An Improved Analysis of the Empirical Convergence Rate of DPO-Mix-R

Claude Opus 4.8, GPT 5.5, Ruizhe Shi

June 30, 2026

Abstract

This note gives a high-probability/moment refinement of the empirical DPO-Mix-R analysis in the tabular bandit setting of *The Crucial Role of Samplers in Online Direct Preference Optimization*. In the exact-gradient setting, the DPO-Mix-R sampler cancels the linear term in the pairwise error update, leaving a quadratic one-step exact-gradient error. With empirical gradients, this quadratic map is perturbed by centered sub-Gaussian noise of scale σ. Under the same moment-stability input used in the empirical proof, we show that empirical DPO-Mix-R reaches the $O(\sigma)$ root-mean-square error floor after a deterministic horizon of order $O(\log\log(1/\sigma))$.

Summary of the refinement. The original theorem [1] proves that, at $T = \lfloor \log(1/\sigma) \rfloor$,

$$\sqrt{\mathbb{E}\,\delta(a, a'; \theta^{(T)})^2} \le O(\sigma).$$

The observation here is that the transient is shorter. On a high-probability event the empirical dynamics obey

$$\|\delta_{t+1}\|_\infty \le K\|\delta_t\|_\infty^2 + \varepsilon, \qquad \varepsilon \asymp \sigma\sqrt{\log(1/\sigma)},$$

so the quadratic phase reaches the ε-window in $O(\log\log(1/\sigma))$ steps. One final fresh empirical update then gives an L^2 contribution of order σ^2, while the deterministic quadratic remainder is only $O(\varepsilon^4)$. The complement of the good event is absorbed by choosing its probability to be of order σ^4 and using a fourth-moment stability bound.

1 Setup and notation

Let $\mathcal{Y}$ be a finite action set and set $A = |\mathcal{Y}|$. Rewards are normalized as

$$r(a) \in [0, 1], \qquad a \in \mathcal{Y}.$$

For a tabular policy π_θ, define the relative logit

$$q_\theta(a) := \log \frac{\pi_\theta(a)}{\pi_{\mathrm{ref}}(a)}.$$

The pairwise error studied in the paper is

$$\delta(a, a'; \theta) := r(a) - r(a') - \beta\big(q_\theta(a) - q_\theta(a')\big). \tag{1}$$

Equivalently, under the common normalization $\pi_{\text{ref}} = \pi_{\theta^{(0)}}$ and centered logits, this is

$$\delta(a, a'; \theta) = r(a) - r(a') - \beta(\theta_a - \theta_{a'}).$$

We also use

$$\Delta(a, a'; \theta) := \sigma\big(r(a) - r(a')\big) - \sigma\big(\beta(q_\theta(a) - q_\theta(a'))\big), \tag{2}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$.

Let

$$I := \mathcal{Y} \times \mathcal{Y}, \qquad P := |I| \le A^2.$$

For $i = (a, a') \in I$, write the time-t pairwise error as

$$\delta_t(i) := \delta(a, a'; \theta^{(t)}), \qquad \|\delta_t\|_\infty := \max_{i \in I} |\delta_t(i)|.$$

The initialization is $\pi_{\theta^{(0)}} = \pi_{\text{ref}}$, so $\|\delta_0\|_\infty \le 1$. Let $\mathcal{F}_t$ be the filtration generated by all empirical-gradient randomness up to time $t - 1$; then $\theta^{(t)}$ and δ_t are $\mathcal{F}_t$-measurable.

2 One exact-gradient step for DPO-Mix-R

DPO-Mix-R uses two components:

$$\text{(i) } \pi^{s1} = \pi^{s2} = \text{Unif}(\mathcal{Y}), \qquad \text{(ii) } \pi^{s1}(a) \propto e^{r(a)}, \quad \pi^{s2}(a) \propto e^{-r(a)}.$$

With the sampling coefficients in the paper, these two parts implement the factor

$$\frac{1}{\sigma'(r(a) - r(b))} = 2 + e^{r(a)-r(b)} + e^{-(r(a)-r(b))}.$$

Thus the mixed exact gradient has the same reweighted form as in Appendix A.2 of the paper. Taylor expansion at the reward difference gives

$$\Delta(a, b; \theta) = \sigma'(r(a) - r(b))\delta(a, b; \theta) - \frac{\sigma''(\xi_R(a, b; \theta))}{2}\delta(a, b; \theta)^2, \tag{3}$$

where $\xi_R(a, b; \theta)$ lies between $r(a) - r(b)$ and $\beta(q_\theta(a) - q_\theta(b))$.

Since $r(a) - r(b) \in [-1, 1]$, define

$$K := \frac{\sup_{x \in \mathbb{R}} |\sigma''(x)|}{\inf_{|x| \le 1} \sigma'(x)} = \frac{1/(6\sqrt{3})}{\sigma'(1)} < \frac{1}{2}. \tag{4}$$

The precise numerical value is unimportant; what matters is that K is an absolute constant smaller than $1/2$.

Lemma 1 (Exact quadratic exact-gradient)**.** *For the exact-gradient DPO-Mix-R update with learning rate $\eta = 1/(\beta^2 A)$, define $\theta_{\text{ex}}^{(t+1)}$ to be the parameter obtained by applying one exact-gradient DPO-Mix-R step from the current empirical iterate $\theta^{(t)}$, with the same learning rate $\eta = 1/(\beta^2 A)$. Define the corresponding one-step exact-gradient pairwise error by*

$$\delta_{t+1}^{\text{ex}}(i) := \delta(i; \theta_{\text{ex}}^{(t+1)}), \qquad i \in I.$$

Then

$$|\delta_{t+1}^{\text{ex}}(i)| \le K\|\delta_t\|_\infty^2, \qquad i \in I. \tag{5}$$

Proof. By the DPO-Mix-R reweighting and (3), the linear term in δ is exactly canceled by the learning rate $\eta = 1/(\beta^2 A)$. For $i = (a, a')$, the one-step exact-gradient pairwise error has the form

$$\delta_{t+1}^{\mathrm{ex}}(a, a') = \frac{1}{2A} \sum_{b \in \mathcal{Y}} \left(\frac{\sigma''(\xi_R(a, b; \theta^{(t)}))}{\sigma'(r(a) - r(b))} \delta_t(a, b)^2 - \frac{\sigma''(\xi_R(a', b; \theta^{(t)}))}{\sigma'(r(a') - r(b))} \delta_t(a', b)^2 \right).$$

Using (4) and $|\delta_t| \le \|\delta_t\|_\infty$ gives

$$|\delta_{t+1}^{\mathrm{ex}}(a, a')| \le \frac{K}{2A} \sum_{b \in \mathcal{Y}} \big(\delta_t(a, b)^2 + \delta_t(a', b)^2 \big) \le K \|\delta_t\|_\infty^2.$$

□

3 Empirical noise and moment stability

In empirical DPO, the exact gradient is replaced by an unbiased empirical gradient $G^{(t)}$. The centered coordinate error is assumed to satisfy the empirical-gradient condition from the paper: conditionally on $\mathcal{F}_t$,

$$\frac{G_a^{(t)} - \mathbb{E}[G_a^{(t)} \mid \mathcal{F}_t]}{\beta A} \quad \text{is sub-Gaussian with proxy variance } \sigma^2. \tag{6}$$

For $i = (a, a')$, define the induced pairwise noise

$$\nu_t(i) := \frac{(G_a^{(t)} - G_{a'}^{(t)}) - \big(\mathbb{E}[G_a^{(t)} \mid \mathcal{F}_t] - \mathbb{E}[G_{a'}^{(t)} \mid \mathcal{F}_t] \big)}{\beta A}. \tag{7}$$

Then the empirical update decomposes into the one-step exact-gradient error plus the centered empirical-gradient noise:

$$\delta_{t+1}(i) = \delta_{t+1}^{\mathrm{ex}}(i) + \nu_t(i), \qquad i \in I. \tag{8}$$

Moreover, by standard closure properties of sub-Gaussian variables, there are absolute constants $c_0, c_2 > 0$ such that, for every $u > 0$,

$$\mathbb{E}[\nu_t(i) \mid \mathcal{F}_t] = 0, \tag{9}$$

$$\mathbb{P}(|\nu_t(i)| > u \mid \mathcal{F}_t) \le 2 \exp\left(-\frac{u^2}{c_0 \sigma^2} \right), \tag{10}$$

$$\mathbb{E}[\nu_t(i)^2 \mid \mathcal{F}_t] \le c_2 \sigma^2. \tag{11}$$

The high-probability argument controls the dynamics only on a good event. To convert it into an unconditional root-mean-square statement, we need the following fourth-moment stability input.

Assumption 1 (Fourth-moment stability)**.** *For the deterministic horizon H under consideration, there exists a constant $B_4 < \infty$, independent of σ, such that*

$$\max_{0 \le t \le H} \max_{i \in I} \mathbb{E}|\delta_t(i)|^4 \le B_4^4. \tag{12}$$

Remark 1 (Where Assumption 1 comes from). This is the same stability input supplied by the L^{2n} moment induction in the empirical proof of the paper. In that proof, one obtains bounds of the form

$$\mathbb{E}|\delta_t(i)|^{2n} \le \big(12\sqrt{n}\, \sigma + 2^{-t} \big)^{2n}$$

whenever $n 2^t \lesssim 1/\sigma$. Taking $n = 2$ gives a uniform fourth-moment bound for every horizon $H = O(\log\log(1/\sigma))$. The analysis below uses only the consequence (12), so the proof is separated from the original moment induction.

4 Good event and log-log hitting time

Fix a deterministic horizon $H \geq 1$ and a failure budget $p \in (0,1)$. Define

$$\rho := \sqrt{c_0 \log \frac{2PH}{p}}, \qquad \varepsilon := \rho\sigma, \tag{13}$$

and the good event

$$\mathcal{G} := \{|\nu_t(i)| \leq \varepsilon \text{ for all } 0 \leq t < H,\ i \in I\}. \tag{14}$$

Lemma 2 (Good-event recursion)**.** *Under* (10)*,* $\mathbb{P}(\mathcal{G}) \geq 1 - p$*. Moreover, on* $\mathcal{G}$*,*

$$\|\delta_{t+1}\|_\infty \leq K\|\delta_t\|_\infty^2 + \varepsilon, \qquad 0 \leq t < H. \tag{15}$$

Proof. For each fixed (t,i), (10) and (13) give

$$\mathbb{P}(|\nu_t(i)| > \varepsilon \mid \mathcal{F}_t) \leq 2\exp(-\rho^2/c_0) = \frac{p}{PH}.$$

A union bound over H times and P ordered pairs gives $\mathbb{P}(\mathcal{G}^c) \leq p$. On $\mathcal{G}$, combine (5) and (8), then maximize over i. □

Set

$$B := \sqrt{\frac{\varepsilon}{K}}. \tag{16}$$

On $\mathcal{G}$, define the hitting time

$$\tau := \inf\{t \geq 0 : \|\delta_t\|_\infty < B\}.$$

Lemma 3 (Log-log hitting time)**.** *Assume* $\|\delta_0\|_\infty \leq 1$ *and* $2\sqrt{K\varepsilon} < 1$*. On* $\mathcal{G}$*,*

$$\tau \leq L(H,p,\sigma) := \left\lceil \log_2 \left(\frac{\log(1/(2\sqrt{K\varepsilon}))}{\log(1/(2K))} \right) \right\rceil. \tag{17}$$

Consequently, if $H \geq L(H,p,\sigma) + 2$*, then* $\tau + 2 \leq H$ *on* $\mathcal{G}$*.*

Proof. As long as $\|\delta_t\|_\infty \geq B$, the noise term is dominated:

$$\varepsilon \leq K\|\delta_t\|_\infty^2.$$

Thus (15) gives $\|\delta_{t+1}\|_\infty \leq 2K\|\delta_t\|_\infty^2$. Let $z_t := 2K\|\delta_t\|_\infty$. During this phase,

$$z_{t+1} = 2K\|\delta_{t+1}\|_\infty \leq (2K\|\delta_t\|_\infty)^2 = z_t^2.$$

Since $z_0 \leq 2K < 1$, induction gives $z_t \leq (2K)^{2^t}$ before the hitting time. The inequality $\|\delta_t\|_\infty < B$ is equivalent to $z_t < 2\sqrt{K\varepsilon}$. Hence it is enough to have

$$(2K)^{2^t} \leq 2\sqrt{K\varepsilon},$$

which is exactly (17). □

5 Reaching the RMS noise floor

The key point is that the second-moment estimate is taken after the good-event pathwise descent. We do not use an unconditional recursion and then try to replace the sup-norm fourth moment by pairwise second moments.

Lemma 4 (Two-step descent to the empirical floor)**.** *Assume $H \geq \tau + 2$ on $\mathcal{G}$ and $4K\varepsilon^2 \leq \varepsilon$. Then, on $\mathcal{G}$,*

$$\|\delta_t\|_\infty \leq 2\varepsilon, \qquad t = \tau + 1, \ldots, H. \tag{18}$$

Furthermore, for every ordered pair $i \in I$,

$$\mathbb{E}[\delta_H(i)^2 \mathbf{1}_{\mathcal{G}}] \leq 32K^2\varepsilon^4 + 2c_2\sigma^2. \tag{19}$$

Proof. At the hitting time, $\|\delta_\tau\|_\infty < B$, and hence

$$\|\delta_{\tau+1}\|_\infty \leq KB^2 + \varepsilon = 2\varepsilon.$$

If $\|\delta_t\|_\infty \leq 2\varepsilon$ for some $t \geq \tau + 1$, then

$$\|\delta_{t+1}\|_\infty \leq K(2\varepsilon)^2 + \varepsilon = 4K\varepsilon^2 + \varepsilon \leq 2\varepsilon.$$

This proves (18). Since $H \geq \tau + 2$, the lock-in bound holds at time $H - 1$. Therefore, on $\mathcal{G}$,

$$|\delta_H^{\mathrm{ex}}(i)| \leq K\|\delta_{H-1}\|_\infty^2 \leq 4K\varepsilon^2.$$

Using $\delta_H(i) = \delta_H^{\mathrm{ex}}(i) + \nu_{H-1}(i)$ and $(x+y)^2 \leq 2x^2 + 2y^2$,

$$\delta_H(i)^2 \mathbf{1}_{\mathcal{G}} \leq 2(4K\varepsilon^2)^2 \mathbf{1}_{\mathcal{G}} + 2\nu_{H-1}(i)^2 \mathbf{1}_{\mathcal{G}}.$$

Taking expectations and applying (11) gives (19). □

Lemma 5 (Bad-event absorption)**.** *Under Assumption 1, for every ordered pair $i \in I$,*

$$\mathbb{E}[\delta_H(i)^2 \mathbf{1}_{\mathcal{G}^c}] \leq B_4^2 \sqrt{p}. \tag{20}$$

Proof. By Cauchy-Schwarz and Assumption 1,

$$\mathbb{E}[\delta_H(i)^2 \mathbf{1}_{\mathcal{G}^c}] \leq \left(\mathbb{E}|\delta_H(i)|^4\right)^{1/2} \mathbb{P}(\mathcal{G}^c)^{1/2} \leq B_4^2 \sqrt{p}.$$

□

6 Main theorem

Theorem 1 (Log-log transient for empirical DPO-Mix-R)**.** *Assume the empirical DPO-Mix-R update satisfies* (8)*, the quadratic exact-gradient bound* (5)*, the noise conditions* (9)–(11)*, and Assumption 1 up to a deterministic horizon H. Suppose $\|\delta_0\|_\infty \leq 1$ and $\sigma < B_4$. Set*

$$p := \frac{\sigma^4}{B_4^4}, \qquad \rho := \sqrt{c_0 \log \frac{2PH}{p}}, \qquad \varepsilon := \rho\sigma. \tag{21}$$

If

$$H \geq L(H, p, \sigma) + 2, \qquad 2\sqrt{K\varepsilon} < 1, \qquad 4K\varepsilon^2 \leq \varepsilon, \tag{22}$$

then for every $a, a' \in \mathcal{Y}$,

$$\sqrt{\mathbb{E}\big[\delta(a, a'; \theta^{(H)})^2\big]} \le C\sigma, \tag{23}$$

where one may take

$$C^2 = 2c_2 + 1 + 32K^2\rho^4\sigma^2. \tag{24}$$

In particular, along the deterministic log-log horizons in Corollary 1, the last term in (24) *is* $o(1)$.

Proof. By Lemma 3, the first condition in (22) implies $\tau + 2 \le H$ on $\mathcal{G}$. Hence Lemma 4 gives

$$\mathbb{E}[\delta_H(i)^2 \mathbf{1}_{\mathcal{G}}] \le 32K^2\varepsilon^4 + 2c_2\sigma^2 = \left(32K^2\rho^4\sigma^2 + 2c_2\right)\sigma^2.$$

By Lemma 5 and the choice of p,

$$\mathbb{E}[\delta_H(i)^2 \mathbf{1}_{\mathcal{G}^c}] \le B_4^2\sqrt{\sigma^4/B_4^4} = \sigma^2.$$

Adding the good-event and bad-event contributions yields

$$\mathbb{E}[\delta_H(i)^2] \le \left(32K^2\rho^4\sigma^2 + 2c_2 + 1\right)\sigma^2.$$

Taking square roots proves the claim. □

Corollary 1 (Existence of an $O(\log\log(1/\sigma))$ deterministic horizon)**.** *Fix finite* A *and constants* K, c_0, c_2, B_4. *There exist constants* $C_H, \sigma_0 > 0$, *depending only on these quantities, such that for every* $0 < \sigma \le \sigma_0$, *the horizon*

$$H = \left\lceil C_H \log\log \frac{e^e}{\sigma} \right\rceil \tag{25}$$

satisfies the hypotheses of Theorem 1. Consequently, for every $a, a' \in \mathcal{Y}$,

$$\sqrt{\mathbb{E}\big[\delta(a, a'; \theta^{(H)})^2\big]} \le C\sigma.$$

Proof. With $p = \sigma^4/B_4^4$,

$$\rho^2 = c_0 \log \frac{2PHB_4^4}{\sigma^4} = O\left(\log\frac{1}{\sigma} + \log H\right) = O\left(\log\frac{1}{\sigma}\right)$$

for $H = O(\log\log(1/\sigma))$. Thus

$$\varepsilon = \rho\sigma = O\left(\sigma\sqrt{\log\frac{1}{\sigma}}\right) \to 0.$$

The two smallness conditions in (22) therefore hold for all sufficiently small σ. Moreover,

$$L(H, p, \sigma) = O\left(\log\log\frac{1}{\sigma}\right).$$

Choosing C_H larger than the implicit constant ensures $H \ge L(H, p, \sigma) + 2$ for all sufficiently small σ. Finally, the fourth-moment stability input holds for such horizons by the original moment induction, since $H = O(\log\log(1/\sigma))$ is much smaller than the range allowed by $n2^H \lesssim 1/\sigma$ with $n = 2$. The result follows from Theorem 1. □

References

[1] Ruizhe Shi, Runlong Zhou, and Simon S. Du. The crucial role of samplers in online direct preference optimization. *International Conference on Learning Representations*, 2025.