We would like to thank all referees for their appreciation of our results and the useful feedback. Below is our reply.

Reviewer 1: Thank you for pointing us to the relevant references. Si et al. was made available only in July 2020 during ICML, while we cited Hu et al. as Reference [17] in our paper. Our setting (including the loss function, the construction of the nominal distribution and the ambiguity set) is clearly different from Faury et al. and Si et al. Results from Section 5.3 of Hu et al. focus solely on loss functions which are linear in Y. On the contrary, our loss function is linear in $T(Y)$, where the sufficient statistics T can be a *non*linear map. Examples of exponential family of distributions with nonlinear T are Gaussian, Gamma and Beta distributions. The results from Hu et al. thus are not directly applicable to robustify MLE problems. Our proof is also novel: it invokes duality results on the expectation parameter space.

To our best knowledge, Faury et al. and Si et al. provide convergence results for the *objective value*, not for the *solution*. We, however, focus on the convergence of our *solution*, and Theorem 4.3 shows that our solution is consistent. Results that provide the rate of convergence for DRO solution is very scarce in the current literature (we are only aware of (arXiv:1906.01614)) because it is significantly harder to prove. This extension goes beyond the scope of this paper.

In many practical applications, the covariate space $\mathcal{X}$ is categorical/ordinal, thus N_c can be greater than 1. Moreover, model (5) needs only $(p+1)C$ parameters, not $C+p^C$ parameters as the reviewer thought (p is the dimension of Θ). We agree that the dependence on N should be made explicit in Proposition 4.2. We update Table 2 below with values of n, N and m. We will change the covariate space from $\mathbb{R}^n$ to $\mathbb{R}^d$ as you kindly recommended. Thank you!

Section 5 demonstrates the generality and power of our proposed approach. Under study are two popular applications of model estimation with exponential family of distributions: Poisson regression (see Example 3.4) is when $Y|X$ follows a Poisson distribution, and logistic regression (see Example 3.5) is when $Y|X$ follows a Bernoulli distribution.

Reviewer 2: We wholeheartedly agree with you that it is interesting to handle the contamination in the X space. We tried it, and we encountered two technical difficulties: first, the log-partition function Ψ is convex; second, there are multiplicative terms between X and Y. Maximizing over the X space to find the worst-case covariate is thus difficult. Even though this is a useful extension, providing sound and rigorous mathematical treatment requires significant breakthroughs in non-convex programming; or we need to approach this problem from a different perspective. One can think of perturbing each $\widehat{x}_c$ in a finite set but that would lead to trivial (and uninteresting) modifications of the constraints. We will discuss about the covariate contamination in the concluding remarks.

We will add relevant references from the robust statistics literature into the paper. Thank you for your suggestion! About the Poisson pmf: we use the *natural* parameter, while you are using the *expectation* parameter to characterize the Poisson distribution. Both are actually equivalent reparametrization. We will clarify this notation in the revised version.

Reviewer 3: 1) Thank you for the suggestion about the joint convergence. For brevity, we present here only the case when the nominal distribution is set using Example 2.2. Complete characterization of Z is also provided.

Lemma (Asymptotic convergence)**.** *Suppose that $|\mathcal{X}| = C$ with $\mathbb{P}(X=\widehat{x}_c)>0$. Let $\theta_c = \lambda(w_0, \widehat{x}_c)$ and $\widehat{\mathbb{P}}$ be defined as in Example 2.2. Let $V_c = D_c \mathrm{Cov}_{f(\cdot|\theta_c)}(T(Y)) D_c^\top$, where $D_c = J(\nabla\Psi)^{-1}(\mathbb{E}_{f(\cdot|\theta_c)}[T(Y)])$ and J denotes the Jacobian operator. Then the following joint convergence holds*

$$\left(N_1 \times \mathrm{KL}(f(\cdot|\theta_1) \parallel f(\cdot|\widehat{\theta}_1)), \ldots, N_C \times \mathrm{KL}(f(\cdot|\theta_C) \parallel f(\cdot|\widehat{\theta}_C))\right)^\top \xrightarrow{d.} Z \qquad \text{as} \qquad N \to \infty,$$

where $Z = (Z_1, \ldots, Z_C)^\top$ with $Z_c = \frac{1}{2} R_c^\top \nabla^2 \Psi(\theta_c) R_c$, R_c are independent and $R_c \sim \mathcal{N}(0, V_c)$.

2) By Proposition 4.1, nonparametric KL is equivalent to setting $\rho_c = 0$ and tune only with $\varepsilon > 0$. We rerun the experiment for logistic regression and update Table 2 as below: the column 'KL' reports the performance of nonparametric KL with $\varepsilon \in [10^{-4}, 10]$ with 10 logarithmic scale points. Intuitively, ρ_c should depend on $\widehat{x}_c$, but the dependence is dictated by the unknown value w_0. Choosing ρ_c based on $\widehat{x}_c$ is thus unrealistic.

Reviewer 4: Regarding the weakness: overfitting is a severe problem when the number of i.i.d. samples is relatively small compared to the dimension. This weakness (the improvement decreases as N gets large) is thus pertinent to all methods aiming to combat overfitting, including regularization and DRO, and not just to our proposed method.

	AUC					CCR				
Dataset	DRO	KL	L_1	L_2	MLE	DRO	KL	L_1	L_2	MLE
australian ($N=690$, $n=14$)	**92.74**	92.62	92.73	92.71	92.61	**85.75**	85.72	85.52	85.60	85.72
banknote ($N=1372$, $n=4$)	**98.46**	**98.46**	98.43	98.45	98.45	94.31	94.32	94.16	**94.35**	94.32
climate ($N=540$, $n=18$)	94.30	82.77	**94.85**	94.13	82.76	**95.04**	93.89	94.85	94.83	93.89
german ($N=1000, n=19$)	**75.75**	75.68	75.74	75.74	75.67	73.86	**74.05**	73.82	73.70	**74.05**
haberman ($N=306, n=3$)	66.86	67.21	**69.19**	68.17	67.20	**73.83**	73.80	73.20	73.18	73.80
housing ($N=506$, $n=13$)	**76.24**	75.73	75.37	75.57	75.73	91.65	91.70	**92.68**	92.65	91.70
ILPD ($N=583$, $n=10$)	**74.01**	73.66	73.56	73.77	73.66	71.11	71.07	71.68	**71.79**	71.07
mammo. ($N=830\, n=5$)	**87.73**	87.72	87.70	87.68	87.71	81.00	**81.20**	80.99	80.94	**81.20**

Table 2: Average area under the curve (AUC) and correct classification rates (CCR) on UCI datasets ($m=1$).