We are grateful to the reviewers for their thoughtful comments which have improved the work. Below are responses to each and every point of the reviewers (where we refer to the same references in the reviewer 2's comments):

Reviewer 1: We thank the reviewer for the positive comments. **1a.** The dependence of the constant C in Theorem 2 on the parameters (such as α) is complicated due to the delicate nature of our chaining method, so we did not write down the constant explicitly. We will carefully address which parameters that the constant C depends on. **1b.** We appreciate the reviewer's excellent suggestions on the writing structure, we will rewrite lines 29-30 and move lines 52-55 (the motivation of introducing distributional robustness) to follow the formal introduction of our model. **1c.** We thank the reviewer for pointing out the typos in our paper. The $\inf$ in line 45 refers to the infimum of a set, so we don't have to add a subscript for it. We will follow the reviewer's advice to correct the other typos. **2a.** Although the Weibull-type condition is already a significant generalization of related assumptions which are standard in convex regression formulations, we agree that weakening to finite moments assumption is a very interesting direction to explore in future work. We also thank the reviewer for bringing up the ideas of robust statistics. In our problem, we care about the shape of the estimator, it is not clear how estimators such as Huber-type M-estimators can be modified to preserve the convex shape, but this is also an interesting direction to explore. **2b.** Please see our response to **4** of Reviewer 2.

Reviewer 2: We thank the reviewer for the insightful comments. **1.** We appreciate the reviewer for bringing [2] to us, and we agree that it is important to illustrate the similarities and differences between [2, Theorem 1] and our result. The main difference is, the loss function in [2, Theorem 1] is required to be a composition of a Lipschitz function $l : \mathbb{R} \to \mathbb{R}$ and a linear function of x, y, due to the non-linear nature of our convex regression model, this result does not apply to our case directly. Besides, our approach also allows l be function maps $\mathbb{R}^2$ to $\mathbb{R}$. However, we agree that the proof of the two results share a similar idea: we apply the dual representation of the standard Wasserstein-based distributionally robust optimization problem, and then lower bound the dual parameter λ using the structure of the loss function l. Moreover, we will follow the reviewer's advice to cite other results similar to Lemma 1. **2.** The new interpretation of equation (2) that relates our problem to covariate shifting is very interesting, we will add this point to our submission. **3.** We thank the reviewer for giving us the chance to illustrate our contribution beyond the results in [4]. First of all, their results only focus on problems with **finite** dimensional decision space (or feasible set $\mathbb{X}$, in the words of [4]), while in our case the decision space $\mathcal{F}_n$ is clearly **infinite** dimensional. Secondly, to the best of our understanding, the concentration result ([4, Theorem 3.4], or [1, 6]) allows us to choose a proper δ_n such that the Wasserstein ball centered at P_n (the empirical measure) with radius δ_n covers the true underlying probability measure P (see [4, Theorem 3.5]) with high probability, and as a consequence we can appropriately choose δ_n to ensure the **consistency** of the estimator of the optimal decision variable (see [4, Theorem 3.6(2)]). However, it is non-trivial to identify the convergence rate of the estimator. For example, for finite dimensional decisions, such convergence rates should match the canonical rate $O(n^{-1/2})$. The choice of δ_n suggested in [4] does not provide the canonical rate, so it doesn't seem that the results in [4] are directly applicable to recover convergence rates for estimators; not even in the finite dimensional case which is the environment of [4], let alone the infinite dimensional case, which is the setting of our paper. **4.** We thank the reviewer for raising this issue, and we are agree with the reviewer's opinion from the optimization point of view. However, there is a technical statistical reason behind our formulation. On the one hand, the penalty term $\delta_n \|\nabla f\|_\infty$ captures the impact of the uncertainty set $\{P : D(P, P_n) \leq \delta_n\}$. On the other hand, the $\log n$ constraint introduced in $\mathcal{F}_n$ is related to a compactification argument applied to the decision space. This issue is particularly important in the current setting of infinite dimensional decisions. The current formulation provides one possible tradeoff of these two effects that guarantee the estimator $\hat{f}_n$ converges to f_* with order $\widetilde{O}(n^{-1/d})$. Furthermore, the $\log n$ constraint already relaxes the typical assumption that $\|\nabla f\|_\infty < C$ (in which C needs to be known apriori). Finally, we will provide a formal proof of the optimality of piecewise affine functions. **5.** We appreciate the reviewer's suggestions of adding experiments on real world data. We consider a public dataset from United States Environmental Protection Agency, which was suggested by [R. Mazumder, A. Choudhury, G. Iyengar, B. Sen, A Computational Framework for Multivariate Convex Regression and its Variants]. The dataset consists of 600 air market data of California in the first quarter of 2019. The response was the amount of heat input with the covariates corresponding to the amounts of emissions of SO2, NOx, CO2 (in tons) and the NOX rate. Empirical evidence suggests that relationship between the response and the log transformation of each individual covariate can be modeled well by a convex fit, so we do the log transformation on covariates and then standardize the data. Since we never know f^* in real data, we can not evaluate our method in the same way as the submitted paper. Instead, we randomly split the dataset into a training set with 400 data and a test set with 200 data, and we implement three different approaches: DRCR (our estimator), LSE (standard convex regression estimator) and LR (linear regression). We repeat the experiment 10 times and then compare the average training l_1 loss and average test l_1 error. We summarize the results in the table on the right, it is clear that our method outperforms both LSE and LR.

Method	Training loss	Test error
DRCR	**0.1238**	**0.1294**
LSE	0.1485	0.1516
LR	0.1691	0.1692

Reviewer 3: We thank the reviewer for the positive comments.