Preference Aligned Visuomotor Diffusion Policies for Deformable Object Manipulation

RKO Convergence Analysis

We provide a concise stability and convergence analysis for RKO by starting from the standard DPO contraction-style analysis in the tabular-softmax bandit setting Shi et al. (2024), first reformulating it for KTO and then extending it to RKO: we first rewrite KTO in KL/log-ratio form with a single uniform sampler, define a centered implied-reward error that vanishes at the KL-regularized optimum, and obtain a stable linear error recursion (linear convergence under exact gradients). RKO then differs only by bounded stop-gradient similarity weights, which preserve the same recursion up to rescaled constants.

Setup and assumptions (bandit abstraction)

We adopt the standard single-context bandit abstraction used in DPO convergence analyses: a finite action set A of size K, a tabular softmax policy

$$\pi_\theta(a) \;=\; \frac{\exp(\theta_a)}{\sum_{b\in A}\exp(\theta_b)},$$

bounded ground-truth rewards $r(a) \in [0,1]$, and a full-support reference policy $\pi_{\mathrm{ref}}(a) > 0$ so that KL/log-ratio terms are well-defined.

Explicit boundedness assumptions. To make stability claims precise, we assume:

1. (**Log-ratio boundedness**) The iterates θ^t remain in a compact set such that

   $$|\log(\pi_\theta(a)/\pi_{\mathrm{ref}}(a))| \le C \quad \forall a, \theta,$$

   which can be enforced by trust-region updates or early stopping.

2. (**Utility regularity**) The utility U used in KTO is smooth and strictly monotone, with derivative bounded on the above compact domain:

   $$0 < m_U \;\le\; U'(x) \;\le\; M_U < \infty.$$

   Here m_U and M_U denote, respectively, lower and upper bounds on the derivative of the utility function U over the bounded range of implied rewards visited during training, which ensures both sensitivity and smoothness of the objective needed for the stability analysis.

3. (**Noise model**) In the empirical setting, gradient estimates are unbiased with coordinate-wise sub-Gaussian noise, matching the "empirical DPO" noise model.

All subsequent statements hold under these assumptions.

KTO implied reward and error variable

KTO defines a per-action implied reward

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

Because KL-regularized optima are invariant to additive reward shifts, we work in centered coordinates:

$$\bar{r}(a) = r(a) - \frac{1}{K}\sum_b r(b), \qquad \bar{r}_\theta(a) = r_\theta(a) - \frac{1}{K}\sum_b r_\theta(b).$$

Here $K := |A|$ is the cardinality of the action set A, i.e., the total number of discrete actions, so $\frac{1}{K}\sum_b(\cdot)$ denotes the uniform average over all actions.

We define the unary error variable

$$\delta_{\mathrm{KTO}}(a;\theta) \;:=\; \bar{r}(a) - \bar{r}_\theta(a).$$

At the KL-regularized optimum, the policy satisfies $\pi^*(a) = \frac{1}{Z}\pi_{\mathrm{ref}}(a)\exp(r(a))$, which implies $r_{\theta^*}(a) = \beta \log \frac{\pi^*(a)}{\pi_{\mathrm{ref}}(a)} = \beta r(a) - \beta \log Z$; since this differs from $r(a)$ only by an action-independent constant and both rewards are centered, the constant cancels, yielding $\delta_{\mathrm{KTO}}(a;\theta^*) = \bar{r}(a) - \bar{r}_{\theta^*}(a) = 0$ for all a.

Relation to loss nonlinearity. Define $\Delta_{\mathrm{KTO}}(a;\theta) = U(\bar{r}(a)) - U(\bar{r}_\theta(a))$. By the mean value theorem, for each a,

$$\Delta_{\mathrm{KTO}}(a;\theta) = U'(\xi_a)\,\delta_{\mathrm{KTO}}(a;\theta),$$

for some ξ_a between $\bar{r}(a)$ and $\bar{r}_\theta(a)$. Under the boundedness assumptions above,

$$m_U\,|\delta_{\mathrm{KTO}}(a;\theta)| \;\le\; |\Delta_{\mathrm{KTO}}(a;\theta)| \;\le\; M_U\,|\delta_{\mathrm{KTO}}(a;\theta)|.$$

Uniform-sampling KTO objective and gradient structure

Under uniform sampling over actions,

$$L_{\mathrm{KTO}}(\theta) \;=\; -\frac{1}{K}\sum_{a\in A} U(\bar{r}_\theta(a)).$$

For the softmax policy $\pi_\theta(a) = \exp(\theta_a)/\sum_b \exp(\theta_b)$, writing $\log \pi_\theta(a) = \theta_a - \log \sum_b e^{\theta_b}$ and differentiating with respect to the parameter θ_k yields

$$\partial_{\theta_k} \log \pi_\theta(a) = \begin{cases} 1 - \pi_\theta(k), & \text{if the differentiated parameter corresponds to action } a, \\ -\pi_\theta(k), & \text{otherwise,} \end{cases}$$

reflecting the fact that increasing θ_k raises the log-probability of action k while decreasing all others through the normalization term.

Using this identity, the gradient of the KTO objective under uniform sampling can be written as

$$\nabla_{\theta_k} L_{\mathrm{KTO}}(\theta) = -\frac{\beta}{K} \sum_{a \in A} g_a(\theta) \begin{cases} 1 - \pi_\theta(k), & a = k, \\ -\pi_\theta(k), & a \neq k, \end{cases}$$

where $g_a(\theta)$ is a scalar coefficient collecting the chain-rule factors (including the centering term) and satisfies bounds $m_U \leq g_a(\theta) \leq M_U$. This shows that each parameter update is a bounded linear combination of softmax mixing directions, which is the same bounded-coefficient "softmax mixing direction" structure exploited in DPO uniform-sampling analyses (Shi et al. (2024)).

Stability and contraction under exact gradients

Now we consider gradient descent $\theta^{t+1} = \theta^t - \eta \nabla L_{\mathrm{KTO}}(\theta^t)$.

Linearizing the induced update on $\bar{r}_\theta$ yields a recursion of the form

$$\delta^{t+1} = \delta^t - \eta\, \beta^2\, H(\theta^t)\, \delta^t,$$

where the linear operator admits a factorization $H(\theta) = J(\theta)^\top D(\theta) J(\theta)$ with $D(\theta) \succcurlyeq 0$, hence $H(\theta)$ is symmetric positive semidefinite matrix depending on π_{θ^t} and $g(\theta^t)$.

Under the compactness assumptions on π_θ and boundedness of $g_a(\theta)$, there exist constants $0 < \mu \leq L < \infty$ such that

$$\mu \|\delta\|_2^2 \;\leq\; \delta^\top H(\theta) \delta \;\leq\; L \|\delta\|_2^2 \quad \forall \theta \text{ in the admissible set.}$$

Choosing $\eta \leq 1/(\beta^2 L)$ yields

$$\|\delta^{t+1}\|_2^2 \;\leq\; (1-c)\|\delta^t\|_2^2,$$

for some $c \in (0,1)$, implying linear convergence of KTO-Unif in δ under exact gradients. In summary, although gradient descent is performed in parameter space, its effect on the implied rewards can be viewed as repeatedly applying a well-behaved linear transformation to the reward error; because this transformation consistently points in directions that reduce the error and has uniformly bounded strength, each update shrinks the error by a fixed fraction, leading to stable and linear convergence of the implied rewards under exact gradients.

Empirical gradients

With unbiased sub-Gaussian gradient noise, the recursion becomes

$$\delta^{t+1} = (I - \eta \beta^2 H_t)\delta^t + \eta \beta E^{(t)},$$

where $E^{(t)}$ is zero-mean with variance proxy σ^2. For sufficiently small η and noise level σ (as in the empirical DPO model), this yields

$$\mathbb{E}\|\delta^t\|_2^2 \;\leq\; (1-c)^t \|\delta^0\|_2^2 + O(\sigma^2),$$

i.e., convergence to a noise-dependent neighborhood whose radius scales with the gradient variance.

Extension to RKO via bounded similarity weights

RKO modifies KTO by introducing per-action similarity weights:

$$L_{\mathrm{RKO}}(\theta) = -\frac{1}{K} \sum_{a \in A} w(a)\, U(\bar{r}_\theta(a)),$$

where the weights are treated as stop-gradient scalars satisfying

$$0 < w_{\min} \le w(a) \le w_{\max} < \infty.$$

The gradient takes the same form as in KTO with coefficients $g_a^{\mathrm{RKO}}(\theta) = w(a) g_a(\theta)$, which remain uniformly bounded. Hence the above stability analysis carries through with constants rescaled by $(w_{\min}, w_{\max})$.

Summary Our analysis follows the standard DPO contraction-style proof in the tabular-softmax bandit abstraction (Shi et al. (2024)). We first cast KTO into a KL/log-ratio form under a single uniform sampler, introduce a centered implied-reward error that is zero at the KL-regularized optimum, and show that gradient descent induces a stable linear recursion on this error: it contracts for a sufficiently small step size under exact gradients, and converges to a noise-controlled neighborhood under unbiased sub-Gaussian gradient noise. We then extend the same argument to RKO by observing that RKO only reweights per-sample terms with bounded stop-gradient similarity weights, which preserves the contraction structure up to rescaled constants.

References

Shi, R., Zhou, R., and Du, S. S. (2024). The crucial role of samplers in online direct preference optimization. *arXiv preprint arXiv:2409.19605*.