^\top \beta_B \neq 0$.
- **Reused flips, Proposition 6.** In a frozen-gradient SGD model, reusing one draw for E epochs multiplies the accumulated flip variance by E, so corrupted-dataset pipelines and per-visit resampling are not interchangeable controls.
- **A predictive crossover.** Together, the two expansions predict from the state alone which excess is larger at a given step size, $\eta < \eta^* = |a_1|/|A_2|$ or not, where $a_1 = \nabla F^\top \beta_B$ and A_2 is the η^2 coefficient of Eq. (4) with the frozen denominator; we test this prospectively at unseen states, together with the live map's own crossover $\eta^\dagger = |a_1|/|B_2|$ and an unbiased interval estimate of a_1 that avoids enumerating flip patterns.

Empirically, exact same-state measurements with registered criteria confirm Theorem 3, its dependence on the margin-Jacobian directions and the first-order bias of live Adam at one state in Section 4.1; in Section 4.2, predictions written before measurement at 24 unseen states of two models on two data panels identify the larger excess, the size of the live one and the effect of freezing the denominator and of an active clip; in Section 4.3 a sampled estimate replaces the enumeration; in Section 4.4 paired multi-epoch runs show a gradient-norm signature consistent with Lemma 2; and in Section 4.5 a loss audit finds a trainer's smoothing option inactive in its native sigmoid loss from one release on. The contribution is a local optimization mechanism and a diagnostic computed from it, evaluated by same-state measurements rather than by downstream task scores.

2 Setup

Let $x = (q, y^+, y^-)$ denote a prompt with a preferred and a dispreferred response. For a policy π_θ and a fixed reference policy π_0, the preference margin is

$$s_\theta(x) = \beta\Big[\log\frac{\pi_\theta(y^+ \mid q)}{\pi_0(y^+ \mid q)} - \log\frac{\pi_\theta(y^- \mid q)}{\pi_0(y^- \mid q)}\Big], \tag{1}$$

where $\beta > 0$ is the DPO inverse temperature. The clean and flipped per-example losses are $\ell_+(s) = -\log\sigma(s)$ and $\ell_-(s) = -\log\sigma(-s)$. For a flip probability $p \in [0,1]$, cDPO minimizes

$$L_{\mathrm{cDPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[(1-p)\ell_+(s_\theta(x)) + p\ell_-(s_\theta(x))\right]. \tag{2}$$

Flip-DPO draws independent $Z_i \sim \mathrm{Bernoulli}(p)$ for the examples of a minibatch $B = \{x_i\}_{i=1}^m$ and uses ℓ_- when $Z_i = 1$ and ℓ_+ otherwise. Write g_{flip} and g_{cDPO} for the two minibatch gradients on the same B, $J_i = \nabla_\theta s_\theta(x_i)$, and $\sigma_i = \sigma(s_\theta(x_i))$.

Proposition 1 (Objective identity). *For fixed θ, reference policy, data distribution and p, $\mathbb{E}[\ell_Z(s_\theta(x))] = L_{\mathrm{cDPO}}(\theta)$, where $\ell_Z = (1-Z)\ell_+ + Z\ell_-$.*

Proposition 1 follows from the cDPO mixture of Mitchell (2023) by taking the flip expectation; it fixes the object that is matched. Everything that follows concerns what the identity leaves unmatched.

3 THEORY

Lemma 2 (Excess conditional covariance). *Conditioned on B, θ and all other sources of randomness such as dropout, $\mathbb{E}[g_{\mathrm{flip}} \mid B] = g_{\mathrm{cDPO}}$ and*

$$g_{\mathrm{flip}} - g_{\mathrm{cDPO}} = \frac{1}{m}\sum_{i=1}^{m}(Z_i - p)\, J_i, \qquad \mathrm{Cov}(g_{\mathrm{flip}} \mid B) = \frac{p(1-p)}{m^2}\sum_{i=1}^{m} J_i J_i^\top, \tag{3}$$

while g_{cDPO} is deterministic given B.

Proof. From $\nabla\ell_+(s) = -(1-\sigma(s))\nabla s$ and $\nabla\ell_-(s) = \sigma(s)\nabla s$, the flip-DPO gradient of example i is $c_i(Z_i)J_i$ with $c_i(z) = -(1-\sigma_i)(1-z) + \sigma_i z$, which is affine in z with slope $c_i(1) - c_i(0) = 1$. Hence $c_i(Z_i) - c_i(p) = Z_i - p$, and the cDPO gradient is $c_i(p)J_i$. Averaging over i and using independence of the Z_i gives Eq. (3). □

Two features distinguish this from generic minibatch noise. First, the perturbation of each example is $(Z_i - p)J_i$, with a coefficient independent of σ_i: saturated margins do not suppress the flip coefficient, although J_i itself can change along training, whereas the cDPO coefficient $c_i(p) = \sigma_i - (1-p)$ vanishes at the finite cDPO target $\sigma_i = 1 - p$. Second, the covariance is rank one per example and lives in the span of the margin Jacobians, so its effect on an update is determined by how the update rule maps those directions.

Update rules. We consider one step $\theta' = \theta - \eta\, U(g)$ from a fixed optimizer state. An update is *affine* if $U(g) = b + Cg$ for a vector b and matrix C that do not depend on the current gradient. SGD has $b = 0, C = I$; heavy-ball momentum and Adam with a frozen denominator D have b given by the momentum buffer and $C = c\, D^{-1}$ with c the bias-corrected first-moment weight $(1-\beta_1)/(1-\beta_1^t)$. Adam and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) are not affine: the current gradient enters the second moment $v \leftarrow \beta_2 v + (1-\beta_2) g \odot g$, and $U_k(g) = \hat{m}_k(g_k)/(\sqrt{\hat{v}_k(g_k)} + \epsilon)$ is a nonlinear scalar function of g_k for each coordinate k. Decoupled weight decay does not depend on g and cancels between the two procedures.

For a smooth evaluation objective F, such as a held-out DPO loss or the local cDPO risk, held *fixed* while p varies, define the one-step excess

$$\Delta_B(\eta, p) = \mathbb{E}\big[F(\theta - \eta U(g_{\mathrm{flip}})) \,\big|\, B\big] - F(\theta - \eta U(g_{\mathrm{cDPO}})).$$

Theorem 3 (Affine updates: second-order excess). *Let $U(g) = b + Cg$ and let F have a locally Lipschitz Hessian near θ. Then*

$$\Delta_B(\eta, p) = \frac{\eta^2 p(1-p)}{2m^2}\sum_{i=1}^{m}(CJ_i)^\top \nabla^2 F(\theta)\,(CJ_i) + O(\eta^3). \tag{4}$$

The coefficient is nonnegative when $\nabla^2 F(\theta)$ is positive semidefinite on the span of $\{CJ_i\}$, and Δ_B is of exact order η^2 when it is nonzero.

Proof. Let $\bar{\theta} = \theta - \eta(b + Cg_{\mathrm{cDPO}})$ and $\delta = g_{\mathrm{flip}} - g_{\mathrm{cDPO}}$. Then $\theta - \eta U(g_{\mathrm{flip}}) = \bar{\theta} - \eta C\delta$. A second-order expansion at $\bar{\theta}$ gives $F(\bar{\theta} - \eta C\delta) - F(\bar{\theta}) = -\eta\nabla F(\bar{\theta})^\top C\delta + \frac{\eta^2}{2}\delta^\top C^\top \nabla^2 F(\bar{\theta}) C\delta + O(\eta^3\|\delta\|^3)$. The linear term has zero conditional mean by Lemma 2; $\mathbb{E}[\delta\delta^\top \mid B]$ is given by Eq. (3); and $\nabla^2 F(\bar{\theta}) = \nabla^2 F(\theta) + O(\eta)$. The support of δ is finite, so the remainder is uniformly $O(\eta^3)$. □

For SGD, $C = I$ recovers the curvature-weighted Jacobian form; for a fixed diagonal preconditioner the relevant geometry is $P^\top \nabla^2 F P$, so preconditioning re-weights which Jacobian directions matter.

At fixed (θ, B, F), the only p-dependence is the Bernoulli factor, which peaks at $p = 1/2$. If F itself is chosen to depend on p, for example the cDPO risk at $\varepsilon = p$, its Hessian changes with p and the peak can move; Appendix F gives an explicit example with a peak at $p = 0.25$.

Proposition 4 (Adaptive second moments: expected-update bias)**.** *Let U be any update map that is bounded on the support of g_{flip}, and F smooth. Then*

$$\Delta_B(\eta, p) = -\eta\, \nabla F(\theta)^\top \beta_B + O(\eta^2), \qquad \beta_B = \mathbb{E}[U(g_{\mathrm{flip}}) \mid B] - U(g_{\mathrm{cDPO}}). \tag{5}$$

If U is the one-step Adam map from a fixed state (m, v, t), acting coordinatewise as $U_k(g) = u_k(g_k)$ with u_k three times differentiable, then β_B vanishes at $p \in \{0, 1\}$, can vanish at interior p, and, coordinatewise,

$$\beta_{B,k} = \frac{u_k''(g_{\mathrm{cDPO},k})}{2} \cdot \frac{p(1-p)}{m^2} \sum_{i=1}^{m} J_{ik}^2 \;+\; R_k, \qquad |R_k| \le \tfrac{1}{6} \sup |u_k'''|\, \mathbb{E}|\delta_k|^3. \tag{6}$$

If the denominator is frozen and no clip is active, u_k is affine, $\beta_B = 0$, and Theorem 3 applies with $C = c\, D^{-1}$.

Proof. For each realization, $F(\theta - \eta U) = F(\theta) - \eta \nabla F(\theta)^\top U + O(\eta^2 \|U\|^2)$; taking the conditional expectation and subtracting the cDPO value gives Eq. (5). Eq. (6) follows from a second-order Taylor expansion of u_k around $g_{\mathrm{cDPO},k}$ with Lagrange remainder, using $\mathbb{E}[\delta_k] = 0$ and $\mathbb{E}[\delta_k^2] = p(1-p)m^{-2} \sum_i J_{ik}^2$ from Lemma 2. At $p \in \{0, 1\}$ the flip gradient is deterministic and equals g_{cDPO}. With a frozen denominator u_k is affine, so the expectation of U equals U at the mean. □

Both terms of Eq. (6) carry a factor $p(1-p)$, since $\mathbb{E}|Z - p|^3 = p(1-p)[p^2 + (1-p)^2]$; the leading term dominates when the per-example perturbations $|J_{ik}|/m$ are small relative to the scale $\sqrt{\hat{v}_k}$ over which u_k bends. The flip term inflates the second moment where the margin Jacobians are large, which changes the expected step length there and F at first order. At a fresh state with $\epsilon \to 0$, Adam reduces to $\mathrm{sign}(g)$. When flips can reverse a coordinate's sign and its mean gradient is nonzero, its expected step is shorter than cDPO's, as a scalar example in Appendix F shows. The proposition gives a first-order Δ_B whenever $\nabla F^\top \beta_B \neq 0$, the $p(1-p)/m^2$ dependence of its leading coefficient and the intervention that removes it; its sign depends on ∇F and u_k'', and the coefficient can vanish: at a fresh state with $s = 0$ and $p = \frac{1}{2}$ the two flip gradients are symmetric and $\beta_B = 0$.

Corollary 5 (Global-norm clipping)**.** *Let $U(g) = c\, g/\|g\|$, a global-norm clip with $\|g\| > c$ at g_{cDPO} and on every segment from g_{cDPO} to a value of g_{flip}, so that these segments avoid the origin. Then, with $\hat{g} = g_{\mathrm{cDPO}}/\|g_{\mathrm{cDPO}}\|$, $P_\perp = I - \hat{g}\hat{g}^\top$ and $\Sigma = \mathrm{Cov}(g_{\mathrm{flip}} \mid B)$,*

$$\hat{g}^\top \beta_B = -\frac{c}{2} \frac{\mathrm{tr}(P_\perp \Sigma)}{\|g_{\mathrm{cDPO}}\|^2} + O\Big(\frac{\mathbb{E}\|\delta\|^3}{\|g_{\mathrm{cDPO}}\|^3}\Big). \tag{7}$$

When the flip perturbation is small relative to $\|g_{\mathrm{cDPO}}\|$, the clipped flip step is thus shorter along the cDPO direction, and by Eq. (5) this gives a first-order Δ_B whenever $\nabla F^\top \beta_B \neq 0$. At fixed Σ and fixed direction $\hat{g}$ the relative shortening grows as $\|g_{\mathrm{cDPO}}\|$ falls, within the range where the expansion holds; the proof is in Appendix F. A clip followed by an affine map keeps this bias, so freezing an Adam denominator does not remove it.

Reused versus resampled flips. Many noise-injection pipelines draw flips once and reuse the corrupted dataset for every epoch, whereas Lemma 2 assumes a fresh draw at each visit. The next result shows that this choice changes the accumulated effect by a factor equal to the number of epochs.

Proposition 6 (Temporal structure of flips)**.** *Run E epochs of SGD with step η and batch size m over n examples in the frozen-gradient model, in which J_i and σ_i are fixed along the run and curvature is not propagated between steps. Let θ_E^{flip} and θ_E^{cDPO} be the final iterates from the same initialization and example order. If flips are redrawn independently at every visit, $\mathrm{Cov}(\theta_E^{\mathrm{flip}} - \theta_E^{\mathrm{cDPO}}) = \eta^2 E\, \frac{p(1-p)}{m^2} \sum_{i=1}^n J_i J_i^\top$; if each example's flip is drawn once and reused, $\mathrm{Cov}(\theta_E^{\mathrm{flip}} - \theta_E^{\mathrm{cDPO}}) = \eta^2 E^2\, \frac{p(1-p)}{m^2} \sum_{i=1}^n J_i J_i^\top$. In both cases the mean deviation is zero, and the leading excess of a smooth F at the end is $\frac{1}{2}\mathrm{tr}\big(\nabla^2 F\, \mathrm{Cov}(\theta_E^{\mathrm{flip}} - \theta_E^{\mathrm{cDPO}})\big)$, E times larger for reused flips.*

Proof. In this model, the visit of example i in epoch e moves the flip iterate by $-\frac{\eta}{m}c_i(Z_i^{(e)})J_i$ and the cDPO iterate by $-\frac{\eta}{m}c_i(p)J_i$. By Lemma 2 the deviation is $-\frac{\eta}{m}\sum_{e=1}^{E}\sum_{i=1}^{n}(Z_i^{(e)}-p)J_i$. Independent draws across epochs add E variances; a reused draw gives $E(Z_i-p)$ and variance $E^2p(1-p)$. A second-order expansion of F around the cDPO iterate gives the final statement. □

Proposition 6 makes the reuse of flips a first-class experimental variable: in this model, at three epochs, a pipeline that reuses flips has three times the accumulated flip variance of one that resamples them, even though both match the cDPO objective at every fixed parameter; along a reused-label trajectory the state depends on the draw, so even this match is lost, as Appendix F shows.

4 Experiments

4.1 Same-state one-step test

The test measures $\Delta_B(\eta, p)$ exactly at a fixed state and compares it with Eq. (4) or Eq. (5), whose coefficients are computed only from local quantities at θ. We use Qwen3-0.6B (Yang et al., 2025) with UltraFeedback preference pairs, prompts of at most 64 tokens, responses truncated to 64 tokens and $\beta = 0.1$; the trainable parameters are the 62.9M parameters of the last four transformer blocks. The state θ is reached by five Adam steps of cDPO with $\varepsilon = 0.1$ on disjoint pairs, so margins are nonzero, 0.015–0.36 in magnitude on the test pairs, and F is the clean DPO loss on eight held-out pairs, held fixed across p. Because the flip-DPO update on a minibatch of size m takes one of 2^m values, $\mathbb{E}[F(\theta - \eta U(g_{\text{flip}})) \mid B]$ is computed exactly by enumerating all flip patterns, with no Monte Carlo error, and every $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ reuses the same evaluations with binomial weights. On three disjoint minibatches, SGD is measured with $m = 2$ and 4, and $m = 8$ on the first, and Adam with a frozen denominator with $m = 2$ and 4, at four step sizes spaced by factors of two, the largest changing F by 0.02 to first order; live and frozen Adam are also measured at the training learning rates $\eta \in \{2, 1, 0.5, 0.25\} \times 10^{-5}$.

Three experiments were registered in sequence; Appendix A gives their details and Tables 11 and 12 their endpoints. The first, E-ONESTEP, with prefix activations, rotary tables and normalization internals in single precision and finite-difference curvature, meets its criteria on every completed SGD and frozen-denominator cell of Table 2. The third, E-ONESTEP3, reported here, runs every operation at one state in double precision, verified by a precision check, and predicts each measurement from the same computation: exact quadratic forms $w^\top \nabla^2 F\, w$ by double backpropagation and the enumerated first-order term $-\eta \nabla F^\top \beta_B$; its criteria are all met, as Table 4 shows.

Results. Affine updates follow Theorem 3 quantitatively. For SGD, all 35 measured-to-predicted ratios at the smallest step lie in $[0.987, 1.057]$, with slopes 1.98–2.24 in η; for frozen-denominator Adam the ratios are 0.990–1.173, 26 of 30 within $\pm 10\%$ and the other four in one $m = 2$ cell, with slopes 1.77–2.13. At the training learning rates the two Adam variants separate by one order in η at the same state, as Figure 2a in Appendix A shows. With a live denominator, Δ_B has slope 0.93–1.18 and matches $-\eta \nabla F^\top \beta_B$ of Proposition 4: the median relative error is 5.1%, 2.5%, 1.3% and 0.63% at $\eta = 2, 1, 0.5, 0.25 \times 10^{-5}$, halving with η as an $O(\eta^2)$ remainder requires, as Figure 2b shows, and all 15 combinations of minibatch and p are within 7% at the smallest step. With the denominator frozen, the slope is 1.99–2.03 and the ratio to Theorem 3 is 0.998–1.010. The live-denominator excess is larger than the frozen one in all 15 combinations for $\eta \le 10^{-5}$, by median factors of 8.8, 18 and 37 at $\eta = 1, 0.5, 0.25 \times 10^{-5}$, and in 14 of 15 by 4.2 at 2×10^{-5}; its sign depends on the minibatch and p, as Proposition 4 allows. At the step sizes used for fine-tuning, the flip-versus-smoothing difference of an Adam update at this state is thus dominated by the expected-update bias of the live denominator, and freezing the denominator reduces it to the second-order term of Theorem 3. The direction of the flip noise, not only its size, sets the affine excess: a registered control, E-ISO9 in Appendix A, replaces each J_i in the SGD flip noise by a random direction of the same norm, keeping the mean update and $\mathrm{tr}\Sigma_B$; both excesses match Eq. (4), ratios 0.990–1.057 and 0.979–1.120 in 15/15 cases each, and the one along the Jacobians is larger by a median factor of 1.2×10^5.

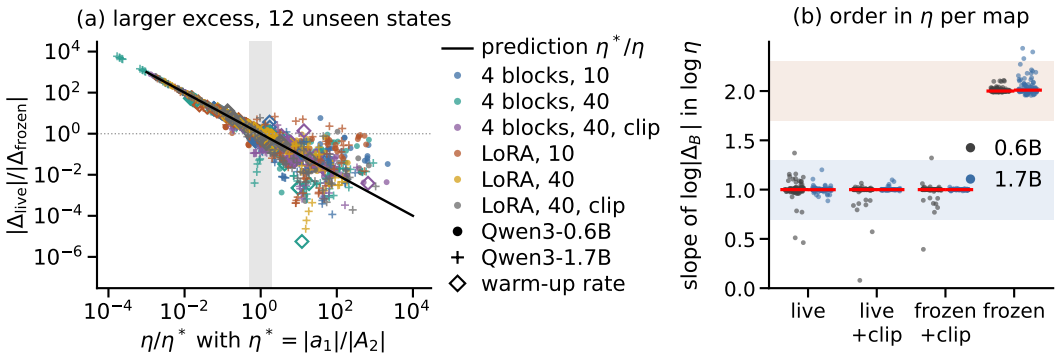


Figure 1: Prospective test at six unseen states of Qwen3-0.6B in E-ONESTEP4, shown as dots, and six of Qwen3-1.7B in E-ONESTEP4b, shown as crosses. (a) Measured $|\Delta_{\text{live}}|/|\Delta_{\text{frozen}}|$ against η/η^* for every state, minibatch, p and step; η^* was computed and recorded before measurement. Line: leading-order prediction η^*/η; grey band: $\eta/\eta^* \in [0.5, 2]$, excluded from the classification; diamonds: warm-up learning rate. (b) Log–log slope of $|\Delta_B|$ over the three smallest steps $\eta^*_{\text{ref}} 4^k$, $k = -3, -2, -1$, of the relative grid, without the added warm-up learning rate, for the four update maps; red: median; shaded: registered ranges for first and second order.

4.2 Prospective prediction at unseen states

For live Adam the leading excess is $-\eta a_1$ with $a_1 = \nabla F^\top \beta_B$ by Proposition 4; with the denominator frozen it is $A_2\eta^2$ by Theorem 3. Both coefficients are local, so the live excess is predicted to be the larger one exactly when $\eta < \eta^* = |a_1|/|A_2|$. E-ONESTEP4, detailed in Appendix B, tests this, and the interventions of Corollary 5, at six states of Qwen3-0.6B that no earlier experiment examined, with new F, warm-up and test pairs: the four-block suffix of 62.9M parameters or LoRA adapters of rank 64 on all its projections with 5.8M parameters, after 10 or 40 AdamW warm-up steps on cDPO with weight decay 0.01, one 40-step state of each kind with the training clip of 1.0 during warm-up. From each state four update maps are compared, all AdamW with decoupled weight decay: live or frozen denominator, each with or without a global-norm clip that is active on every flip pattern. A first phase computes, without evaluating F after any tested update, a_1 for every map by enumerating the 2^m update directions at $m = 4$, A_2 from exact Hessian quadratic forms, and η^* for every batch and p; it fixes the steps $\eta^*_{\text{ref}} 4^k$, $k = -3, \ldots, 3$, plus the warm-up learning rate, and its output file is recorded in the registration before a second phase measures Δ_B exactly in 72 cells over 3 minibatches and $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$.

Results. The measured ratio of the two excesses follows the predicted η^*/η where the live excess is predicted larger; elsewhere it is below one in 295 of 299 cases outside the band, and above the leading-order line for $\eta \gg \eta^*$, where higher-order terms enter, as Figure 1a shows. The predicted larger excess is the measured one in 599 of the 607 combinations of state, minibatch, p and η outside the band, 98.7%, with 95% interval $[97.1, 99.8]\%$ when the 18 state–minibatch groups are resampled as in Appendix B; the band leaves 607 of the 720 cases classified, and the prediction holds in 304 of 308 where live was predicted larger. At the warm-up learning rate all 67 classifications outside the band, among 90 cases, are correct, and the larger term there differs between states: the median $|\Delta_{\text{live}}/\Delta_{\text{frozen}}|$ is 9.5, 1.2 and 0.05 for the four-block states and 7.5, 2.3 and 2.2 for LoRA, as Table 5 lists. The size of the live excess is predicted as well: at $\eta^*_{\text{ref}}/64$, $\Delta_B/(-\eta a_1)$ lies in $[0.9, 1.1]$ for 89 of the 90 combinations of state, minibatch and p, with median 1.000. The interventions separate the two mechanisms at shared states, as Figure 1b shows. Freezing the denominator without a clip gives slopes 1.99–2.11, second order in 90/90 cases. With an active clip, the frozen-denominator excess is first order when $\nabla F^\top \beta_B \neq 0$: slope in $[0.7, 1.3]$ for 88/90 with median 1.00, and within 10% of $-\eta a_1$ for 89/90; for the live map with clip the ratio is within 10% for 90/90.

Second model. E-ONESTEP4b, registered before any analysed E-ONESTEP4 measurement, repeats the protocol unchanged at Qwen3-1.7B, with a suffix of 201.3M and LoRA adapters of 10.0M parameters, analysed separately. The predicted larger excess is the measured one in 582 of the 598

cases classified outside the band among 720, 97.3% with interval $[95.6, 98.8]\%$, and in 57 of the 58 classified among 90 at the warm-up learning rate, where the frozen-denominator excess is now the larger one in four of six states, with median live-to-frozen ratios of 0.25–4.8 in Table 5. The live excess is within 10% of $-\eta a_1$ at $\eta^*_{\mathrm{ref}}/64$ in 89/90 cases and in 90/90 with a clip; with an active clip the frozen-denominator excess is first order by ratio and by slope in 90/90 cases each, and without a clip second order in 87/90.

Independent data panel. E-ONESTEP7, registered before any of its output, repeats the protocol with unchanged code and thresholds on pairs from the training split of UltraFeedback that no earlier experiment used, with a new objective, new warm-up and test pairs and six new states per model, detailed in Appendix B. The predicted larger excess is the measured one in 582 of 617 classified cases at Qwen3-0.6B and 590 of 604 at Qwen3-1.7B, with 95% intervals over state–minibatch groups of $[90.9, 97.1]\%$ and $[95.5, 99.5]\%$, and at the warm-up learning rate in 76 of 76 and 63 of 63, where the median live-to-frozen ratio ranges from 0.31 to 27.6 over the twelve states. The live excess is within 10% of $-\eta a_1$ in 87/90 and 90/90 cases; with an active clip the frozen-denominator excess is first order by slope in 86/90 and 87/90 cases, and without a clip second order in 81/90 and 89/90.

4.3 Within-live order and an enumeration-free estimate

E-ONESTEP5, registered before any of its output, compares terms within the live map and removes the 2^m cost of a_1 at the twelve states of Section 4.2; Appendix C gives the details. Three predictions are distinct: the larger of the live and frozen excesses of Section 4.2; the larger part of the live excess, treated below; and Eq. (6), which no prediction here uses. *Within the live update*, the next order of Proposition 4 is $\Delta_B = -\eta a_1 + \eta^2 B_2 + O(\eta^3)$, with $B_2 = \frac{1}{2}(\mathbb{E}[U^\top \nabla^2 F\, U] - U(g_{\mathrm{cDPO}})^\top \nabla^2 F\, U(g_{\mathrm{cDPO}}))$ computed from exact quadratic forms by Eq. (8), so, up to an $O(\eta^3)$ remainder that is checked empirically, the first-order term is the larger part of the live excess when $\eta < \eta^\dagger = |a_1|/|B_2|$. At the six Qwen3-0.6B states this prediction, written before measurement, is correct in 661 of 721 rows outside the band $\eta/\eta^\dagger \in [0.5, 2]$; $\eta^2 B_2$ matches the measured remainder $\Delta_B + \eta a_1$ within 20% at the smallest step in 155/180 cases; the two-term expansion is closer to the measurement than the first-order term in 340/360 rows at the two smallest steps; and its own remainder falls with median slope 2.96 in η, as an $O(\eta^3)$ term requires; the criterion compares finite-order terms rather than bounding the remainder. At Qwen3-1.7B the prediction is correct in at least 90% of the rows of the live map without clip at five of the six states; its endpoints by state and map are in Tables 12 and 13. *Without enumeration*, a_1 is estimated by sampling flip patterns stratified by their number of flips, with the full rule in Appendix C, and with the per-example main effects of the update map as a control variate whose conditional mean is exact; the estimate is unbiased with a nominal 95% interval, accepted by the heuristic rule half-width $\leq 20\%$ of the estimate. Tested at $m = 8$ on a new objective F_2 of eight unused pairs against enumeration of all 256 patterns, including the training clip $c = 1.0$, and at $m = 16$ for LoRA against all 65,536, as summarized in Table 1, the interval covers the enumerated a_1 in 94–95% of cases, 64% are accepted, and 229 of 231 accepted estimates are within 20%. E-ONESTEP6, registered after these outcomes, re-estimates each of the 360 fixed cases 1,000 times from stored projections of all 256 patterns, Table 7: averaged over these fixed cases, the interval covers a_1 in 95.1% of repetitions at 0.6B and 94.8% at 1.7B, and in at least 90% in 357 of the 360 cases, with 94.9% and 94.5% among accepted estimates; acceptance is a precision rule on the half-width, and the per-case coverage, whose lowest values Table 1 lists, is measured by these repetitions; and at exactly the same number of update-map evaluations the root-mean-square error is smaller than that of plain Monte Carlo in 161 and 170 of 180 cases, by a median factor of 16. One p costs 76–102 map evaluations at $m = 8$ against 256 for enumeration, and in time, including all per-example gradients, a median 0.46 of it at the four-block and 0.86–0.87 at the LoRA states; a sweep over five p costs about 246 against 261 at $m = 8$ and 736 against 65,541 at $m = 16$. Predictions from the estimate alone hold on F_2: the larger of the live and frozen excesses is predicted correctly in 69 of 69 and 60 of 65 classified cases, and the accepted estimates predict the live excess within 20% in all 43 cases, with the frozen excess second order in all 60; these are the registered endpoints M4–M7 of Table 12.

Table 1: Enumeration-free estimate of a_1 in E-ONESTEP5, Part B, per model: six states, three minibatches, five p, live map with and without the training clip, new objective F_2. Accepted: half-width of the nominal 95% interval $\leq 20\%$ of the estimate. Repetitions: E-ONESTEP6, Table 7. Cost: update-map evaluations per specified p, relative to exact enumeration. At $m = 16$ the estimate is compared with the enumerated coefficient.

	Qwen3-0.6B	Qwen3-1.7B
$m = 8$: interval covers the enumerated a_1	169/180	171/180
$m = 8$: accepted; accepted and within 20% of a_1	116/180; 115/116	115/180; 114/115
$m = 16$, LoRA: covers; accepted; accepted and within 20%	15/15; 6/15; 6/6	14/15; 3/15; 3/3
$m = 8$, 1,000 repetitions per case: coverage; lowest case	95.1%; 92.9%	94.8%; 84.4%
Median cost per p, $m = 8$; $m = 16$	0.36; 0.0022	0.36; 0.0022

4.4 Gradient norms in multi-epoch training

Setup. We fine-tune Qwen3.5-2B and Qwen3.5-4B (Qwen Team, 2026) with LoRA (Hu et al., 2022) of rank 64 and $\alpha = 128$ on all attention and MLP projections, on 3,000 UltraFeedback training pairs (Cui et al., 2024) for three epochs with AdamW, $\beta = 0.1$ and $p = \varepsilon = 0.1$, over learning rates $\{1, 2\} \times 10^{-5}$, effective batches (EB) $\{16, 32\}$, three 2B seeds and two 4B cells, 14 paired cells in total. Both arms use one loss implementation: the cDPO arm the smoothed loss with $\varepsilon = 0.1$, the flip arm the clean loss on labels redrawn for every example at every epoch, as Lemma 2 assumes. The arms have identical budgets of 9,000 pair visits and 8.32 million tokens and identical hyperparameters, including a global gradient-norm clip at 1.0, and both learn in every cell: accuracy on 500 held-out pairs is 74–80%, a check that training proceeds, detailed in Appendix D.

Gradient-norm signature. At a common state, Lemma 2 gives $\mathbb{E}\|g_{\text{flip}}\|^2 = \|g_{\text{cDPO}}\|^2 + \text{tr}\Sigma_B$. At zero margin the cDPO coefficient of every example is $p - \frac{1}{2}$, whereas a flipped or unflipped example has coefficient $\pm\frac{1}{2}$; if the per-example Jacobians are mutually orthogonal, the norm ratio is $\sqrt{1 + p(1-p)/(p - \frac{1}{2})^2} = 1.25$ at $p = 0.1$, with no free parameter; as the cDPO coefficients $\sigma_i - (1 - p)$ vanish near their target, the ratio grows as $\|g_{\text{cDPO}}\|$ falls. The two runs of a cell follow different trajectories, so the norms in Figure 3 of Appendix D are a descriptive check of this structure rather than a same-state measurement of Σ_B. At the first logged step 10, where the mean margin satisfies $|\bar{s}| \leq 0.074$ in all 28 runs, the ratio has median 1.23 and range 1.14–1.61, 13 of 14 cells within 1.14–1.36. In the first epoch, where every pair is seen once, the per-cell median ratio stays at 1.23–1.39. It rises at the first epoch boundary, when pairs are revisited, to per-cell medians of 2.7–5.4 in the second epoch and 3.1–6.1 in the third, with medians over cells of 3.3 and 4.7 in Table 8.

4.5 The executed loss defines the compared procedure

We evaluate the loss of eight fixed pairs under a perturbed Qwen3-0.6B policy with nonzero margins through the trainer's own loss path across TRL releases (Hugging Face, 2026). In the newest executed release the native sigmoid loss is bitwise invariant to the smoothing parameter, whereas older executed releases and an explicit cDPO loss shift by up to 4.7147×10^{-4}, matching the analytic cDPO shift εz to within 5.2×10^{-8}, and the robust variant its own correction, as Table 9 in Appendix E lists. A scan of the loss code of 16 releases finds the smoothing mixture in the sigmoid branch up to the second newest executed release and in no later one. For the native sigmoid branch, this option therefore configures clean DPO in all later releases; this finding does not apply to every loss branch. The multi-epoch runs of Section 4.4 use one explicit smoothed loss for both arms.

5 Discussion

When does the choice between flips and smoothing matter? With affine updates, the difference per step is second order and governed by $p(1 - p)$, the batch size and the curvature along the preconditioned margin-Jacobian directions. With a live adaptive denominator or an active clip, the expected update itself is biased. Which excess is larger at a given learning rate depends on the state, and $\eta^* = |a_1|/|A_2|$, computed before any step, predicts it, as Section 4.2 shows; at the

warm-up learning rates of our 24 unseen states the median live-to-frozen ratio ranges from 0.05 to 27.6. Freezing or sharing the denominator and avoiding an active clip keeps the stochasticity of flips without this bias; cDPO removes the flip covariance at no extra cost.

Reporting checklist. A flip-versus-smoothing comparison should report the executed loss of Section 4.5, flip reuse as in Proposition 6, the update rule with its denominators and clips as in Proposition 4 and Corollary 5, and the step size and batch of Theorem 3.

Scope. The theory concerns one update; Proposition 6 and Appendix G accumulate it in a frozen-gradient model. Tests cover 13 Qwen3-0.6B and 12 Qwen3-1.7B states, two disjoint data panels, four eight-pair objectives, minibatches of 2 to 8 pairs, 16 for coefficient estimation, and responses of at most 64 tokens; the within-live crossover is established mainly at Qwen3-0.6B. Resampling over state–minibatch groups is conditional on each panel's shared pairs, objectives and warm-up histories. The predictions tested in Sections 4.2 and 4.3 use Eq. (5) with enumerated or sampled coefficients; Eqs. (6) and (7) give the mechanism and the leading form of these coefficients, and Corollary 5 assumes a clip active on whole segments, checked at every flip pattern and at the cDPO gradient. The multi-epoch runs, including Qwen3.5-4B, are used for the gradient-norm structure.

6 Related Work

DPO, IPO and conservative DPO (Rafailov et al., 2023; Azar et al., 2024; Mitchell, 2023) define the losses studied here. Robust variants change the target, the loss, or the per-example weights (Chowdhury et al., 2024; Liang et al., 2025; Wu et al., 2025; Nishimori et al., 2026; Liu et al., 2026), and the effect of preference noise on alignment has been measured (Gao et al., 2024); we hold the target fixed. DisturbLabel (Xie et al., 2016) shows empirically that per-iteration random labels and their soft mean train differently despite equal expected gradients, without stating when or at what order the difference appears, and soft-label DPO has been run with flipped and averaged soft labels (Furuta et al., 2024). Label-noise SGD theory (Blanc et al., 2020; HaoChen et al., 2021; Damian et al., 2021; Li et al., 2022) uses the same covariance: in the classification extension of Damian et al. (2021), labels flipped at every step give noise along ∇f_i with variance $p(1-p)$ for the logistic loss, and SGD and momentum converge near zero-loss solutions to stationary points of a flatness-regularized loss. For preference pairs $\ell'_- - \ell'_+ \equiv 1$, so the covariance of Lemma 2 holds exactly at every state; our statements are local one-step expansions that hold from any state, cover every affine update, and identify the expected-update bias of adaptive denominators and clipping. For adaptive methods, Adam shortens steps where the relative gradient variance is high, and stochastic update directions are compared through their one-step expected improvement, a first-order mean term plus a second-order quadratic form (Balles & Hennig, 2018); we specialize this comparison to preference flips, where the perturbation $(Z_i - p)J_i$ has a margin-independent coefficient and the first-order term arises only through the update map, with closed-form leading terms for Adam and clipping and a crossover tested prospectively. SDE limits describe noise over many steps (Malladi et al., 2022; Li et al., 2021; Smith & Le, 2018); under label noise with $1 - \beta_2 = \Theta(\eta^2)$ Adam minimizes $\mathrm{tr}(\mathrm{Diag}(H)^{1/2})$ near the minimizer manifold (Li et al., 2025), where a single gradient's contribution to the second moment, which Proposition 4 isolates at standard β_2, is negligible; and mini-batch noise changes the long-run implicit bias of Adam through its memory (Cattaneo & Shigida, 2026). The expectation of a clipped stochastic gradient differs from the clipped expectation (Koloskova et al., 2023); Corollary 5 gives the leading term of this bias for flips in closed form. For equal-mean preference flips, η^* compares the leading live and frozen excesses at a state, and Section 4.2 tests it prospectively at two model sizes.

7 Conclusion

Flip-DPO and cDPO share an objective but not an optimizer. Affine updates turn the flip covariance into a second-order change set by the curvature along the margin Jacobians; a live adaptive second moment or an active clip can bias the expected update, which gives a first-order change whenever $\nabla F^\top \beta_B \neq 0$. The crossover $\eta^* = |a_1|/|A_2|$ predicts which excess is larger at 24 unseen states of two models on two data panels, and a sampled estimate, with intervals calibrated on average over the tested cases, replaces the enumeration of a_1. In LoRA training, the flip-to-cDPO gradient-norm ratio rises several-fold once pairs are revisited.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to generate synthetic data sets, and qualitative and thematic data analysis, surveys, interviews and transcription are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and collects no personal data. All experiments use the publicly released Qwen3 and Qwen3.5 language models and the public UltraFeedback preference data under their licenses; no new data set is created or released, and no text generated by the trained models is released. The results concern how an optimizer responds to label flips and label smoothing that have the same expected objective; they do not provide a method for manipulating preference data, and the practical recommendations, to check the executed loss and to report flip reuse, the update rule and the step size, aim at more reliable experiments. The loss audit concerns the behaviour of public releases of an open-source library and does not involve any private system.

REPRODUCIBILITY STATEMENT

The objective identity, the covariance lemma and all expansions are stated with their assumptions in Section 3, with proofs and examples in Appendix F and the accumulated schedule in Appendix G. The one-step experiments, with data selection, trainable parameters, warm-up states, step grids, numerical precision and curvature computation, and the random-direction control are specified in Section 4.1 and Appendix A; the prospective tests on both data panels, their prediction and measurement phases and the full registered specification in Section 4.2 and Appendix B, and the grouped resampling intervals in Appendix B; the within-live expansion and the sampled estimator, including its sampling rule, interval and cost accounting, in Section 4.3 and Appendix C; the multi-epoch runs in Section 4.4 and Appendix D; and the loss audit in Section 4.5 and Appendix E. Every registered endpoint and its outcome is listed in Tables 11 and 12. The anonymous supplementary material contains the code of the double-precision one-step test and its random-direction control, the prospective tests on both data panels and the estimator, all their per-case outputs, the registered analyses including the grouped resampling, a script that recomputes the reported numbers from these outputs, the pre-specified protocol of every one-step experiment, the order of the prediction and measurement events, and the selection record of the independent panel.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513. PMLR, 2020. URL https://proceedings.mlr.press/v125/blanc20a.html.

Matias D. Cattaneo and Boris Shigida. The effect of mini-batch noise on the implicit bias of Adam. *arXiv preprint arXiv:2602.01642*, 2026. URL https://arxiv.org/abs/2602.01642.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42258–42274. PMLR, 2024. URL https://proceedings.mlr.press/v235/ray-chowdhury24a.html.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27449–27461, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e6af401c28c1790eaef7d55c92ab6ab6-Abstract.html.

Hiroki Furuta, Kuang-Huei Lee, Shixiang Shane Gu, Yutaka Matsuo, Aleksandra Faust, Heiga Zen, and Izzeddin Gur. Geometric-averaged preference optimization for soft preference labels. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57076–57114, 2024. doi: 10.52202/079017-1819. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/688c7a82e31653e7c256c6c29fd3b438-Abstract-Conference.html.

Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=nMAaCsCTCI.

Jeff Z. HaoChen, Colin Wei, Jason D. Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2315–2357. PMLR, 2021. URL https://proceedings.mlr.press/v134/haochen21a.html.

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

Hugging Face. TRL DPO trainer documentation, 2026. URL https://huggingface.co/docs/trl/dpo_trainer. Software documentation.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363. PMLR, 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=kCUDzyKQ7G.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In *Advances in Neural Information Processing Systems*, volume 34, pp. 12712–12725, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/69f629564298659092 1fa916d61c1f8-Abstract.html.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=siCt4xZn5Ve.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Hanzhu Chen, Feng Wu, and Jieping Ye. ROPO: Robust preference optimization for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37131–37161. PMLR, 2025. URL https://proceedings.mlr.press/v267/liang25d.html.

Jilong Liu, Yonghui Yang, Pengyang Shao, Wenjian Tao, Hao Zhan, Haokai Ma, Wei Qin, and Richang Hong. CompassDPO: Dynamics-controlled direct preference optimization for robust safety alignment. *arXiv preprint arXiv:2603.07211*, 2026. URL https://arxiv.org/abs/2603.07211.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6448–6458. PMLR, 2020. URL https://proceedings.mlr.press/v119/lukasik20a.html.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7697–7711, 2022. doi: 10.52202/068431-0559. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/32ac710102f0620d0f28d5d05a44fe08-Abstract-Conference.html.

Eric Mitchell. A note on DPO with noisy preferences & relationship to IPO. Technical note, November 2023. URL https://ericmitchell.ai/cdpo.pdf.

Soichiro Nishimori, Yu-Jie Zhang, Thanawat Lodkaew, and Masashi Sugiyama. On symmetric losses for policy optimization with noisy preferences. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=cBWGLmSeao.

Qwen Team. Qwen3.5: Towards native multimodal agents. Model card and technical blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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, pp. 53728–53741, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJij4yg0Z.

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 *The Thirteenth International Conference on Learning Representations*, pp. 16123–16150, 2025. URL https://openreview.net/forum?id=CbfsKHiWEn.

Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian. DisturbLabel: Regularizing CNN on the loss layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4753–4762, 2016. doi: 10.1109/CVPR.2016.514. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Xie_DisturbLabel_Regularizing_CNN_CVPR_2016_paper.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

A ONE-STEP TEST: DETAILS

Data, model and trainable set. Pairs come from the test preferences of the binarized release of UltraFeedback (Cui et al., 2024), in a fixed shuffled order. Prompts are rendered with the Qwen3 chat template with thinking disabled; pairs whose rendered prompt exceeds 64 tokens, or whose truncated responses coincide, are skipped. In order, the first eight kept pairs define F, the next twenty form five warm-up minibatches of four, and the next twenty-four form three disjoint test minibatches of eight, of which the $m = 4$ cells use the first four pairs. Response log-probabilities are sums over response tokens. The trainable set is the last four of 28 transformer blocks; embeddings, lower blocks, the final normalization and the tied output layer are frozen, so the activations entering the trainable blocks are computed once. The reference policy is the base model evaluated through the same double-precision computation, so all margins are exactly zero before warm-up.

Warm state. Five Adam steps with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, learning rate 2×10^{-5} and no weight decay on the cDPO loss with $\varepsilon = 0.1$ produce the evaluation point θ, at which $F(\theta) = 0.6787$ and $\|\nabla F(\theta)\| = 22.6$. The margins of the test pairs at θ lie between 0.015 and 0.36 in magnitude.

Step sizes and curvature estimates. For each minibatch, $\eta_{\mathrm{ref}} = 0.02/|\nabla F(\theta)^\top U(g_{p=0.1})|$ and $\eta \in \eta_{\mathrm{ref}} \cdot \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$, so that the largest step changes F by 0.02 to first order; the registered $m = 8$ SGD cell uses the two smallest steps. Quadratic forms $w^\top \nabla^2 F(\theta) w$ use central differences along $w/\|w\|$ with step h chosen so that $|\nabla F^\top w| h/\|w\| = 10^{-3}$, and Richardson extrapolation over h and $h/2$. Repeated evaluations of F at the same parameters are bitwise identical, and the truncated-network log-probabilities agree with the full model to 5.2×10^{-5}. The registration also covers Adam with a live denominator from the warm state and from a fresh state; Table 11 in Appendix H lists all registered endpoints on every completed cell.

Table 2: First registered one-step experiment, E-ONESTEP, every completed cell of the affine update rules, with finite-difference curvature; prefix activations, rotary tables and normalization internals are computed in single precision, and the other operations of the trainable suffix, including attention, in double precision. Ratio: measured over predicted at the smallest step of each cell; slope: log–log slope in η; both ranges over $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$; R^2: fit of Δ_B at the smallest step to $p(1-p)$ through the origin.

Update rule	Batch	m	Steps	Ratio	Slope in η	R^2
SGD	1	2	4	0.96–0.98	2.01–2.07	0.9998
SGD	1	4	4	0.94–1.01	2.03–2.13	0.9977
SGD	1	8	2	0.94–0.94	2.00–2.00	1.0000
SGD	2	2	4	0.87–0.88	1.99–2.01	1.0000
SGD	2	4	4	0.86–0.87	1.98–2.00	0.9999
SGD	3	2	4	1.01–1.02	1.98–2.00	1.0000
SGD	3	4	4	0.98–1.00	2.12–2.24	0.9997
Adam, frozen D	1	4	4	1.03–1.05	2.09–2.13	0.9997
Adam, frozen D	2	4	4	0.90–0.96	1.88–1.93	0.9970
Adam, frozen D	3	4	4	0.94–0.95	1.95–1.98	1.0000

Second experiment, E-ONESTEP2. E-ONESTEP2 was registered after the E-ONESTEP outcomes were known and uses the same model, pairs, trainable set, F and warm-up. It computes every normalization layer of the trainable blocks and the final normalization in double precision and uses an attention computation for which double backpropagation is available; the attention softmax and the rotary tables remain in single precision in this experiment, and E-ONESTEP3 below removes both. The recomputed warm state has $F(\theta) = 0.6786871$ and $\|\nabla F(\theta)\| = 22.57353$, against 0.6786867 and 22.57351 in E-ONESTEP. Exact quadratic forms $w^\top \nabla^2 F\, w$ are computed for the 16 SGD and 12 frozen-denominator directions of Table 2; three separate computations reproduce them bitwise. In double precision, Richardson extrapolation of central differences with the E-ONESTEP step rule is within 3.5% of the exact value for 27 of the 28 directions, and single central differences at h_0 and $h_0/2$ are within 3.0% and 2.6% for all but one. Part C uses the absolute step sizes $\{2, 1, 0.5\} \times 10^{-5}$ for live, frozen, live-after-clip and fresh Adam; Part D re-measures the $m = 4$ SGD and frozen cells

and runs the $m = 2$ Adam cells registered for E-ONESTEP. Table 3 summarizes Parts C and D; Table 11 lists every registered endpoint, including those of the active-clip cells of Part B. Because E-ONESTEP2 differs from E-ONESTEP in its computation and warm state, where $F(\theta)$ differs by 3.9×10^{-7}, its A1 and A2 endpoints, which combine E-ONESTEP deltas with E-ONESTEP2 curvature, are cross-implementation diagnostics.

Table 3: Second registered one-step experiment, E-ONESTEP2: the state recomputed with double-precision normalization layers and with attention softmax and rotary tables in single precision, with exact Hessian quadratic forms by double backpropagation for Theorem 3 and the exactly enumerated first-order term $-\eta\nabla F^\top \beta_B$ for live Adam. Ratio at the smallest step of each grid, 5×10^{-6} for the training grid. Scaled: the E-ONESTEP step grid; training: $\eta \in \{2, 1, 0.5\}\times 10^{-5}$, the learning rates of Section 4.4 and half the smaller one. Ranges over the three batches and $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$; R^2: smallest over cells.

Update rule	Grid	Cells	Prediction	Ratio	Slope in η	R^2
SGD	$m = 4$, scaled	3	Thm. 3	0.99–1.06	1.99–2.24	0.9977
Adam, frozen D	$m = 4$, scaled	3	Thm. 3	0.99–1.07	1.88–2.13	0.9970
Adam, frozen D	$m = 2$, scaled	3	Thm. 3	0.99–1.17	1.77–1.97	0.9953
Adam, frozen D	$m = 4$, training	3	Thm. 3	0.96–1.04	1.99–2.04	0.9997
Adam, live	$m = 4$, training	3	Eq. (5)	0.96–1.14	0.90–1.22	–

Third experiment, E-ONESTEP3. E-ONESTEP3 was registered after E-ONESTEP2 was known. It evaluates the whole model in double precision, including the attention softmax, every normalization layer and the rotary tables, so that prefix activations, references, warm-up and all parameter-dependent operations are in double precision. A precision check before any measurement confirms that the output of every module of the trainable blocks and every attention softmax is in double precision in all three computations, and the truncated-network log-probabilities agree with the full model to 1.3×10^{-13}. The warm state has $F(\theta) = 0.678687$ and $\|\nabla F(\theta)\| = 22.5735$; the three computations for the SGD, frozen and training-scale cells reproduce these values and all 28 exact quadratic forms bitwise. Every ratio in Table 4 pairs a measured Δ_B with a prediction from the same computation. Central differences along the 28 directions at $h_0 2^{-k}$, $k = 0, \dots, 6$, approach the exact values as h decreases: the median relative error is 7.5×10^{-5} at h_0, falls to 3.0×10^{-7} and 2.9×10^{-7} at $h_0/16$ and $h_0/32$, and is 6.2×10^{-7} at $h_0/64$, where the maximum is 2.6×10^{-4}.

Table 4: Same-state one-step test in double precision throughout, E-ONESTEP3, with predictions from the same computation: exact Hessian quadratic forms for Theorem 3 or the exactly enumerated first-order term $-\eta\nabla F^\top \beta_B$ for live Adam. Grid "scaled": the E-ONESTEP step sizes of each cell; "training": $\eta \in \{2, 1, 0.5, 0.25\} \times 10^{-5}$. Ratio: measured over predicted *at the smallest step of the grid*; slope: log–log slope over all steps of the grid. Ranges over cells and $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$; R^2: smallest over cells.

Update rule	m	Grid	Cells	Prediction	Ratio	Slope in η	R^2
SGD	$m = 2, 4, 8$	scaled	7	Thm. 3	0.987–1.057	1.98–2.24	0.9977
Adam, frozen D	$m = 4$	scaled	3	Thm. 3	0.990–1.071	1.88–2.13	0.9970
Adam, frozen D	$m = 2$	scaled	3	Thm. 3	0.995–1.173	1.77–1.97	0.9953
Adam, frozen D	$m = 4$	training	3	Thm. 3	0.998–1.010	1.99–2.03	1.0000
Adam, live	$m = 4$	training	3	Eq. (5)	0.982–1.067	0.93–1.18	–

Random-direction control, E-ISO9. E-ISO9 was registered after the E-ONESTEP3 outcomes were known and before any of its own output existed. It recomputes the E-ONESTEP3 state, whose $F(\theta)$ agrees to 2×10^{-15}, and, for SGD with the first four pairs of each test minibatch, compares two noises with the same mean update g_{cDPO} and the same Bernoulli weights: the flip noise $(Z_i - p)J_i/m$, and $(Z_i - p)\|J_i\|r_i/m$, where each r_i is a random unit vector drawn once with a fixed seed, with $|\cos(J_i, r_i)| \le 2.5 \times 10^{-4}$. The per-example norms are equal, so $\mathrm{tr}\Sigma_B$ is the same. Both excesses are computed exactly by enumerating the 16 flip patterns at the four SGD steps of E-ONESTEP3 and compared with Eq. (4), evaluated with exact quadratic forms along J_i or $\|J_i\|r_i$. At the smallest step

the measured-to-predicted ratio lies in $[0.990, 1.057]$ for the Jacobian directions and in $[0.979, 1.120]$ for the random directions in 15/15 cases each, the registered endpoints I1 and I2, with slopes 1.99–2.24 and 1.93–2.15 in η. The curvature sum $\sum_i J_i^\top \nabla^2 F\, J_i$ exceeds $\sum_i \|J_i\|^2 r_i^\top \nabla^2 F\, r_i$ by factors of 2.3×10^4, 1.2×10^5 and 3.5×10^5 on the three minibatches, and the measured excess along the Jacobians is larger by a median factor of 1.2×10^5, with range 2.4×10^4–3.5×10^5. The one-step excess of an affine update is therefore set by the curvature of F along the margin Jacobians, as Eq. (4) states, and noise of the same trace in random directions changes F four to five orders of magnitude less.

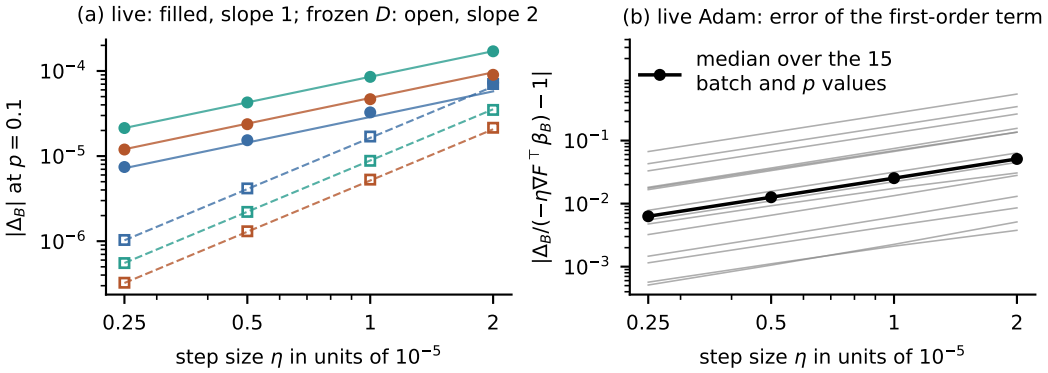


Figure 2: Live versus frozen Adam denominators in E-ONESTEP3 at the same state and training-scale step sizes. (a) $|\Delta_B|$ at $p = 0.1$ for the three minibatches in colors; markers: measured; lines: predicted, $-\eta \nabla F^\top \beta_B$ for live and Theorem 3 for frozen. (b) Relative error of the first-order prediction for live Adam, one grey line per minibatch and p; it halves with each halving of η.

B Prospective Test at Unseen States: Details

Registration. E-ONESTEP4 was registered after E-ONESTEP3 was known and before any of its own output existed. The registration fixes the following before any analysed output: the step grid is set in the first phase relative to the predicted crossover, so that both sides of it are tested; two of the six states use the training clip during warm-up; the endpoints on Eq. (6) use the criterion $\hat{R}$ of Appendix H; the enumerated a_1 is the primary input of η^*; the double-precision forward pass is bitwise reproducible. The first phase writes the prediction file with the predictions, the step grid and a fingerprint of the parameters and Adam moments; this file was recorded in the registration before the second phase of that state started, and the second phase stops unless both the recorded file and the state fingerprint match. The supplementary material lists this order of events for every state. E-ONESTEP4b is part of the same registration and was fixed before any analysed E-ONESTEP4 measurement existed; it fixes the same protocol, rules and thresholds for Qwen3-1.7B, and its endpoints are evaluated on the 1.7B states alone.

States, maps and clip level. Data follow the rule of Appendix A, skipping the 52 pairs used earlier: 8 pairs define F, 40 form ten warm-up minibatches of four, used cyclically, and 24 form three test minibatches of which the first four pairs are used; both models use the same pairs. Warm-up is cDPO with $\varepsilon = 0.1$ and AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and weight decay 0.01, at learning rate 2×10^{-5} for the four-block suffix and 10^{-4} for LoRA, whose A is seeded uniform and $B = 0$. All maps include the decoupled weight-decay shift. The clip level is the training value $c = 1.0$ when every flip-pattern and cDPO gradient of the minibatch has norm above one, in 13 of 18 state–minibatch pairs at 0.6B and 12 of 18 at 1.7B, and half the smallest such norm otherwise, so that the clip is active on the whole support. Numerics are those of E-ONESTEP3: double precision throughout with the same precision check, the suffix-versus-full check and bitwise repeats of F.

Per-state results and consistency checks. Table 5 lists the per-state results of both models. At 0.6B the crossovers η^* of individual minibatches and values of p span 3×10^{-8}–6.6×10^{-3}. Each second phase recomputes the enumerated a_1 before measuring; the registered comparison is the maximum

relative difference over the four maps, 4–10.5 at 0.6B, which is attained by the frozen-denominator map, whose a_1 is zero analytically and at most 3.6×10^{-12} in magnitude numerically. Per map, the maximum relative differences are 1.4×10^{-9} for live, 6.7×10^{-9} for live with clip and 2.0×10^{-7} for frozen with clip over the 90 combinations of state, minibatch and p of each map, and for the frozen map absolute differences of at most 3.0×10^{-12} with relative differences that reproduce the registered 4–10.5. At each model all 72 cells contain every combination of step and p exactly once, with 2,115 evaluations of F per state.

Table 5: Prospective one-step test at unseen states, per state. η^*: median over batches and p of the crossover $|a_1|/|A_2|$ predicted before measurement; the step grid is $\eta^* 4^k$, $k = -3, \ldots, 3$, plus the warm-up learning rate η_{w}. Live/frozen: median measured $|\Delta_{\mathrm{live}}/\Delta_{\mathrm{frozen}}|$ at η_{w}. Dominance: predicted order of $|\Delta_{\mathrm{live}}|$ and $|\Delta_{\mathrm{frozen}}|$ correct, over all steps and at η_{w}, outside the band $\eta/\eta^* \in [0.5, 2]$. Live: $\Delta/(-\eta a_1) \in [0.9, 1.1]$ at $\eta^*/64$. Interventions: frozen denominator with active clip, ratio in $[0.9, 1.1]$ and slope in $[0.7, 1.3]$; frozen without clip, slope in $[1.7, 2.3]$; slopes over the three smallest relative-grid steps, without η_{w}. Counts over 3 batches $\times$ 5 values of p. The two models are analysed separately.

Model	State	η^*	η_{w}	Live/frozen at η_{w}	Dominance	at η_{w}	Live	Interventions
0.6B	4 blocks, 10 steps	1.9×10^{-4}	2.0×10^{-5}	9.51	99/101	11/11	14/15	15/15, 14/15, 15/15
0.6B	4 blocks, 40 steps	2.2×10^{-5}	2.0×10^{-5}	1.20	98/98	8/8	15/15	15/15, 15/15, 15/15
0.6B	4 blocks, 40 steps, clip	9.6×10^{-7}	2.0×10^{-5}	0.05	104/104	14/14	15/15	15/15, 15/15, 15/15
0.6B	LoRA, 10 steps	7.8×10^{-4}	1.0×10^{-4}	7.46	95/99	9/9	15/15	15/15, 15/15, 15/15
0.6B	LoRA, 40 steps	2.3×10^{-4}	1.0×10^{-4}	2.26	101/102	12/12	15/15	15/15, 15/15, 15/15
0.6B	LoRA, 40 steps, clip	2.2×10^{-4}	1.0×10^{-4}	2.22	102/103	13/13	15/15	14/15, 14/15, 15/15
1.7B	4 blocks, 10 steps	3.0×10^{-5}	2.0×10^{-5}	0.88	92/94	4/4	15/15	15/15, 15/15, 15/15
1.7B	4 blocks, 40 steps	1.5×10^{-5}	2.0×10^{-5}	0.73	100/100	10/10	15/15	15/15, 15/15, 15/15
1.7B	4 blocks, 40 steps, clip	3.7×10^{-6}	2.0×10^{-5}	0.25	101/105	14/15	15/15	15/15, 15/15, 12/15
1.7B	LoRA, 10 steps	5.1×10^{-4}	1.0×10^{-4}	4.77	95/103	13/13	14/15	15/15, 15/15, 15/15
1.7B	LoRA, 40 steps	1.1×10^{-4}	1.0×10^{-4}	1.07	96/96	6/6	15/15	15/15, 15/15, 15/15
1.7B	LoRA, 40 steps, clip	8.3×10^{-5}	1.0×10^{-4}	0.48	98/100	10/10	15/15	15/15, 15/15, 15/15

Independent data panel, E-ONESTEP7. E-ONESTEP7 was registered after the E-ONESTEP4 to E-ONESTEP6 outcomes were known and before any of its own output existed. Its pairs come from the training preferences of the binarized UltraFeedback release (Cui et al., 2024), whereas all earlier one-step experiments use the test preferences, in a fixed shuffled order and without the prompts of the multi-epoch runs of Section 4.4. The selection rule above is applied unchanged, with no skipped pairs: the first 8 accepted pairs define a new objective F, the next 40 the warm-up minibatches and the next 24 the three test minibatches, so no pair is shared with the first panel; both models use these pairs. The code, the six state definitions, the four maps, the clip rule, the step rule, the classification band, the numerics, the two-phase protocol and all thresholds are those of E-ONESTEP4, and the prediction file of each state was recorded before its measurement started. All twelve states completed both phases, as the registered completion rule requires. Table 6 lists the per-state results and Table 12 the endpoints P1–P5.

Grouped resampling, E-BOOT8. The cases of a model share states and minibatches, so E-BOOT8, registered after all earlier outcomes and computed only from recorded outputs, takes a pair of state and test minibatch as the unit: all 18 groups of a model are resampled with replacement 10,000 times, a group without an eligible case entering with zero weight, and the pooled proportion is recomputed for a 95% percentile interval. For the larger-excess prediction of Section 4.2 the intervals are $[97.1, 99.8]\%$ at Qwen3-0.6B and $[95.6, 98.8]\%$ at Qwen3-1.7B, and $[97.5, 99.7]\%$ and $[95.0, 99.3]\%$ when the six states are resampled instead; at the warm-up learning rate $[100, 100]\%$ and $[94.3, 100]\%$; for the live excess within 10% of $-\eta a_1$, $[96.7, 100]\%$ for both models; for accepted $m = 8$ estimates within 20%, $[97.2, 100]\%$ and $[97.0, 100]\%$; for the pooled coverage of Table 7, $[94.9, 95.2]\%$ and $[94.5, 95.1]\%$. On the independent panel of E-ONESTEP7 the intervals for the larger-excess prediction are $[90.9, 97.1]\%$ and $[95.5, 99.5]\%$, or $[91.5, 96.8]\%$ and $[96.3, 99.0]\%$ with the states resampled, and for the live excess within 10% of $-\eta a_1$ $[93.3, 100]\%$ and $[100, 100]\%$. These intervals describe the variation over the state–minibatch groups of the two fixed panels, with their objectives and warm-up histories held fixed; the batches of a panel recur across its states, and the independent panel is the evidence for a new objective and new states.

Table 6: Prospective one-step test on the independent data panel, E-ONESTEP7, per state: new objective, warm-up and test pairs; same code, rules and thresholds as Table 5. η^*: median over batches and p of the crossover $|a_1|/|A_2|$ predicted before measurement; the step grid is $\eta^* 4^k$, $k = -3, \ldots, 3$, plus the warm-up learning rate η_{w}. Live/frozen: median measured $|\Delta_{\mathrm{live}}/\Delta_{\mathrm{frozen}}|$ at η_{w}. Dominance: predicted order of $|\Delta_{\mathrm{live}}|$ and $|\Delta_{\mathrm{frozen}}|$ correct, over all steps and at η_{w}, outside the band $\eta/\eta^* \in [0.5, 2]$. Live: $\Delta/(-\eta a_1) \in [0.9, 1.1]$ at $\eta^*/64$. Interventions: frozen denominator with active clip, ratio in $[0.9, 1.1]$ and slope in $[0.7, 1.3]$; frozen without clip, slope in $[1.7, 2.3]$; slopes over the three smallest relative-grid steps, without η_{w}. Counts over 3 batches $\times$ 5 values of p. The two models are analysed separately.

Model	State	η^*	η_{w}	Live/frozen at η_{w}	Dominance	at η_{w}	Live	Interventions
0.6B	4 blocks, 10 steps	4.2×10^{-4}	2.0×10^{-5}	21.35	91/103	12/12	14/15	15/15, 15/15, 10/15
0.6B	4 blocks, 40 steps	7.7×10^{-5}	2.0×10^{-5}	3.33	96/101	11/11	14/15	15/15, 15/15, 15/15
0.6B	4 blocks, 40 steps, clip	9.5×10^{-6}	2.0×10^{-5}	0.31	96/101	11/11	15/15	15/15, 15/15, 11/15
0.6B	LoRA, 10 steps	2.8×10^{-3}	1.0×10^{-4}	27.60	97/105	15/15	15/15	11/15, 11/15, 15/15
0.6B	LoRA, 40 steps	1.2×10^{-3}	1.0×10^{-4}	11.57	99/103	13/13	14/15	15/15, 15/15, 15/15
0.6B	LoRA, 40 steps, clip	4.0×10^{-4}	1.0×10^{-4}	3.01	103/104	14/14	15/15	15/15, 15/15, 15/15
1.7B	4 blocks, 10 steps	3.1×10^{-4}	2.0×10^{-5}	11.81	105/105	15/15	15/15	14/15, 14/15, 15/15
1.7B	4 blocks, 40 steps	5.2×10^{-5}	2.0×10^{-5}	2.88	99/101	11/11	15/15	15/15, 15/15, 15/15
1.7B	4 blocks, 40 steps, clip	1.3×10^{-5}	2.0×10^{-5}	3.70	98/99	8/8	15/15	15/15, 15/15, 14/15
1.7B	LoRA, 10 steps	4.2×10^{-4}	1.0×10^{-4}	4.02	103/105	15/15	15/15	14/15, 13/15, 15/15
1.7B	LoRA, 40 steps	1.7×10^{-4}	1.0×10^{-4}	2.02	94/99	9/9	15/15	15/15, 15/15, 15/15
1.7B	LoRA, 40 steps, clip	1.4×10^{-4}	1.0×10^{-4}	1.54	91/95	5/5	15/15	15/15, 15/15, 15/15

C Within-Live Order and Enumeration-Free Estimate: Details

Registration. E-ONESTEP5 was registered after the E-ONESTEP4 and E-ONESTEP4b outcomes were known and before any of its own output existed. It reuses the twelve saved states of Section 4.2, each reloaded and checked against its fingerprint, and adds new quantities and new measurements; the two models are analysed separately. The registration, fixed before any output of Part B existed, specifies the stratified estimator described here and a separate $m = 8$ enumeration. As before, the prediction file of each state was written and recorded in the registration before that state's measurements started, and every measurement checks it.

Part A: second-order coefficient of the live update. Expanding $F(\theta - \eta U)$ to second order for every realization and taking the conditional expectation, as in the proof of Proposition 4, gives for any update map bounded on the support of g_{flip}

$$\Delta_B(\eta, p) = -\eta\, a_1 + \eta^2 B_2 + O(\eta^3),$$
$$B_2 = \tfrac{1}{2}\Big(\mathbb{E}\big[U(g_{\mathrm{flip}})^\top \nabla^2 F\, U(g_{\mathrm{flip}}) \mid B\big] - U(g_{\mathrm{cDPO}})^\top \nabla^2 F\, U(g_{\mathrm{cDPO}})\Big), \tag{8}$$

so the first-order term is the larger part of the live excess, up to $O(\eta^3)$, when $\eta < \eta^\dagger = |a_1|/|B_2|$. For the live map and the live map with the registered clip, B_2 is computed from exact Hessian quadratic forms of the 2^m pattern updates and the cDPO update, with $m = 4$ and the E-ONESTEP4 objective, minibatches and clip levels, before any measurement. The steps are $\eta^\dagger_{\mathrm{ref}} 4^k$, $k = -2, \ldots, 2$, with $\eta^\dagger_{\mathrm{ref}}$ the median of $\eta^\dagger$ over p; rows with $\eta/\eta^\dagger \in [0.5, 2]$ are excluded from the classification. Δ_B is then measured exactly and $r = \Delta_B + \eta a_1$ is the measured remainder of the first-order term.

Part B: stratified estimate of a_1 with an interval. Let $\phi(Z) = \nabla F^\top U(g_Z)$, so $a_1 = \mathbb{E}\phi(Z) - \phi(g_{\mathrm{cDPO}})$. The per-example main effects $s_i^+ = \phi(g_{\mathrm{cDPO}} + (1-p)J_i/m) - \phi(g_{\mathrm{cDPO}})$ and $s_i^- = \phi(g_{\mathrm{cDPO}} - pJ_i/m) - \phi(g_{\mathrm{cDPO}})$ cost $2m + 1$ map evaluations, and $c(Z) = \sum_i s_i^{Z_i}$ is used as a control variate. Patterns are stratified by their number of flips k, with weight $\binom{m}{k} p^k (1-p)^{m-k}$; given k, $P(Z_i = 1) = k/m$, so the control variate has an exact conditional mean in every stratum. The complete rule for one p is:

1. evaluate $\phi(g_{\mathrm{cDPO}})$ and the $2m$ main effects; set $c_k = \sum_i \big(\frac{k}{m} s_i^+ + (1 - \frac{k}{m}) s_i^-\big)$, the exact mean of $c(Z)$ given k flips, and $w_k = \binom{m}{k} p^k (1-p)^{m-k}$;

2. take all $\binom{m}{k}$ patterns of stratum k if $\binom{m}{k} \le m$; otherwise draw $n_k = \min\{\binom{m}{k}, \max(2, \lceil 64\, w_k/W \rceil)\}$ distinct patterns with k flips uniformly without replacement, where W is the total weight of these strata, so a stratum whose allocation reaches its size is evaluated in full;
3. $\mu_k = c_k +$ the mean of $\phi(Z) - c(Z)$ over the patterns of stratum k, and $\hat{a}_1 = \sum_k w_k \mu_k - \phi(g_{\mathrm{cDPO}})$;
4. $\hat{V} = \sum_k w_k^2 S_k^2\, n_k^{-1}\big(1 - n_k/\binom{m}{k}\big)$ over the strata with $n_k < \binom{m}{k}$, with S_k^2 the sample variance of $\phi(Z) - c(Z)$, so that fully evaluated strata contribute zero;
5. report the half-width $t_{0.975,\nu}\hat{V}^{1/2}$ and the cost $2m+1$ plus the number of patterns evaluated, where, with $v_k = w_k^2 S_k^2 n_k^{-1}(1 - n_k/\binom{m}{k})$ for the sampled strata and $z = 1.959964$,

$$\nu = \Big(\sum_k v_k\Big)^2 \Big/ \sum_k \frac{v_k^2}{n_k - 1},$$

$$t_{0.975,\nu} \approx z + \frac{z^3 + z}{4\nu} + \frac{5z^5 + 16z^3 + 3z}{96\nu^2} + \frac{3z^7 + 19z^5 + 17z^3 - 15z}{384\nu^3}.$$

At $m = 8$ the cap binds in the stratum $k = 2$ of 28 patterns for $p \le 0.2$, which is then evaluated in full, and one p costs 76, 81, 92, 102 and 102 map evaluations for $p = 0.05, 0.1, 0.2, 0.3, 0.5$. At $m = 4$ every stratum has at most six patterns and is evaluated in full, so the half-width is zero and the estimate is the enumerated value up to rounding: over the 360 cases of both models it differs from the a_1 computed in E-ONESTEP4 and E-ONESTEP4b by at most 2.9×10^{-9} relative and 2.7×10^{-8} absolute, and the six cases of row M1b in Table 12 not counted as covered are those, all at one Qwen3-0.6B state, whose relative difference of 1.0–2.9×10^{-9} exceeds the registered comparison tolerance of 10^{-9}. The estimate is unbiased, and it is *accepted* when the half-width is at most 20% of $|\hat{a}_1|$; an accepted interval that covers a_1 implies a relative error of at most $0.2/0.8 = 25\%$, so the relative error is reported separately. It is evaluated at $m = 8$, with all eight pairs of each test minibatch, on a new objective F_2, the next eight admissible pairs, never used before and shared by the six states of a model, for the live map and the live map with the training clip $c = 1.0$, against exact enumeration of the 256 patterns; and, for the LoRA states, at $m = 16$, with the first eight pairs of two test minibatches, against enumeration of the 65,536 patterns. From the estimate, $\eta^* = |\hat{a}_1|/|A_2|$ on F_2, with A_2 from exact quadratic forms, fixes the steps $\eta^*_{\mathrm{ref}}/8$, $8\eta^*_{\mathrm{ref}}$ and the training learning rate; a row is excluded when the interval of η/η^* meets $[0.5, 2]$. Δ_B on F_2 is then measured exactly over all 256 patterns for the live, frozen and training-clip maps.

Completeness and further quantities. All twelve states completed both parts, the $m = 8$ enumerations and, for the six LoRA states, the $m = 16$ comparison. For one specified p, the estimate at $m = 8$ used 76–102 map evaluations against 256 for enumeration, and at $m = 16$ about 147 against 65,536; for the five values of p of one minibatch at $m = 16$, 736 against 65,541 for enumeration, without sharing patterns across p. The control variate shortens the median half-width by a factor of 5.0 at 0.6B and 4.9 at 1.7B relative to the same samples without it. At the training learning rate the median $|\Delta_{\mathrm{live}}/\Delta_{\mathrm{frozen}}|$ on F_2 ranges from 0.20 to 2.7 over the 0.6B states and from 0.06 to 9.2 over the 1.7B states. On the E-ONESTEP4 and E-ONESTEP4b measurements, adding $\eta^2 B_2$ to the first-order term reduces the error in 1,300 of 1,440 rows at 0.6B and 1,272 of 1,440 at 1.7B. All registered endpoints are listed in Table 12.

Repeated-sampling calibration and equal-budget baselines, E-ONESTEP6. E-ONESTEP6 was registered after the E-ONESTEP5 outcomes were known and before any of its own output existed. From each of the twelve saved states, reloaded and verified to be unchanged, it stores for every $m = 8$ minibatch and both estimator maps, live and live with the training clip, the projections $\phi(Z)$ of all 256 flip patterns on F_2, and for every p the $2m + 1$ main-effect evaluations; no objective is evaluated at any other parameter. Enumeration over the stored values reproduces the $m = 8$ values of a_1 in all 360 cases, and the estimator run on them with the recorded random sequence reproduces every recorded estimate, half-width and evaluation count. Each of the 360 cases, one per state, minibatch, map and p, is then re-estimated $R = 1{,}000$ times with independent random sequences. Two baselines receive, in each repetition, exactly the number of map evaluations used by the estimator: plain Monte Carlo,

with $\phi(g_{\mathrm{cDPO}})$, independent patterns from $\mathrm{Bernoulli}(p)^m$ and a t interval; and stratified sampling without the control variate, over the same strata, where the budget left after the fully evaluated strata is allocated in proportion to w_k and what a complete stratum cannot absorb is passed on to the others. Both baselines thus use exactly the estimator's number of evaluations in every repetition of every case. Table 7 gives the results; all registered endpoints are in Table 12. In every case of the clipped map the training clip is active on all 256 flip patterns, so the registered subset with the clip active on only part of the patterns is empty; at the LoRA states the smoothed gradient itself lies below the clip level for one of the five p in 17 of the 18 combinations of model, state and minibatch, so these cases include updates on both sides of the clip boundary. At a common state all methods need the same m per-example margin gradients and one gradient of F_2; the stored projections make the cost of a sweep over several p explicit: sharing the evaluated patterns across the five p of one state, minibatch and map, the estimator needs on average 246 map evaluations, 233–262, against 261 for enumeration, and 453 without sharing.

Pooled and per-case calibration. The pooled coverage averages over the fixed cases; resampling the 18 state–minibatch groups of a model gives 95% intervals of $[94.9, 95.2]\%$ at Qwen3-0.6B and $[94.5, 95.1]\%$ at Qwen3-1.7B for it. Per case, the coverage is at least 0.90 in all 180 cases at Qwen3-0.6B, with lowest 0.929, and in 177 of 180 at Qwen3-1.7B. The three cases below 0.90, at 0.895, 0.876 and 0.844, share one minibatch and the live map without clip at the two unclipped LoRA states, at $p = 0.05$ and 0.1; 98–100% of their repetitions are accepted. In each of them the sampled stratum with the largest weight, $k = 3$ with 7 or 12 of its 56 patterns sampled, has a strongly skewed residual $\phi(Z) - c(Z)$, with skewness -1.9 and excess kurtosis 3.9–4.2 over all 56 patterns, for which a t interval from a few draws is short whenever the draws miss the tail; the median quantile $t_{0.975,\nu}$ used there is 2.2–2.4. The per-case coverages of all 360 cases and the stratum diagnostics of these three cases are in the supplementary material.

Time including the prerequisite derivatives. The same runs recorded the time of the per-example margin gradients, of the gradient of F_2, and of the 256 map evaluations with their projections, all measured within one run per state. In these runs the gradients of the first four pairs of each minibatch were already held from the preceding phase, so the recorded gradient time covers four of the eight pairs; the full prerequisite cost is reconstructed by counting it twice, since all pairs obey the same length limits. A descriptive comparison, registered before it was computed, counts for one p the prerequisite derivatives plus the recorded number of map evaluations of the estimate against the derivatives plus 256 evaluations for enumeration. With all eight gradients counted, at the four-block states, where a map evaluation over all trainable parameters dominates and the derivatives take a median 16% of the enumeration time, the estimate takes a median 0.46 of the enumeration time, with range 0.36–0.53 over both models; at the LoRA states the derivatives take 78–79% of it and the estimate takes medians of 0.86 and 0.87, with range 0.79–0.91. The directly measured increment with four gradients held gives medians of 0.44 and 0.84. The saving in time thus grows with the share of the enumeration cost that the map evaluations take; at $m = 16$ about 147 evaluations replace 65,536. Both procedures hold the same m per-example gradients in memory.

Table 7: Repeated-sampling calibration of the $m = 8$ estimate in E-ONESTEP6: 1,000 independent repetitions of each of the 180 cases per model. Coverage: nominal 95% interval contains the enumerated a_1. SE ratio: standard deviation of the estimates over the root-mean-square reported standard error, median over cases. RMSE ratio: root-mean-square error of the baseline over that of the estimator at the same number of map evaluations, median over cases, followed by the number of cases with ratio > 1. Small $|a_1|$: the 18 cases below the model's 10th percentile of $|a_1|$.

	Qwen3-0.6B	Qwen3-1.7B
Coverage, pooled; lowest case	95.1%; 92.9%	94.8%; 84.4%
Accepted fraction; coverage and within 20% when accepted	64.3%; 94.9%, 99.6%	63.0%; 94.5%, 99.5%
Median SE ratio; mean within 3 Monte Carlo SE of a_1	1.00; 180/180	1.00; 179/180
Small $\lvert a_1\rvert$: coverage; accepted fraction	95.3%; 49.3%	94.2%; 40.3%
Plain Monte Carlo: RMSE ratio; accepted fraction	15.8 (161/180); 4.1%	16.0 (170/180); 10.0%
Stratified, no control variate: RMSE ratio; accepted fraction	3.3 (143/180); 45.1%	2.9 (158/180); 41.5%

D Multi-Epoch Comparison: Details

Both arms use the same loss implementation and the same 3,000 training and 500 held-out pairs, with no overlap. The cDPO arm logs the applied smoothing value at every logged optimizer step; it equals $\varepsilon = 0.1$ throughout. The flip arm applies the clean loss, $\varepsilon = 0$, to labels assigned by independent Bernoulli draws for every example at the start of every epoch, with a fresh seed per epoch; in the first 2B run the three epochs flip 291, 314 and 301 of the 3,000 pairs, and the observed per-epoch flip rates are 0.088–0.112 across the nine epoch draws of the three seeds. Each run has 564 optimizer updates at EB 16 and 282 at EB 32. From the first to the last quarter of training the cDPO gradient norm falls by a factor of 3.6–8.3 and the flip norm by 1.07–2.1, as Table 8 lists. Every logged pre-clipping norm exceeds 1.11, so the clip at 1.0 is active at every logged step of both arms. Effective batches of 16 and 32 are accumulated from micro-batches of two pairs; the global gradient-norm clip is 1.0 in both arms. Each run processes 8,317,380 tokens, identical between the two arms of every cell; evaluation on the 500 held-out pairs requires one forward pass per pair and adapter.

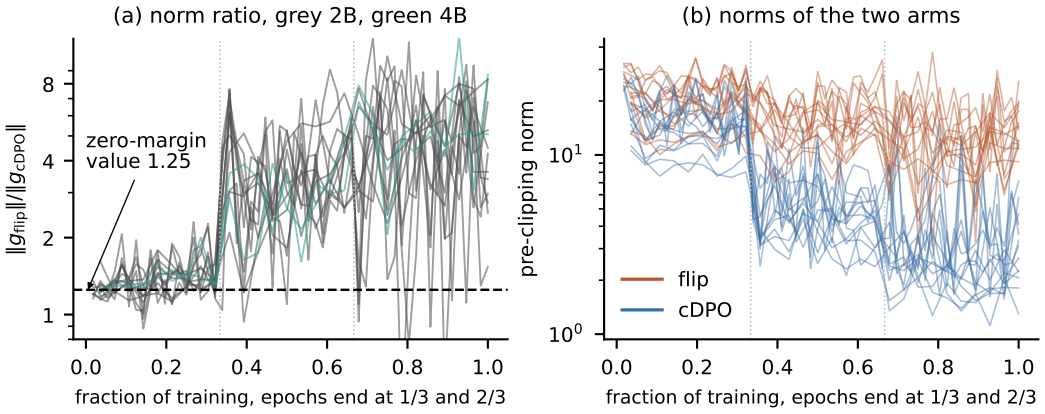


Figure 3: Pre-clipping gradient norms in the 14 paired cells, logged every 10 optimizer steps, against the training fraction by optimizer step, so epochs end at $1/3$ and $2/3$. (a) Ratio of the flip and cDPO norms at the same optimizer step, one line per cell; dashed: the zero-margin value of Lemma 2 for orthogonal Jacobians. (b) The norms of the two arms.

Gradient norms. The trainer logs the global gradient norm before clipping every 10 optimizer steps. Table 8 lists, for each paired cell, the flip-to-cDPO ratio at the first logged step, the per-epoch medians of the ratio at matched steps, and the factor by which each arm's norm falls from the first to the last quarter of training.

E Loss-Implementation Audit

All cells use the same eight preference pairs and Qwen3-0.6B in single precision. The policy is deterministically perturbed and the reference is fixed, so margins are nonzero; a policy-equals-reference control has zero margins and cannot detect smoothing. Each cell evaluates $\varepsilon = 0$, 0.1 and 0 again, and both zero-smoothing evaluations agree exactly. For the temperature-scaled margin z, the cDPO loss shift is εz and the robust shift is $-\varepsilon z/(1-2\varepsilon)$ (Chowdhury et al., 2024). The "formula error" column is the deviation from the cDPO formula for native and explicit arms and from the robust formula for robust arms; for the inactive native path of the newest release it equals the full expected shift. Results are in Table 9, with the executed releases ordered from newest to oldest.

F Derivations and Examples

Gradients of the two preference losses. For a scalar margin $s_\theta(x)$,

$$\nabla_\theta \ell_+(s_\theta(x)) = -\big(1-\sigma(s_\theta(x))\big)\,\nabla_\theta s_\theta(x),$$
$$\nabla_\theta \ell_-(s_\theta(x)) = \sigma(s_\theta(x))\,\nabla_\theta s_\theta(x).$$

Table 8: Flip-to-cDPO pre-clipping gradient-norm ratio in the 14 paired cells. First: ratio at the first logged step, step 10. Ep. 1–3: median ratio over the logged steps of each epoch. Fall: first-quarter over last-quarter median norm of each arm.

Model / seed / LR / EB	First	Ep. 1	Ep. 2	Ep. 3	Fall, cDPO	Fall, flip
2B/42/10^{-5}/16	1.28	1.37	3.02	4.10	4.68	1.33
2B/42/10^{-5}/32	1.24	1.36	2.84	4.46	4.74	1.07
2B/42/2×10^{-5}/16	1.35	1.31	3.57	3.06	4.21	1.58
2B/42/2×10^{-5}/32	1.32	1.36	5.44	4.20	4.90	1.33
2B/43/10^{-5}/16	1.20	1.24	2.98	5.40	7.07	1.70
2B/43/10^{-5}/32	1.23	1.23	2.90	6.05	7.81	1.48
2B/43/2×10^{-5}/16	1.21	1.31	3.33	3.60	3.59	2.03
2B/43/2×10^{-5}/32	1.22	1.28	3.38	4.92	8.30	1.74
2B/44/10^{-5}/16	1.16	1.33	3.64	5.72	7.11	1.60
2B/44/10^{-5}/32	1.14	1.23	3.31	5.30	6.13	1.84
2B/44/2×10^{-5}/16	1.20	1.39	4.16	3.69	6.04	1.91
2B/44/2×10^{-5}/32	1.61	1.36	3.57	4.51	7.74	2.11
4B/42/10^{-5}/32	1.21	1.35	2.69	5.10	4.93	1.38
4B/44/10^{-5}/32	1.24	1.32	3.27	5.02	5.80	1.67

Table 9: Loss-implementation audit: maximum absolute loss shift at $\varepsilon = 0.1$ over eight preference pairs, and deviation from the analytic cDPO formula for native and explicit losses or from the robust formula. For the inactive native path of the newest release the deviation equals the full expected shift.

TRL release	Loss	Shift	Formula error
newest	native sigmoid	0	4.71×10^{-4}
newest	explicit cDPO	4.71×10^{-4}	2.69×10^{-8}
newest	robust	5.89×10^{-4}	3.60×10^{-8}
second newest	native sigmoid	4.71×10^{-4}	5.13×10^{-8}
second newest	robust	5.89×10^{-4}	1.01×10^{-7}
third newest	native sigmoid	4.71×10^{-4}	2.69×10^{-8}
third newest	robust	5.89×10^{-4}	3.60×10^{-8}
oldest	native sigmoid	4.71×10^{-4}	2.69×10^{-8}
oldest	robust	5.89×10^{-4}	3.60×10^{-8}

Their difference is $\nabla\ell_-(s) - \nabla\ell_+(s) = \nabla_\theta s_\theta(x)$ for every value of s, which is why the flip noise in Lemma 2 is $(Z_i - p)J_i$ with a coefficient that does not decay with margin saturation.

Momentum and frozen-denominator Adam are affine. With first-moment buffer m, step count t and a denominator D that does not depend on the current gradient, the Adam direction is

$$U(g) = \frac{\beta_1 m + (1-\beta_1)g}{(1-\beta_1^{t+1})D} = b + Cg,$$
$$b = \frac{\beta_1 m}{(1-\beta_1^{t+1})D}, \qquad C = c\,D^{-1}, \qquad c = \frac{1-\beta_1}{1-\beta_1^{t+1}}.$$

The term b is shared by both procedures and does not enter Eq. (4).

The flip-rate peak requires a fixed evaluation objective. If the evaluation objective is the cDPO risk at $\varepsilon = p$, its Hessian depends on p. Take a scalar parameter with $s(\theta) = \theta - \theta^2$ at $\theta = 0$, so $J = 1$ and $s'' = -2$. Then

$$F_p''(0) = \sigma(0)\big(1-\sigma(0)\big)\cdot 1 + \big(p - \tfrac{1}{2}\big)(-2) = 1.25 - 2p,$$

and the coefficient of Eq. (4) is proportional to $p(1-p)(1.25-2p)$, which is maximized on $[0, \frac{1}{2}]$ at $p = 0.25$. With F held fixed, as in all our measurements, the p-dependence is exactly $p(1-p)$.

Remainder terms. Given B and θ, the flip-DPO gradient takes at most 2^m values, so all conditional moments of $\delta = g_{\text{flip}} - g_{\text{cDPO}}$ are finite and the Taylor remainders in Theorem 3 and Proposition 4 are uniform in the realization. The coordinatewise expansion in Eq. (6) requires u_k to be smooth on the segment between $g_{\text{cDPO},k}$ and each realization; with $\epsilon > 0$ and $v_k > 0$ this holds. Since $\mathbb{E}|Z - p|^3 = p(1-p)[p^2 + (1-p)^2]$, the remainder R_k also carries a factor $p(1-p)$; it is small relative to the leading term when $|J_{ik}|/m$ is small relative to the scale $\sqrt{\hat{v}_k}$ on which u_k'' varies, not because p is small. At a fresh state with $\epsilon \to 0$ the map is a sign function, and Eq. (5) still holds with β_B computed exactly, as in the scalar example.

A scalar example of the expected-update bias. Take one example with $s = 0$, $J = 1$ and $p = 0.1$. The two possible gradients are -0.5 without and $+0.5$ with a flip, and the cDPO gradient is -0.4. From a fresh Adam state with $m = v = 0$ and $\epsilon \to 0$ the step is $\operatorname{sign}(g)$: the cDPO step has length η, whereas the expected flip step is

$$0.9\,(-\eta) + 0.1\,(+\eta) = -0.8\,\eta.$$

The expected updates therefore differ by 0.2η, a first-order difference. With a frozen denominator the step is linear in g and the expected updates coincide.

Reused labels along a trajectory. Along a reused-label trajectory the state depends on the draw, so the conditional match of Proposition 1 is lost: with one example, $s(\theta) = \theta$ and $\theta_0 = 0$, one SGD step gives

$$\mathbb{E}\big[\ell_Z(\theta_1) - L_{\text{cDPO}}(\theta_1)\big] = -\eta\, p(1-p).$$

Proof of Corollary 5. Write $u(g) = g/\|g\|$ and $r = \|g\|$. Then

$$Du(g)[\delta] = \frac{\delta}{r} - \frac{g\,(g^\top \delta)}{r^3},$$
$$D^2u(g)[\delta,\delta] = -\frac{2\delta\,(g^\top\delta)}{r^3} - \frac{g\,\|\delta\|^2}{r^3} + \frac{3g\,(g^\top\delta)^2}{r^5}.$$

Projecting on $\hat{g} = g/r$ gives

$$\hat{g}^\top D^2u(g)[\delta,\delta] = \frac{(\hat{g}^\top\delta)^2 - \|\delta\|^2}{r^2} = -\frac{\|P_\perp\delta\|^2}{r^2}.$$

A second-order expansion of $U = c\,u$ at g_{cDPO} with $\mathbb{E}[\delta \mid B] = 0$ and $\mathbb{E}[\delta\delta^\top \mid B] = \Sigma$ from Lemma 2 yields Eq. (7); the remainder is bounded by the third derivative of u, which scales as r^{-3}, times $\mathbb{E}\|\delta\|^3$. The component of β_B orthogonal to $\hat{g}$ is $-c\,P_\perp\Sigma\hat{g}/r^2$ to the same order and vanishes when $\hat{g}$ is an eigenvector of Σ. For a clip that is active only on part of the support, U is not differentiable at the threshold, but Eq. (5) still holds because U is bounded.

Summary of the local predictions. Table 10 collects the leading-order predictions of Section 3.

G Accumulated Flip Excess over a Schedule

Noise curricula and robustness ablations that compare stochastic flips against smoothing measure two things at once: the change of objective, which Proposition 1 removes when $\varepsilon = p$, and the change of optimizer noise, which Lemma 2 and Proposition 4 quantify. For a schedule (η_t, p_t, m_t) with flips redrawn at every visit and a *memoryless* affine update $U_t(g) = b_t + C_t g$ with b_t, C_t fixed in advance, such as SGD with a fixed preconditioner, the deviations injected at different steps are independent in the frozen-gradient model of Proposition 6, and the leading flip excess of F at the final iterate is

$$\sum_t \frac{\eta_t^2\, p_t(1-p_t)}{2m_t^2} \sum_{i \in B_t} (C_tJ_i)^\top \nabla^2 F\,(C_tJ_i), \tag{9}$$

a *covariance budget* that a curriculum spends independently of the objective it targets. Two schedules with the same time-averaged p can therefore have the same expected objective but different budgets: annealing p from 0.4 to 0 spends most of the budget early, when learning rates after warm-up

Table 10: Local predictions for the one-step excess Δ_B at fixed (θ, B, F) and optimizer state, to leading order in η.

Quantity varied	Affine update: SGD, momentum, frozen D	Adaptive second moment: Adam, AdamW
Step size η	$\Delta_B = O(\eta^2)$, Eq. (4)	$\Delta_B = O(\eta)$, first order iff $\nabla F^\top \beta_B \neq 0$, Eq. (5)
Flip rate p	η^2 coefficient exactly $\propto p(1-p)$	vanishes at $p \in \{0,1\}$; $\propto p(1-p)$ with coefficient $u_k''/2$ when $\lvert J_{ik}\rvert/m \ll \sqrt{\hat{v}_k}$
Batch size m	$\propto m^{-2} \sum_i (CJ_i)^\top \nabla^2 F (CJ_i)$	$\propto m^{-2} \sum_i J_{ik}^2$ per coordinate
Freezing the denominator	no change	removes β_B without an active clip
Active global-norm clip	first order, Cor. 5	first order, Cor. 5
Larger of the two excesses	live denominator iff $\eta < \eta^* = \lvert a_1\rvert/\lvert A_2\rvert$, Section 4.2	
Margin saturation	flip coefficient not suppressed	flip coefficient not suppressed
Reused flips over E epochs	accumulated excess $\times E$, Prop. 6	–

are largest and, for Adam, when the current gradient contributes most to the second moment. A curriculum that "helps" may do so by changing the target objective, by changing the budget in Eq. (9), or both. The matched cDPO control with $\varepsilon_t = p_t$ holds the first fixed and removes the second, which is the comparison needed to attribute a gain to the noise itself. An update with memory carries each deviation forward, and then Eq. (9) is not the final-iterate excess: with heavy-ball momentum μ and a constant step, the deviation injected at step s reaches the final iterate T multiplied by $\kappa_s = (1-\mu^{T-s+1})/(1-\mu)$, so its term is multiplied by $\kappa_s^2 \leq (1-\mu)^{-2}$; for an exponential moving average with weight $1-\beta_1$, which is Adam's first moment without bias correction, $\kappa_s = 1-\beta_1^{T-s+1} \leq 1$.

H Registered Endpoints

Tables 11 and 12 list every registered endpoint of the one-step experiments on every completed cell, in the order of registration; E-ONESTEP4 and E-ONESTEP4b are evaluated separately. The endpoints Q1a, Q1b, Q2b and the secondary part of Q3 concern the coordinatewise expansion Eq. (6) and its next-order criterion

$$\hat{R} = \frac{\left|\nabla F^\top \left(u'''\, \mathbb{E}\delta^3/6 + u''''\, \mathbb{E}\delta^4/24\right)\right|}{\left|\nabla F^\top \beta^{(6)}\right|},$$

where $\beta^{(6)}$ is the leading term of Eq. (6), the moments of δ_k are closed-form in p and J_{ik}, and u''', u'''' are central differences of the analytic u''; its cost is $O(mD)$. $\hat{R}$ compares two finite-order terms and is an indicator, not a bound on the remainder.

Table 13 decomposes the within-live endpoints W1–W3 of E-ONESTEP5 by state and update map for both models, and W1 by the distance of the step from the crossover $\eta^\dagger$, as a descriptive breakdown.

Table 11: All registered endpoints of the one-step experiments on every completed cell. Ratios count pairs of cell and $p > 0$ at the smallest step unless a step is given; slopes count pairs of cell and p, or cells at $p = 0.1$ where marked *. A count must reach 80% unless stated. †Cross-implementation diagnostic: E-ONESTEP deltas with curvature from the E-ONESTEP2 computation.

Exp.		Update rule	Criterion	Result	Met
E-ONESTEP	H1	SGD, $m = 2, 4, 8$	ratio in $[0.8, 1.2]$; median $\lvert\text{ratio}-1\rvert$ falls with η	35/35; 0.129→0.032	yes
	H2	SGD	$R^2 \geq 0.95$; ratio rule per m	7/7; 15/15, 15/15, 5/5	yes
	H3	Adam, live, warm	slope* in $[0.7, 1.3]$; ratio in $[0.8, 1.2]$	2/3; 8/15	no
	H3	Adam, live, fresh	as H3	1/3; 11/15	no
	H4	Adam, frozen D	slope* in $[1.7, 2.3]$; ratio in $[0.8, 1.2]$	3/3; 15/15	yes
E-ONESTEP2	A1†	SGD, frozen D	E-ONESTEP ratio with E-ONESTEP2 exact curvature in $[0.9, 1.1]$	50/50	yes
	A2†	SGD, batch 2	as A1, in $[0.95, 1.05]$ for all p	2/2 cells	yes
	B1	clip, SGD	slope* in $[0.7, 1.3]$ in at least 2 of 3; ratio in $[0.8, 1.2]$	2/3; 8/15	no
	B2	clip, frozen D	as B1	2/3; 9/15	no
	B3	clip, SGD	Cor. 5 over exact in $[0.8, 1.2]$; $\hat{g}^\top \beta_B < 0$ in all	4/15; 15/15	no
	C1	Adam, live, training	ratio at 5×10^{-6} in $[0.8, 1.2]$; slope in $[0.7, 1.3]$	15/15; 15/15	yes
	C2	Adam, frozen D, training	slope in $[1.7, 2.3]$; ratio in $[0.8, 1.2]$; $\lvert\Delta_{\text{frozen}}\rvert < \lvert\Delta_{\text{live}}\rvert$ at 2×10^{-5}	15/15; 15/15; 14/15	yes
	D	SGD, $m = 4$	as H1	15/15	yes
	D	frozen D, $m = 4$; $m = 2$	as H4	3/3, 15/15; 3/3, 15/15	yes
	D	Adam, live, $m = 2$	as H3	2/3; 6/15	no
	D	Adam, fresh, $m = 2$	as H3	2/3; 11/15	no
E-ONESTEP3	R1	SGD; frozen D	ratio in $[0.9, 1.1]$; slope in $[1.7, 2.3]$; $R^2 \geq 0.95$	35/35, 35/35; 26/30, 30/30	yes
	R2	SGD, batch 2	ratio in $[0.95, 1.05]$ for all p	2/2 cells	yes
	R3	Adam, live, training	ratio at 2.5×10^{-6} in $[0.9, 1.1]$; slope in $[0.7, 1.3]$; median error falls with η	15/15; 15/15; yes	yes
	R4	Adam, frozen D, training	slope in $[1.7, 2.3]$; ratio at 2.5×10^{-6} in $[0.9, 1.1]$; $\lvert\Delta_{\text{frozen}}\rvert < \lvert\Delta_{\text{live}}\rvert$ at 5×10^{-6}	15/15; 15/15; 15/15	yes
E-ONESTEP4	Q1a	Eq. (6)	where $\hat{R} \leq 0.1$: relative error ≤ 0.1	0/2	no
six states	Q1b	Eq. (6)	where $\hat{R} > 1$: relative error > 0.1	67/74	yes
	Q2a	Adam, live	ratio to enumerated first-order term at $\eta^*_{\text{ref}}/64$ in $[0.9, 1.1]$	89/90	yes
	Q2b	Adam, live	where $\hat{R} \leq 0.1$: ratio to Eq. (6) term in $[0.8, 1.2]$	0/2	no
	Q3	live vs frozen D	predicted larger term correct outside $\eta/\eta^* \in [0.5, 2]$, at least 90%; secondary, η^* from Eq. (6) where $\hat{R} \leq 0.1$	599/607; 13/13	yes
	Q4a	clip, live	ratio at $\eta^*_{\text{ref}}/64$ in $[0.9, 1.1]$	90/90	yes
	Q4b	clip, frozen D	ratio at $\eta^*_{\text{ref}}/64$ in $[0.9, 1.1]$	89/90	yes
	Q4c	clip, frozen D	slope over three smallest relative-grid steps, without the warm-up learning rate, in $[0.7, 1.3]$	88/90	yes
	Q4d	Adam, frozen D	slope over the same steps in $[1.7, 2.3]$	90/90	yes
E-ONESTEP4b	Q1a	Eq. (6)	as E-ONESTEP4	0/1	no
Qwen3-1.7B	Q1b	Eq. (6)	as E-ONESTEP4	75/84	yes
	Q2a	Adam, live	as E-ONESTEP4	89/90	yes
	Q2b	Adam, live	as E-ONESTEP4	0/1	no
	Q3	live vs frozen D	as E-ONESTEP4; primary; secondary	582/598; 6/7	yes; no
	Q4a–d	clip; frozen D	as E-ONESTEP4	90/90, 90/90, 90/90, 87/90	yes

Table 12: Registered endpoints, continued from Table 11: the within-live comparison and the sampled estimate of E-ONESTEP5 and E-ONESTEP6, the independent data panel of E-ONESTEP7, grouped resampling in E-BOOT8 and the random-direction control of E-ISO9. Entries for the two models are separated by a semicolon, Qwen3-0.6B first.

Exp.		Update rule	Criterion	Result	Met
E-ONESTEP5	W1	Adam, live, $\pm$clip	first-order term larger iff $\eta < \eta^\dagger$, outside $\eta/\eta^\dagger \in [0.5, 2]$, at least 90%	661/721; 606/722	yes; no
0.6B; 1.7B	W2	Adam, live, $\pm$clip	$(\Delta_B + \eta a_1)/(\eta^2 B_2) \in [0.8, 1.2]$ at $\eta^\dagger_{\mathrm{ref}}/16$	155/180; 139/180	yes; no
	W3	Adam, live, $\pm$clip	two-term error below one-term error, two smallest steps, at least 90%	340/360; 322/360	yes; no
	M1	estimate, $m = 8$	95% interval covers enumerated a_1, at least 90%	169/180; 171/180	yes; yes
	M1b	estimate, $m = 4$	as M1, against registered a_1; descriptive, as all strata are evaluated in full with zero half-width, Appendix C	174/180; 180/180	–
	M2	estimate, $m = 8$	accepted: relative error ≤ 0.2, at least 90%	115/116; 114/115	yes; yes
	M4	live vs frozen D, F_2	larger term from $\hat{a}_1$ correct outside band, at least 90%	69/69; 60/65	yes; yes
	M5	Adam, live, F_2	accepted: $\Delta_B/(-\eta\hat{a}_1) \in [0.8, 1.2]$ at $\eta^*_{\mathrm{ref}}/8$	18/18; 13/13	yes; yes
	M6	live, clip 1.0, F_2	as M5, where clip 1.0 is active on the whole support	7/7; 5/5	yes; yes
	M7	Adam, frozen D, F_2	$\Delta_B/(\eta^2 A_2) \in [0.8, 1.2]$ at $\eta^*_{\mathrm{ref}}/8$	30/30; 30/30	yes; yes
	S1	estimate, $m = 16$	coverage, at least 90%; accepted relative error ≤ 0.2, at least 90%	15/15, 6/6; 14/15, 3/3	yes; yes
E-ONESTEP6	C1	estimate, $m = 8$,	pooled coverage ≥ 0.93; cases with coverage ≥ 0.90, at least 90%	0.951, 180/180; 0.948, 177/180	yes; yes
0.6B; 1.7B	C2	1,000 repetitions	accepted: coverage ≥ 0.93; relative error ≤ 0.2, at least 95%	0.949, 0.996; 0.945, 0.995	yes; yes
	C3	per case	median SE ratio in $[0.8, 1.25]$	0.997; 1.001	yes; yes
	C4		mean within 3 Monte Carlo SE of a_1, at least 95% of cases	180/180; 179/180	yes; yes
	B1	plain Monte Carlo	equal-budget RMSE ratio > 1, at least 80% of cases	161/180; 170/180	yes; yes
	B2	stratified, no c.v.	as B1, with the equal-budget allocation of Appendix C	143/180; 158/180	no; yes
E-ONESTEP7	P1	live vs frozen D	as Q3 primary, independent panel	582/617; 590/604	yes; yes
0.6B; 1.7B	P2	Adam, live	as Q2a	87/90; 90/90	yes; yes
	P3	clip, live; clip, frozen D	as Q4a; as Q4b	86/90, 86/90; 89/90, 88/90	yes; yes
	P4	clip, frozen D	as Q4c	86/90; 87/90	yes; yes
	P5	Adam, frozen D	as Q4d	81/90; 89/90	yes; yes
E-BOOT8	BT1	grouped resampling	lower end of the 95% interval of Q3/P1 accuracy > 0.5 for E-ONESTEP4, 4b and E-ONESTEP7 at 0.6B, 1.7B	0.971, 0.956; 0.909, 0.955	yes
E-ISO9	I1	SGD, $m = 4$, noise along J_i	ratio to Eq. (4) at the smallest step in $[0.9, 1.1]$	15/15	yes
	I2	SGD, $m = 4$, random directions	as I1, noise along $\|J_i\| r_i$, ratio in $[0.8, 1.2]$	15/15	yes

Table 13: Within-live endpoints of E-ONESTEP5 by state and update map: a descriptive decomposition of the registered counts W1–W3 of Table 12; live: without clip; clip: with the registered clip. W1: the first-order term is correctly predicted to be the larger part of the live excess, outside $\eta/\eta^\dagger \in [0.5, 2]$; W2: $(\Delta_B + \eta a_1)/(\eta^2 B_2) \in [0.8, 1.2]$ at the smallest step; W3: the two-term expansion is closer than the first-order term, at the two smallest steps.

Model	State	W1 live	W1 clip	W2 live	W2 clip	W3 live	W3 clip
0.6B	4 blocks, 10 steps	58/60	56/60	14/15	8/15	29/30	21/30
0.6B	4 blocks, 40 steps	58/60	48/60	15/15	13/15	30/30	28/30
0.6B	4 blocks, 40 steps, clip	56/60	47/61	10/15	11/15	29/30	26/30
0.6B	LoRA, 10 steps	60/60	55/60	15/15	10/15	30/30	27/30
0.6B	LoRA, 40 steps	55/60	59/60	15/15	15/15	30/30	30/30
0.6B	LoRA, 40 steps, clip	58/60	51/60	15/15	14/15	30/30	30/30
0.6B	W1 by $\lvert \log_2(\eta/\eta^\dagger) \rvert$ in $(1, 2]$, $(2, 3]$, $(3, \infty)$: 187/211, 116/134, 358/376						
1.7B	4 blocks, 10 steps	60/60	55/61	15/15	7/15	27/30	26/30
1.7B	4 blocks, 40 steps	54/60	35/60	15/15	15/15	30/30	30/30
1.7B	4 blocks, 40 steps, clip	41/60	38/60	15/15	15/15	30/30	30/30
1.7B	LoRA, 10 steps	60/60	51/61	9/15	5/15	30/30	20/30
1.7B	LoRA, 40 steps	58/60	55/60	13/15	10/15	30/30	30/30
1.7B	LoRA, 40 steps, clip	59/60	40/60	15/15	5/15	20/30	19/30
1.7B	W1 by $\lvert \log_2(\eta/\eta^\dagger) \rvert$ in $(1, 2]$, $(2, 3]$, $(3, \infty)$: 175/213, 123/140, 308/369						