SquareχPO: Differentially Private and Robust χ^2-Preference Optimization in Offline Direct Alignment

Xingyu Zhou [1] **Yulian Wu** [2] **Wenqian Weng** [1] **Francesco Orabona** [2]

Abstract

We theoretically study the offline alignment of language models with human preference feedback, under both preference label corruption and privacy protections. To this end, we propose SquareχPO, a simple one-line change to χPO where the standard log-loss is replaced by a new square loss over probability. This allows us to advance the state-of-the-art of differentially private and robust offline direct alignment. Specifically, for the local model of label privacy, SquareχPO is the first algorithm that attains an optimal rate based on *single-policy concentrability*, even with general function approximations. On the robustness side against Huber label corruption, SquareχPO is the first alignment method that has a meaningful theoretical guarantee under general function approximations. More importantly, SquareχPO can address privacy protection and corruption *simultaneously*, where an interesting separation is observed, implying that the order of privacy and corruption matters.

1. Introduction

Aligning large language models (LLMs) to human values is crucial for their responsible deployment. Two primary paradigms have emerged: *indirect alignment*, where a reward model is learned before the policy optimized via Reinforcement Learning (RL) (Christiano et al., 2017; Ouyang et al., 2022), and *direct alignment*, an RL-free approach leveraging reparametrization techniques like Direct Preference Optimization (DPO) (Rafailov et al., 2023). Very recently, a variant of DPO, called χPO (Huang et al., 2024), addresses the overoptimization issue in direct alignment by relying on a significantly weaker condition – single-policy concentrability – making it the first offline direct alignment method with such a guarantee.

[1]Wayne State University [2]King Abdullah University of Science and Technology.

Meanwhile, privacy and robustness concerns in the preference datasets of the alignment process have gained significant attention. Membership inference attacks expose privacy vulnerabilities (Feng et al., 2024), while data poisoning undermines label integrity (Casper et al., 2023). Recent efforts have addressed these challenges separately, providing theoretical guarantees for privacy or robustness. On the privacy side, existing theoretical work has primarily focused on simple linear function approximations (Chowdhury et al., 2024b; Korkmaz & Brown-Cohen, 2024), which are insufficient for practical scenarios involving non-linear reward or policy function classes (e.g., neural networks).

Q1. For general function approximations, can we achieve optimal (or better) rates under privacy constraints?

Contribution 1. We answer **Q1** affirmatively by introducing SquareχPO, a simple variant of χPO which replaces the log loss with a new square loss over probabilities. For preference label privacy under the local model of Differential Privacy (DP) (Kasiviswanathan et al., 2011; Chaudhuri & Hsu, 2011), SquareχPO achieves the optimal privacy cost, even with general function approximations.

Moving now to the robustness side, Mandal et al. (2024) takes an indirect approach, focusing on the linear setting, while Chowdhury et al. (2024a) follows a DPO-style direct method, which however only achieves a suboptimal rate for the linear case and suffers from a non-vanishing suboptimality gap for general function approximations.

Q2. Can we improve these results under label corruption, even for general function approximations?

Contribution 2. Our SquareχPO provides an affirmative answer to **Q2**. Specifically, it not only preserves the favorable single-policy concentrability property of χPO, but also achieves the optimal $\mathcal{O}(1/\sqrt{n})$ rate for general function approximations under the same random-flipping corruption setting as in Chowdhury et al. (2024a). Furthermore, due to the inherent boundedness of our new loss, SquareχPO is the first alignment method to provide meaningful guarantees under stronger Huber label corruption (Huber, 1964), matching the best-known results in the non-preference feedback offline RL setting (Zhang et al., 2022).

Instead of studying privacy protection and robustness to corruption separately, there is growing interest in understanding their interplay, driven by both practical scenarios and theoretical insights, for example, in bandits (Zhou & Zhang, 2024; Wu et al., 2024b; Charisopoulos et al., 2023) or general statistical tasks; please refer to Kamath (2024) for a wonderful recent survey.

Q3. Can we achieve privacy protection and robustness simultaneously, and what are the interplays between them?

Contribution 3. Our SquareχPO simultaneously addresses privacy and robustness in offline direct alignment, uncovering interesting interplays between the two. SquareχPO is adaptive, as it does not require prior knowledge of the specific setting while providing sharp rates. Notably, our results reveal that corruption following privacy leads to worse bounds.

2. Preliminaries

2.1. Offline Alignment

In the offline alignment problem, there exists a pre-collected preference dataset $\mathcal{D}_{\mathsf{pref}} = \{(x_i, a_i^0, a_i^1, y_i)\}_{i=1}^n$, where each context/prompt x_i is i.i.d. sampled from a distribution ρ, and two responses a_i^0 and a_i^1 are i.i.d sampled from a reference policy π_{ref}, i.e., $a_i^0 \sim \pi_{\mathsf{ref}}(\cdot \mid x_i)$ and $a_i^1 \sim \pi_{\mathsf{ref}}(\cdot \mid x_i)$, and finally the preference label $y_i \in \{0, 1\}$ is generated according to some probability distribution, i.e., $y_i \sim \mathrm{Ber}(\mathcal{P}^\star(a_i^1 \succ a_i^0 \mid x_i))$, where $\mathcal{P}^\star(a_i^1 \succ a_i^0 \mid x_i) \in [0, 1]$ is the probability that given x_i, a_i^1 is preferred over a_i^0 and $\mathrm{Ber}(\cdot)$ denotes a Bernoulli distribution. Without loss of generality, we assume that $\rho(x) > 0$ for all x and $\pi_{\mathsf{ref}}(a \mid x) > 0$ for all x and a. Depending on the modeling assumption of the preference probability $\mathcal{P}^\star(a_i^1 \succ a_i^0 \mid x_i)$, the (offline) alignment is often categorized into the following two settings.

Bradley-Terry (BT) preference model (Bradley & Terry, 1952). In this setting, there exists an unknown true reward function $r^\star : \mathcal{X} \times \mathcal{A} \to [0, R_{\max}]$ that induces the preference probability as follows

$$\mathcal{P}^\star(a_i^1 \succ a_i^0 \mid x_i) = \frac{\exp(r^\star(x_i, a_i^1))}{\exp(r^\star(x_i, a_i^1)) + \exp(r^\star(x_i, a_i^0))}.$$

With the preference dataset $\mathcal{D}_{\mathsf{pref}}$, the goal under this setting is to learn a policy $\widehat{\pi}$ that minimizes the suboptimality gap:

$$\mathsf{SG}(\widehat{\pi}; \pi^\star) := J(\pi^\star) - J(\widehat{\pi}), \tag{1}$$

where $J(\pi) := \mathbb{E}_{x\sim\rho, a\sim\pi(\cdot|x)}[r^\star(x, a)]$ and $\pi^\star$ is any comparator policy (e.g., it could be the optimal policy maximizing $J(\pi)$ or any other policy). For notation simplicity, we will abbreviate $\mathbb{E}_\pi[\cdot] := \mathbb{E}_{x\sim\rho, a\sim\pi(\cdot|x)}[\cdot]$.

2.2. χPO

To address the overoptimization issue in DPO, Huang et al. (2024) recently proposed a simple variant of DPO by introducing an additional χ^2-regularization term, which leads to the following optimization[1]

$$\widehat{\pi}_{\chi\mathsf{PO}} = \operatorname*{argmax}_{\pi\in\Pi} \sum_{(x,a_+,a_-)\in\mathcal{D}_{\mathsf{pref}}} \log[\sigma(\beta h_{\chi\mathsf{PO}}(x, a_+, a_-))],$$

where $h_{\chi\mathsf{PO}}(x, a_+, a_-) := \phi\left(\frac{\pi(a_+|x)}{\pi_{\mathrm{ref}}(a_+|x)}\right) - \phi\left(\frac{\pi(a_-|x)}{\pi_{\mathrm{ref}}(a_-|x)}\right)$ and $\phi(u) := u + \log u$. Compared to DPO, there is an additional linear term in $\phi(z)$ that introduces *pessimism* (Jin et al., 2021b), which enables a suboptimality gap that only depends on *single policy concentrability* (Rashidinejad et al., 2021). On the other hand, DPO could only achieve a suboptimality gap in terms of *all-policy concentrability coefficient* (Chen & Jiang, 2019) due to the lack of pessimism. Given the stronger performance of χPO, we will mainly focus on it when we consider robustness and privacy in offline alignment, as discussed below.

2.3. Robustness and Privacy in Preference Data

Label corruption. In practice, the preference label y_i may not be sampled from the clean distribution $\mathrm{Ber}(\mathcal{P}^\star(a_i^1 \succ a_i^0 \mid x_i))$. To characterize this, we borrow the classic *Huber corruption* model from robust statistics.

Definition 2.1 (α-Huber corruption (Huber, 1964))**.** We consider the following α-Huber corruption: each label is independently sampled from $(1-\alpha)G_i + \alpha B_i$, where G_i is the clean distribution $\mathrm{Ber}(\mathcal{P}^\star(a_i^1 \succ a_i^0 \mid x_i))$ and B_i is some arbitrary unknown Bernoulli distribution. That is, with probability $\alpha \in [0, 1/2]$, each label is sampled from some bad distribution.

Label privacy in the local model. The preference label is often collected via human feedback, which could potentially reveal each person's private information, as discussed before. To this end, a strong privacy protection is to ensure *Local Differential Privacy* (LDP) via a local randomizer. Given the binary data of the preference label, it is natural to consider the classic *randomized response* mechanism.

Definition 2.2 (Randomized response and ε-LDP (Warner, 1965))**.** Let $\varepsilon > 0$ be the privacy parameter and $y \in \{0, 1\}$ be the true label. The randomized response (RR) mechanism $\mathcal{R}$ flips y and outputs private $\widetilde{y}$ based on the following distribution

$$\mathbb{P}[\widetilde{y} = y] = \frac{e^\varepsilon}{1 + e^\varepsilon} \text{ and } \mathbb{P}[\widetilde{y} \neq y] = \frac{1}{1 + e^\varepsilon}. \tag{2}$$

Interplay between corruption and LDP. In practice, corruption and LDP protection can exist together, which motivates us to consider their interplay in the following settings.

[1] We ignore the clipping operation for the ease of presentation.

Algorithm 1 SquareχPO for CTL and LTC

input Locally private and corrupted preference dataset $\widetilde{\mathcal{D}}_{\mathsf{pref}} = \{(x_i, a_i^0, a_i^1, z_i)\}_{i=1}^n$ under CTL and LTC, privacy parameter $\varepsilon > 0$, regularization coefficient $\beta > 0$, reference policy π_{ref}

1: Define

$$\phi(u) := u + \log u \tag{3}$$

$$h_{\chi\mathsf{PO},i} := \phi\left(\frac{\pi(a_i^1 \mid x_i)}{\pi_{\mathsf{ref}}(a_i^1 \mid x_i)}\right) - \phi\left(\frac{\pi(a_i^0 \mid x_i)}{\pi_{\mathsf{ref}}(a_i^0 \mid x_i)}\right) \tag{4}$$

2: Optimize the following objective:

$$\widehat{\pi} \leftarrow \operatorname*{argmin}_{\pi \in \Pi} \sum_{i \in [n]} \left[2\sigma\left(\mathrm{clip}_{2R_{\max}}\left[\beta h_{\chi\mathsf{PO},i}\right]\right) - 1 - c(\varepsilon)\bar{z}_i\right]^2,$$

where $c(\varepsilon) := \frac{e^\varepsilon + 1}{e^\varepsilon - 1}$ and $\bar{z}_i = 2z_i - 1$

3: **output:** $\widehat{\pi}$

Definition 2.3 (CTL and LTC)**.** Given a raw preference dataset $\mathcal{D}_{\mathsf{pref}} = \{(x_i, a_i^0, a_i^1, y_i)\}_{i=1}^n$ and two parameters $\alpha \in [0, 1/2]$, $\varepsilon > 0$, we consider the following two settings that differ in the order of corruption and label privacy protection in the local model:

Corruption-then-LDP (CTL)**.** The raw label y_i is first corrupted by the α-Huber model, which is then further privatized by ε-LDP RR mechanism, leading to the final preference dataset given by $\widetilde{\mathcal{D}}_{\mathsf{pref}} = \{(x_i, a_i^0, a_i^1, z_i)\}_{i=1}^n$.

LDP-then-Corruption (LTC)**.** The raw label y_i is first privatized by ε-LDP RR mechanism, which is then further corrupted by the α-Huber model, leading to the final preference dataset given by $\widetilde{\mathcal{D}}_{\mathsf{pref}} = \{(x_i, a_i^0, a_i^1, z_i)\}_{i=1}^n$.

3. Locally Private and Robust Alignment

In this section, we study offline alignment in the BT-preference model under privacy constraints and corruption. We will focus on the interplay between corruption and the label LDP (i.e., CTL and LTC).

Our proposed algorithm, SquareχPO in Algorithm 1, is the same for both CTL and LTC, i.e., adaptive. The key modification compared with χPO is to use a square loss instead of the log loss, plus an additional $c(\varepsilon)$ factor for the private case. We will dive into the intuition about the choice of our loss function in the sequel. Before that, we remark that the clipping $\mathrm{clip}_R(u) = \max\{\min\{u, R\}, -R\}$ with $R = 2R_{\max}$ is adopted in χPO as well, mainly used for a slightly tighter theoretical bound.

3.1. Intuition behind SquareχPO

We now discuss our new loss function in SquareχPO, highlighting the intuition on how it helps to handle corruption and privacy protection. It is worth noting that our new loss function could be of its own interest even in the standard scenario, i.e., non-private and non-corrupted cases, with DPO-type (rather than χPO-type) reparameterization.

1. Square loss over probability. Without privacy protection ($c(\varepsilon) = 1$), our new loss function essentially reduces to

$$\sum_{i \in [n]} (p_i(\pi) - z_i)^2, \tag{5}$$

where we define $p_i(\pi) := \sigma\left(\mathrm{clip}_{2R_{\max}}\left[\beta h_{\chi\mathsf{PO},i}\right]\right)$, while DPO and χPO essentially adopts the standard log-loss, i.e.,

$$-z_i \log p_i(\pi) - (1 - z_i)\log(1 - p_i(\pi)). \tag{6}$$

The loss in (5) is also often referred to as *Brier score* (Brier, 1950) in probabilistic predictions. One direct observation here is that the Brier score is always upper bounded by 1 while the log-loss can be unbounded, which implies that label corruption under log-loss may have a larger impact than that under the Brier score.

2. Converting to ± 1 with $c(\varepsilon)$ scaling. Instead of working with $z_i \in \{0, 1\}$, we convert it to $\bar{z}_i = 2z_i - 1 \in \{1, -1\}$ and we similarly update the probability part. The reason is, from (2) of RR, we can easily see that the private mean (under ± 1) is $1/c(\varepsilon)$ of the true mean (probability). This implies that the $c(\varepsilon)$ factor in front of the private data leads to an unbiased estimate of the true probability, which essentially follows from the same intuition as in private mean estimation under RR, since the empirical average mean estimator can also be written as the solution to a square loss.

Remark 3.1. We mention in passing that many alignment algorithms draw inspiration from binary classification for their loss functions, in the non-private non-corrupted cases. For instance, in addition to log-loss in DPO and χPO, SLiC (Zhao et al., 2023) leverages the hinge loss while IPO (Azar et al., 2024) adopts the standard square loss. The key conceptual difference between our square loss and that of IPO lies in the fact that the latter takes the square over the raw log-ratio (i.e., implicit reward) while ours is a square over probability (i.e., an additional sigmoid step is applied). More recently, Tang et al. (2024) proposed a family of loss functions for alignment based on standard supervised learning, including *exponential loss*, *truncated quadratic loss*, and *savage loss*. To the best of our knowledge, our SquareχPO is the first one that proposes to use the Brier score as the loss. In the next section, we will demonstrate its strong theoretical guarantees.

3.2. Theoretical Guarantees

In this section, our aim is to establish the suboptimality gap (cf. (1)) of SquareχPO (Algorithm 1), under both CTL and LTC, without knowledge of the setting in advance.

We start with the same assumptions as in χPO (Huang et al., 2024), i.e., policy realizability and bounded range.

Assumption 3.2 (Policy realizability)**.** Fix $\beta > 0$. The policy class Π satisfies $\pi^\star_\beta \in \Pi$, where $\pi^\star_\beta$ is the optimal policy of the following mixed χ^2-regularized objective:

$$J^{\chi^{\mathsf{mix}}}_\beta(\pi) := \mathbb{E}_\pi[r^\star(x,a)] - \beta \cdot [D_{\chi^2}(\pi \| \pi_{\mathsf{ref}}) + D_{\mathsf{KL}}(\pi \| \pi_{\mathsf{ref}})].$$

The $J^{\chi^{\mathsf{mix}}}_\beta(\pi)$ in χPO mixes χ^2-regularization with the standard KL-regularization in DPO, which in turn leads to the new reward reparameterization using optimal solution $\pi^\star_\beta$:

$$r^\star(x,a) = \beta\phi\left(\frac{\pi^\star_\beta(a|x)}{\pi_{\mathsf{ref}}(a|x)}\right) + Z_{\beta,r^\star}(x),$$

where we recall that $\phi(u) = u + \log u$ and $Z_{\beta,r^\star}(x)$ is some action-independent normalization term. Thus, Assumption 3.2 essentially implies the implicit reward realizability under the above parameterization.

As in χPO (Huang et al., 2024), the next assumption asserts that the *implicit reward difference* under any policy in Π is upper bounded by some constant.

Assumption 3.3 (Bounded implicit reward difference)**.** For a parameter $V_{\max} \geqslant R_{\max}$, it holds that for all $\pi \in \Pi$, $x \in \mathcal{X}$, and $a, b \in \mathcal{A}$,

$$\left|\beta\phi\left(\frac{\pi(a \mid x)}{\pi_{\mathsf{ref}}(a \mid x)}\right) - \beta\phi\left(\frac{\pi(b \mid x)}{\pi_{\mathsf{ref}}(b \mid x)}\right)\right| \leqslant V_{\max}.$$

Finally, we will measure the theoretical performance using the same type of *single-policy concentrability* as in χPO.

Definition 3.4 (L_1-Concentrability)**.** The single-policy L_1-concentrability coefficient for a policy π is given by

$$\mathcal{C}^\pi := \mathbb{E}_\pi\left[\frac{\pi(a|x)}{\pi_{\mathsf{ref}}(a|x)}\right],$$

where we recall that $\mathbb{E}_\pi[\cdot] := \mathbb{E}_{x\sim\rho, a\sim\pi(\cdot|x)}[\cdot]$.

Our main result is the following suboptimality bound.

Theorem 3.5. *For any given comparator policy $\pi^\star$, there exists a proper choice of $\beta > 0$ such that when Assumptions 3.2 and 3.3 hold, with probability at least $1-\zeta$, the output of Algorithm 1 satisfies the following suboptimality gaps under* CTL *and* LTC*:*

$$\mathsf{SG}_{\mathsf{CTL}}(\widehat{\pi}; \pi^\star) \lesssim \kappa(\pi^\star)\left(c(\varepsilon)\sqrt{\frac{\log(|\Pi|/\zeta)}{n}} + \sqrt{\alpha}\right),$$

$$\mathsf{SG}_{\mathsf{LTC}}(\widehat{\pi}; \pi^\star) \lesssim \kappa(\pi^\star)\left(c(\varepsilon)\sqrt{\frac{\log(|\Pi|/\zeta)}{n}} + \sqrt{\alpha \cdot c(\varepsilon)}\right),$$

where $a \lesssim b$ as shorthand for $a = \mathcal{O}(b)$, $c(\varepsilon) = \frac{e^\varepsilon+1}{e^\varepsilon-1}$ and $\kappa(\pi^\star) := e^{2R_{\max}} \cdot \frac{V_{\max}}{R_{\max}}\sqrt{\mathcal{C}^{\pi^\star}}$ is the single-policy concentrability related term.

Remark 3.6. Thanks to the use of RR in CTL and LTC, our algorithm is ε-LDP. Setting $\varepsilon = \infty$ and $\alpha = 0$ in the above utility bounds, leads to the same bound as in χPO. Moreover, as a by-product, the above theorem also directly gives results for privacy-only and corruption-only settings. Furthermore, it can be easily leveraged to establish bounds for the setting where corruption happens both before and after local privacy with a simple summation of the two bounds above. We stress that, as in Huang et al. (2024), we consider a finite policy class Π for the ease of presentation. The extension to an infinite function class can be easily achieved via the standard covering number argument. For example, for a linear reward model in $\mathbb{R}^d$ (or equivalently, a log-linear policy class), $\log|\Pi|$ will roughly be $\widetilde{\mathcal{O}}(d)$.

Interplay between local privacy and corruption. Under CTL, the impact of local privacy parameter ε (i.e., the first term) and corruption parameter α (i.e., the second term) is *separable* (additive), while LTC introduces a multiplicative term, adding an extra factor $\sqrt{c(\varepsilon)} \geqslant 1$. While these are upper bound results, we believe the different interplay (additive vs. multiplicative) is intrinsic, supported by recent tight results in mean estimation (Zhou & Zhang, 2024).

Comparison with prior private alignment. To the best of our knowledge, Chowdhury et al. (2024b) is the only related work that studies label privacy protection in offline alignment. However, it considers the standard RL-based approach where a reward model is explicitly learned before the policy optimization, rather than our RL-free direct optimization method. More importantly, while their work is limited to the linear reward setting, our method is the first to provide formal guarantees under *general function approximation* settings with the same (optimal) privacy cost of $c(\varepsilon)$ and a similar single-policy concentrability dependence.

Comparison with prior robust alignment. Formal bounds on robust DPO exist only in Chowdhury et al. (2024a) under *random-flipping* corruption, where labels flip with a *known* probability. This model is weaker than our Huber corruption and equivalent to label privacy noise under RR via reparameterization. In this context, our main result has two significant improvements over Chowdhury et al. (2024a): (i) Even under the linear model, Chowdhury et al. (2024a) only archives a $\mathcal{O}(1/n^{1/4})$ rate with worse *all-policy concentrability* dependence while ours is the optimal $\mathcal{O}(1/n^{1/2})$ rate with *single-policy concentrability*; (ii) For the general function approximation setting, Chowdhury et al. (2024a) fails to achieve a vanish suboptimality gap as $n \to \infty$ while ours maintains the optimal $\mathcal{O}(1/n^{1/2})$ rate. Another related work is Mandal et al. (2024), which only considers RL-based alignment with linear function approximations under adversary corruption of both prompt (responses) and labels. In contrast, our main focus is RL-free alignment for general function approximations while under label-corruption only.

References

Abdin, M. I., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H. S., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C. C. T., Chen, W., Chaudhary, V., Chopra, P., Giorno, A. D., de Rosa, G., Dixon, M., Eldan, R., Iter, D., Garg, A., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Huynh, J., Javaheripi, M., Jin, X., Kauffmann, P., Karampatziakis, N., Kim, D., Khademi, M., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Liang, C., Liu, W., Lin, E., Lin, Z., Madan, P., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Song, X., Tanaka, M., Wang, X., Ward, R., Wang, G., Witte, P., Wyatt, M., Xu, C., Xu, J., Yadav, S., Yang, F., Yang, Z., Yu, D., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Amortila, P., Foster, D. J., Jiang, N., Sekhari, A., and Xie, T. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024a.

Amortila, P., Foster, D. J., and Krishnamurthy, A. Scalable online exploration via coverability. *arXiv preprint arXiv:2403.06571*, 2024b.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Bagnell, J., Kakade, S. M., Schneider, J., and Ng, A. Policy search by dynamic programming. *Advances in neural information processing systems*, 16, 2003.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

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

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Charisopoulos, V., Esfandiari, H., and Mirrokni, V. Robust and private stochastic linear bandits. In *International Conference on Machine Learning*, pp. 4096–4115. PMLR, 2023.

Chaudhuri, K. and Hsu, D. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186. JMLR Workshop and Conference Proceedings, 2011.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chhor, J. and Sentenac, F. Robust estimation of discrete distributions under local differential privacy. In *International Conference on Algorithmic Learning Theory*, pp. 411–446. PMLR, 2023.

Chowdhury, S. R., Kini, A., and Natarajan, N. Provably robust DPO: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024a.

Chowdhury, S. R., Zhou, X., and Natarajan, N. Differentially private reward estimation with preference feedback. In *International Conference on Artificial Intelligence and Statistics*, pp. 4843–4851. PMLR, 2024b.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Feng, Q., Kasa, S. R., Yun, H., Teo, C. H., and Bodapati, S. B. Exposing privacy gaps: Membership inference attack on preference data for LLM alignment. *arXiv preprint arXiv:2407.06443*, 2024.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Gabbianelli, G., Neu, G., and Papini, M. Importance-weighted offline learning done right. In *International Conference on Algorithmic Learning Theory*, pp. 614–634. PMLR, 2024.

Georgiev, K. and Hopkins, S. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in neural information processing systems*, 35:6829–6842, 2022.

Hopkins, S. B., Kamath, G., Majid, M., and Narayanan, S. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 497–506, 2023.

Huang, A., Zhan, W., Xie, T., Lee, J. D., Sun, W., Krishnamurthy, A., and Foster, D. J. Correcting the mythos of KL-regularization: Direct alignment without overparameterization via Chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kamath, G. The broader landscape of robustness in algorithmic statistics, 2024. URL https://arxiv.org/abs/2412.02670.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Korkmaz, E. and Brown-Cohen, J. Learning differentially private rewards from human feedback. *https://openreview.net/pdf?id=reBq1gmlhS*, 2024.

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Li, M., Berrett, T. B., and Yu, Y. On robustness and local differential privacy. *The Annals of Statistics*, 51(2):717–737, 2023.

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Liu, Z., Lu, M., Zhang, S., Liu, B., Guo, H., Yang, Y., Blanchet, J., and Wang, Z. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. *arXiv preprint arXiv:2405.16436*, 2024.

Ma, J. Y., Yan, J., Jayaraman, D., and Bastani, O. Offline goal-conditioned reinforcement learning via f-advantage regression. *Advances in neural information processing systems*, 35:310–323, 2022a.

Ma, Y. J., Shen, A., Jayaraman, D., and Bastani, O. Smodice: Versatile offline imitation learning via state occupancy matching. *arXiv preprint arXiv:2202.02433*, 1(2):3, 2022b.

Mandal, D., Nika, A., Kamalaruban, P., Singla, A., and Radanović, G. Corruption robust offline reinforcement learning with human feedback. *arXiv preprint arXiv:2402.06734*, 2024.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

OpenAI, T. ChatGPT: Optimizing language models for dialogue. OpenAI, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and

Lowe, R. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Ross, S. and Bagnell, J. A. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

Rosset, C., Cheng, C.-A., Mitra, A., Santacroce, M., Awadallah, A., and Xie, T. Direct Nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shin, D., Dragan, A. D., and Brown, D. S. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. Hybrid RL: Using both offline and online data can make RL efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Song, Y., Swamy, G., Singh, A., Bagnell, D., and Sun, W. The importance of online data: Understanding preference fine-tuning via coverage. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., Liu, Z., Liu, Y., Wang, Y., Zhang, Z., Vidgen, B., Kailkhura, B., Xiong, C., Xiao, C., Li, C., Xing, E., Huang, F., Liu, H., Ji, H., Wang, H., Zhang, H., Yao, H., Kellis, M., Zitnik, M., Jiang, M., Bansal, M., Zou, J., Pei, J., Liu, J., Gao, J., Han, J., Zhao, J., Tang, J., Wang, J., Vanschoren, J., Mitchell, J., Shu, K., Xu, K., Chang, K.-W., He, L., Huang, L., Backes, M., Gong, Neil Zhenqiang Yu, P. S., Chen, P.-Y., Gu, Q., Xu, R., Ying, R., Ji, S., Jana, S., Chen, T., Liu, T., Zhou, T., Wang, W., Li, X., Zhang, X., Wang, X., Xie, X., Chen, X., Wang, X., Liu, Y., Ye, Y., Cao, Y., Chen, Y., and Yue, Z. TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024a.

Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., and Gan, C. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36, 2024b.

Tang, Y., Guo, Z. D., Zheng, Z., Calandriello, D., Munos, R., Rowland, M., Richemond, P. H., Valko, M., Pires, B. Á., and Piot, B. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023.

Wang, K., Kallus, N., and Sun, W. The central role of the loss function in reinforcement learning. *arXiv preprint arXiv:2409.12799*, 2024a.

Wang, L., Krishnamurthy, A., and Slivkins, A. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 766–774. PMLR, 2024b.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, 60(309):63–69, 1965.

Wu, Y., Sun, Z., Yuan, H., Ji, K., Yang, Y., and Gu, Q. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024a.

Wu, Y., Zhou, X., Tao, Y., and Wang, D. On private and robust bandits. *Advances in Neural Information Processing Systems*, 36, 2024b.

Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., and Su, W. J. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024.

Xiao, T. and Zhu, J. Foundations of large language models. *arXiv preprint arXiv:2501.09223*, 2025.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Xie, T., Foster, D. J., Krishnamurthy, A., Rosset, C., Awadallah, A., and Rakhlin, A. Exploratory preference optimization: Harnessing implicit Q*-approximation for sample-efficient RLHF. *arXiv preprint arXiv:2405.21046*, 2024.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., and Zhang, H. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5757–5773. PMLR, 2022.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Zhou, X. and Zhang, W. Locally private and robust multi-armed bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Zhu, H. and Zhang, A. Provably efficient offline goal-conditioned reinforcement learning with general function approximation and single-policy concentrability. *Advances in Neural Information Processing Systems*, 36, 2024.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Additional Related Work

The alignment problem has been extensively studied in the previous literature (Yu et al., 2021; Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022a; Shin et al., 2023; Zhan et al., 2023; Mandal et al., 2024). Besides the private or robust alignment related work we mentioned in the main text, we refer the readers to Sun et al. (2024a) for more general trustworthiness in large language models and to Xiao & Zhu (2025); Touvron et al. (2023) for comprehensive surveys on large language models. Here, we discuss some additional related work.

Alignment with Human Feedback. The most fundamental method to align LLM is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), which has been practically used in OpenAI (2022); Sun et al. (2024b); Bai et al. (2022a;b). Instead of fine-tuning models by training a reward model from human feedback and optimizing policy using Reinforcement Learning (e.g., Proximal policy optimization (PPO) (Schulman et al., 2017)), Direct Preference Optimization (DPO) (Rafailov et al., 2023) simplifies alignment by directly optimizing the policy using human preference data. This approach bypasses the need for a reward model and reinforcement learning method, resulting in a more stable and efficient training process (Abdin et al., 2024). In the following, we divide related work on alignment with human feedback based on different perspectives:

- **Extended works from DPO.** Taking DPO as a starting point, many preference optimization variants have emerged to improve efficiency, stability, adaptability, or other properties. Relevant examples are Chi-Squared Preference Optimization (χPO) (Huang et al., 2024), Rejection Sampling Optimization (RSO) (Liu et al., 2023), Identity Preference Optimization (IPO) (Azar et al., 2024), ΨPO (Azar et al., 2024), generalized preference optimization (GPO) (Tang et al., 2024), Direct Nash Optimization (DNO) (Rosset et al., 2024), Self-Play Preference Optimization (SPPO) (Wu et al., 2024a), and Exploratory Preference Optimization (XPO) (Xie et al., 2024). Our SquareχPO is a variant of χPO, where the main difference is in the loss function—more on this in the next bullet point.

- **The role of loss function.** Our SquareχPO is mainly different from the original χPO in the loss function used to estimate the policy, changed from log-loss to least square loss over probabilities. Compared to the log-loss, the square loss provides a more interpretable measure of error, avoids extreme gradient values for small probability estimates, and ensures numerical stability. Wang et al. (2024a) explores how different loss functions affect the sample efficiency and adaptivity in classification and RL problems. We remark that the use of the square loss is not by any means new in RL. For example, we have temporal-difference (TD) learning with squared loss for regression (Jin et al., 2021a; Xie et al., 2022) and Fitted Q-Iteration (FQI) (Munos & Szepesvári, 2008; Chen & Jiang, 2019), which uses least-squares to approximate the Bellman backup. Thus, we believe that our new generalization error bound can be useful when one aims to extend those problems to private and robust scenarios.

- **Type of regularization divergence.** The objective function of preference optimization can be generally written as *(reward) loss* + (regularization) *penalty* (Xiao & Zhu, 2025). A number of different regularizers have been proposed in the literature. Wang et al. (2023) proposes a generalized approach, f-DPO, by using f-divergences for the regularization term, to integrate a variety of popular divergences. Our mixed χ^2 divergence in SquareχPO can be viewed as a special case of f-DPO, and it can provably alleviate overoptimization and achieve sample-complexity guarantees based on single-policy concentrability (Huang et al., 2024). Notably, χ^2-regularization has been used in a number of RL works to derive single-policy concentrability guarantee (Wang et al., 2024b; Gabbianelli et al., 2024; Duan et al., 2020; Zhan et al., 2022; Amortila et al., 2024b; Zhu & Zhang, 2024; Lee et al., 2021; Ma et al., 2022a;b). Xiao et al. (2024) introduces a new regularizer called preference matching divergence which helps the LLM balance response diversification and reward maximization. Moreover, Liu et al. (2024) shows that the SFT Loss is implicitly an adversarial regularizer in RLHF, that provably mitigates overoptimization.

- **Coverage coefficients (or concentrability coefficients).** Coverage, a concept that captures how the training data "covers" the test distribution, has played a fundamental role in offline RL (Munos & Szepesvári, 2008; Xie et al., 2021a; Uehara & Sun, 2021; Zhan et al., 2022), offline-online (hybrid) RL (Ross & Bagnell, 2012; Xie et al., 2021b; Song et al., 2022; Amortila et al., 2024a; Song et al., 2024), and online RL (Kakade & Langford, 2002; Bagnell et al., 2003; Xie et al., 2022). The sub-optimality guarantees of SquareχPO obtained under the BT-preference model are based on the *single-policy concentrability*, that is, the data only needs to have a good cover over the chosen comparator policy. This is the gold standard in offline reinforcement learning due to being more effective compared with *all-policy concentrability*, which requires the offline data distribution to provide good coverage over the state distributions induced by *all* candidate policies.

Privacy and robustness interplay. The interaction of privacy and robustness has been investigated in many machine learning tasks. In the multi-arm bandits problem, the interaction of central DP and Huber corruption on rewards is investigated in Wu et al. (2024b), while the different orders of LDP and Huber corruption of rewards feedback of bandits have been studied in Zhou & Zhang (2024). Charisopoulos et al. (2023) study the problem of linear bandits problem, where the rewards are under LDP and Huber model. In statistical learning, there are many works that studied the interaction of privacy and robustness in different tasks (e.g., Kamath, 2024; Li et al., 2023; Chhor & Sentenac, 2023). Other works have studied the possibility of privacy might imply robustness or vice-versa. For example, Georgiev & Hopkins (2022) concludes that private mechanisms are automatically robust in many statistics problems. In contrast, Hopkins et al. (2023) shows adversarial robustness implies differential privacy in statistical estimation. In this paper, we investigate both central DP and local DP interacting with Huber contamination model in the offline alignment problem.

B. Generalization Bounds of Least-Square Regression under Privacy and Corruption

In this section, we provide a detailed version of our main techniques, i.e., generalization error bound of least-square regression under privacy constraints and corruption. We mainly focus on the case where the response variable is binary, given its immediate application in our scenarios. However, it can be easily generalized to the continuous case via random rounding, see Zhou & Zhang (2024).

Lemma B.1. *Let $\{(u_i, y_i')\}_{i=1}^n$ be a clean dataset of n points where each point is independently sampled from $u_i \sim \rho'$ and $y_i' \sim p(\cdot|u_i) := h^*(u_i) + \eta_i$, where $\{\eta_i\}_{i=1}^n$ are independent random variables such that $\mathbb{E}[y_i'|u_i] = h^*(u_i)$ and $y_i' \in \{-1, 1\}$. Let $\mathcal{H} : \mathcal{U} \to [-1, 1]$ be a class of real valued functions such that $h^* \in \mathcal{H}$, i.e., we assume realizability. Define the generalization error bounds for a learning algorithm's output $\widehat{h}$ as*

$$\mathsf{err}_{\mathsf{gen}}^2 := \mathbb{E}_{u\sim\rho'}[(\widehat{h}(u) - h^*(u))^2] \,.$$

Then, we have the following results across different settings:

1. *Under* CTL *or* LTC *where the observed dataset is $\{(u_i, z_i')\}_{i=1}^n$ (with $z_i' \in \{-1, 1\}$) that is generated according to* CTL *or* LTC *(Definition 2.3), the least-square regression solution $\widehat{h} = \operatorname{argmin}_{h\in\mathcal{H}} \sum_{i=1}^n (h(u_i) - c(\varepsilon)z_i')^2$ (with $c(\varepsilon) = \frac{e^\varepsilon+1}{e^\varepsilon-1}$) satisfies with probability at least $1 - \zeta$*

$$\mathsf{err}_{\mathsf{gen,CTL}}^2 \lesssim c(\varepsilon)^2 \cdot \frac{\log(|\mathcal{H}|/\zeta)}{n} + \alpha,$$
$$\mathsf{err}_{\mathsf{gen,LTC}}^2 \lesssim c(\varepsilon)^2 \cdot \frac{\log(|\mathcal{H}|/\zeta)}{n} + \alpha \cdot c(\varepsilon) \,.$$

Remark B.2. This result can be viewed as a nontrivial generalization of the standard one in Song et al. (2022) to the private and corrupted scenarios.

A key lemma in our proof is the following form of Freedman's inequality.

Lemma B.3 (Lemma A.2 in (Foster et al., 2021))**.** *Let $\{u_i\}_{i\le n}$ be a real-valued martingale difference sequence adapted to a filtration $\{\mathcal{F}_i\}_{i\le n}$. If $|u_i| \le R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1 - \zeta$,*

$$\sum_{i=1}^n u_i \le \eta \sum_{i=1}^n \mathbb{E}_{i-1}[u_i^2] + \frac{\log(1/\zeta)}{\eta},$$

where $\mathbb{E}_{i-1}[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_{i-1}]$.

Now we are ready to prove our generalization bound.

Proof of Lemma B.1. We start with CTL and the other one is similar. For any fixed $h \in \mathcal{H}$, we define

$$U_i^h := (h(u_i) - c(\varepsilon)z_i')^2 - (h^*(u_i) - c(\varepsilon)z_i')^2.$$

If we define the filtration $\mathcal{F}_i = \sigma(u_1, z_1', \ldots, u_i, z_i')$ and let $\mathbb{E}_{i-1}[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{i-1}]$, then we have that $\{D_i^h\}_{i\le n}$ where

$$D_i^h := \mathbb{E}_{i-1}[U_i^h] - U_i^h$$

is a martingale difference sequence adapted to $\{\mathcal{F}_i\}_{i\le n}$. We further notice that

$$\begin{aligned}\mathbb{E}_{i-1}[(D_i^h)^2] \le \mathbb{E}_{i-1}[(U_i^h)^2] &= \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2(h(u_i)+h^*(u_i)-2c(\varepsilon)z_i')^2] \\ &\lesssim c(\varepsilon)^2\cdot \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2],\end{aligned}$$

where the last step holds by the boundedness of z_i' and $h\in\mathcal{H}$. Moreover, let $\bar{y}_i$ be the intermediate corrupted label, we have

$$\begin{aligned}\mathbb{E}_{i-1}[U_i^h] &= \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(h(u_i)+h^*(u_i)-2c(\varepsilon)z_i')] \\ &= \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(h(u_i)+h^*(u_i)-2c(\varepsilon)z_i'+2\bar{y}_i-2\bar{y}_i+2y_i'-2y_i')] \\ &= \underbrace{\mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(-2c(\varepsilon)z_i'+2\bar{y}_i)]}_{\mathcal{T}_{\text{privacy}}} + \underbrace{\mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(2y_i'-2\bar{y}_i)]}_{\mathcal{T}_{\text{corruption}}} \\ &\quad + \underbrace{\mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(h(u_i)+h^*(u_i)-2y_i')]}_{\mathcal{T}_{\text{standard}}}.\end{aligned}$$

We are going to bound each of them. For $\mathcal{T}_{\text{privacy}}$, due to the generation process of z_i' via random response over $\bar{y}_i$ and the fact that each privacy noise in random response is independent of all other randomness, we have $\mathcal{T}_{\text{privacy}}=0$. For $\mathcal{T}_{\text{standard}}$, due to the fact that $\mathbb{E}_{i-1}[y_i'|u_i]=h^*(u_i)$, we have

$$\mathcal{T}_{\text{standard}} = \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2].$$

Combining all three terms, yields that

$$\mathbb{E}_{i-1}[U_i^h] = \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] + \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(2y_i'-2\bar{y}_i)].$$

Then, applying Lemma B.3 to $\{D_i^h\}_{i\le n}$ with a proper choice of η, we have

$$\begin{aligned}&\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] + \sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(2y_i'-2\bar{y}_i)] \\ &\lesssim \sum_i U_i^h + \frac{1}{2}\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] + c(\varepsilon)^2\cdot\log(1/\zeta).\end{aligned}$$

Re-arranging it leads to

$$\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] \lesssim \sum_i U_i^h + c(\varepsilon)^2\cdot\log(1/\zeta) + \sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(2\bar{y}_i-2y_i')].$$

Using a union bound over all $h\in\mathcal{H}$, we have that

$$\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] \lesssim \sum_i U_i^h + c(\varepsilon)^2\cdot\log(|\mathcal{H}|/\zeta) + \sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(2\bar{y}_i-2y_i')],\ \forall h\in\mathcal{H}.$$

Let's now use this result for $\widehat{h}$, noting that $\sum_i U_i^{\widehat{h}} \le 0$. So, we have

$$\begin{aligned}\sum_i \mathbb{E}_{i-1}[(\widehat{h}(u_i)-h^*(u_i))^2] &\lesssim c(\varepsilon)^2\cdot\log(|\mathcal{H}|/\zeta) + \sum_i \mathbb{E}_{i-1}[(\widehat{h}(u_i)-h^*(u_i))(2\bar{y}_i-2y_i') \\ &\lesssim c(\varepsilon)^2\cdot\log(|\mathcal{H}|/\zeta) + \alpha n,\end{aligned}$$

where the last step follows from α-Huber corruption. Finally, given the independent corruption, we can safely change from conditional expectation to unconditional one and divide both sides by n, leading to

$$\mathbb{E}_{u\sim\rho}[(\widehat{h}(u)-h^*(u))^2] \lesssim c(\varepsilon)^2\cdot\frac{\log(|\mathcal{H}|/\zeta)}{n} + \alpha,$$

which completes the proof for CTL.

LTC **case.** It follows the same proof flow as above and we highlight the different steps only. Now, let $\widetilde{y}_i$ be the intermediate privatized label, we have

$$\begin{aligned}\mathbb{E}_{i-1}[U_i^h] &= \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(h(u_i)+h^*(u_i)-2c(\varepsilon)z_i')] \\ &= \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))((h(u_i)+h^*(u_i))-2c(\varepsilon)(z_i'-\widetilde{y}_i+\widetilde{y}_i))] \\ &= \underbrace{\mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(-2c(\varepsilon)(z_i'-\widetilde{y}_i))]}_{\mathcal{T}_{\text{corruption}}} + \underbrace{\mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(-2c(\varepsilon)\widetilde{y}_i+h(u_i)+h^*(u_i))]}_{\mathcal{T}_{\text{privacy}}}.\end{aligned}$$

By the unbiased property of $c(\varepsilon)\widetilde{y}_i$ due to randomized response, we have

$$\mathcal{T}_{\text{privacy}} = \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2].$$

Then, again, applying Lemma B.3 to $\{D_i^h\}_{i\le n}$ with a proper choice of η, we have

$$\begin{aligned}&\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] + \sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(-2c(\varepsilon)(z_i'-\widetilde{y}_i))] \\ &\lesssim \sum_i U_i^h + \frac{1}{2}\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] + c(\varepsilon)^2\cdot\log(1/\zeta).\end{aligned}$$

Re-arranging it leads to

$$\sum_i \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))^2] \lesssim \sum_i U_i^h + c(\varepsilon)^2\cdot\log(1/\zeta) + \mathbb{E}_{i-1}[(h(u_i)-h^*(u_i))(2c(\varepsilon)(z_i'-\widetilde{y}_i))],$$

where the last term is the key difference with an additional $c(\varepsilon)$ factor. Following the same argument as in CTL, we have that under LTC

$$\mathbb{E}_{u\sim\rho}[(\widehat{h}(u)-h^*(u))^2] \lesssim c(\varepsilon)^2\cdot\frac{\log(|\mathcal{H}|/\zeta)}{n} + \alpha c(\varepsilon).$$

□

C. Additional Details on Section 3

In this section, we provide the proof of our main results in Section 3, which directly follows from Theorem C.1 and Lemma C.2 below. As we already mentioned, our proof is modular once we have the generalization error bounds. To provide more intuition on this, we first present the following meta theorem, which is a simple adaptation from the proof in Huang et al. (2024) to our `Square`χ`PO`.

Theorem C.1 (Meta Theorem for `Square`χ`PO` under BT)**.** *Under the BT-preference model, let Assumptions 3.2 and 3.3 hold. Define $\widehat{r}(x,a) := \beta\phi\left(\frac{\widehat{\pi}(a|x)}{\pi_{\mathsf{ref}}(a|x)}\right)$ for any output policy of `Square`χ`PO` (Algorithm 1). Then, we have*

$$J(\pi^\star) - J(\widehat{\pi}) \leqslant \frac{2V_{\max}}{R_{\max}}\sqrt{\mathcal{C}^{\pi^\star}\cdot\mathsf{err}_{\text{stat}}^2} + \beta\cdot\mathcal{C}^{\pi^\star} + 2\beta^{-1}\cdot\frac{V_{\max}^2\mathsf{err}_{\text{stat}}^2}{R_{\max}^2},$$

where

$$\mathsf{err}_{\mathsf{stat}}^2 = \mathbb{E}_{\pi_{\mathsf{ref}},\pi_{\mathsf{ref}}}\left[\left(\text{clip}_{2R_{\max}}[\widehat{\Delta}] - \text{clip}_{2R_{\max}}[\Delta^\star]\right)^2\right],$$

with $\widehat{\Delta} := \widehat{r}(x,a)-\widehat{r}(x,b)$ and $\Delta^\star := r^\star(x,a)-r^\star(x,b)$. Furthermore, by taking $\beta = \sqrt{\frac{2}{\mathcal{C}^{\pi^\star}}}\cdot\frac{V_{\max}\mathsf{err}_{\text{stat}}}{R_{\max}}$, we have

$$J(\pi^\star) - J(\widehat{\pi}) \lesssim \frac{V_{\max}}{R_{\max}}\sqrt{\mathcal{C}^{\pi^\star}\cdot\mathsf{err}_{\text{stat}}^2}.$$

Proof. The above result largely follows from the proof of Theorem E.1 in Huang et al. (2024). The key in their proof is the translation from working with policy to working with the implicit reward $\widehat{r}$ define above, i.e., Lemma E.2 in Huang et al. (2024). With this, one can follow the standard proof for RLHF to arrive at the above result by relying on the fact that $\mathcal{C}^{\pi} = 2D_{\chi^2}(\pi \| \pi_{\mathsf{ref}}) + 1$. Note that since our SquareχPO uses the same re-parametrization function ϕ as in χPO, so the above argument via their Lemma E.2 still works. □

With this meta theorem, all we need to do is to bound $\mathsf{err}^2_{\mathsf{stat}}$ under CTL and LTC, respectively, which will directly lead to our main results in Theorem 3.5. At a high level, without clipping, $\mathsf{err}^2_{\mathsf{stat}}$ can be bounded by directly leveraging our generalization error bound under realizability (Lemma B.1) and mean-value theorem to handle the non-linearity of $\sigma(\cdot)$ function. Here, the main reason for us to do the clipping is to ensure that the cost due to non-linearity is $O(e^{cR_{\max}})$ (for some constant $c > 0$) rather than the worse bound $O(e^{cV_{\max}})$. Due to this additional clipping, we have to carefully show that clipping will not impact our analysis, by showing that *realizability* is still satisfied. This should not be a surprise given the boundedness of r^* and all we need in the analysis is the *reward difference*.

Formally, we have the following bounds on $\mathsf{err}^2_{\mathsf{stat}}$ under CTL and LTC, respectively.

Lemma C.2. *Under the same conditions of Theorem C.1,* $\mathsf{err}^2_{\mathsf{stat}}$ *for* SquareχPO *in Algorithms 1 satisfies the following bounds:*

$$\mathsf{err}^2_{\mathsf{stat},\mathsf{CTL}} \lesssim e^{4R_{\max}} \left(c(\varepsilon)^2 \cdot \frac{\log(|\Pi|/\zeta)}{n} + \alpha \right),$$
$$\mathsf{err}^2_{\mathsf{stat},\mathsf{LTC}} \lesssim e^{4R_{\max}} \left(c(\varepsilon)^2 \cdot \frac{\log(|\Pi|/\zeta)}{n} + \alpha \cdot c(\varepsilon) \right).$$

Proof. **Local model.** By using the implicit reward function, we can re-write Step 3 in Algorithm 1 as

$$\widehat{r} = \operatorname*{argmin}_{r \in \mathcal{R}_\Pi} \sum_{i \in [n]} \left[2\sigma \left(\mathrm{clip}_{2R_{\max}} \left[r(x_i, a_i^1) - r(x_i, a_i^0) \right] \right) - 1 - c(\varepsilon) \bar{z}_i \right]^2,$$

where

$$\mathcal{R}_\Pi := \left\{ r(x,a) = \beta \cdot \phi \left(\frac{\pi(a \mid x)}{\pi_{\mathsf{ref}}(a \mid x)} \right) : \pi \in \Pi \right\},$$

and $\bar{z}_i = 2z_i - 1 \in \{1, -1\}$. In order to apply our generalization error bound in Lemma B.1, we can do the following mappings: for any $r \in \mathcal{R}_\Pi$, we map it to a function $h \in \mathcal{H}$ with $|\mathcal{H}| \le |\Pi|$ via $h(u_i) := 2\sigma \left(\mathrm{clip}_{2R_{\max}} \left[r(x_i, a_i^1) - r(x_i, a_i^0) \right] \right) - 1 \in [-1, 1]$ with $u_i = (x_i, a_i^1, a_i^0)$. Moreover, the label $\bar{z}_i$ is mapped to z_i' and the distribution over prompts and actions is mapped to ρ' in Lemma B.1. With such a mapping, all we need to check is the realizability, i.e., there exists an $h^* \in \mathcal{H}$ defined below such that for the true clean preference label $y_i \in \{0, 1\}$

$$\mathbb{E}[y_i' | u_i] = \mathbb{E}[2y_i - 1 | u_i] = h^*(u_i) := 2\sigma \left(\mathrm{clip}_{2R_{\max}} \left[\widetilde{r}^*(x_i, a_i^1) - \widetilde{r}^*(x_i, a_i^0) \right] \right) - 1, \tag{7}$$

where h^* is mapped from $\widetilde{r}^* := \beta \cdot \phi \left(\frac{\pi_\beta^*(a|x)}{\pi_{\mathsf{ref}}(a|x)} \right)$, which satisfies $\widetilde{r}^* \in \mathcal{R}_\Pi$ (hence $h^* \in \mathcal{H}$), because of policy realizability $\pi_\beta^* \in \Pi$. To verify that (7) indeed holds, we note that

$$\mathrm{clip}_{2R_{\max}} \left[\widetilde{r}^\star(x,a) - \widetilde{r}^\star(x,b) \right] = \mathrm{clip}_{2R_{\max}} \left[r^\star(x,a) - r^\star(x,b) \right] = r^\star(x,a) - r^\star(x,b),$$

where the first equality holds by the folklore fact that $\widetilde{r}^*$ is equivalent to r^* up to an action-independent normalization factor, which gets canceled in the reward difference, and the second equality holds by the boundedness of true reward $r^* \in [0, R_{\max}]$. Applying σ function to both sides and noting that under the BT-preference model $\mathbb{E}[y_i | u_i] = \sigma(r^*(x_i, a_i^1) - r^*(x_i, a_i^0))$, yields the realizability condition in (7).

Thus, we can now safely apply Lemma B.1 to obtain results for the local model. In particular, for CTL, we have

$$\mathbb{E}_{u \sim \rho}[(\widehat{h}(u) - h^*(u))^2] = \mathbb{E}_{\pi_{\mathsf{ref}}, \pi_{\mathsf{ref}}} \left[\left(\sigma(\mathrm{clip}_{2R_{\max}}[\widehat{\Delta}]) - \sigma(\mathrm{clip}_{2R_{\max}}[\Delta^\star]) \right)^2 \right] \lesssim c(\varepsilon)^2 \cdot \frac{\log(|\Pi|/\zeta)}{n} + \alpha,$$

which directly leads to our conclusion by a standard mean-value theorem argument to get rid of σ function. The same argument applies to LTC case.

□