SINKHORN-DPO: DISTRIBUTIONALLY ROBUST PREFERENCE ALIGNMENT VIA ENTROPIC OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning from human feedback (RLHF) is a standard approach for aligning large language models (LLMs) with human preferences, while direct preference optimization (DPO) provides a simpler offline alternative that avoids explicit reward modeling. However, DPO remains sensitive to distribution shifts in preference data, motivating robust formulations based on distributionally robust optimization (DRO). Wasserstein-DPO (WDPO) methods provided a principled treatment of preference distribution shift, but their exact robust objectives involved nonsmooth dual structures that complicate gradient-based analysis and finite-sample estimation. We propose Sinkhorn-DPO (SDPO), a distributionally robust preference optimization framework based on entropically regularized optimal transport. Its dual formulation yields a smooth log-sum-exp envelope, enabling explicit derivative characterizations of the robust objective. Under mild regularity and a data-coverage condition imposed on the fixed reference measure, we prove smoothness and strong convexity of the profiled objective. Our finite-sample analysis establishes a parameter estimation rate of near-$n^{-1/2}$ with sample size n, which is a faster sample-size dependence than the $\mathcal{O}(n^{-1/4})$ bound of WDPO. The robustification is implemented through Sinkhorn-induced sample weights, so the policy is updated using a weighted first-order DPO gradient. We evaluate SDPO against DPO, Kullback-Leibler DPO (KLDPO), χPO, and WDPO in linear experiments and in computational profiling with LLMs. The linear experiments examine radius calibration, parameter stability, and behavior under an induced within-corpus shift, while the LLM study reports runtime, peak GPU-memory usage and validation accuracy.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) is an important framework for aligning language models with human preferences. Early work learned policies directly from pairwise human comparisons Christiano et al. (2017), while later RLHF systems combined supervised fine-tuning, reward modeling, and policy optimization to align LLMs at scale Ouyang et al. (2022); Bai et al. (2022). This multi-stage pipeline relies on the quality and coverage of preference data and reward signals. Subsequent studies showed overoptimization and other forms of mismatch between offline supervision and the behavior induced during alignment Liu et al. (2024); Xu et al. (2025a); Huang et al. (2025).

DPO simplified this pipeline by exploiting the connection between KL-regularized reward maximization and pairwise preference learning Rafailov et al. (2023). Under the Bradley–Terry preference model Bradley & Terry (1952), DPO optimized the policy through a classification-style loss without separately fitting a reward model or performing online reinforcement learning. Subsequent work investigated robustness to imperfect

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.

preference supervision and offline overoptimization. Chowdhury et al. (2024) developed a robust DPO objective for policy optimization in the presence of random preference flips. Huang et al. (2025) proposed χPO, which modified the logarithmic link in DPO and induced χ^2-based pessimistic regularization. These approaches addressed robustness through noise correction or alternative regularization, which motivated a question: how should preference optimization account directly for uncertainty in the preference distribution itself?

DRO provides a natural framework for this purpose by optimizing against a family of distributions around a nominal data-generating law Ben-Tal et al. (2013); Duchi & Namkoong (2021). Wasserstein ambiguity sets are particularly attractive because they incorporate the geometry of the sample space and admit tractable dual reformulations with finite-sample guarantees Mohajerin Esfahani & Kuhn (2018); Blanchet & Murthy (2019); Blanchet et al. (2022). Related work applied distributional robustness to reinforcement learning and offline decision making, including Wasserstein-robust Markov decision processes and model-free methods Carmona et al. (2023); Neufeld & Sester (2024); Xu et al. (2023); Panaganti et al. (2022). Recent work also applied distributional robustness directly to RLHF and DPO. Mandal et al. (2025) formulated distributionally robust RLHF and DPO under distribution-shifted prompts, while Xu et al. (2025b) developed WDPO and KLDPO for distribution shift in offline preference optimization and established statistical guarantees for their robust policy parameters.

The most direct precursor to our work is the WDPO framework of Xu et al. (2025b). Their Wasserstein ambiguity set provided a treatment of preference distribution shift within the offline DPO framework, but the resulting robust objective involved a nonsmooth worst-case transport envelope whose direct derivative analysis was more difficult than that of the corresponding non-robust DPO problem. This motivates the following question: can we find a smoother structure that preserves the distributionally robust interpretation while enabling explicit derivative and finite-sample analyses?

We address this question using entropic optimal transport. Entropic regularization provided computationally efficient Sinkhorn formulations of optimal transport Cuturi (2013); Genevay et al. (2016); Peyr'e & Cuturi (2019). Wang et al. (2026) developed Sinkhorn distributionally robust optimization (SDRO), showing that entropy regularization yielded a smoothed dual representation with tractable first-order optimization. Related applications also used entropy-regularized transport as a semantic-aware policy regularizer in online, PPO-based language-model alignment Na et al. (2026). We specialize SDRO to offline preference optimization and show that the resulting entropic envelope induces explicit DPO sample weights, which preserves a first-order policy update while changing the analytical structure of the robust objective.

Our starting point is the standard DPO loss under a log-linear policy class. We define a Sinkhorn ambiguity set around the nominal preference distribution and maximize the DPO loss over this set. Using the dual representation of Sinkhorn DRO, the robust objective reduces to a scalar optimization over a dual variable and a smooth log-sum-exp envelope. The resulting tilted distribution induces explicit sample weights, so the policy update can be written as a weighted first-order DPO gradient.

The main theoretical benefit is that the resulting smooth structure enables direct differentiation with respect to both the policy parameter and the dual variable. We derive explicit derivative expressions, establish uniform regularity and bounds for the dual minimizer, and prove strong convexity of the profiled objective under a data-coverage condition imposed only on a fixed reference measure, rather than uniformly over the entire ambiguity set. These properties allow us to control the empirical profiled gradient directly and obtain the parameter estimation rate $\mathcal{O}(\sqrt{\log n/n})$ with sample size n, with faster sample-size dependence than the $\mathcal{O}(n^{-1/4})$ bound of WDPO in Xu et al. (2025b).

Our contributions are threefold. First, we formulate SDPO, combining the offline preference-learning structure of DPO with an entropically regularized transport ambiguity set for preference distribution shift. Second, we characterize the resulting smooth dual envelope, and establish strong convexity of the profiled robust objective under a coverage condition on a fixed reference measure. Third, we prove a finite-sample parameter

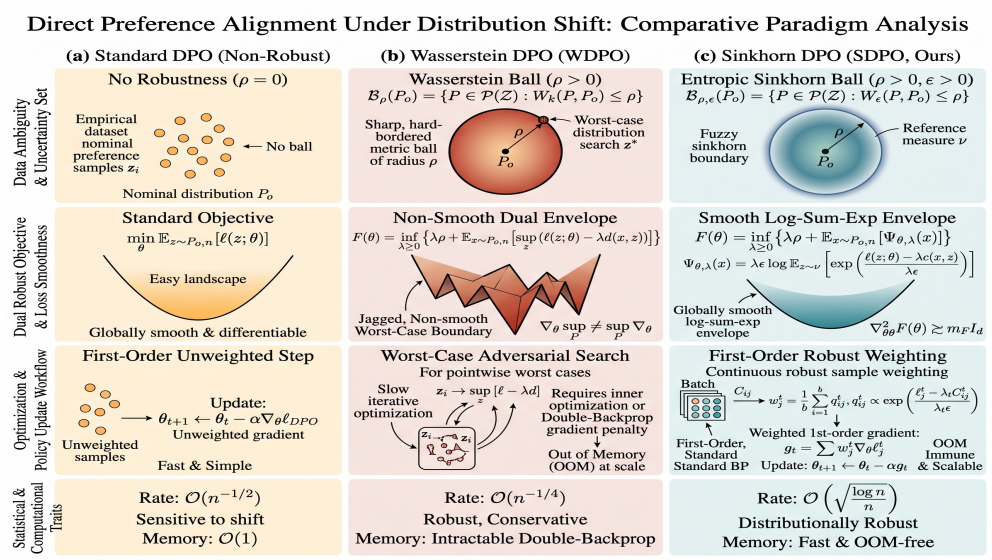

Figure 1: Overview of preference alignment under distribution shift. Standard DPO optimizes the empirical preference objective, WDPO robustifies it through an exact transport ambiguity set, and SDPO introduces entropic regularization to obtain a smooth log-sum-exp robust envelope while preserving first-order policy optimization.

estimation bound of order $\mathcal{O}(\sqrt{\log n/n})$ with sample size n and develop a first-order implementation based on Sinkhorn-induced sample weights. We complement the theory with experiments on calibration, parameter stability, induced within-corpus distribution shift, computational cost and validation accuracy.

2 Problem Formulation

We begin by formalizing the data generation process and the preference optimization objective. Let $z = (s, a_1, a_2, y) \in \mathcal{Z} := \mathcal{S} \times \mathcal{A} \times \mathcal{A} \times \{0, 1\}$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is a finite discrete action space, and $y \in \{0, 1\}$ is the binary preference label satisfying

$$y = 1 \iff a_1 \succ a_2 \mid s, \qquad y = 0 \iff a_2 \succ a_1 \mid s.$$

Let $P_o \in \mathcal{P}(\mathcal{Z})$ denote the nominal preference distribution on the sample space $\mathcal{Z}$, and let $P_{o,n} := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ be the empirical distribution based on i.i.d. samples $z_1, \dots, z_n \sim P_o$.

To distinguish the roles of the probability measures used below, we adopt the following terminology. P_o denotes the nominal preference distribution and serves as the center of the ambiguity set, while $P_{o,n}$ denotes its empirical counterpart. The measure ν denotes the fixed reference probability measure used in the entropic Sinkhorn dual, and $P \in \mathcal{B}_{\rho,\epsilon}(P_o)$ denotes a candidate distribution in the ambiguity set. We use "reference policy" exclusively for π_{ref}.

We consider the DPO pointwise loss following the standard Bradley–Terry preference model Bradley & Terry (1952); Zhu et al. (2023): $\ell(z;\theta) := -y \log \sigma\big(\beta h_\theta(s, a_1, a_2)\big) - (1 - y) \log \sigma\big(-\beta h_\theta(s, a_1, a_2)\big)$, where $z = (s, a_1, a_2, y) \in \mathcal{Z}$, $\beta > 0$, $\sigma(t) = (1 + e^{-t})^{-1}$, π_θ is a parameterized policy and $h_\theta(s, a_1, a_2) := \log \frac{\pi_\theta(a_1|s)}{\pi_{\text{ref}}(a_1|s)} - \log \frac{\pi_\theta(a_2|s)}{\pi_{\text{ref}}(a_2|s)}$. For comparison, exact Wasserstein ambiguity sets provide

the standard distributionally robust baseline; their definition is recalled in Appendix A.1. In this paper, we use the following entropically regularized Sinkhorn uncertainty set. For $\rho \geq 0$ and $\epsilon > 0$, define the Sinkhorn uncertainty set by $\mathcal{B}_{\rho,\epsilon}(P_o) := \{P \in \mathcal{P}(\mathcal{Z}) : W_\epsilon(P, P_o) \leq \rho\}$, where the Sinkhorn distance is given by $W_\epsilon(P,Q) = \min_{\pi \in \Pi(P,Q)} \{\int c(x,z)\, d\pi(x,z) + \epsilon H(\pi \,\|\, P \otimes Q)\}$. Here $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ is a measurable transport cost, $\Pi(P,Q)$ denotes the set of couplings with marginals P and Q, and $H(\pi \,\|\, P \otimes Q) = \mathbb{E}_{(x,z)\sim\pi}\left[\log\left(\frac{d\pi}{d(P\otimes Q)}(x,z)\right)\right]$ is the relative entropy. The Sinkhorn-robust DPO objective is defined by

$$F(\theta) := \sup_{P \in \mathcal{B}_{\rho,\epsilon}(P_o)} \mathbb{E}_{z\sim P}\big[\ell(z;\theta)\big]. \tag{1}$$

Assumption 2.1 (Reference measure and bounded transport cost)**.** *Let $\mathcal{Z}$ be a Polish space, and let $\nu \in \mathcal{P}(\mathcal{Z})$ be the reference probability measure. We assume that the transport cost $c : \mathcal{Z} \times \mathcal{Z} \to [0, \bar{c}]$ is measurable, lower semicontinuous, and bounded above by some constant $\bar{c} > 0$.*

Under Assumption 2.1, we obtain the following sinkhorn dual formulation.

Proposition 2.2 (Sinkhorn dual reformulation)**.** *(Wang et al., 2026, Theorem 1) Let $\hat{\mathbb{P}} \in \mathcal{P}(\mathcal{Z})$, $\rho \geq 0$, and $\mathcal{B}_{\rho,\epsilon}(\hat{\mathbb{P}}) := \{P \in \mathcal{P}(\mathcal{Z}) : W_\epsilon(P, \hat{\mathbb{P}}) \leq \rho\}$. Under Assumption 2.1, for any non-negative measurable loss function $\ell(z;\theta) \geq 0$,*

$$\sup_{P \in \mathcal{B}_{\rho,\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_P[\ell(z;\theta)] = \inf_{\lambda \geq 0} \left\{ \lambda\rho + \lambda\epsilon\, \mathbb{E}_{x\sim\hat{\mathbb{P}}}\left[\log \mathbb{E}_{z\sim\nu}\left[\exp\left(\frac{\ell(z;\theta) - \lambda c(x,z)}{\lambda\epsilon}\right)\right]\right]\right\}. \tag{2}$$

By convention, the value at $\lambda = 0$ is understood as the corresponding limit as $\lambda \downarrow 0$, which equals the ν-essential supremum of ℓ.

Define the dual log-sum-exp envelope component in (2) as: $\Psi_{\theta,\lambda}(x) := \lambda\epsilon \log \mathbb{E}_{z\sim\nu}\left[\exp\left(\frac{\ell(z;\theta)-\lambda c(x,z)}{\lambda\epsilon}\right)\right]$. Applying Proposition 2.2 in (1), we define the profiled objective as follows:

$$F(\theta) = \inf_{\lambda \geq 0} \underbrace{\big\{\lambda\rho + \mathbb{E}_{x\sim P_o}\big[\Psi_{\theta,\lambda}(x)\big]\big\}}_{=:\mathcal{L}(\theta,\lambda)}. \tag{3}$$

The bounded-cost and integrability conditions required by Proposition 2.2 are verified for the present DPO setting in Remark A.1.

3 Theoretical properties of the profiled objective

In this section, we analyze the properties of the robust objective $F(\theta)$. In contrast to the WDPO setting Xu et al. (2025b), the Sinkhorn ball yields a smooth log-sum-exp envelope through entropic regularization in the dual representation Wang et al. (2026). This structure permits the derivative-based analysis and the objective's strong convexity developed below. To establish properties such as the boundedness of gradients and Hessians of $\mathcal{L}(\theta,\lambda)$, we specify the parametric form of the policy and introduce the following standard regularity conditions.

Assumption 3.1 (Log-linear policy class and reference policy)**.** *Let $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a known feature map with $\max_{s,a} \|\psi(s,a)\|_2 \leq 1$. Assume that the policy parameter θ belongs to the compact set $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$ with a given $B > 0$, and that the log-linear policy class is defined as*

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \psi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \psi(s,a'))}. \tag{4}$$

Assume further that the reference policy π_{ref} is fixed, strictly positive, and also log-linear, parameterized by a known vector $\theta_{\mathrm{ref}} \in \mathbb{R}^d$ satisfying $\|\theta_{\mathrm{ref}}\|_2 \leq B$.

More discussion about Assumption 3.1 is given in Remark A.2. To ensure the dual problem exhibits nice properties (e.g., possessing bounded, non-degenerate dual minimizers), we introduce two additional non-degeneracy conditions.

Assumption 3.2 (Non-degeneracy of the loss)**.** *Assume that for any $\theta \in \Theta$, the loss function $z \mapsto \ell(z;\theta)$ is not ν-almost surely constant.*

Assumption 3.3 (Strictly Active Sinkhorn Constraint)**.** *Let $A_\theta := \{z \in \mathcal{Z} : \ell(z;\theta) = \text{ess sup}_\nu \ell(\cdot;\theta)\}$ denote the set of exact maximizers of the loss function. We define the critical effective radius as $\bar{\rho}_{crit}(\theta) := \epsilon \mathbb{E}_{x\sim P_o}\left[\log\left(\frac{1}{\mathbb{Q}_{x,\epsilon}(A_\theta)}\right)\right]$, where $d\mathbb{Q}_{x,\epsilon}(z) := \frac{e^{-c(x,z)/\epsilon}}{\mathbb{E}_{u\sim\nu}[e^{-c(x,u)/\epsilon}]} d\nu(z)$. We assume that the user-specified robust radius ρ is chosen such that the effective radius $\bar{\rho} := \rho + \epsilon \mathbb{E}_{x\sim P_o}\left[\log \mathbb{E}_{z\sim\nu}[e^{-c(x,z)/\epsilon}]\right]$ satisfies the strictly active condition:*

$$0 < \bar{\rho} < \inf_{\theta\in\Theta} \bar{\rho}_{crit}(\theta). \tag{5}$$

Remark 3.4 (Calibration of the effective radius)**.** *Under Assumption 3.3, if $\nu(A_\theta) = 0$ for all $\theta \in \Theta$ (for instance, if the loss has no flat maximizer region under ν), then $\mathbb{Q}_{x,\epsilon}(A_\theta) = 0$ because $\mathbb{Q}_{x,\epsilon} \ll \nu$. Consequently, the critical effective radius degenerates to $\bar{\rho}_{crit}(\theta) \equiv +\infty$, which trivially satisfies the upper bound constraint in* (5)*. Moreover, the boundedness of the transport cost in Assumption 2.1 implies $e^{-\bar{c}/\epsilon} \le e^{-c(x,z)/\epsilon} \le 1$. Taking expectations, we obtain $-\bar{c} \le \epsilon \mathbb{E}_{x\sim P_o}\left[\log \mathbb{E}_{z\sim\nu}\left[e^{-c(x,z)/\epsilon}\right]\right] \le 0$. Thus, choosing $\rho > \bar{c}$ is a conservative sufficient condition for $\bar{\rho} > 0$, provided that the resulting $\bar{\rho}$ remains $\bar{\rho} < \inf_{\theta\in\Theta} \bar{\rho}_{crit}(\theta)$. In practical implementations, however, a much smaller ρ is typically sufficient, as the shift term can be precomputed offline via Monte Carlo integration to empirically calibrate the exact effective radius before solving the optimization problem.*

Next, we establish that the dual minimizer of (3), denoted by $\lambda^\star(\theta)$, remains strictly bounded away from zero and infinity uniformly across the parameter space.

Lemma 3.5 (Smoothness and uniform bounds of the dual minimizer)**.** *Following the notations of Proposition A.4, under Assumptions 2.1, 3.1, 3.2, 3.3, the map $\theta \mapsto \lambda^\star(\theta)$ is locally continuously differentiable (C^1) and globally continuous on Θ. Furthermore, there exist constants $0 < \underline{\lambda} \le \overline{\lambda} < \infty$ independent of θ such that $\lambda^\star(\theta) \in [\underline{\lambda}, \overline{\lambda}]$ for all $\theta \in \Theta$.*

The proof of Lemma 3.5 is given in Appendix A.3.

Assumption 3.6 (Reference-Measure Data Coverage)**.** *There exists a constant $\lambda_0 > 0$ such that the feature covariance matrix under the reference measure satisfies: $\Sigma_\nu := \mathbb{E}_{z\sim\nu}\left[\Delta\psi(z)\Delta\psi(z)^\top\right] \succeq \lambda_0 I_d$, where $\Delta\psi(z) := \psi(s,a_1) - \psi(s,a_2)$ for $z = (s,a_1,a_2,y)$.*

Remark 3.7 (Relaxation of the Data Coverage Assumption)**.** *In offline RL, recent studies Agarwal et al. (2019); Wang et al. (2020); Xu et al. (2025b) analyzing global convergence often require a feature covariance lower bound uniformly over the entire ambiguity set (e.g., $\Sigma_P \succeq \lambda I$ for all $P \in \mathcal{P}_{W_k}(\rho; P_o)$ as seen in Xu et al. (2025b)), which is typically used to establish the strong convexity of the expected loss under the worst-case measure. In our Sinkhorn-robust setting, the strong-convexity argument in Theorem 3.8 only needs the covariance lower bound under the fixed reference measure ν, i.e. Assumption 3.6. This is a weaker and more localized requirement than imposing a uniform lower bound across an infinite family of adversarially perturbed distributions.*

Theorem 3.8 (Intrinsic strong convexity of the profiled objective)**.** *Recall the profiled objective $F(\theta) = \inf_{\lambda>0} \mathcal{L}(\theta,\lambda)$ defined in* (3)*, and let $\lambda^\star(\theta)$ denote its strictly positive minimizer established in Lemma 3.5, such that $F(\theta) = \mathcal{L}(\theta, \lambda^\star(\theta))$. Under Assumptions 2.1, 3.1, 3.2, 3.3, and 3.6, the map F is twice continuously differentiable on Θ. Moreover, its Hessian matrix preserves the strong convexity of the expected DPO loss, satisfying $\nabla^2_\theta F(\theta) \succeq m_F I_d$, where $m_F := \gamma_0 \lambda_0 \exp\left(-\frac{\overline{\ell} + \underline{\lambda}\,\bar{c}}{\underline{\lambda}\,\epsilon}\right) > 0$. Consequently, $F(\theta)$ is globally m_F-strongly convex over the compact parameter space Θ.*

We provide the proof of Theorem 3.8 in Appendix A.3.

4 FINITE-SAMPLE STATISTICAL GUARANTEES

Having obtained the smoothness and strong convexity of the profiled objective $\mathcal{L}(\theta, \lambda)$, we now turn to the empirical estimation problem. By Lemma 3.5, the dual minimizer $\lambda^\star(\theta)$ satisfies $\lambda^\star(\theta) \in [\underline{\lambda}, \overline{\lambda}]$ for all $\theta \in \Theta$ with fixed $\underline{\lambda}, \overline{\lambda} \in \mathbb{R}$. Let $\Lambda := [\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$. Thus, the objective $F(\theta)$ defined in (3) can be rewritten as $F(\theta) = \inf_{\lambda>0} \mathcal{L}(\theta, \lambda) = \min_{\lambda\in\Lambda} \mathcal{L}(\theta, \lambda)$. To establish finite-sample guarantees, we constrain the empirical dual search space to the same compact interval Λ. Recalling that the empirical distribution is $P_{o,n} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, we define the empirical profiled loss $\mathcal{L}_n(\theta, \lambda) := \lambda\rho + \mathbb{E}_{x\sim P_{o,n}}\big[\Psi_{\theta,\lambda}(x)\big]$, and the corresponding constrained empirical profiled objective as $F_n(\theta) := \min_{\lambda\in\Lambda} \mathcal{L}_n(\theta, \lambda)$. To establish the uniform convergence of F_n and its gradient, we define the joint parameter space $\mathcal{W} := \Theta \times \Lambda \subset \mathbb{R}^{d+1}$. Let $R_{\mathcal{W}} > 0$ denote the radius of the Euclidean ball containing $\mathcal{W}$. Next, we provide the finite-sample guarantee of SDPO.

Theorem 4.1 (Finite-Sample Near-Parametric Rate for SDPO)**.** *Suppose Assumptions 2.1, 3.1, 3.2, 3.3 and 3.6 hold. Following the notation of Lemma A.6. Let $\hat{\theta}_n \in \arg\min_{\theta\in\Theta} F_n(\theta)$ and $\theta^\star \in \arg\min_{\theta\in\Theta} F(\theta)$. Then, for any $n \geq 1$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_n - \theta^\star\|_2 \leq \frac{1}{m_F}\left(\mathcal{E}_{n,2}(\delta) + \frac{C_{\theta\lambda}}{m_\lambda}\mathcal{E}_{n,1}(\delta)\right), \tag{6}$$

where $m_F > 0$ is the strong convexity modulus from Theorem 3.8. In particular, $\|\hat{\theta}_n - \theta^\star\|_2 = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$.

We present the complete proof of Theorem 4.1 in Appendix A.3.

Remark 4.2. *Theorem 4.1 establishes a parameter estimation error bound $\|\hat{\theta}_n - \theta^\star\|_2 = \mathcal{O}\big(\sqrt{\log n/n}\big)$, which is closer to the $\mathcal{O}(n^{-1/2})$ rate typical of standard DPO Chowdhury et al. (2024); Xu et al. (2025b). In contrast, Xu et al. (2025b) obtains an $\mathcal{O}(n^{-1/4})$ rate for WDPO and KLDPO. This difference stems from the dual structure: the exact Wasserstein formulation involves a pointwise infimum through the Moreau–Yosida envelope, inducing nonsmoothness in θ and preventing direct interchange of the gradient and worst-case expectation (i.e., $\nabla_\theta \sup_P \neq \sup_P \nabla_\theta$). Thus, WDPO relies on uniform concentration of objective values, whose $\mathcal{O}(n^{-1/2})$ rate translates through strong convexity into the $\mathcal{O}(n^{-1/4})$ parameter rate (i.e., $\|\hat{\theta}_n - \theta^\star\|_2 \lesssim \sqrt{|\mathcal{L}_n - \mathcal{L}|}$). In contrast, entropic regularization yields the smooth log-sum-exp envelope in Proposition 2.2, which permits derivative interchange and direct concentration of the profiled gradient (Proposition A.4; Lemma A.6). This avoids the square-root bottleneck and leads to the stated near-$n^{-1/2}$ parameter rate.*

5 SINKHORN-DPO ALGORITHM

Let $\mathcal{B}_t = \{z_{t,1}, \ldots, z_{t,b}\}$ denote the support used at iteration t, where $b := |\mathcal{B}_t|$ is its cardinality. For $z_{t,j} \in \mathcal{B}_t$, let $\ell_j^t := \ell(z_{t,j}; \theta_t)$ be the pointwise DPO loss and $C_{ij}^t := c(z_{t,i}, z_{t,j})$ the corresponding ground cost. The entropically regularized Sinkhorn dual evaluated at θ_t is $D_t(\lambda) = \lambda\rho + \frac{\lambda\epsilon}{b}\sum_{i=1}^b \log\Big[\frac{1}{b}\sum_{j=1}^b \exp\Big(\frac{\ell_j^t - \lambda C_{ij}^t}{\lambda\epsilon}\Big)\Big]$, for $\lambda > 0$. Here, $\rho > 0$ is the Sinkhorn robustness radius and $\epsilon > 0$ is the entropic regularization parameter. The scalar dual variable is restricted in the implementation to the numerical search interval $\Lambda_{\text{num}} = [\lambda_{\min}, \lambda_{\max}] \subset (0, \infty)$. For a given dual minimizer λ_t, the normalized transport weights are

$$q_{ij}^t := \exp\left(\frac{\ell_j^t - \lambda_t C_{ij}^t}{\lambda_t\epsilon}\right)\left(\sum_{k=1}^b \exp\left(\frac{\ell_k^t - \lambda_t C_{ik}^t}{\lambda_t\epsilon}\right)\right)^{-1}, \qquad i, j = 1, \ldots, b. \tag{7}$$

Thus, $\sum_{j=1}^{b} q_{ij}^t = 1$ for every $i \in \{1, ..., b\}$. The resulting robust weight assigned to support point $z_{t,j}$ is the column marginal, which satisfy $w_j^t := \frac{1}{b} \sum_{i=1}^{b} q_{ij}^t$ and $\sum_{j=1}^{b} w_j^t = 1$. Thus, the weighted policy gradient is $g_t = \sum_{j=1}^{b} w_j^t \nabla_\theta \ell(z_{t,j}; \theta_t)$. We propose a Sinkhorn-DPO method shown in Algorithm 1. In Algorithm 1,

Algorithm 1 Sinkhorn-DPO with entropic robust weighting

Require: Preference dataset $\mathcal{D} = \{z_i\}_{i=1}^M$, initial parameters $\theta_0 \in \Theta$, fixed feature map ψ, robustness radius $\rho > 0$, entropic regularization $\epsilon > 0$, batch size b, learning rate α, number of iterations T, and $\Lambda_{\text{num}} = [\lambda_{\min}, \lambda_{\max}] \subset (0, \infty)$. Pre-compute the ground cost c over $\mathcal{D}$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: Sample $\mathcal{B}_t = \{z_{t,1}, \ldots, z_{t,b}\} \subseteq \mathcal{D}$ and evaluate $\ell_j^t = \ell(z_{t,j}; \theta_t)$ and $C_{ij}^t = c(z_{t,i}, z_{t,j})$.
3: Solve $\lambda_t \in \arg\min_{\lambda \in \Lambda_{\text{num}}} D_t(\lambda)$.
4: Compute q_{ij}^t from (7) and set $w_j^t = b^{-1} \sum_{i=1}^{b} q_{ij}^t$.
5: Compute $g_t = \sum_{j=1}^{b} w_j^t \nabla_\theta \ell(z_{t,j}; \theta_t)$.
6: Update policy $\theta_{t+1} \leftarrow \theta_t - \alpha g_t$ (via projected gradient descent or Adam optimizer).
7: **end for**

Ensure: π_{θ_T}

$\mathcal{D} = \{z_i\}_{i=1}^M$ is the observed preference dataset and induces the empirical nominal distribution $P_{o,M}$, so $M = |\mathcal{D}|$, while b is the size of the support used in the current optimization step. For the full-support linear fits, $b = M$ and the same empirical support is used throughout optimization. In the following linear experiments, the corresponding support is likewise fixed for each fit. At LLM scale, the support is the current mini-batch and is resampled across optimization iterations. The experiments use the same empirical support for the linear full-support setting and mini-batch outer/reference supports at LLM scale; the corresponding support conventions and stochastic approximation details are given in Appendix B.3.

6 Numerical Experiments

We evaluate SDPO on real human-preference data in a log-linear setting, followed by an induced within-corpus shift and LLM-scale computational profiling. The experiments focus on calibration, parameter stability, distributional sensitivity, and runtime.

6.1 Experimental Setup

We use the *harmless-base* split of the Anthropic HH-RLHF corpus Bai et al. (2022), retaining $1,200$ canonical preference pairs and the two oriented atoms from each pair, for $M = 2,400$ samples. We use a normalized rank-24 TF–IDF/SVD representation for each prompt–response pair; the detailed feature construction is given in Appendix B.1. We use the squared feature-distance transport cost with a unit label-mismatch penalty. We set $\theta_{\text{ref}} = 0$, $B = 3$, $\beta = 2$, and fix the entropic regularization at $\epsilon = 0.05$. We compare SDPO with DPO, WDPO, KLDPO Xu et al. (2025b), and χPO Huang et al. (2025); the optimization protocol and baseline hyperparameters are reported in Appendix B.2. The SDPO robustness radius ρ is selected by three-fold cross-validation (CV).

6.2 Calibration and Finite-Sample Parameter Stability

We select the SDPO robustness radius by three-fold CV at the canonical-pair level, while enforcing the empirical feasibility condition associated with Assumption 3.3. The smallest candidate feasible in all folds is $\rho_{\text{cv}} = 0.15$; the mean held-out nominal DPO loss varies only slightly across the feasible candidates. Detailed

candidate-grid results and dual-minimizer diagnostics are reported in Appendix B.4 and Appendix B.5. Next, we study finite-sample parameter stability using an oriented-atom bootstrap, which is presented in Figure 2 (a). Across the tested sample sizes, SDPO and WDPO remain closer to their full-pool reference parameters than DPO, KLDPO, and χPO; at $n = 2{,}000$, their mean deviations are both below 0.05, whereas the other methods remain above 0.9. The complete sample-size grid and numerical values are reported in Appendix B.6.

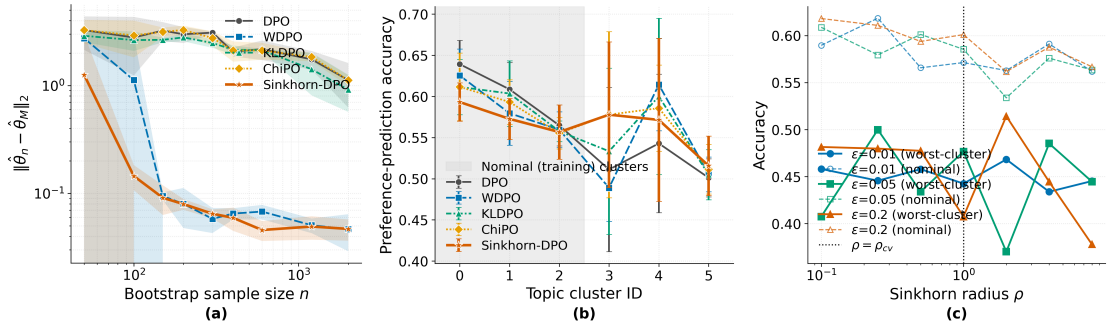

Figure 2: Finite-sample stability and within-corpus distribution-shift results. **(a)** Parameter deviation from the full-pool reference across bootstrap sample sizes. **(b)** Cluster-wise preference accuracy under the induced distribution shift. **(c)** SDPO accuracy over the tested (ρ, ϵ).

6.3 Within-Corpus Distribution Shift

We examine an induced within-corpus distribution shift using six K-Means clusters of the fixed dialogue-prefix representations. Clusters 0–2 form the nominal training pool, while clusters 3–5 are held out for evaluation; the full cluster composition and sampling protocol are given in Appendix B.7. Figure 2 (b)-(c) report the resulting cluster-wise accuracies and the SDPO sensitivity over the tested (ρ, ϵ) grid. The (ρ, ϵ) sensitivity analysis shows no clear monotone trend, with nominal accuracy remaining relatively stable over the tested grid.

6.4 Computational Profiling

In the linear benchmark (see Table 6 in Appendix B.12), SDPO requires 4.50–40.34,s per fit over $n = 50$–400, compared with 0.87–9.58,s for WDPO, while the remaining baselines are time faster. At LLM scale, we profile GPT-2, LLaMA-3.2-1B, LLaMA-3.2-3B, and LLaMA-3.1-8B on NVIDIA V100 GPUs under matched batch sizes at each model scale (see Figure 3). SDPO incurs additional computation from its entropic weighting and scalar dual optimization. This time overhead is acceptable in practice, because SDPO successfully scales to the 8B model within the 32GB V100 memory limit. In contrast, WDPO relies on an input-gradient penalty approximation Xu et al. (2025b) that necessitates double backpropagation. This leads to WDPO failing at the 8B scale under the 32GB setting. Conversely, for DPO, KLDPO, χPO and SDPO, whose memory is dominated by static weights rather than activations, exhibit normal monotonic memory growth across scales. More discussion, detailed sequence lengths, batch settings, and profiling measurements are reported in Appendix B.12.

6.5 LLM-Scale Optimization Convergence

We examine optimization dynamics at the LLaMA-3.1-8B scale on the held-out validation split. All methods use batch size 2 and are evaluated every 25 steps over 150 steps. Table 1 reports the corresponding validation

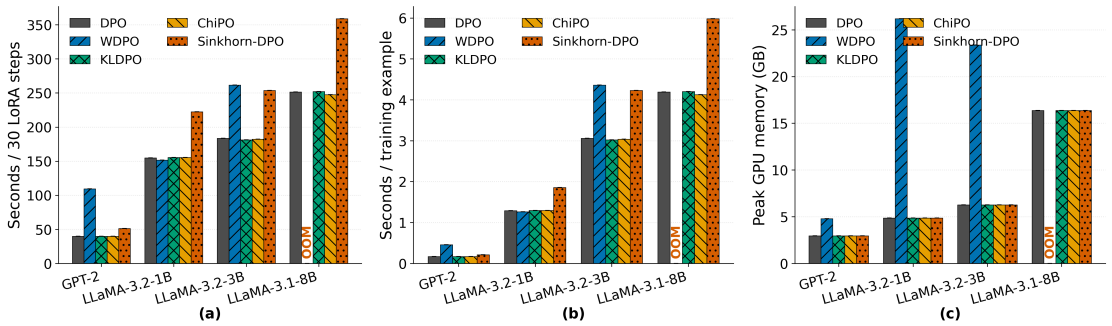

Figure 3: LLM computational profiling. **(a)** Wall-clock runtime for the timed LoRA steps. **(b)** Wall-clock runtime for per training example. **(c)** Peak GPU-memory usage. The tested are model scales include GPT-2, LLaMA-3.2-1B, LLaMA-3.2-3B, and LLaMA-3.1-8B. The tested 8B WDPO configuration is incomplete because of an out-of-memory (labeled "OOM") error under the 32GB setting.

preference accuracy. All methods improve during optimization, with different convergence patterns. Under this 8B training configuration, SDPO exhibits a gradual improvement over the optimization trajectory and reaches the relatively high observed validation accuracy at an intermediate checkpoint.

Table 1: LLaMA-3.1-8B validation preference accuracy over optimization steps. Darker blue indicates higher validation accuracy within each step column.

Method	**25**	**50**	**75**	**100**	**125**	**150**
DPO	0.570	0.645	0.640	0.630	0.635	0.645
KLDPO	0.555	0.610	0.620	0.620	0.625	0.625
χPO	0.565	0.630	0.650	0.650	0.650	0.645
SDPO	0.575	0.605	0.645	0.650	0.665	0.650

7 Conclusion

We proposed SDPO, a distributionally robust preference optimization framework that combines DPO with an entropically regularized transport ambiguity set. Its smooth dual log-sum-exp structure induces explicit transport-based sample weights and enables derivative-based analysis of the profiled objective. Under the stated regularity and reference-measure coverage assumptions, we established smoothness, strong convexity, and a finite-sample parameter estimation bound of order $\mathcal{O}(\sqrt{\log n/n})$, which is faster than the $\mathcal{O}(n^{-1/4})$ guarantee obtained in WDPO Xu et al. (2025b). In the experiments, SDPO and WDPO showed similar parameter deviations in the finite-sample linear study. At the LLM scale, although SDPO incurs additional preprocessing time compared to baseline methods, it demonstrates superior memory efficiency and scalability. Crucially, it successfully optimizes larger models under constrained GPU resources where WDPO becomes intractable due to memory exhaustion. These results characterize the statistical and computational trade-offs of the proposed entropic robustification. Extending the analysis beyond the log-linear setting and evaluating broader alignment tasks remain directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, readability, and presentation of the manuscript. We also used generative AI tools as an auxiliary aid for identifying basic syntax errors and debugging during the development of our numerical experiments. All AI-assisted text and code were reviewed and verified by the authors, and all mathematical results, experimental procedures, and reported findings were independently checked by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve new human-subject experiments, animal experiments, or the collection of personal data. The empirical study uses existing datasets and pretrained language models for research purposes, with the corresponding sources described and cited in the paper. To the best of our knowledge, this work does not introduce additional ethical, privacy, or safety concerns beyond those associated with the underlying data and models.

REPRODUCIBILITY STATEMENT

We make efforts to ensure the reproducibility of both the theoretical and empirical results. Our work comprises both theoretical results and numerical experiments. The main paper and appendices provide complete descriptions of the proposed method, its assumptions, and the corresponding theoretical results and proofs. The experimental setup and supplementary discussions document the datasets, preprocessing procedures, model configurations, hyperparameters, evaluation protocols, and computational settings used in the reported experiments and the corresponding appendices. All numerical results were obtained following the procedures described in the paper and were checked by the authors for consistency and correctness.

REFERENCES

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Charalambos D Aliprantis and Owen Burkinshaw. *Principles of real analysis*. Gulf Professional Publishing, 1998.

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

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization. *Mathematics of Operations Research*, 47(2):1500–1529, 2022.

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

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *The Annals of Applied Probability*, 33(6B):5334–5381, 2023.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

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

Donald L Cohn. *Measure theory*, volume 2. Springer, 2013.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport Distances. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

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

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $mathcalV$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 2022.

Aude Genevay, Marco Cuturi, Gabriel Peyr'e, and Francis Bach. Stochastic optimization for large-scale optimal transport. *arXiv preprint arXiv:1605.08527*, 2016.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason Lee, Wen Sun, Akshay Krishnamurthy, and Dylan Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 92647–92701, 2025.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *Advances in Neural Information Processing Systems*, 37:138663–138697, 2024.

Debmalya Mandal, Paulius Sasnauskas, and Goran Radanovic. Distributionally robust reinforcement learning with human feedback. *arXiv preprint arXiv:2503.00539*, 2025.

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

Byeonghu Na, Hyungho Na, Yeongmin Kim, Suhyeon Jo, HeeSun Bae, Mina Kang, and Il-Chul Moon. Semantic-aware wasserstein policy regularization for large language model alignment. *arXiv preprint arXiv:2602.01685*, 2026.

Ariel Neufeld and Julian Sester. Robust q-learning algorithm for markov decision processes under wasserstein uncertainty. *Automatica*, 168:111825, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32211–32224, 2022.

Gabriel Peyr'e and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Now Publishers, 2019.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *Operations Research*, 74(3): 1581–1603, 2026.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Yinglun Xu, Hangoo Kang, Tarun Suresh, Yuxuan Wan, and Gagandeep Singh. Learning a pessimistic reward model in rlhf. *arXiv preprint arXiv:2505.20556*, 2025a.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 9728–9754. PMLR, 2023.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30747–30784, 2025b. doi: 10.52202/085713-0912.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

A Additional theoretical details

Before proceeding, we specify our notation: for a vector v, $\|v\|_2$ denotes the standard Euclidean norm. For a matrix A, $\|A\|_{\mathrm{op}}$ denotes the spectral operator norm (i.e., its maximum singular value $\|A\|_{\mathrm{op}} = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$). Note that for a rank-1 symmetric positive semi-definite matrix of the form $vv^\top$, its operator norm is $\|vv^\top\|_{\mathrm{op}} = \|v\|_2^2$.

A.1 Wasserstein background

For any $P, Q \in \mathcal{P}(\mathcal{Z})$, the k-Wasserstein distance is defined by

$$W_k(P,Q) := \left(\inf_{\pi\in\Pi(P,Q)} \int_{\mathcal{Z}\times\mathcal{Z}} \mathsf{d}(z,z')^k \, d\pi(z,z') \right)^{1/k},$$

where $\mathsf{d}(z,z')$ is a ground metric on $\mathcal{Z}$. For any nominal preference distribution $P_o \in \mathcal{P}(\mathcal{Z})$ and radius $\rho \geq 0$, we write

$$\mathcal{P}_{W_k}(\rho; P_o) := \{P \in \mathcal{P}(\mathcal{Z}) : W_k(P, P_o) \leq \rho\}.$$

In distributionally robust optimization literature, ambiguity sets are commonly constructed as metric balls around a nominal distribution, with Wasserstein balls providing a data-geometric formulation that admits tractable dual reformulations and statistical guarantees Mohajerin Esfahani & Kuhn (2018); Blanchet & Murthy (2019); Blanchet et al. (2022); Xu et al. (2025b). While the exact Wasserstein robustification provides a powerful and intuitive framework for hedging against distribution shifts, its dual reformulation often introduces nonsmoothness in the profiled objective. To induce structural smoothness while preserving the core robustification principles, we leverage entropic regularization and consider the Sinkhorn uncertainty set Cuturi (2013); Genevay et al. (2016); Peyr'e & Cuturi (2019); Wang et al. (2026).

A.2 TECHNICAL VERIFICATION OF THE REGULARITY ASSUMPTIONS

Remark A.1. *In the present DPO setting, the data space is structured as $\mathcal{Z} := \mathcal{S} \times \mathcal{A} \times \mathcal{A} \times \{0,1\}$, where $\mathcal{A}$ is a finite discrete action space and the preference labels are binary. Given that the policy features are bounded (as formally stated in Assumption 3.1, $\max_{s,a} \|\psi(s,a)\|_2 \leq 1$), the effective metric space inherently possesses a finite maximum diameter. In this finite/discrete and bounded-feature regime, any sensible continuous distance metric on this product space is strictly bounded. Moreover, since $c(x,z) \geq 0$ and ν is a probability measure, we have $\mathbb{E}_{z\sim\nu}\left[e^{-c(x,z)/\epsilon}\right] \leq 1 < \infty$, which means that the integrability condition appearing in (Wang et al., 2026, Assumption 1) is satisfied.*

Remark A.2 (Discrete Action Space and Policy Parameterization)**.** *Assumption 3.1 is a standard condition in RLHF alignment Zhu et al. (2023); Chowdhury et al. (2024); Xu et al. (2025b). While we assume the policy parameter θ is continuously optimized over a compact set Θ, the action space $\mathcal{A}$ is assumed to be finite and discrete. In the context of LLM alignment, $\mathcal{A}$ corresponds to a finite vocabulary or a fixed set of candidate responses. From a theoretical perspective, this discrete nature ensures that $\sum_{a'\in\mathcal{A}} \exp(\theta^\top \psi(s,a'))$ in the log-linear policy formulation* (4) *is well-defined and tractable. It circumvents the integration complexities and potential unboundedness issues that arise over continuous action spaces, thereby preserving the uniform boundedness of the loss gradients and Hessians required for subsequent analyses. Extending the results to continuous or very large action spaces would require a different treatment (such as neural network based methods) and is left for future work.*

A.3 PROOF OF THEORY

Lemma A.3 (Pointwise properties of the DPO loss)**.** *Xu et al. (2025b) Under Assumption 3.1, define $\Delta\psi(z) := \psi(s,a_1) - \psi(s,a_2)$ for $z = (s,a_1,a_2,y) \in \mathcal{Z}$. For every $z \in \mathcal{Z}$ and every $\theta \in \Theta$, the DPO loss $\ell(z;\theta)$ is twice continuously differentiable in θ, with derivatives:*

$$\nabla_\theta \ell(z;\theta) = \beta\big(\sigma(\beta h_\theta(s,a_1,a_2)) - y\big)\Delta\psi(z),$$
$$\nabla^2_\theta \ell(z;\theta) = \beta^2 \sigma(\beta h_\theta(s,a_1,a_2))\sigma(-\beta h_\theta(s,a_1,a_2))\,\Delta\psi(z)\Delta\psi(z)^\top,$$

where $\sigma(t) = (1+e^{-t})^{-1}$. The loss and its derivatives are uniformly bounded over $\mathcal{Z}\times\Theta$:

$$0 \leq \ell(z;\theta) \leq \log\big(1+e^{4\beta B}\big) =: \bar{\ell}, \quad \|\nabla_\theta \ell(z;\theta)\|_2 \leq 2\beta, \quad \|\nabla^2_\theta \ell(z;\theta)\|_{\mathrm{op}} \leq \beta^2.$$

Furthermore, the Hessian satisfies the following matrix inequality uniformly:

$$\gamma_0 \Delta\psi(z)\Delta\psi(z)^\top \preceq \nabla^2_\theta \ell(z;\theta) \preceq \gamma_1 \Delta\psi(z)\Delta\psi(z)^\top,$$

where $\gamma_0 := \frac{\beta^2 e^{4\beta B}}{(1+e^{4\beta B})^2}$ and $\gamma_1 := \frac{\beta^2}{4}$.

The analytical expressions for the derivatives and the positive semi-definite lower bound of the Hessian (γ_0) were established in Xu et al. (2025b) to prove strong convexity. Lemma A.3 introduces additional pointwise guarantees required for our Sinkhorn framework: the uniform upper bounds on the loss value ($\bar{\ell}$), the gradient norm, and the Hessian operator norm (γ_1). Although the core derivation shares underlying similarities with Xu et al. (2025b), we provide a complete and self-contained proof in the following for rigorous clarity and completeness.

Proof of Lemma A.3. Recall that the pointwise DPO loss is given by:

$$\ell(z;\theta) = -y \log\sigma\big(\beta h_\theta(s,a_1,a_2)\big) - (1-y)\log\sigma\big(-\beta h_\theta(s,a_1,a_2)\big),$$

where $y \in \{0,1\}$ and $z = (s,a_1,a_2,y) \in \mathcal{Z}$. By definition of the preference score under a parameterized policy π_θ and a reference policy π_{ref}:

$$h_\theta(s,a_1,a_2) = \log\pi_\theta(a_1 \mid s) - \log\pi_\theta(a_2 \mid s) - \Big(\log\pi_{\mathrm{ref}}(a_1 \mid s) - \log\pi_{\mathrm{ref}}(a_2 \mid s)\Big).$$

By the definition of the log-linear policies in Assumption 3.1, the preference score simplifies to:

$$h_\theta(s, a_1, a_2) = (\theta - \theta_{\text{ref}})^\top \big(\psi(s, a_1) - \psi(s, a_2)\big) = (\theta - \theta_{\text{ref}})^\top \Delta\psi(z). \tag{8}$$

Hence, the derivatives with respect to θ are independent of the partition function:

$$\nabla_\theta h_\theta(s, a_1, a_2) = \Delta\psi(z), \qquad \nabla^2_\theta h_\theta(s, a_1, a_2) = \mathbf{0}. \tag{9}$$

Next, let $u = \beta h_\theta(s, a_1, a_2)$. Using $\frac{d}{du} \log \sigma(u) = \sigma(-u)$ and $\frac{d}{du} \log \sigma(-u) = -\sigma(u)$, we compute the derivative of $\ell(z; \theta)$ with respect to u:

$$\frac{d\ell}{du} = -y\sigma(-u) + (1 - y)\sigma(u) = \sigma(u) - y,$$

where we used the identity $\sigma(-u) = 1 - \sigma(u)$. Applying the multivariate chain rule:

$$\nabla_\theta \ell(z; \theta) = \frac{d\ell}{du} \nabla_\theta u = \beta\big(\sigma(\beta h_\theta(s, a_1, a_2)) - y\big)\Delta\psi(z).$$

Differentiating once more with respect to θ, and noting that $\Delta\psi(z)$ is constant in θ:

$$\begin{aligned}
\nabla^2_\theta \ell(z; \theta) &= \beta \nabla_\theta \big[\sigma(\beta h_\theta(s, a_1, a_2))\big] \Delta\psi(z)^\top \\
&= \beta^2 \sigma'(\beta h_\theta(s, a_1, a_2)) \Delta\psi(z) \Delta\psi(z)^\top \\
&= \beta^2 \sigma(\beta h_\theta(s, a_1, a_2)) \sigma(-\beta h_\theta(s, a_1, a_2)) \Delta\psi(z) \Delta\psi(z)^\top,
\end{aligned}$$

where we utilized the derivative of the sigmoid function $\sigma'(t) = \sigma(t)(1 - \sigma(t)) = \sigma(t)\sigma(-t)$.

Now, we prove the uniform bounds over the joint space $\mathcal{Z} \times \Theta$:

1. Under Assumption 3.1, $\|\theta\|_2 \le B$, $\|\theta_{\text{ref}}\|_2 \le B$, and $\max_{s,a} \|\psi(s, a)\|_2 \le 1$. By the Cauchy-Schwarz inequality, for all $z \in \mathcal{Z}$ and $\theta \in \Theta$:

$$|h_\theta(s, a_1, a_2)| = \big|(\theta - \theta_{\text{ref}})^\top \Delta\psi(z)\big| \le \|\theta - \theta_{\text{ref}}\|_2 \|\Delta\psi(z)\|_2 \le (2B)(2) = 4B.$$

2. The loss can be written as:

$$\ell(z; \theta) = \begin{cases} -\log \sigma\big(\beta h_\theta(s, a_1, a_2)\big), & y = 1, \\ -\log \sigma\big(-\beta h_\theta(s, a_1, a_2)\big), & y = 0. \end{cases}$$

Since $-\log \sigma(x) = \log(1 + e^{-x})$ and $|\beta h_\theta| \le 4\beta B$, we immediately obtain the uniform bound:

$$0 \le \ell(z; \theta) \le \log(1 + e^{4\beta B}) =: \overline{\ell}.$$

3. Because $\|\Delta\psi(z)\|_2 \le \|\psi(s, a_1)\|_2 + \|\psi(s, a_2)\|_2 \le 2$ and $|\sigma(t) - y| \le 1$ for all $y \in \{0, 1\}$, $t \in \mathbb{R}$, we obtain a uniform bound on the gradient norm:

$$\|\nabla_\theta \ell(z; \theta)\|_2 \le \beta |\sigma(\beta h_\theta(s, a_1, a_2)) - y| \|\Delta\psi(z)\|_2 \le 2\beta.$$

4. Consider the scalar multiplier function $f(u) := \sigma(u)\sigma(-u) = \frac{e^u}{(1+e^u)^2}$. The function $f(u)$ is positive, symmetric, and bounded by its unique global maximum $f(0) = 1/4$. Since $|u| = |\beta h_\theta(s, a_1, a_2)| \le 4\beta B$ for all (z, θ), its minimum over this domain is attained at the boundary:

$$f(u) \ge f(4\beta B) = \frac{e^{4\beta B}}{(1 + e^{4\beta B})^2}.$$

Therefore, we obtain the positive semi-definite pointwise lower bound for the Hessian matrix, valid uniformly for all $(z, \theta) \in \mathcal{Z} \times \Theta$:

$$\nabla^2_\theta \ell(z;\theta) \succeq \beta^2 \frac{e^{4\beta B}}{(1+e^{4\beta B})^2} \Delta\psi(z)\Delta\psi(z)^\top.$$

For the upper bound, utilizing $f(u) \leq 1/4$ globally, we have:

$$\nabla^2_\theta \ell(z;\theta) \preceq \beta^2 \frac{1}{4} \Delta\psi(z)\Delta\psi(z)^\top.$$

Finally, to compute the spectral operator norm $\|\nabla^2_\theta \ell(z;\theta)\|_{\mathrm{op}}$, we observe that $\Delta\psi\Delta\psi^\top$ is a rank-1 symmetric positive semi-definite matrix. Its operator norm is exactly its unique non-zero eigenvalue, which is $\|\Delta\psi\|_2^2 \leq 4$. Therefore, we obtain

$$\|\nabla^2_\theta \ell(z;\theta)\|_{\mathrm{op}} = \beta^2 f(u) \|\Delta\psi\|_2^2 \leq \beta^2 \cdot \frac{1}{4} \cdot 4 = \beta^2.$$

This completes the proof. □

Equipped with the pointwise bounds and regularity conditions, we proceed to calculate the derivatives of the objective $\mathcal{L}(\theta, \lambda)$. This reveals the smoothness induced by the entropic regularization.

Proposition A.4 (Analytical derivatives of the Sinkhorn envelope)**.** *Under Assumptions 2.1, 3.1 and the notations of Lemma A.3, for any $\theta \in \Theta$ and any $\lambda \in (0, \infty)$, the joint objective function $\mathcal{L}(\theta, \lambda)$ is twice continuously differentiable. The exact analytical forms of its partial and mixed derivatives are given by:*

$$\nabla^2_{\theta\theta} \mathcal{L}(\theta,\lambda) = \mathbb{E}_{P_o}\left[\mathbb{E}_{Q_{\theta,\lambda}}\left[\nabla^2_\theta \ell(z;\theta)\right] + \frac{1}{\lambda\epsilon}\operatorname{Cov}_{Q_{\theta,\lambda}(\cdot|x)}\big(\nabla_\theta \ell(z;\theta)\big)\right] = \mathbb{E}_{P_o}\left[\nabla^2_{\theta\theta}\Psi_{\theta,\lambda}\right], \tag{10}$$

$$\partial^2_{\lambda\lambda} \mathcal{L}(\theta,\lambda) = \frac{1}{\lambda^3\epsilon}\mathbb{E}_{P_o}\left[\operatorname{Var}_{Q_{\theta,\lambda}(\cdot|x)}\big(\ell(z;\theta)\big)\right] = \mathbb{E}_{P_o}\left[\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}\right], \tag{11}$$

$$\partial^2_{\theta\lambda} \mathcal{L}(\theta,\lambda) = -\frac{1}{\lambda^2\epsilon}\mathbb{E}_{P_o}\left[\operatorname{Cov}_{Q_{\theta,\lambda}(\cdot|x)}\big(\nabla_\theta \ell(z;\theta), \ell(z;\theta)\big)\right] = \mathbb{E}_{P_o}\left[\partial^2_{\theta\lambda}\Psi_{\theta,\lambda}\right], \tag{12}$$

where the tilted probability measure $Q_{\theta,\lambda}(\cdot \mid x)$ is defined via its Radon-Nikodym derivative with respect to ν:

$$\frac{dQ_{\theta,\lambda}(z \mid x)}{d\nu(z)} = \frac{1}{Z_{\theta,\lambda}(x)} \exp\left(\frac{\ell(z;\theta) - \lambda c(x,z)}{\lambda\epsilon}\right), \tag{13}$$

and $Z_{\theta,\lambda}(x) := \mathbb{E}_{z\sim\nu}\left[\exp\left(\frac{\ell(z;\theta)-\lambda c(x,z)}{\lambda\epsilon}\right)\right]$ is the normalizing partition function.

Proof. Let the energy function be $E(x, z; \theta, \lambda) = \frac{\ell(z;\theta)-\lambda c(x,z)}{\lambda\epsilon}$, and define the partition function $Z_{\theta,\lambda}(x) := \mathbb{E}_{z\sim\nu}[\exp(E)]$.

Step 1: Justification for differentiating under the integral sign. To rigorously justify exchanging the differential operators (∂_λ, ∇_θ, and their higher-order combinations) with the expectation over ν, we rely on the Leibniz integral rule (Dominated Convergence Theorem) (see (Aliprantis & Burkinshaw, 1998, Theorem 24.5) (Billingsley, 2017, Theorem 16.8) (Cohn, 2013, Example 2.4.6)). This requires demonstrating that all relevant partial derivatives of the integrand $f(z;\theta,\lambda) := \exp(E)$ are bounded by an integrable function.

For any given $\lambda > 0$, we construct a compact local neighborhood $I_\lambda = [\lambda/2, 3\lambda/2] \subset (0, \infty)$. Within this neighborhood, the reciprocal $1/\lambda \leq 2/\lambda$ is strictly bounded. By Lemma A.3, the loss is uniformly bounded ($0 \leq \ell(z;\theta) \leq \overline{\ell}$). Thus, the exponential term itself is uniformly bounded over $I_\lambda \times \Theta \times \mathcal{Z}$:

$$\exp(E) \leq \exp\left(\frac{\overline{\ell}}{(\lambda/2)\epsilon}\right) =: E_{\max}.$$

Applying the chain and product rules, any first or second-order partial derivative of $f(z;\theta,\lambda)$ takes the analytical form of $\exp(E)$ multiplied by algebraic combinations of ℓ, $\nabla_\theta \ell$, $\nabla^2_\theta \ell$, and polynomial powers of $1/\lambda$. For instance:

$$|\partial_\lambda f| = \exp(E)\frac{\ell}{\lambda^2\epsilon} \le E_{\max}\frac{\overline{\ell}}{(\lambda/2)^2\epsilon},$$
$$\|\nabla_\theta f\|_2 = \exp(E)\frac{\|\nabla_\theta \ell\|_2}{\lambda\epsilon} \le E_{\max}\frac{2\beta}{(\lambda/2)\epsilon}.$$

By Lemma A.3, all necessary quantities ($\ell \le \overline{\ell}$, $\|\nabla_\theta \ell\|_2 \le 2\beta$, $\|\nabla^2_\theta \ell\|_{\mathrm{op}} \le \beta^2$) are uniformly bounded. Consequently, every partial derivative of $f(z;\theta,\lambda)$ up to the second order is strictly bounded by some finite constant $C_{\text{local}} > 0$ that depends solely on the chosen neighborhood I_λ and the established upper bounds, but is independent of z.

Because ν is a probability measure with $\nu(\mathcal{Z}) = 1$ (Assumption 2.1), this uniform constant serves as an integrable dominating function (i.e., $\int_{\mathcal{Z}} C_{\text{local}}\, d\nu = C_{\text{local}} < \infty$). This secures the condition for the Leibniz integral rule, allowing us to pass any differential operator up to the second order inside the expectation $\mathbb{E}_\nu$.

Step 2: Derivation of the λ derivatives. Having justified the exchange, we apply the ∂_λ operator directly to the integrand of the partition function to obtain $\partial_\lambda Z_{\theta,\lambda} = \mathbb{E}_\nu\left[\exp(E)\left(-\frac{\ell}{\lambda^2\epsilon}\right)\right]$. Thus, for $\Psi_{\theta,\lambda}(x) = \lambda\epsilon \log Z_{\theta,\lambda}(x)$:

$$\partial_\lambda \Psi_{\theta,\lambda}(x) = \epsilon \log Z_{\theta,\lambda} + \lambda\epsilon\frac{\partial_\lambda Z_{\theta,\lambda}}{Z_{\theta,\lambda}} = \epsilon \log Z_{\theta,\lambda} - \frac{1}{\lambda}\mathbb{E}_{Q_{\theta,\lambda}}[\ell(z;\theta)].$$

Differentiating again and grouping terms yields:

$$\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x) = \epsilon\frac{\partial_\lambda Z_{\theta,\lambda}}{Z_{\theta,\lambda}} + \frac{1}{\lambda^2}\mathbb{E}_{Q_{\theta,\lambda}}[\ell] - \frac{1}{\lambda}\partial_\lambda \mathbb{E}_{Q_{\theta,\lambda}}[\ell].$$

The first two terms exactly cancel out since $\epsilon\frac{\partial_\lambda Z_{\theta,\lambda}}{Z_{\theta,\lambda}} = -\frac{1}{\lambda^2}\mathbb{E}_{Q_{\theta,\lambda}}[\ell]$. To evaluate the third term, we apply the quotient rule to $\mathbb{E}_{Q_{\theta,\lambda}}[\ell] = N/D$, where $N = \mathbb{E}_\nu[\ell\exp(E)]$ and $D = Z_{\theta,\lambda} = \mathbb{E}_\nu[\exp(E)]$.

Crucially, by Lemma A.3 the loss is uniformly bounded ($0 \le \ell \le \overline{\ell}$), and by Assumption 2.1 the transport cost is bounded ($0 \le c \le \bar{c}$) and ν is a probability measure. This guarantees that the partition function $Z_{\theta,\lambda}(x)$ is strictly positive and finite, $Q_{\theta,\lambda}$ is a well-defined probability measure, and all moments of ℓ (including $\mathbb{E}_{Q_{\theta,\lambda}}[\ell^2]$) are strictly finite, validating the following algebraic expansion:

$$\begin{aligned}
\partial_\lambda\left(\frac{N}{D}\right) &= \frac{(\partial_\lambda N)D - N(\partial_\lambda D)}{D^2} \\
&= \frac{\left(-\frac{1}{\lambda^2\epsilon}\mathbb{E}_\nu[\ell^2\exp(E)]\right)D - N\left(-\frac{1}{\lambda^2\epsilon}\mathbb{E}_\nu[\ell\exp(E)]\right)}{D^2} \\
&= -\frac{1}{\lambda^2\epsilon}\left[\frac{\mathbb{E}_\nu[\ell^2\exp(E)]}{D} - \left(\frac{N}{D}\right)^2\right] \\
&= -\frac{1}{\lambda^2\epsilon}\left(\mathbb{E}_{Q_{\theta,\lambda}}[\ell^2] - \left(\mathbb{E}_{Q_{\theta,\lambda}}[\ell]\right)^2\right) = -\frac{1}{\lambda^2\epsilon}\operatorname{Var}_{Q_{\theta,\lambda}}(\ell).
\end{aligned}$$

Substituting this back yields the precise identity $\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x) = \frac{1}{\lambda^3\epsilon}\operatorname{Var}_{Q_{\theta,\lambda}(\cdot|x)}(\ell)$.

Step 3: Derivation of the θ derivatives. Applying the Leibniz rule with respect to the vector θ, we have $\nabla_\theta Z_{\theta,\lambda} = \mathbb{E}_\nu\left[\exp(E)\frac{\nabla_\theta \ell}{\lambda\epsilon}\right]$. Thus, the gradient of Ψ is:

$$\nabla_\theta \Psi_{\theta,\lambda}(x) = \lambda\epsilon \frac{\nabla_\theta Z_{\theta,\lambda}}{Z_{\theta,\lambda}} = \mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell].$$

To find the Hessian $\nabla^2_{\theta\theta}\Psi_{\theta,\lambda}(x)$, we differentiate this vector again. Let $\mathbf{N}_\theta = \mathbb{E}_\nu[\nabla_\theta \ell \exp(E)]$ and $D = Z_{\theta,\lambda}$. By the product and chain rules, the Jacobian matrix of $\mathbf{N}_\theta is$:

$$\nabla_\theta \mathbf{N}_\theta = \mathbb{E}_\nu\left[\nabla^2_\theta \ell \exp(E) + \nabla_\theta \ell \exp(E)\frac{(\nabla_\theta \ell)^\top}{\lambda\epsilon}\right].$$

Applying the multivariate quotient rule for the Jacobian of $\mathbf{N}_\theta / D$:

$$\begin{aligned}
\nabla_\theta\left(\frac{\mathbf{N}_\theta}{D}\right) &= \frac{\nabla_\theta \mathbf{N}_\theta}{D} - \frac{\mathbf{N}_\theta(\nabla_\theta D)^\top}{D^2} \\
&= \mathbb{E}_{Q_{\theta,\lambda}}\left[\nabla^2_\theta \ell + \frac{1}{\lambda\epsilon}\nabla_\theta \ell(\nabla_\theta \ell)^\top\right] - \frac{\mathbf{N}_\theta}{D}\left(\frac{1}{\lambda\epsilon}\mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell]^\top\right) \\
&= \mathbb{E}_{Q_{\theta,\lambda}}[\nabla^2_\theta \ell] + \frac{1}{\lambda\epsilon}\left(\mathbb{E}_{Q_{\theta,\lambda}}\left[\nabla_\theta \ell(\nabla_\theta \ell)^\top\right] - \mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell]\mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell]^\top\right) \\
&= \mathbb{E}_{Q_{\theta,\lambda}}[\nabla^2_\theta \ell] + \frac{1}{\lambda\epsilon}\operatorname{Cov}_{Q_{\theta,\lambda}}(\nabla_\theta \ell).
\end{aligned}$$

Step 4: Derivation of the mixed derivative. We rigorously derive the mixed derivative $\partial^2_{\theta\lambda}\Psi_{\theta,\lambda}(x) = \partial_\lambda\big(\mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell]\big)$. Using the vector numerator $\mathbf{N}_\theta = \mathbb{E}_\nu[\nabla_\theta \ell \exp(E)]$ and the denominator $D = Z_{\theta,\lambda}$, the derivative with respect to λ is $\partial_\lambda \mathbf{N}_\theta = -\frac{1}{\lambda^2\epsilon}\mathbb{E}_\nu[\nabla_\theta \ell \cdot \ell \exp(E)]$. Applying the quotient rule component-wise:

$$\begin{aligned}
\partial_\lambda\left(\frac{\mathbf{N}_\theta}{D}\right) &= \frac{(\partial_\lambda \mathbf{N}_\theta)D - \mathbf{N}_\theta(\partial_\lambda D)}{D^2} \\
&= \frac{\left(-\frac{1}{\lambda^2\epsilon}\mathbb{E}_\nu[\nabla_\theta \ell \cdot \ell \exp(E)]\right)D - \mathbf{N}_\theta\left(-\frac{1}{\lambda^2\epsilon}\mathbb{E}_\nu[\ell \exp(E)]\right)}{D^2} \\
&= -\frac{1}{\lambda^2\epsilon}\left[\mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell \cdot \ell] - \mathbb{E}_{Q_{\theta,\lambda}}[\nabla_\theta \ell]\mathbb{E}_{Q_{\theta,\lambda}}[\ell]\right] \\
&= -\frac{1}{\lambda^2\epsilon}\operatorname{Cov}_{Q_{\theta,\lambda}}\big(\nabla_\theta \ell, \ell\big).
\end{aligned}$$

Step 5: Uniform bounding and the outer expectation over P_o. To rigorously exchange these derived analytical forms with the outer expectation over $x \sim P_o$, we bound the resulting pointwise derivatives. For any random variable $X \in [0, \overline{\ell}]$, Popoviciu's inequality guarantees $\operatorname{Var}_{Q_{\theta,\lambda}}(\ell) \le \frac{\overline{\ell}^2}{4}$. Thus, $|\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x)| \le \frac{\overline{\ell}^2}{4\lambda^3\epsilon}$. For the gradient-related terms, since $\|\nabla_\theta \ell\|_2 \le 2\beta$ and $\|\nabla^2_\theta \ell\|_{\text{op}} \le \beta^2$, the spectral norm of the covariance matrix is bounded by $4\beta^2$, yielding $\|\nabla^2_{\theta\theta}\Psi_{\theta,\lambda}(x)\|_{\text{op}} \le \beta^2 + \frac{4\beta^2}{\lambda\epsilon}$. Similarly, applying the triangle inequality to the mixed covariance:

$$\left\|\operatorname{Cov}_{Q_{\theta,\lambda}}(\nabla_\theta \ell, \ell)\right\|_2 = \|\mathbb{E}_Q[\nabla_\theta \ell \cdot \ell] - \mathbb{E}_Q[\nabla_\theta \ell]\mathbb{E}_Q[\ell]\|_2 \le 2\beta\overline{\ell} + (2\beta)\overline{\ell} = 4\beta\overline{\ell}.$$

Because these norms are bounded by absolute constants independent of x, they constitute integrable dominating functions in $L^1(P_o)$. This validates the final application of the Leibniz integral rule to distribute $\mathbb{E}_{P_o}$ across the derived components, completing the proof for (10), (11), and (12). □

Proof of Lemma 3.5. We define the gradient mapping associated with the first-order optimality condition of the dual problem:

$$H(\theta, \lambda) := \partial_\lambda \mathcal{L}(\theta, \lambda). \tag{14}$$

First, we establish the existence and strict positivity of the dual minimizer $\lambda^\star(\theta)$. By Lemma A.3, the loss function is uniformly bounded by a finite constant, i.e., $\ell(z;\theta) \le \overline{\ell}$. Consequently, for any $\lambda > 0$, the exponential integrability condition (Condition 1 in Wang et al. (2026)) is inherently satisfied, as $\mathbb{E}_{z\sim\mathbb{Q}_{x,\epsilon}}[e^{\ell(z;\theta)/(\lambda\epsilon)}] \le e^{\overline{\ell}/(\lambda\epsilon)} < \infty$. Furthermore, Assumption 3.3 postulates that the effective radius is strictly positive, i.e., $\bar{\rho} > 0$. Given that $\bar{\rho} > 0$ and Condition 1 of Wang et al. (2026) is satisfied, (Wang et al., 2026, Lemma EC.4) guarantees the existence of a dual minimizer $\lambda^\star(\theta)$ which either equals 0 or satisfies Condition 1 in Wang et al. (2026).

To rule out the boundary case $\lambda^\star(\theta) = 0$, we rely on Lemma EC.5 and Theorem 1(IV) of Wang et al. (2026). These results dictate that, under the premises of $\bar{\rho} > 0$ and Condition 1 in Wang et al. (2026), the dual minimizer attains $\lambda^\star(\theta) = 0$ if and only if the loss function is essentially bounded from above and $\bar{\rho} \ge \bar{\rho}_{\text{crit}}(\theta)$. While our loss function is bounded, Assumption 3.3 enforces $\bar{\rho} < \bar{\rho}_{\text{crit}}(\theta)$ for all $\theta \in \Theta$. This strict inequality falsifies the necessary condition for $\lambda^\star(\theta) = 0$. Thus, we must have $\lambda^\star(\theta) > 0$. As a strictly positive interior extremum on $(0, \infty)$, the minimizer $\lambda^\star(\theta)$ necessarily satisfies the first-order stationarity condition $H(\theta, \lambda^\star(\theta)) = 0$.

Taking the partial derivative of H with respect to λ yields the second derivative derived in Proposition A.4:

$$\partial_\lambda H(\theta, \lambda^\star(\theta)) = \partial^2_{\lambda\lambda} \mathcal{L}(\theta, \lambda^\star(\theta)) = \frac{1}{(\lambda^\star(\theta))^3 \epsilon} \mathbb{E}_{P_o}\left[\mathrm{Var}_{Q_{\theta,\lambda^\star(\theta)}(\cdot|x)}\big(\ell(z;\theta)\big)\right]. \tag{15}$$

Under Assumption 3.2, the loss has strictly positive variance under the reference measure, i.e., $\mathrm{Var}_\nu(\ell) > 0$. By definition, the Radon-Nikodym derivative of the tilted exponential measure $Q_{\theta,\lambda}$ with respect to the reference measure ν is proportional to the term $\exp(E) := \exp\big(\frac{\ell(z;\theta) - \lambda c(x,z)}{\lambda\epsilon}\big)$. Using Lemma A.3, both the loss function $\ell(z;\theta)$ and the transport cost $c(x,z)$ are uniformly bounded over their domains. Because both terms are finite, this exponential density $\exp(E)$ is strictly positive and finite everywhere (bounded away from zero and infinity). Consequently, the probability measures $Q_{\theta,\lambda}$ and ν share identical null sets, making them mutually absolutely continuous. This absolute continuity guarantees that $\mathrm{Var}_\nu(\ell) > 0 \iff \mathrm{Var}_{Q_{\theta,\lambda}}(\ell) > 0$. Thus, $\partial_\lambda H(\theta, \lambda^\star(\theta)) > 0$.

Because $H(\theta, \lambda^\star(\theta)) = 0$ and the Jacobian $\partial_\lambda H(\theta, \lambda^\star(\theta))$ is strictly non-zero, the implicit function theorem guarantees that around each $\theta \in \Theta$, there exists an open neighborhood wherein the unique solution $\lambda^\star(\theta)$ is a C^1 function. Therefore, the mapping $\theta \mapsto \lambda^\star(\theta)$ is continuous on the entire domain Θ.

Finally, by the assumed compactness of the parameter space Θ and the continuity of the mapping $\theta \mapsto \lambda^\star(\theta)$, its image $\lambda^\star(\Theta)$ constitutes a compact set within the strictly positive domain $(0, \infty)$. By the Extreme Value Theorem, this image attains a strictly positive minimum $\underline{\lambda} > 0$ and a finite maximum $\overline{\lambda} < \infty$, establishing the uniform global bounds. □

Proof of Theorem 3.8. By Lemma 3.5, $\lambda^\star(\theta)$ is a continuously differentiable interior minimizer for any $\theta \in \Theta$. The multivariate Envelope Theorem implies that the gradient of the profiled objective coincides with the partial gradient of $\mathcal{L}$ evaluated at the optimum, i.e., $\nabla_\theta F(\theta) = \nabla_\theta \mathcal{L}(\theta, \lambda^\star(\theta))$.

To obtain the Hessian of $F(\theta)$, we apply the multivariate chain rule to compute the total derivative of $\nabla_\theta F(\theta)$ with respect to θ:

$$\nabla^2_\theta F(\theta) = \frac{d}{d\theta}\big[\nabla_\theta \mathcal{L}(\theta, \lambda^\star(\theta))\big] = \nabla^2_{\theta\theta} \mathcal{L}(\theta, \lambda^\star) + \partial^2_{\theta\lambda} \mathcal{L}(\theta, \lambda^\star)\big(\nabla_\theta \lambda^\star(\theta)\big)^\top. \tag{16}$$

Differentiating the first-order optimality condition $\partial_\lambda \mathcal{L}(\theta, \lambda^\star(\theta)) \equiv 0$ with respect to θ yields:

$$\partial^2_{\lambda\theta}\mathcal{L}(\theta,\lambda^\star) + \partial^2_{\lambda\lambda}\mathcal{L}(\theta,\lambda^\star)\nabla_\theta\lambda^\star(\theta) = 0 \implies \nabla_\theta\lambda^\star(\theta) = -\frac{\partial^2_{\lambda\theta}\mathcal{L}(\theta,\lambda^\star)}{\partial^2_{\lambda\lambda}\mathcal{L}(\theta,\lambda^\star)}, \tag{17}$$

where Assumption 3.2 guarantees $\partial^2_{\lambda\lambda}\mathcal{L}(\theta,\lambda^\star) > 0$. Substituting (17) into (16) gives:

$$\nabla^2_\theta F(\theta) = \nabla^2_{\theta\theta}\mathcal{L}(\theta,\lambda^\star) - \frac{\left(\partial^2_{\theta\lambda}\mathcal{L}(\theta,\lambda^\star)\right)\left(\partial^2_{\theta\lambda}\mathcal{L}(\theta,\lambda^\star)\right)^\top}{\partial^2_{\lambda\lambda}\mathcal{L}(\theta,\lambda^\star)}. \tag{18}$$

To establish a lower bound for $\nabla^2_\theta F(\theta)$, we must control the negative term in (18) to ensure it does not compromise the positive semi-definiteness. We achieve this by constructing the joint covariance matrix of the gradient vector $\nabla_\theta\ell$ and the scalar loss ℓ under the tilted measure $Q_{\theta,\lambda^\star}$:

$$\mathbf{M} := \mathbb{E}_{P_o}\left[\mathrm{Cov}_{Q_{\theta,\lambda^\star}}\begin{pmatrix}\nabla_\theta\ell\\ \ell\end{pmatrix}\right] = \begin{pmatrix}\mathbf{A} & \mathbf{b}\\ \mathbf{b}^\top & c\end{pmatrix},$$

where $\mathbf{A} = \mathbb{E}_{P_o}[\mathrm{Cov}_{Q_{\theta,\lambda^\star}}(\nabla_\theta\ell)] \in \mathbb{R}^{d\times d}$, $\mathbf{b} = \mathbb{E}_{P_o}[\mathrm{Cov}_{Q_{\theta,\lambda^\star}}(\nabla_\theta\ell,\ell)] \in \mathbb{R}^{d\times 1}$, and $c = \mathbb{E}_{P_o}[\mathrm{Var}_{Q_{\theta,\lambda^\star}}(\ell)] \in \mathbb{R}$. These blocks match the exact terms derived in Proposition A.4.

Since any covariance matrix is positive semi-definite, its expectation over the nominal distribution P_o is also positive semi-definite ($\mathbf{M} \succeq 0$). Furthermore, by Assumption 3.2 and the mutual absolute continuity between $Q_{\theta,\lambda^\star}$ and ν, the variance of the loss is strictly positive, which implies that $c = \mathbb{E}_{P_o}[\mathrm{Var}_{Q_{\theta,\lambda^\star}}(\ell)] > 0$. Given $\mathbf{M} \succeq 0$ and $c > 0$, a standard property of block positive semi-definite matrices is that the Schur complement of the lower-right block c is positive semi-definite. This implies $\mathbf{A} - \frac{1}{c}\mathbf{b}\mathbf{b}^\top \succeq 0$.

Substituting the analytical derivatives from Proposition A.4 into Equation (18), we obtain:

$$\begin{aligned}\nabla^2_\theta F(\theta) &= \mathbb{E}_{P_o}\left[\mathbb{E}_{Q_{\theta,\lambda^\star}}[\nabla^2_\theta\ell]\right] + \frac{1}{\lambda^\star\epsilon}\mathbf{A} - \frac{\left(-\frac{1}{(\lambda^\star)^2\epsilon}\mathbf{b}\right)\left(-\frac{1}{(\lambda^\star)^2\epsilon}\mathbf{b}\right)^\top}{\frac{1}{(\lambda^\star)^3\epsilon}c}\\ &= \mathbb{E}_{P_o}\left[\mathbb{E}_{Q_{\theta,\lambda^\star}}[\nabla^2_\theta\ell]\right] + \frac{1}{\lambda^\star\epsilon}\left(\mathbf{A} - \frac{1}{c}\mathbf{b}\mathbf{b}^\top\right).\end{aligned}$$

Because $\mathbf{A} - \frac{1}{c}\mathbf{b}\mathbf{b}^\top \succeq 0$, the negative term in the expression is absorbed by the covariance block $\mathbf{A}$. This leaves a lower bound determined solely by the expected Hessian of the loss:

$$\nabla^2_\theta F(\theta) \succeq \mathbb{E}_{x\sim P_o}\left[\mathbb{E}_{z\sim Q_{\theta,\lambda^\star}(\cdot|x)}\left[\nabla^2_\theta\ell(z;\theta)\right]\right].$$

To relate this back to the reference measure ν and apply the data coverage condition (Assumption 3.6), we establish a uniform lower bound on the Radon-Nikodym derivative $dQ_{\theta,\lambda^\star}/d\nu$. From Proposition A.4, this derivative is proportional to the exponential term $\exp\left(\frac{\ell(z;\theta)-\lambda^\star c(x,z)}{\lambda^\star\epsilon}\right)$, normalized by the partition function $Z_{\theta,\lambda^\star}(x)$. Since $0 \le \ell(z;\theta) \le \bar{\ell}$ and $0 \le c(x,z) \le \bar{c}$, the exponential numerator is bounded from below by $\exp\left(-\frac{\bar{c}}{\epsilon}\right)$, and the denominator is bounded from above by $Z_{\theta,\lambda^\star}(x) \le \exp\left(\frac{\bar{\ell}}{\lambda^\star\epsilon}\right)$. Therefore, the density ratio satisfies:

$$\frac{dQ_{\theta,\lambda^\star}(z \mid x)}{d\nu(z)} \ge \exp\left(-\frac{\bar{c}}{\epsilon}\right)\exp\left(-\frac{\bar{\ell}}{\lambda^\star\epsilon}\right) = \exp\left(-\frac{\bar{\ell}+\lambda^\star(\theta)\bar{c}}{\lambda^\star(\theta)\epsilon}\right).$$

By Lemma 3.5, $\lambda^\star(\theta)$ has a global lower bound $\lambda^\star(\theta) \ge \underline{\lambda} > 0$. Since the exponential term is monotonically increasing with respect to $\lambda^\star(\theta)$, it attains its minimum at this boundary, providing a uniform constant lower bound for the density ratio:

$$\frac{dQ_{\theta,\lambda^\star}(z \mid x)}{d\nu(z)} \ge \exp\left(-\frac{\bar{\ell}+\underline{\lambda}\,\bar{c}}{\underline{\lambda}\,\epsilon}\right) =: c_{\min} > 0.$$

Applying the pointwise Hessian lower bound from Lemma A.3 ($\nabla^2_\theta \ell \succeq \gamma_0 \Delta\psi \Delta\psi^\top$), we integrate over the tilted measure $Q_{\theta,\lambda^\star}$ to deduce:

$$\begin{aligned}\nabla^2_\theta F(\theta) &\succeq \gamma_0 \, \mathbb{E}_{x\sim P_o}\left[\mathbb{E}_{z\sim Q_{\theta,\lambda^\star}(\cdot|x)}\left[\Delta\psi(z)\Delta\psi(z)^\top\right]\right] \\ &\succeq \gamma_0 \, c_{\min} \, \mathbb{E}_{x\sim P_o}\left[\mathbb{E}_{z\sim\nu}\left[\Delta\psi(z)\Delta\psi(z)^\top\right]\right].\end{aligned} \tag{19}$$

The inner expectation over the reference measure ν in (19) is independent of the empirical sample x. Thus, evaluating the outer expectation over P_o yields:

$$\nabla^2_\theta F(\theta) \succeq \gamma_0 \, c_{\min} \, \mathbb{E}_{z\sim\nu}\left[\Delta\psi(z)\Delta\psi(z)^\top\right] = \gamma_0 \, c_{\min} \Sigma_\nu. \tag{20}$$

Finally, applying the data coverage condition $\Sigma_\nu \succeq \lambda_0 I_d$ (Assumption 3.6) yields $\nabla^2_\theta F(\theta) \succeq (\gamma_0\lambda_0 c_{\min}) I_d$. This yields the positive strong convexity modulus $m_F = \gamma_0\lambda_0 c_{\min} > 0$, concluding the proof. □

Lemma A.5 (Derived uniform bounds for the profiled loss, dual derivative, and profiled gradient)**.** *Suppose Assumptions 2.1, 3.1, 3.2 and 3.3 hold. Let $\Lambda = [\underline{\lambda}, \overline{\lambda}] \subset (0,\infty)$ be the compact interval from Lemma 3.5, and define $\mathcal{W} := \Theta \times \Lambda$. For $x \in \mathcal{Z}$, we define three evaluation functions based on Proposition A.4:*

$$h_0(\theta,\lambda;x) := \Psi_{\theta,\lambda}(x), \qquad h_1(\theta,\lambda;x) := \partial_\lambda \Psi_{\theta,\lambda}(x), \qquad h_2(\theta,\lambda;x) := \nabla_\theta \Psi_{\theta,\lambda}(x).$$

Then there exist finite constants $M_0, M_1, M_2 > 0$ such that

$$\sup_{(\theta,\lambda)\in\mathcal{W}, x\in\mathcal{Z}} |h_0(\theta,\lambda;x)| \le M_0, \tag{21}$$

$$\sup_{(\theta,\lambda)\in\mathcal{W}, x\in\mathcal{Z}} |h_1(\theta,\lambda;x)| \le M_1, \tag{22}$$

$$\sup_{(\theta,\lambda)\in\mathcal{W}, x\in\mathcal{Z}} \|h_2(\theta,\lambda;x)\|_2 \le M_2. \tag{23}$$

Moreover, there exist finite constants $L_0, L_1, L_2 > 0$ such that for all $(\theta,\lambda),(\theta',\lambda') \in \mathcal{W}$ and all $x \in \mathcal{Z}$,

$$|h_0(\theta,\lambda;x) - h_0(\theta',\lambda';x)| \le L_0 \|(\theta,\lambda) - (\theta',\lambda')\|_2, \tag{24}$$

$$|h_1(\theta,\lambda;x) - h_1(\theta',\lambda';x)| \le L_1 \|(\theta,\lambda) - (\theta',\lambda')\|_2, \tag{25}$$

$$\|h_2(\theta,\lambda;x) - h_2(\theta',\lambda';x)\|_2 \le L_2 \|(\theta,\lambda) - (\theta',\lambda')\|_2. \tag{26}$$

Proof. We derive the bounds directly from the formulas in Proposition A.4 and the pointwise estimates from Lemma A.3.

Step 1: A bound for $h_0 = \Psi_{\theta,\lambda}$. Since $0 \le \ell(z;\theta) \le \overline{\ell}$ and $0 \le c(x,z) \le \bar{c}$, for every $(\theta,\lambda,x,z) \in \Theta \times \Lambda \times \mathcal{Z} \times \mathcal{Z}$ we have

$$e^{-\bar{c}/\epsilon} \le \exp\left(\frac{\ell(z;\theta) - \lambda c(x,z)}{\lambda\epsilon}\right) \le e^{\overline{\ell}/(\underline{\lambda}\epsilon)}.$$

Therefore,

$$e^{-\bar{c}/\epsilon} \le Z_{\theta,\lambda}(x) := \mathbb{E}_{z\sim\nu}\left[\exp\left(\frac{\ell(z;\theta) - \lambda c(x,z)}{\lambda\epsilon}\right)\right] \le e^{\overline{\ell}/(\underline{\lambda}\epsilon)}.$$

Using $\Psi_{\theta,\lambda}(x) = \lambda\epsilon \log Z_{\theta,\lambda}(x)$, we obtain $-\overline{\lambda}\bar{c} \le \Psi_{\theta,\lambda}(x) \le \overline{\ell}$. which implies that (21) holds with $M_0 := \max\{\overline{\ell}, \overline{\lambda}\bar{c}\}$.

Step 2: A bound for $h_2 = \nabla_\theta \Psi_{\theta,\lambda}$. Proposition A.4 gives $\nabla_\theta \Psi_{\theta,\lambda}(x) = \mathbb{E}_{Q_{\theta,\lambda}(\cdot|x)}[\nabla_\theta \ell(z;\theta)]$. Thus, we obtain

$$\|h_2(\theta,\lambda;x)\|_2 \le \mathbb{E}_{Q_{\theta,\lambda}(\cdot|x)}\left[\|\nabla_\theta \ell(z;\theta)\|_2\right] \le 2\beta,$$

where the last inequality follows from Lemma A.3. Thus, (23) holds with $M_2 := 2\beta$.

Step 3: A bound for $h_1 = \partial_\lambda \Psi_{\theta,\lambda}$. Following the proof of Proposition A.4 yields

$$\partial_\lambda \Psi_{\theta,\lambda}(x) = \epsilon \log Z_{\theta,\lambda}(x) - \frac{1}{\lambda} \mathbb{E}_{Q_{\theta,\lambda}(\cdot|x)}[\ell(z;\theta)].$$

From Step 1 and $0 \le \ell \le \bar{\ell}$, we have

$$-\bar{c} \le \epsilon \log Z_{\theta,\lambda}(x) \le \frac{\bar{\ell}}{\underline{\lambda}}, \qquad 0 \le \frac{1}{\lambda} \mathbb{E}_{Q_{\theta,\lambda}(\cdot|x)}[\ell(z;\theta)] \le \frac{\bar{\ell}}{\underline{\lambda}}.$$

which means that $|\partial_\lambda \Psi_{\theta,\lambda}(x)| \le \bar{c} + \frac{\bar{\ell}}{\underline{\lambda}}$. Thus, (22) holds with $M_1 := \bar{c} + \frac{\bar{\ell}}{\underline{\lambda}}$.

Step 4: Lipschitz continuity of h_0. Since $\Psi_{\theta,\lambda}$ is C^1 on $\mathcal{W}$ by Proposition A.4, the mean value theorem gives

$$|h_0(\theta,\lambda;x) - h_0(\theta',\lambda';x)| \le \sup_{(\vartheta,\eta)\in\mathcal{W}} \|\nabla_{(\theta,\lambda)} \Psi_{\vartheta,\eta}(x)\|_2 \|(\theta,\lambda) - (\theta',\lambda')\|_2.$$

Since

$$\nabla_{(\theta,\lambda)} \Psi_{\theta,\lambda}(x) = \begin{pmatrix} \nabla_\theta \Psi_{\theta,\lambda}(x) \\ \partial_\lambda \Psi_{\theta,\lambda}(x) \end{pmatrix},$$

combining Steps 2 and 3 yields $\|\nabla_{(\theta,\lambda)} \Psi_{\theta,\lambda}(x)\|_2 \le \sqrt{M_2^2 + M_1^2}$. Thus, (24) holds with $L_0 := \sqrt{M_2^2 + M_1^2}$.

Step 5: Lipschitz continuity of h_1. Again by the mean value theorem,

$$|h_1(\theta,\lambda;x) - h_1(\theta',\lambda';x)| \le \sup_{(\vartheta,\eta)\in\mathcal{W}} \|\nabla_{(\theta,\lambda)} \partial_\lambda \Psi_{\vartheta,\eta}(x)\|_2 \|(\theta,\lambda) - (\theta',\lambda')\|_2.$$

From Proposition A.4, we know that

$$\nabla_{(\theta,\lambda)} \partial_\lambda \Psi_{\theta,\lambda}(x) = \begin{pmatrix} \partial^2_{\theta\lambda} \Psi_{\theta,\lambda}(x) \\ \partial^2_{\lambda\lambda} \Psi_{\theta,\lambda}(x) \end{pmatrix}.$$

where the mixed derivative are calculated by

$$\partial^2_{\theta\lambda} \Psi_{\theta,\lambda}(x) = -\frac{1}{\lambda^2 \epsilon} \operatorname{Cov}_{Q_{\theta,\lambda}(\cdot|x)}(\nabla_\theta \ell, \ell), \quad \partial^2_{\lambda\lambda} \Psi_{\theta,\lambda}(x) = \frac{1}{\lambda^3 \epsilon} \operatorname{Var}_{Q_{\theta,\lambda}(\cdot|x)}(\ell).$$

Thus, using $\|\nabla_\theta \ell\|_2 \le 2\beta$ and $\ell \le \bar{\ell}$, we obtain $\|\partial^2_{\theta\lambda} \Psi_{\theta,\lambda}(x)\|_2 \le \frac{4\beta\bar{\ell}}{\underline{\lambda}^2 \epsilon}$. Since $\ell \in [0, \bar{\ell}]$, Popoviciu's inequality gives $\operatorname{Var}_{Q_{\theta,\lambda}(\cdot|x)}(\ell) \le \frac{\bar{\ell}^2}{4}$, which implies that $|\partial^2_{\lambda\lambda} \Psi_{\theta,\lambda}(x)| \le \frac{\bar{\ell}^2}{4\underline{\lambda}^3 \epsilon}$. Thus, (25) holds with $L_1 := \sqrt{\left(\frac{4\beta\bar{\ell}}{\underline{\lambda}^2\epsilon}\right)^2 + \left(\frac{\bar{\ell}^2}{4\underline{\lambda}^3\epsilon}\right)^2}$.

Step 6: Lipschitz continuity of h_2. By the mean value theorem, we get

$$\|h_2(\theta,\lambda;x) - h_2(\theta',\lambda';x)\|_2 \le \sup_{(\vartheta,\eta)\in\mathcal{W}} \|\nabla_{(\theta,\lambda)} \nabla_\theta \Psi_{\vartheta,\eta}(x)\|_{\mathrm{op}} \|(\theta,\lambda) - (\theta',\lambda')\|_2.$$

The Jacobian of h_2 with respect to (θ,λ) has blocks $\nabla^2_{\theta\theta} \Psi_{\theta,\lambda}(x)$ and $\partial^2_{\lambda\theta} \Psi_{\theta,\lambda}(x)$. Proposition A.4 gives $\|\nabla^2_{\theta\theta} \Psi_{\theta,\lambda}(x)\|_{\mathrm{op}} \le \beta^2 + \frac{4\beta^2}{\underline{\lambda}\epsilon}$, and the mixed derivative bound from Step 5 yields $\|\partial^2_{\lambda\theta} \Psi_{\theta,\lambda}(x)\|_2 \le \frac{4\beta\bar{\ell}}{\underline{\lambda}^2\epsilon}$. Hence, (26) holds with $L_2 := \sqrt{\left(\beta^2 + \frac{4\beta^2}{\underline{\lambda}\epsilon}\right)^2 + \left(\frac{4\beta\bar{\ell}}{\underline{\lambda}^2\epsilon}\right)^2}$. This proves all bounds. □

Lemma A.6 (Finite-sample uniform concentration of the profiled objective and gradient)**.** *Suppose Assumptions 2.1, 3.1, 3.2, 3.3 and 3.6 hold. Let* $\Lambda = [\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$*, and* $\mathcal{W} := \Theta \times \Lambda$*. Define* $m_\lambda := \inf_{(\theta,\lambda)\in\mathcal{W}} \partial^2_{\lambda\lambda}\mathcal{L}(\theta, \lambda)$*, and* $C_{\theta\lambda} := \sup_{(\theta,\lambda)\in\mathcal{W}} \|\partial^2_{\theta\lambda}\Psi_{\theta,\lambda}\|_2$*. Then,* $m_\lambda > 0$*. Let* $M_0, M_1, M_2, L_0, L_1, L_2$ *be the constants from Lemma A.5. Let* $\delta \in (0, 1)$ *and* $n \geq 1$*. Then, with probability at least* $1 - \delta$*,*

$$\sup_{\theta\in\Theta} |F_n(\theta) - F(\theta)| \leq \mathcal{E}_{n,0}(\delta), \tag{27}$$

$$\sup_{\theta\in\Theta} \|\nabla_\theta F_n(\theta) - \nabla_\theta F(\theta)\|_2 \leq \mathcal{E}_{n,2}(\delta) + \frac{C_{\theta\lambda}}{m_\lambda}\mathcal{E}_{n,1}(\delta). \tag{28}$$

where

$$\mathcal{E}_{n,0}(\delta) := M_0\sqrt{\frac{2}{n}\left((d+1)\log\big(3R_{\mathcal{W}}\sqrt{n}\big) + \log\frac{6}{\delta}\right)} + \frac{2L_0}{\sqrt{n}}, \tag{29}$$

$$\mathcal{E}_{n,1}(\delta) := M_1\sqrt{\frac{2}{n}\left((d+1)\log\big(3R_{\mathcal{W}}\sqrt{n}\big) + \log\frac{6}{\delta}\right)} + \frac{2L_1}{\sqrt{n}}, \tag{30}$$

$$\mathcal{E}_{n,2}(\delta) := M_2\sqrt{\frac{2d}{n}\left((d+1)\log\big(3R_{\mathcal{W}}\sqrt{n}\big) + \log\frac{6d}{\delta}\right)} + \frac{2L_2}{\sqrt{n}}. \tag{31}$$

Proof. We first show that $m_\lambda > 0$, and then prove the concentration bounds. By Proposition A.4, we know that

$$\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x) = \frac{1}{\lambda^3\epsilon}\operatorname{Var}_{Q_{\theta,\lambda}(\cdot|x)}\big(\ell(z;\theta)\big).$$

The proof of Lemma 3.5 shows that $Q_{\theta,\lambda}(\cdot \mid x)$ and ν have the same null sets, so the Assumption 3.2 implies that $\operatorname{Var}_{Q_{\theta,\lambda}(\cdot|x)}(\ell) > 0$. Thus, $\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x) > 0$ for all $(\theta, \lambda, x) \in \mathcal{W} \times \mathcal{Z}$. Since $\mathcal{L}(\theta, \lambda) = \lambda\rho + \mathbb{E}_{x\sim P_o}[\Psi_{\theta,\lambda}(x)]$, we get $\partial^2_{\lambda\lambda}\mathcal{L}(\theta, \lambda) = \mathbb{E}_{x\sim P_o}\left[\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x)\right] > 0$ for all $(\theta, \lambda) \in \mathcal{W}$. Because both the objectives $\mathcal{L}$ and $\mathcal{L}_n$ are constructed via linear expectations of $\Psi_{\theta,\lambda}(x)$ over P_o and $P_{o,n}$ respectively, their second derivatives are the expectations of these strictly positive variances:

$$\partial^2_{\lambda\lambda}\mathcal{L}(\theta, \lambda) = \mathbb{E}_{P_o}\left[\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x)\right] > 0, \qquad \partial^2_{\lambda\lambda}\mathcal{L}_n(\theta, \lambda) = \mathbb{E}_{P_{o,n}}\left[\partial^2_{\lambda\lambda}\Psi_{\theta,\lambda}(x)\right] > 0.$$

This establishes that for any fixed $\theta \in \Theta$, both $\lambda \mapsto \mathcal{L}(\theta, \lambda)$ and $\lambda \mapsto \mathcal{L}_n(\theta, \lambda)$ are strictly convex on Λ. Furthermore, the map $(\theta, \lambda) \mapsto \partial^2_{\lambda\lambda}\mathcal{L}(\theta, \lambda)$ is continuous on the compact set $\mathcal{W}$ by Proposition A.4. Thus, it attains a minimum on $\mathcal{W}$, and this minimum is strictly positive, which proves $m_\lambda > 0$.

We now prove the concentration bounds. The argument is the same covering argument, but we write it separately for h_0, h_1, h_2.

Step 1: A scalar covering argument. Let $h : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ be measurable. Assume that

$$\sup_{w\in\mathcal{W}, x\in\mathcal{Z}} |h(w; x)| \leq M, \qquad |h(w; x) - h(w'; x)| \leq L\|w - w'\|_2$$

for all $w, w' \in \mathcal{W}$ and all $x \in \mathcal{Z}$. Let $\varepsilon \in (0, R_{\mathcal{W}})$, and let $\mathcal{W}_\varepsilon$ be an ε-net of $\mathcal{W}$ with $N_\varepsilon := |\mathcal{W}_\varepsilon| \leq \left(\frac{3R_{\mathcal{W}}}{\varepsilon}\right)^{d+1}$. For every $w \in \mathcal{W}$, choose $v \in \mathcal{W}_\varepsilon$ such that $\|w - v\|_2 \leq \varepsilon$. Then

$$\begin{aligned}\big|(P_{o,n} - P_o)h(w)\big| &\leq \big|(P_{o,n} - P_o)h(v)\big| + \big|P_{o,n}(h(w) - h(v))\big| + \big|P_o(h(w) - h(v))\big| \\ &\leq \big|(P_{o,n} - P_o)h(v)\big| + 2L\varepsilon.\end{aligned}$$

Thus, we obtain

$$\sup_{w\in\mathcal{W}}\big|(P_{o,n}-P_o)h(w)\big| \le \max_{v\in\mathcal{W}_\varepsilon}\big|(P_{o,n}-P_o)h(v)\big| + 2L\varepsilon.$$

Now fix $v \in \mathcal{W}_\varepsilon$. Since $|h(v;x)| \le M$, Hoeffding's inequality gives

$$\mathbb{P}\left(\big|(P_{o,n}-P_o)h(v)\big| \ge s\right) \le 2\exp\left(-\frac{ns^2}{2M^2}\right).$$

By the union bound over the finite set $\mathcal{W}_\varepsilon$, we obtain that

$$\mathbb{P}\left(\max_{v\in\mathcal{W}_\varepsilon}\big|(P_{o,n}-P_o)h(v)\big| \ge s\right) \le \sum_{v\in\mathcal{W}_\varepsilon}\mathbb{P}\left(\big|(P_{o,n}-P_o)h(v)\big| \ge s\right) \le 2N_\varepsilon\exp\left(-\frac{ns^2}{2M^2}\right).$$

Choosing s such that $2N_\varepsilon\exp\left(-\frac{ns^2}{2M^2}\right) = \frac{\delta}{3}$, we get $s = M\sqrt{\frac{2}{n}\left(\log(2N_\varepsilon)+\log\frac{3}{\delta}\right)}$. Using $N_\varepsilon \le (3R_{\mathcal{W}}/\varepsilon)^{d+1}$ yields

$$s \le M\sqrt{\frac{2}{n}\left((d+1)\log\big(3R_{\mathcal{W}}/\varepsilon\big)+\log\frac{6}{\delta}\right)}.$$

Thus, we conclude that

$$\sup_{w\in\mathcal{W}}\big|(P_{o,n}-P_o)h(w)\big| \le M\sqrt{\frac{2}{n}\left((d+1)\log\big(3R_{\mathcal{W}}/\varepsilon\big)+\log\frac{6}{\delta}\right)} + 2L\varepsilon \tag{32}$$

with probability at least $1-\delta/3$. Setting $\varepsilon = n^{-1/2}$ into (32) yields

$$\sup_{w\in\mathcal{W}}\big|(P_{o,n}-P_o)h(w)\big| \le M\sqrt{\frac{2}{n}\left((d+1)\log\big(3R_{\mathcal{W}}\sqrt{n}\big)+\log\frac{6}{\delta}\right)} + \frac{2L}{\sqrt{n}}.$$

Step 2: Application to h_0 and h_1. Using Lemma A.5, the functions $h_0 = \Psi_{\theta,\lambda}$ and $h_1 = \partial_\lambda\Psi_{\theta,\lambda}$ satisfy the hypotheses of Step 1 with constants (M_0, L_0) and (M_1, L_1), respectively. Thus, we obtain

$$\sup_{(\theta,\lambda)\in\mathcal{W}}\big|(P_{o,n}-P_o)h_0(\theta,\lambda;\cdot)\big| \le \mathcal{E}_{n,0}(\delta), \qquad \sup_{(\theta,\lambda)\in\mathcal{W}}\big|(P_{o,n}-P_o)h_1(\theta,\lambda;\cdot)\big| \le \mathcal{E}_{n,1}(\delta),$$

with probability at least $1-\delta/3$ for each bound.

Step 3: Application to the vector-valued map h_2. Write

$$h_2(\theta,\lambda;x) = \big(h_{2,1}(\theta,\lambda;x),\dots,h_{2,d}(\theta,\lambda;x)\big)^\top.$$

Using Lemma A.5, we know that

$$\sup_{w\in\mathcal{W},x\in\mathcal{Z}}\|h_2(w;x)\|_2 \le M_2, \qquad \|h_2(w;x)-h_2(w';x)\|_2 \le L_2\|w-w'\|_2.$$

Thus, for each coordinate $m \in \{1,\dots,d\}$, we obtain

$$\sup_{w\in\mathcal{W},x\in\mathcal{Z}}|h_{2,m}(w;x)| \le M_2, \qquad |h_{2,m}(w;x)-h_{2,m}(w';x)| \le L_2\|w-w'\|_2.$$

Applying Step 1 to each coordinate with threshold $s/\sqrt{d}$ yields

$$\mathbb{P}\left(\big|(P_{o,n}-P_o)h_{2,m}(v)\big| \ge \frac{s}{\sqrt{d}}\right) \le 2\exp\left(-\frac{ns^2}{2dM_2^2}\right).$$

Hence, for any fixed $v \in \mathcal{W}_\varepsilon$,

$$\mathbb{P}\left(\|(P_{o,n} - P_o)h_2(v)\|_2 \geq s\right) \leq \sum_{m=1}^{d} \mathbb{P}\left(\left|(P_{o,n} - P_o)h_{2,m}(v)\right| \geq \frac{s}{\sqrt{d}}\right) \leq 2d \exp\left(-\frac{ns^2}{2dM_2^2}\right).$$

A union bound over $v \in \mathcal{W}_\varepsilon$ gives

$$\mathbb{P}\left(\max_{v \in \mathcal{W}_\varepsilon} \|(P_{o,n} - P_o)h_2(v)\|_2 \geq s\right) \leq 2dN_\varepsilon \exp\left(-\frac{ns^2}{2dM_2^2}\right).$$

Choosing s such that $2dN_\varepsilon \exp\left(-\frac{ns^2}{2dM_2^2}\right) = \frac{\delta}{3}$, yields $s = M_2\sqrt{\frac{2d}{n}\left(\log(2dN_\varepsilon) + \log\frac{3}{\delta}\right)}$. Thus, using $N_\varepsilon \leq (3R_{\mathcal{W}}/\varepsilon)^{d+1}$ and letting $\varepsilon = n^{-1/2}$, we obtain

$$\sup_{w \in \mathcal{W}} \|(P_{o,n} - P_o)h_2(w;\cdot)\|_2 \leq M_2\sqrt{\frac{2d}{n}\left((d+1)\log\big(3R_{\mathcal{W}}\sqrt{n}\big) + \log\frac{6d}{\delta}\right)} + \frac{2L_2}{\sqrt{n}},$$

with probability at least $1 - \delta/3$.

Step 4: Objective concentration. Since $\mathcal{L}_n(\theta,\lambda) - \mathcal{L}(\theta,\lambda) = (P_{o,n} - P_o)\Psi_{\theta,\lambda}$, Step 2 with h_0 yields

$$\sup_{(\theta,\lambda)\in\mathcal{W}} |\mathcal{L}_n(\theta,\lambda) - \mathcal{L}(\theta,\lambda)| \leq \mathcal{E}_{n,0}(\delta).$$

Because $F_n(\theta) = \min_{\lambda\in\Lambda} \mathcal{L}_n(\theta,\lambda)$ and $F(\theta) = \min_{\lambda\in\Lambda} \mathcal{L}(\theta,\lambda)$, we have

$$|F_n(\theta) - F(\theta)| \leq \sup_{\lambda\in\Lambda} |\mathcal{L}_n(\theta,\lambda) - \mathcal{L}(\theta,\lambda)| \leq \sup_{(\theta,\lambda)\in\mathcal{W}} |\mathcal{L}_n(\theta,\lambda) - \mathcal{L}(\theta,\lambda)|.$$

Taking the supremum over $\theta \in \Theta$ yields (27).

Step 5: Deviation of the dual minimizer. Since $m_\lambda > 0$, $\mathcal{L}(\theta,\lambda)$ is globally m_λ-strongly convex in λ. Because both $\lambda_n^\star(\theta)$ and $\lambda^\star(\theta)$ minimize their respective objectives over the convex compact set Λ, they satisfy the first-order necessary variational inequalities:

$$\left\langle \partial_\lambda \mathcal{L}(\theta,\lambda^\star(\theta)), \lambda - \lambda^\star(\theta)\right\rangle \geq 0, \qquad \left\langle \partial_\lambda \mathcal{L}_n(\theta,\lambda_n^\star(\theta)), \lambda - \lambda_n^\star(\theta)\right\rangle \geq 0, \qquad \forall \lambda \in \Lambda.$$

Substitute $\lambda = \lambda_n^\star(\theta)$ into the first inequality and $\lambda = \lambda^\star(\theta)$ into the second inequality. Utilizing the m_λ-strong monotonicity of $\partial_\lambda \mathcal{L}$, we obtain:

$$\begin{aligned} m_\lambda \left|\lambda_n^\star(\theta) - \lambda^\star(\theta)\right|^2 &\leq \left\langle \partial_\lambda \mathcal{L}(\theta,\lambda_n^\star(\theta)) - \partial_\lambda \mathcal{L}(\theta,\lambda^\star(\theta)), \lambda_n^\star(\theta) - \lambda^\star(\theta)\right\rangle \\ &= \left\langle \partial_\lambda \mathcal{L}(\theta,\lambda_n^\star(\theta)) - \partial_\lambda \mathcal{L}_n(\theta,\lambda_n^\star(\theta)), \lambda_n^\star(\theta) - \lambda^\star(\theta)\right\rangle \\ &\quad + \left\langle \partial_\lambda \mathcal{L}_n(\theta,\lambda_n^\star(\theta)), \lambda_n^\star(\theta) - \lambda^\star(\theta)\right\rangle \\ &\quad - \left\langle \partial_\lambda \mathcal{L}(\theta,\lambda^\star(\theta)), \lambda_n^\star(\theta) - \lambda^\star(\theta)\right\rangle \\ &\leq \left|\partial_\lambda \mathcal{L}(\theta,\lambda_n^\star(\theta)) - \partial_\lambda \mathcal{L}_n(\theta,\lambda_n^\star(\theta))\right| \cdot \left|\lambda_n^\star(\theta) - \lambda^\star(\theta)\right|. \end{aligned} \tag{33}$$

Thus, using (33) and Step 2 with h_1, we obtain

$$\sup_{\theta\in\Theta} \left|\lambda_n^\star(\theta) - \lambda^\star(\theta)\right| \leq \frac{1}{m_\lambda} \sup_{(\theta,\lambda)\in\mathcal{W}} \left|\partial_\lambda \mathcal{L}_n(\theta,\lambda) - \partial_\lambda \mathcal{L}(\theta,\lambda)\right| \leq \frac{1}{m_\lambda}\mathcal{E}_{n,1}(\delta). \tag{34}$$

Step 6: Deviation of the profiled gradient. As proven in Step 1, both $\mathcal{L}(\theta, \lambda)$ and $\mathcal{L}_n(\theta, \lambda)$ are strictly convex in λ over the compact set Λ. Consequently, their respective minimizers $\lambda^\star(\theta) = \arg\min_{\lambda\in\Lambda} \mathcal{L}(\theta, \lambda)$ and $\lambda_n^\star(\theta) = \arg\min_{\lambda\in\Lambda} \mathcal{L}_n(\theta, \lambda)$ are unique. Because the minimizers are unique and the functions are differentiable, using Danskin's theorem, we get

$$\nabla_\theta F(\theta) = \nabla_\theta \mathcal{L}(\theta, \lambda^\star(\theta)), \qquad \nabla_\theta F_n(\theta) = \nabla_\theta \mathcal{L}_n(\theta, \lambda_n^\star(\theta)).$$

Thus, we obtain

$$\begin{aligned}\|\nabla_\theta F_n(\theta) - \nabla_\theta F(\theta)\|_2 \le \|\nabla_\theta \mathcal{L}_n(\theta, \lambda_n^\star(\theta)) - \nabla_\theta \mathcal{L}_n(\theta, \lambda^\star(\theta))\|_2 \\ + \|\nabla_\theta \mathcal{L}_n(\theta, \lambda^\star(\theta)) - \nabla_\theta \mathcal{L}(\theta, \lambda^\star(\theta))\|_2.\end{aligned} \tag{35}$$

In the right hand side of (35), the first term is controlled by the mixed derivative bound:

$$\|\nabla_\theta \mathcal{L}_n(\theta, \lambda_n^\star(\theta)) - \nabla_\theta \mathcal{L}_n(\theta, \lambda^\star(\theta))\|_2 \le C_{\theta\lambda} \big|\lambda_n^\star(\theta) - \lambda^\star(\theta)\big|, \tag{36}$$

and the second term is bounded by Step 3:

$$\sup_{\theta\in\Theta} \|\nabla_\theta \mathcal{L}_n(\theta, \lambda^\star(\theta)) - \nabla_\theta \mathcal{L}(\theta, \lambda^\star(\theta))\|_2 \le \mathcal{E}_{n,2}(\delta). \tag{37}$$

Combining (34), (35), (36) and (37) yields (28). □

Proof of Theorem 4.1. Since Θ is convex and F_n is differentiable, the minimizer $\hat\theta_n$ satisfies

$$\big\langle \nabla_\theta F_n(\hat\theta_n), \theta - \hat\theta_n \big\rangle \ge 0, \qquad \forall \theta \in \Theta. \tag{38}$$

The minimizer $\theta^\star$ satisfies the analogous inequality for F:

$$\big\langle \nabla_\theta F(\theta^\star), \theta - \theta^\star \big\rangle \ge 0, \qquad \forall \theta \in \Theta. \tag{39}$$

Because F is m_F-strongly convex on Θ, with $\Delta_n := \hat\theta_n - \theta^\star$, we obtain

$$m_F \|\Delta_n\|_2^2 \le \big\langle \nabla_\theta F(\hat\theta_n) - \nabla_\theta F(\theta^\star), \Delta_n \big\rangle. \tag{40}$$

We decouple the right-hand side of Inequality (40) by introducing $\nabla_\theta F_n(\hat\theta_n)$:

$$\begin{aligned} m_F \|\Delta_n\|_2^2 &\le \big\langle \nabla_\theta F(\hat\theta_n) - \nabla_\theta F_n(\hat\theta_n), \Delta_n \big\rangle + \langle \nabla_\theta F_n(\hat\theta_n), \Delta_n \rangle - \langle \nabla_\theta F(\theta^\star), \Delta_n \rangle \\ &\le \big\langle \nabla_\theta F(\hat\theta_n) - \nabla_\theta F_n(\hat\theta_n), \Delta_n \big\rangle + 0 + 0 \\ &\le \|\nabla_\theta F(\hat\theta_n) - \nabla_\theta F_n(\hat\theta_n)\|_2 \cdot \|\Delta_n\|_2 \\ &\le \left(\sup_{\theta\in\Theta} \|\nabla_\theta F_n(\theta) - \nabla_\theta F(\theta)\|_2 \right) \|\Delta_n\|_2, \end{aligned}$$

where the second inequality utilizes the variational inequalities (noting that $\langle \nabla_\theta F_n(\hat\theta_n), \Delta_n \rangle = -\langle \nabla_\theta F_n(\hat\theta_n), \theta^\star - \hat\theta_n \rangle \le 0$), and the third inequality applies the Cauchy-Schwarz inequality.

Dividing both sides by $m_F \|\Delta_n\|_2$ (the bound (6) holds trivially if $\Delta_n = 0$) yields:

$$\|\hat\theta_n - \theta^\star\|_2 \le \frac{1}{m_F} \sup_{\theta\in\Theta} \|\nabla_\theta F_n(\theta) - \nabla_\theta F(\theta)\|_2. \tag{41}$$

Substituting the uniform gradient bound (28) from Lemma A.6 into (41) establishes the final result (6). The rate follows immediately. □

B Additional Experimental Details

B.1 Data and feature construction

We use the *harmless-base* split of the Anthropic HH-RLHF corpus Bai et al. (2022). We combine up to 15,000 training and 3,000 test records and randomly retain $N_{\text{pair}} = 1,200$ canonical preference pairs. For each canonical pair, let s denote the dialogue prefix, a^+ the chosen response, and a^- the rejected response. Each pair generates the two oriented atoms

$$z_i^+ = (s_i, a_i^+, a_i^-, 1), \qquad z_i^- = (s_i, a_i^-, a_i^+, 0),$$

so the selected pool contains $M = 2,400$ oriented atoms.

For each prompt–response pair (s, a), we form a TF–IDF vector $x(s, a)$ from the concatenated text $s \parallel a$. The vocabulary uses at most 20,000 terms, with unigrams and bigrams, minimum document frequency 2, and sublinear term-frequency scaling. A rank-d truncated SVD with $d = 24$ gives $\phi(s, a) = x(s, a)V_d^\top$. The SVD is fitted once to the selected chosen and rejected response texts, and one global normalization over the selected pool is then applied:

$$s_\psi = \max_{(s,a) \text{ in the selected pool}} \|\phi(s, a)\|_2, \qquad \psi(s, a) = \phi(s, a)/s_\psi, \qquad \Delta\psi(z) = \psi(s, a_1) - \psi(s, a_2).$$

Thus, $\max_{s,a} \|\psi(s, a)\|_2 \le 1$. The same fitted TF–IDF/SVD map and global scaling are used for the dialogue-prefix representations in the K-Means analysis.

The transport cost is

$$c(z_i, z_j) = \|\Delta\psi(z_i) - \Delta\psi(z_j)\|_2^2 + \kappa \mathbf{1}\{y_i \neq y_j\},$$

with $\kappa = 1$ in the experiments. Since $\|\psi(s, a)\|_2 \le 1$, we have $\|\Delta\psi(z_i) - \Delta\psi(z_j)\|_2 \le 4$ and hence $c(z_i, z_j) \le 17$.

B.2 Implementation and baseline settings

We set $\theta_{\text{ref}} = 0$, $\|\theta\|_2 \le B$ with $B = 3$, $\beta = 2$, and $\kappa = 1$, giving $c(z_i, z_j) = \|\Delta\psi(z_i) - \Delta\psi(z_j)\|_2^2 + \kappa\mathbf{1}\{y_i \neq y_j\}$. All linear fits start from $\theta_0 = 0$ and use 60 projected-Adam iterations. We compare SDPO with DPO, WDPO, KLDPO Xu et al. (2025b), and χPO Huang et al. (2025). For SDPO, we fix the entropic regularization at $\epsilon = 0.05$ and select the robustness radius ρ by cross-validation below. For the baseline robust methods, $\rho_{\text{W}} = 0.10$ denotes the WDPO robustness radius, $\rho_{\text{KL}} = 0.03$ the KLDPO robustness parameter, and $R_{\max} = 3$ the χPO clipping parameter.

B.3 Reference-measure protocol for CV and bootstrap

Let I_{out} denote the outer support of the empirical nominal measure and I_{ref} the support used for the reference measure ν. The implementation uses

$$C_{\text{sub}} = C[I_{\text{out}}, I_{\text{ref}}], \qquad \ell_{\text{ref}} = \ell[I_{\text{ref}}], \qquad \nu_j = 1/|I_{\text{ref}}|, \qquad \log \nu_j = -\log |I_{\text{ref}}|$$

for a uniform empirical reference. Similarly, for an outer atom $i \in I_{\text{out}}$, the entropic transport weights are normalized over $j \in I_{\text{ref}}$ and the robust weights assigned to the outer atoms are obtained from the corresponding column marginals. In three-fold cross-validation, the fitting stage uses $I_{\text{out}} = I_{\text{ref}} = I_{\text{train}}$. For held-out evaluation, the outer support is changed to $I_{\text{out}} = I_{\text{test}}$ while the reference support remains $I_{\text{ref}} = I_{\text{train}}$. Thus, the reference measure used for evaluation is constructed entirely from the observations available to the fitted model. In the bootstrap stability experiment, the bootstrap support is used for both I_{out} and I_{ref}. Consequently, the empirical reference measure is re-estimated together with the outer empirical measure for each bootstrap sample and varies with the bootstrap sample size n.

Let I_t^{out} and I_t^{ref} denote the supports of the outer empirical measure and the reference measure ν in (2), respectively. For uniform weights, the corresponding empirical Sinkhorn dual can be written as

$$D_t(\lambda) = \lambda\rho + \lambda\epsilon \sum_{i \in I_t^{\mathrm{out}}} \widehat{P}_t(i) \log \left[\sum_{j \in I_t^{\mathrm{ref}}} \widehat{\nu}_t(j) \exp\left(\frac{\ell(z_j;\theta_t) - \lambda c(z_i, z_j)}{\lambda\epsilon} \right) \right],$$

where $\widehat{P}_t(i) = 1/|I_t^{\mathrm{out}}|$ and $\widehat{\nu}_t(j) = 1/|I_t^{\mathrm{ref}}|$. When I_t^{out} and I_t^{ref} coincide, this reduces to the batch dual above. The support conventions used in the cross-validation and bootstrap experiments are specified as the above. At LLM scale, the mini-batch formulation provides a stochastic approximation of the full empirical Sinkhorn objective. In particular, the empirical average is evaluated inside the nonlinear log-sum-exp term of (2), so the resulting batch-based objective is generally biased relative to the corresponding full empirical objective; biased stochastic estimators also arise in Sinkhorn-DRO Wang et al. (2026).

B.4 Radius calibration details

We first calibrate the SDPO robustness radius ρ by three-fold CV. The folds are formed at the canonical-pair level, so the two orientations generated from the same preference pair remain in the same fold. We evaluate the empirical feasibility condition associated with Assumption 3.3 at the optimized parameter of each fold. The candidate grid and the complete CV results are reported in the following; Figure 4 shows the resulting loss and feasibility diagnostics.

The candidate grid is $\{0.01, 0.02, 0.04, 0.07, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.70\}$. The mean CV results are shown in Table 2. The selected value is $\rho_{\mathrm{cv}} = 0.15$, the smallest candidate with full-fold feasibility.

Table 2: Cross-validation results and empirical feasibility diagnostics across candidate robustness radii ρ.

ρ	**Mean CV Risk**	**Feasible Fraction**	$(\bar{\rho}_{\mathrm{emp}}, \bar{\rho}_{\mathrm{crit}})$
0.01	0.671477	0	$(-0.078240, 0.502696)$
0.02	0.671477	0	$(-0.068240, 0.502696)$
0.04	0.671477	0	$(-0.048240, 0.502696)$
0.07	0.671477	0	$(-0.018240, 0.502696)$
0.10	0.692918	0	$(0.011758, 0.533036)$
0.15	0.692990	1	$(0.061758, 0.525812)$
0.20	0.693144	1	$(0.111758, 0.460966)$
0.25	0.693038	1	$(0.161758, 0.543555)$
0.30	0.692839	1	$(0.211758, 0.539320)$
0.40	0.693139	1	$(0.311758, 0.601638)$
0.50	0.693062	1	$(0.411758, 0.505343)$
0.70	0.692284	0	$(0.611758, 0.569712)$

The effective-radius quantity and critical-radius quantity above are empirical diagnostics evaluated at the fitted parameters. Theoretically, Assumption 3.3 requires $\bar{\rho} < \inf_{\theta\in\Theta} \bar{\rho}_{\mathrm{crit}}(\theta)$ to guarantee global dual smoothness. Since evaluating this uniform infimum over the high-dimensional continuous parameter space is computationally intractable, we implement a rigorous post-hoc empirical diagnostic. For each cross-validation fold k, after the optimization converges to the parameter $\hat{\theta}_k$, we explicitly evaluate the critical effective radius $\bar{\rho}_{\mathrm{crit}}(\hat{\theta}_k)$ by calculating the log-sum-exp ratio of the loss maximizers under the Sinkhorn transport costs. A candidate ρ is marked as theoretically feasible (denoted by red circles in Figure 4) if and only if $\bar{\rho} < \bar{\rho}_{\mathrm{crit}}(\hat{\theta}_k)$

holds strictly across all K folds. Candidates failing this condition on any fold (red crosses) exhibit dual degeneracy and are automatically excluded.

B.5 Dual-minimizer diagnostic

The feasibility screen identifies $\rho = 0.15$–0.50 as feasible in all folds, and we select $\rho_{\mathrm{cv}} = 0.15$, the smallest fully feasible candidate. Across the feasible candidates, the mean held-out nominal DPO loss changes only slightly, remaining in $[0.69284, 0.69314]$. The right panel of Figure 4 shows the corresponding numerical dual minimizers observed along the CV trajectories. The diagnostic interval is constructed from the empirical 5th and 95th percentiles of these observed values and then expanded by 25%. This quantile-based construction is used to summarize the central range of the trajectory and its numerical scale; values outside the central quantile range are retained in the trajectory and are not discarded as outliers. Thus, the resulting interval is a diagnostic summary rather than a statistical filtering rule.

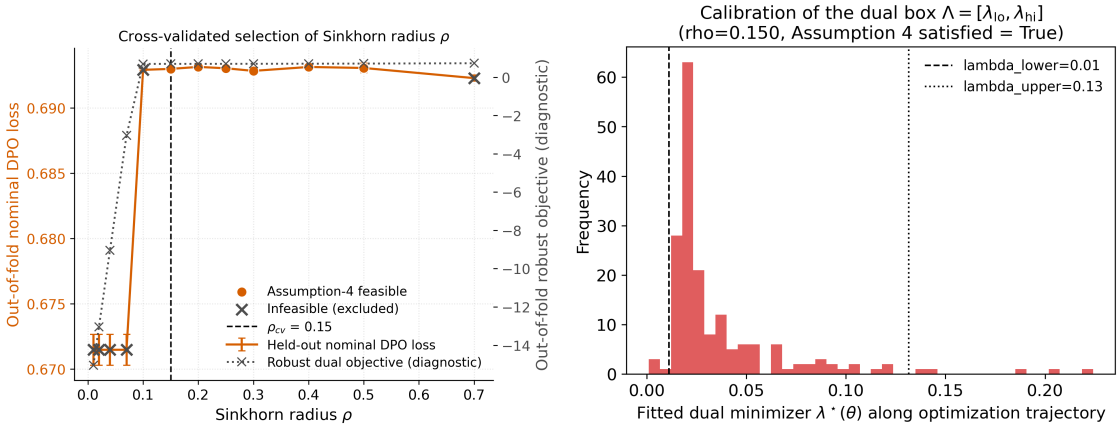

Figure 4: SDPO radius and dual-minimizer calibration diagnostics. **(Left)** Cross-validation nominal DPO loss and empirical feasibility across candidate robustness radii. **(Right)** Numerical dual minimizers recorded along the cross-validation optimization trajectories.

B.6 Finite-sample stability details

We study finite-sample parameter stability using an oriented-atom bootstrap with three replications at each sample size. Let $\hat{\theta}_n$ denote the parameter fitted from the bootstrap sample of size n, and let $\hat{\theta}_M$ be the corresponding full-pool fit. We report $\|\hat{\theta}_n - \hat{\theta}_M\|_2$ in the left panel of Figure 2. The bootstrap uses $n \in \{50, 100, 150, 200, 300, 400, 600, 1200, 2000\}$ and three replications per n. At $n = 2,000$, the mean deviations are presented in Table 3. The deviations of SDPO and WDPO decrease rapidly for small and intermediate n and then become relatively flat at larger sample sizes. Over this finite sample-size grid, SDPO and WDPO show smaller deviations from their full-pool reference parameters than the other three methods.

B.7 Topic-shift and sensitivity details

The six K-Means clusters contain 174, 468, 840, 36, 28, and 854 oriented atoms. Clusters 0–2 provide the 1,482-atom nominal pool, and five replications sample 80 training atoms without replacement. Mean SDPO accuracies for clusters 0–5 are $(0.5931, 0.5726, 0.5567, 0.5778, 0.5714, 0.5157)$. The SDPO ablation uses

Table 3: Mean parameter deviation $\|\hat{\theta}_n - \hat{\theta}_M\|_2$ from full-pool reference fits at bootstrap sample size $n = 2,000$.

Method	Mean Deviation
DPO	1.1235
WDPO	0.0466
KLDPO	0.9149
χPO	1.1204
SDPO	0.0468

$\epsilon \in \{0.01, 0.05, 0.20\}$ and $\rho/\rho_{\text{cv}} \in \{0.1, 0.25, 0.5, 1, 2, 4, 8\}$, with three replications and 80 training atoms per configuration. The maximum nominal accuracy is 0.6181 at $(0.20, 0.015)$ and the maximum worst-cluster accuracy is 0.5142 at $(0.20, 0.30)$.

The cluster-wise results reflect differences in sample size: clusters 3 and 4 contain only 36 and 28 oriented atoms, leading to larger accuracy fluctuations, whereas cluster 5 contains 854 atoms and gives SDPO its highest observed cluster accuracy (0.5157). SDPO is also closer to the other methods on the larger cluster 2 than on clusters 0 and 1, suggesting that both sample size and sub-population characteristics may affect its observed accuracy in the current linear setting. The (ρ, ϵ) sensitivity analysis shows no clear monotone trend, with nominal accuracy remaining relatively stable over the tested grid.

B.8 Radius-sensitivity details

The auxiliary radius sweep uses multipliers $\{0.1, 0.3, 0.6, 1.0, 1.5, 2.5\}$. For SDPO, these correspond to $\rho \in \{0.015, 0.045, 0.09, 0.15, 0.225, 0.375\}$. The maximum observed worst-cluster accuracy is 0.5172 at $\rho = 0.045$, with nominal accuracy 0.5961; at $\rho_{\text{cv}} = 0.15$, these values are 0.4425 and 0.5713.

B.9 Additional distribution-shift summaries

The mean cluster-wise accuracies are given in Table 4. The corresponding cross-cluster DPO-loss ranges are

Table 4: Mean cluster-wise accuracy across six K-Means clusters (C_0–C_5) under topic shift.

Method	C_0	C_1	C_2	C_3	C_4	C_5
DPO	0.6391	0.6085	0.5648	0.5111	0.5429	0.5012
WDPO	0.6253	0.5795	0.5595	0.4889	0.6143	0.5054
KLDPO	0.6115	0.6034	0.5576	0.5333	0.6000	0.5049
χPO	0.6115	0.5932	0.5552	0.5778	0.5857	0.5115
SDPO	0.5931	0.5726	0.5567	0.5778	0.5714	0.5157

0.05502 (DPO), 0.03967 (WDPO), 0.05904 (KLDPO), 0.05094 (χPO), and 0.00279 (SDPO).

B.10 Radius-sensitivity table

The auxiliary radius sweep gives the following nominal and worst-cluster accuracies, shown in Table 5.

Table 5: Nominal and worst-cluster accuracy sensitivity across various robustness radii ρ for WDPO, KLDPO, and SDPO.

Method	ρ	Nominal	Worst
WDPO	0.01	0.5894	0.4696
	0.03	0.5933	0.4259
	0.06	0.5934	0.3968
	0.10	0.5936	0.4630
	0.15	0.5740	0.4485
	0.25	0.5424	0.4491
KLDPO	0.003	0.6117	0.4610
	0.009	0.6144	0.4209
	0.018	0.5822	0.3736
	0.030	0.6089	0.4388
	0.045	0.5869	0.4392
	0.075	0.5913	0.4206
SDPO	0.015	0.5863	0.3889
	0.045	0.5961	0.5172
	0.090	0.5791	0.4920
	0.150	0.5713	0.4425
	0.225	0.5582	0.3836
	0.375	0.5715	0.4021

B.11 Runtime details

The linear runtime study uses two repetitions at each $n \in \{50, 100, 150, 200, 300, 400\}$ and 60 Adam iterations.

Table 6: Linear runtime comparison (in seconds) per fit across sample sizes n over 60 Adam iterations.

Method	50	100	150	200	300	400
DPO	0.0196	0.0202	0.0195	0.0196	0.0205	0.0204
χPO	0.0342	0.0341	0.0331	0.0341	0.0376	0.0347
KLDPO	0.1318	0.1251	0.1273	0.1260	0.1288	0.1336
WDPO	0.8709	2.7723	3.9378	5.0381	7.5117	9.5805
SDPO	4.4970	10.9901	16.1043	20.5245	30.0958	40.3426

B.12 Computational profiling and LLM profiling details

In the linear benchmark, SDPO takes 4.50–40.34 s per fit as n increases from 50 to 400, compared with 0.87–9.58 s for WDPO. DPO, KLDPO, and χPO remain near 0.02, 0.13, and 0.03–0.04 s, respectively. Thus, the present linear implementation makes SDPO roughly four to five times slower than WDPO over the tested range; the additional cost comes from the entropic log-sum-exp evaluation and scalar dual search.

The LLM profiling uses GPT-2, LLaMA-3.2-1B, LLaMA-3.2-3B, and LLaMA-3.1-8B on Stanford Human Preferences dataset (SHP) Ethayarajh et al. (2022) with NVIDIA V100 GPUs. Timed runs use 30 LoRA steps. To accommodate varying model sizes while ensuring fair comparisons, the maximum sequence lengths are configured as 512, 384, 192, and 512, respectively, and all evaluated methods strictly share the same batch size at each scale: 8, 4, 2, and 2. The 8B WDPO configuration triggers an Out-Of-Memory (OOM) error and is not completed under the tested memory setting. Timing uses explicit CUDA synchronization to capture accurate wall-clock durations, and peak memory is the maximum allocated GPU memory during the timed interval.

Our empirical results reveal a divergence between time complexity and memory usage. SDPO generally exhibits a higher total time cost than DPO, KLDPO and χPO. This is primarily driven by an upfront, algorithm-specific "warm-up" calibration (requiring numerous forward passes to estimate global feature scales without gradient tracking) alongside solving the scalar dual per mini-batch. However, this time overhead is acceptable in practice. SDPO successfully scales to the 8B model within the 32GB V100 memory limit. In contrast, WDPO relies on an input-gradient penalty approximation Xu et al. (2025b) that necessitates double backpropagation. This dominates the intermediate activation footprint, leading to fatal Out-Of-Memory (OOM) failures at the 8B scale. Furthermore, the decoupling of time and memory is distinctly illustrated by WDPO's peak memory actually *decreasing* from the 1B to the 3B model. Because the 3B configuration halves both the sequence length and batch size (reducing the context volume by 75%), the massive drop in the double-backward computation graphs completely offsets the increase in static model weights. Conversely, for DPO, KLDPO, χPO and SDPO, whose memory is dominated by static weights rather than activations, exhibit normal monotonic memory growth across scales.