Automatic Outlier Rectification via Optimal Transport

Jose Blanchet
Dept. of Management Science & Engineering
Stanford University
jose.blanchet@stanford.edu

Jiajin Li
Sauder School of Business
University of British Columbia
jiajin.li@sauder.ubc.ca

Markus Pelger
Dept. of Management Science & Engineering
Stanford University
mpelger@stanford.edu

Greg Zanotti
Dept. of Management Science & Engineering
Stanford University
gzanotti@stanford.edu

Abstract

In this paper, we propose a novel conceptual framework to detect outliers using optimal transport with a concave cost function. Conventional outlier detection approaches typically use a two-stage procedure: first, outliers are detected and removed, and then estimation is performed on the cleaned data. However, this approach does not inform outlier removal with the estimation task, leaving room for improvement. To address this limitation, we propose an automatic outlier rectification mechanism that integrates rectification and estimation within a joint optimization framework. We take the first step to utilize the optimal transport distance with a concave cost function to construct a rectification set in the space of probability distributions. Then, we select the best distribution within the rectification set to perform the estimation task. Notably, the concave cost function we introduced in this paper is the key to making our estimator effectively identify the outlier during the optimization process. We demonstrate the effectiveness of our approach over conventional approaches in simulations and empirical analyses for mean estimation, least absolute regression, and the fitting of option implied volatility surfaces.

1 Introduction

Outlier removal has a long tradition in statistics, both in theory and practice. This is because it is common to have (for example, due to collection errors) data contamination or corruption. Directly applying a learning algorithm to corrupted data can, naturally, lead to undesirable out-of-sample performance. Our goal in this paper is to provide a single-step optimization mechanism based on optimal transport for automatically removing outliers.

The challenge of outlier removal has been documented for centuries: early work is introduced in e.g. Gergonne [1821] and Peirce [1852]. Yet the outlier removal problem continues to interest practitioners and researchers alike due to the danger of distorted model estimation. A natural family of approaches followed in the literature takes the form of "two-stage" methods, which involve an outlier removal step followed by an estimation step. Methods within this family range from rules of thumb, such as removing outliers beyond a particular threshold based on a robust measure of scale like the interquartile range [Tukey, 1977], to various model-based or parametric distributional assumption-based tests [Thompson, 1985]. While the two-stage approach can be useful by separating the estimation task from the outlier removal objective, it is not without potential pitfalls. For example, outlier detection which relies on a specific fitted model may tend to overfit to a particular type of outlier, which can mask the effect of other types [Rousseeuw and Leroy, 1987]. Conversely, the

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1113

outlier detection step may be overly conservative if it is not informed by the downstream estimation task; this can lead to a significant reduction in the number of observations available for model estimation in the next step, resulting in a loss of statistical efficiency [He and Portnoy, 1992].

Robust statistics, pioneered by Box [1953], Tukey [1960], Huber [1964] and others such as Hampel [1968, 1971] offers alternative approaches for obtaining statistical estimators in the presence of outliers without removing them, particularly in parametric estimation problems such as linear regression. Beyond these, one closely related approach is the minimum distance functionals-based estimator, introduced by Parr and Schucany [1980], Millar [1981], Donoho and Liu [1988a,b], Park et al. [1995], Zhu et al. [2022], Jaenada et al. [2022]. This method involves projecting the corrupted distribution onto a family of distributions using a distribution metric and selecting the optimal estimator for the resulting distribution. However, this projection mechanism is not informed by the estimation task (such as fitting a linear regression or neural network). Thus, without additional information about the contamination model or estimation task, it can be challenging to choose an appropriate family of distributions for projection, which may lead to limitations similar to those outlined above in practice.

In this paper, we propose a novel approach to integrate both outlier detection and estimation in a joint optimization framework. A key observation is that statisticians aim to "clean" data **before** decisions are made; thus, ideal robust statistical estimators tend to be optimistic. To address this, we introduce the *rectification set*: a ball centered around the empirical (contaminated) distribution and defined by the optimal transport distance (see Villani [2009], Peyré et al. [2019]) in probability space. This rectification set aims to exclude potential outliers and capture the true underlying distribution, allowing us to minimize the expectation over the "best-case" scenarios, leading to a min-min problem. The study in Jiang and Xie [2024] also explores connections with min-min-type problems, but concentrates on the construction of artificially constructed rectification sets to cover certain existing estimators from the robust statistics literature. However, our primary focus is on introducing a novel rectification set, which is based on the optimal transport approach with a concave cost function.

To automatically detect outliers during the estimation process, one of our main contributions is the use of a concave cost function for the optimal transport distance. This function encourages what we refer to as "long haul" transportation, in which the optimal transport plan moves only a portion of the data to a distant location, while leaving other parts unchanged. This strategic approach effectively repositions identified outliers closer to positions that better align with the clean data. Our novel formulation then involves minimizing the expected loss under the optimally rectified empirical distribution. This rectification is executed through the application of an optimal transport distance with a concave cost function, thereby correcting outliers to enhance performance within a fixed estimation task. More importantly, our method distinguishes itself from distributionally robust optimization (DRO) [Ben-Tal et al., 2013, Bayraksan and Love, 2015, Wiesemann et al., 2014, Delage and Ye, 2010, Gao and Kleywegt, 2022, Blanchet and Kang, 2017, Shafieezadeh Abadeh et al., 2015, Shafieezadeh-Abadeh et al., 2019, Sinha et al., 2018, Kuhn et al., 2019] due to the distinctive correction mechanisms initiated by the robust formulation employing a min-min strategy.

As we discuss in Section 2 below, the timing of error generation differs crucially between DRO and robust statistics. DRO approach employs a min-max game strategy to control the worst-case loss over potential post-decision distributional shifts. In contrast, the robust estimator acts after the pre-decision distributional contamination materializes. Thus the approach of robust statistics can be motivated as being closer to a max-min game against nature. As a consequence, in robust statistics, the adversary moves first, and therefore the statistician can be more optimistic that they can rectify the contamination applied by the nature thus motivating the min-min strategy suggested above. However, it is also worth mentioning that Kang and Bansal [2023], Nietert et al. [2023] are also trying to formulate the outlier-robust problem in a min-max form, incorporating additional information about the contamination model to further shrink the ambiguity set.

It is useful to contrast our approach with more traditional approaches to achieving robustness in statistics, such as M-estimators and random sample consensus methods such as RANSAC. The class of M-estimators [Lehmann and Casella, 2006, Van der Vaart, 2000, Hayashi, 2000] is quite broad. This class includes methods which are both robust and not robust. M-estimators in the context of traditional robust statistics (as in introduced by Huber in the 1960s) attempt to achieve robustness by using specific loss functions which make the estimation procedure less sensitive to outliers. Common examples include the Huber loss function and least absolute deviations regression. Our approach is more fundamental as it builds on top of a given (general) loss and the statistician has full control of

the "outlier removal" modeling task (via the specification of optimal transport theory using a concave optimal transport cost) and the statistical task (via the chosen loss). In addition, the statistician also obtains the optimal "rectifier" (i.e. transporter) while the amount of rectification can be estimated via a cross-validation approach. Our approach is thus conceptually different from M-estimators. The benefits of our approach over M estimators (including the Huber loss) are illustrated in the experimental sections of our paper. RANSAC, first proposed in Fischler and Bolles [1981], handles outliers by iteratively selecting random data subsets, fitting models, and evaluating models on inliers to reach a best model. In contrast, our method optimally selects the set of points for rectification using a single optimization procedure. Though work on random sample consensus methods which guide sampling in more intelligent ways, such as Chin et al. [2011], may be seen as connected to our approach, our work is notably different. For example, random sample consensus methods only identify inliers, while the approach proposed in this paper identifies inliers and outliers and rectifies outliers. Our approach scales to high dimensions, while RANSAC encounters fundamental issues with sampling in such dimensions. Finally, our approach is based on optimal transport theory rather than heuristic or random sampling.

We summarize our main contributions as follows:

(i) Novel statistically robust estimator. We propose a new statistically robust estimator that incorporates a novel rectification set constructed using the optimal transport distance. By employing a concave cost function within the optimal transport distance, our estimator enables automatic outlier rectification. We prove that the optimal rectified distribution can be found via a finite convex program, and show that we can determine the optimal rectification set uncertainty budget via cross-validation.

(ii) Connection to adaptive quantile regression. For mean estimation and least absolute regression, we demonstrate that our robust estimator is equivalent to an adaptive quantile (regression) estimator, with the quantile controlled by the budget parameter δ. Furthermore, we prove that the optimal rectified distribution exhibits a long haul structure that facilitates outlier detection.

(iii) Effectiveness in estimating the option implied volatility surface. We evaluate our estimator on various tasks, including mean estimation, least absolute regression, and option implied volatility surface estimation. Our experimental results demonstrate that our estimator produces surfaces that are 30.4% smoother compared to baseline estimators, indicating success in outlier removal, as surfaces should be smooth based on structural financial properties. Additionally, it achieves an average reduction of 6.3% in mean average percent error (MAPE) across all estimated surfaces, providing empirical evidence of the effectiveness of our rectifying optimal transporter.

2 DRO and Robust Statistics as Adversarial Games

In this section, we summarize conceptually how robust statistics is different from DRO (for more details, we refer the reader to e.g. Blanchet et al. [2024]). To lay a solid mathematical foundation, we begin by investigating a generic stochastic optimization problem. Here, we assume that Z is a random vector in a space $\mathcal{Z} \subseteq \mathbb{R}^d$ that follows the distribution $\mathbb{P}_\star$. The set of feasible model parameters is denoted Θ (assumed to be finite-dimensional to simplify). Given a realization z and a model parameter $\theta \in \Theta$ the corresponding loss is $\ell(\theta, z)$. A standard expected loss minimization decision rule is obtained by solving

$$\min_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta,\xi)] = \int_{\Xi} \ell(\theta,\xi)\, \mathrm{d}\mathbb{P}_\star(\xi). \tag{1}$$

Since $\mathbb{P}_\star$ is generally unknown, to approximate the objective function in (1), we often gather n independent and identically distributed (i.i.d.) samples $\{z_i\}_{i=1}^n$ from the unknown data-generating distribution $\mathbb{P}_\star$ and consider the empirical risk minimization counterpart,

$$\min_{\theta\in\Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, Z)] = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, z_i), \tag{2}$$

where $\hat{\mathbb{P}}_n$ denotes the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and δ_z is the Dirac measure centered at z. These problems are solved within the context of the general data-driven decision-making cycle below:

In the cycle depicted in Figure 1, we usually collect n i.i.d samples from the unknown data-generating distribution $\mathbb{P}'$, which may be identical to or distinct from the clean distribution $\mathbb{P}_\star$. Subsequently, we make a decision (e.g., parameter estimation) based on a model, $\mathbb{P}'_n$, built from these samples. Such a model could be parametric or non-parametric. These decisions are then put into action within the out-of-sample environment $\tilde{\mathbb{P}}$, which may or may not conform to the distribution $\mathbb{P}_\star$. In this general cycle, the sample average method (2) may lead to poor out-of-sample guarantees. This motivates

https://doi.org/10.52202/079017-1113

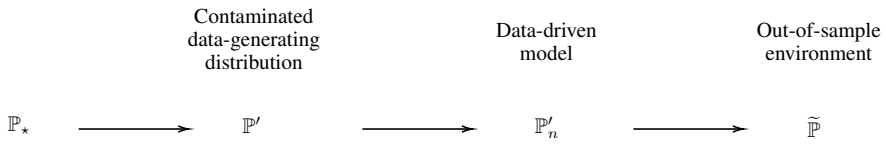


Figure 1: Data-Driven Decision Making Cycle

the use of alternative approaches. In the following paragraphs, we describe two of these approaches, DRO and robust statistics, by treating them as adversarial games. The crucial distinction lies in the **timing** of the contamination or attacks.

(i) (**DRO**: $\widetilde{\mathbb{P}} \neq \mathbb{P}_\star = \mathbb{P}'$) The attack occurs in the **post-decision** stage. In this scenario, the out-of-sample environment $\widetilde{\mathbb{P}}$ diverges from the data-generating distribution $\mathbb{P}_\star$, and no contamination occurs before the decision, implying that $\mathbb{P}_\star = \mathbb{P}'$. For example, in adversarial deployment scenarios, malicious actors can deliberately manipulate the data distribution to compromise the performance of trained models. With full access to our trained machine learning model, the adversary endeavors to create adversarial examples specifically designed to provoke errors in the model's predictions.

To ensure good performance in terms of the optimal expected population loss over the out-of-sample distribution, the DRO framework introduces an uncertainty set $\mathcal{B}(\mathbb{P}'_n)$ to encompass discrepancies between the in-sample-distribution $\mathbb{P}'_n$ and the out-of-sample distribution $\widetilde{\mathbb{P}}$. Subsequently, the DRO formulation minimizes the worst-case loss within this uncertainty set, thereby aiming to solve

$$\min_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{B}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, Z)]. \tag{3}$$

(ii) (**Robust Statistics**: $\widetilde{\mathbb{P}} = \mathbb{P}_\star \neq \mathbb{P}'$) The contamination occurs in the **pre-decision** stage. Many real-world datasets exhibit outliers or measurement errors at various stages of data generation and collection. In such scenarios, the observed samples are generated by a contaminated distribution $\mathbb{P}'$, which differs from the underlying uncontaminated distribution $\mathbb{P}_\star$. But, the out-of-sample distribution equals the original clean distribution. In contrast with DRO, the adversary corrupts the clean data according to a contamination model prior to training. Our objective is to clean the data during the training phase to achieve a robust classifier. It is noteworthy that the attacker does not have access to the specific model to be selected by the learner.

Given that the statistician knows that the data has been contaminated, a natural policy class to consider involves rectifying/correcting the contamination, and, for this, we introduce a rectification set $\mathcal{R}(\mathbb{P}'_n)$ which models a set of possible pre-contamination distributions based on the knowledge of the empirical measure $\mathbb{P}'_n$. To ensure good performance in terms of the optimal expected population loss over the clean distribution, the rectification/decontamination approach naturally induces the following min-min strategy:

$$\min_{\theta\in\Theta} \min_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}\left[\ell(\theta, Z)\right]. \tag{4}$$

Our goal in this paper is to develop (4), which is completely distinct from DRO.

3 Automatic Outlier Rectification Mechanism

We now introduce our primary contribution by delineating (4). Our estimator is crafted to incorporate outlier rectification and parameter estimation within a unified optimization framework, facilitating automatic outlier rectification. A natural question arises from (4): how do we construct the rectification set in the space of probability distributions to correct the observed data set? In this paper, we employ an optimal transport distance to create a ball centered at the contaminated empirical distribution $\mathbb{P}'_n$.

Definition 3.1 (Rectification Set)**.** The optimal transport-based rectification set is defined as

$$\mathcal{R}(\mathbb{P}'_n) = \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : \mathbb{D}_c(\mathbb{Q}, \mathbb{P}'_n) \leqslant \delta\}, \tag{5}$$

where $\delta > 0$ is a radius, and $\mathbb{D}_c$ is a specific optimal transport distance defined in Definition 3.2.

Definition 3.2. (Optimal Transport Distance [Peyré et al., 2019, Villani, 2009]) Suppose that $c(\cdot,\cdot) : \mathcal{Z}\times\mathcal{Z} \to [0,\infty]$ is a lower semi-continuous cost function such that $c(z,z') = 0$ for all $z, z' \in \mathcal{Z}$ satisfying $z = z'$. The optimal transport distance between two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ is

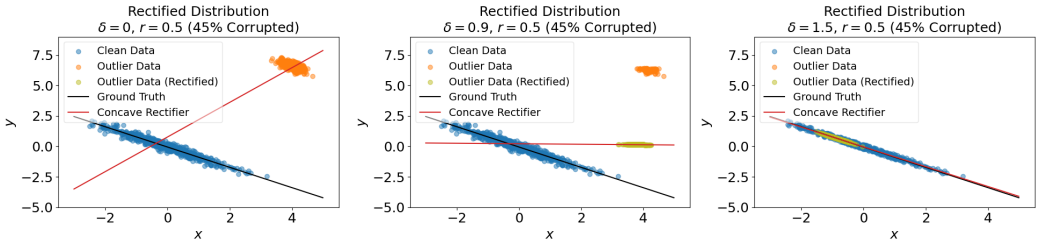
$$\mathbb{D}_c(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \Big\{\mathbb{E}_\pi[c(Z,Z')] : \pi_Z = \mathbb{P},\ \pi_{Z'} = \mathbb{Q}\Big\}.$$

https://doi.org/10.52202/079017-1113

Here, $\mathcal{P}(\mathcal{Z} \times \mathcal{Z})$ is the set of joint probability distribution π of (Z, Z') supported on $\mathcal{Z} \times \mathcal{Z}$ while π_Z and $\pi_{Z'}$ respectively refer to the marginals of Z and Z' under the joint distribution π.

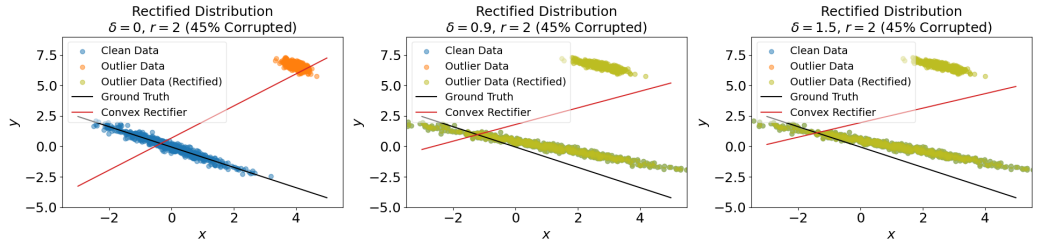
If we consider $c(z, z') = \|z - z'\|^r$ as a metric defined on $\mathbb{R}^d$, where $r \in [1, \infty)$, then the distance metric $\mathbb{D}_c^{1/r}(\mathbb{P}, \mathbb{Q})$ corresponds to the r-th order Wasserstein distance [Villani, 2009]. In this paper, we pioneer the utilization of *concave* cost functions, exemplified by $c(z, z') = \|z - z'\|^r$ where $r \in (0, 1)$ in statistical robust estimation. We note that $\mathbb{D}_c(\mathbb{P}, \mathbb{Q})$ (as opposed to $\mathbb{D}_c^{1/r}(\mathbb{P}, \mathbb{Q})$) is also a metric for $r \in [0, 1)$. The rationale behind selecting a concave cost function is intuitive: it promotes what we colloquially refer to as *long haul* transportation plans, enabling outliers to be automatically moved significant distances back towards the central tendency of data distribution. This, in turn, facilitates automatic outlier rectification. Concave costs promote long hauls due to their characteristic of exhibiting decreasing marginal increases in transportation cost. In other words, if an adversary decides to transport a data point by $\|\Delta\|$ units, it becomes cheaper to continue transporting the same point by an additional ε distance compared to moving another point from its initial location. We make further remarks about the connection between our estimator and prior work in Appendix B.1.

Illustrative Example. We provide empirical evidence showcasing the efficacy of the concave cost function on simulated data in Figure 2 and 3. This example shows that our concave cost function is critical for moving the outliers (orange) properly to the bulk of the clean distribution (blue). For the concave cost, only the outliers are rectified (green), resulting in the proper line of best fit.



(a) Budget $\delta = 0$ (b) Budget $\delta = 0.9$ (c) Budget $\delta = 1.5$

Figure 2: The rectified data generated by our estimator with **concave** cost function ($r = 0.5$).



(a) Budget $\delta = 0$ (b) Budget $\delta = 0.9$ (c) Budget $\delta = 1.5$

Figure 3: The rectified data generated by our estimator with **convex** cost function ($r = 2$).

However, for the convex cost, regardless of the budget, all points both clean and corrupted are rectified towards each other instead, which results in an incorrect line of best fit. This illustrates succinctly the importance of our novel concave cost. More details on this example are given in Appendix F.1, and an additional example with non-clustered outliers is given in Appendix F.1.1.

4 Reformulation Results

With this intuition in hand, we focus on the derivation of equivalent reformulations for the infinite-dimensional optimization problem over probability measures in (4). These reformulations provide us with a fresh perspective on adaptive quantile regression or estimation, where our introduction of an efficient transporter rectifies the empirical distribution, eliminating the influence of outliers.

To begin with, we can transform problem (4) into an equivalent finite-dimensional problem by leveraging the following strong duality theorem.

Proposition 1 (Strong Duality)**.** Suppose that $\ell(\theta, \cdot)$ is lower semicontinuous and integrable under $\mathbb{P}'_n$ for any $\theta \in \Theta$. Then, the strong duality holds, i.e.,

https://doi.org/10.52202/079017-1113

$$\inf_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;Z)] = \max_{\lambda\geqslant 0} \mathbb{E}_{\mathbb{P}'_n}\left[\min_{z\in\mathcal{Z}} \ell(\theta;z) + \lambda c(z,Z')\right] - \lambda\delta.$$

The proof is essentially based on the strong duality results developed for Wasserstein distributionally robust optimization problems [Zhang et al., 2022, Li et al., 2022, Blanchet and Murthy, 2019, Gao and Kleywegt, 2022, Mohajerin Esfahani and Kuhn, 2018], which allows us to rewrite the original problem (4) as $\inf_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;Z)] = -\sup_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[-\ell(\theta;Z)]$.

We proceed to examine several representative examples to better understand the proposed statistically robust estimator (4). We begin with mean estimation to showcase our estimator's performance on one of the most classic problems of point estimation which can be easily understood. We then give an example for least absolute deviations (LAD) regression, one of the most important problems of robust statistics. LAD regression builds a conceptual foundation which leads into a discussion of our framework in more general cases and problem domains.

4.1 Mean Estimation

The mean estimation task is widely recognized as a fundamental problem in robust statistics, making it an essential example to consider. In this context, we define the loss function as $\ell(\theta;z) = \|\theta - z\|$. It is worth noting that when $\delta = 0$, Problem (4) is equivalent to the median, which has already been proven effective in the existing literature. However, beyond the equivalence to the median, there are additional benefits to be explored regarding the proposed rectification set. By deriving the equivalent reformulation and analyzing the optimal rectified distribution, we can gain further insights into how the proposed statistically robust estimator operates. This analysis provides valuable intuition into the workings of the estimator and its advantages.

Theorem 2 (Mean Estimation)**.** Suppose that $\mathcal{Z} = \mathbb{R}^d$, $\ell(\theta;z) = \|\theta - z\|$ and the cost function is defined as $c(z,z') = \|z - z\|^r$ where $r \in (0,1)$. Without loss of generality, suppose that the following condition holds

$$\|\theta - z'_1\| \leqslant \|\theta - z'_2\| \leqslant \ldots \leqslant \|\theta - z'_n\|. \tag{6}$$

Then, we have the inner minimization of (4) as

$$\max\left(\frac{1}{n}\sum_{i=1}^{k(\theta)-1}\|\theta - z'_i\| + \frac{1}{n}\left(1 - \frac{n\delta - \sum_{i=k(\theta)+1}^{n}\|\theta - z'_i\|^r}{\|\theta - z'_{k(\theta)}\|^r}\right)\|\theta - z'_{k(\theta)}\|, 0\right). \tag{7}$$

where $k(\theta) := \max_{k\in[n]}\left\{k : \frac{1}{n}\sum_{i=k}^{n}\|\theta - z'_i\|^r \geqslant \delta\right\}$.

We give the proof in Appendix C.1.

Remark 4.1. The resulting reformulation problem can be viewed as finding a quantile of $\|\theta - z'\|$ controlled by the budget δ. If we have a sufficient budget δ such that $\mathbb{E}_{\mathbb{P}'_n}[\|\theta - Z'\|^r] \leqslant \delta$, it implies that all data points have been rectified to the value of θ. Consequently, the minimum value of $\min_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta,Z)]$ will be equal to zero. In nontrivial cases, when given a budget $\delta > 0$ and the current estimator θ, our objective is to identify and rectify the outliers in the observed data. To achieve this, we start by sorting the data points based on their loss value $\|\theta - z'_i\|$. We relocate the data points starting with the one having the largest loss value, z'_n. The goal is to move each data point towards the current mean estimation θ until the entire budget is fully utilized.

Building upon the proof of Theorem 2, we can establish the following characterization of the rectified distribution. The concave cost function $\|z - z'\|^r$ (with $r \in (0,1)$) plays a pivotal role in this context by endowing the perturbation with a distinctive long haul structure. Intuitively, for each data point, we can only observe two possible scenarios: either the perturbation is zero, indicating no movement or the data point is adjusted to eliminate the loss. In this process, the rectified data points are automatically identified as outliers and subsequently rectified.

Proposition 3 (Characterization of Rectified Distribution)**.** Assuming the same conditions as stated in Theorem 2, we can conclude that the optimal distribution $\mathbb{Q}^\star$

$$\mathbb{Q}^\star(dz) = \frac{1}{n}\sum_{i=1}^{k(\theta)-1}\delta_{z'_i}(dz) + \frac{\eta}{n}\delta_{z'_{k(\theta)}} + \frac{n - k(\theta) + 1 - \eta}{n}\delta_\theta(dz)$$

where $\eta = 1 - \frac{n\delta - \sum_{i=k(\theta)+1}^{n}\|\theta - z'_i\|^r}{\|\theta - z'_{k(\theta)}\|^r}$.

Remark 4.2. The existence of optimal solutions follows directly from Yue et al. [2022, Theorem 2].

4.2 Least Absolute Deviation (LAD) Regression and More General Forms

Following the same technique, we can also derive the least absolute deviation case.

Theorem 4 (LAD Regression)**.** Suppose that $\mathcal{Z} := \mathcal{X} \times \mathcal{Y} = \mathbb{R}^{d+1}$, $\ell(\theta, z) = \|y - \theta^T x\|$ and the cost function is defined as $c(z, z') = \|z - z\|^r$ where $r \in (0, 1)$ and $\|\cdot\|$ is the ℓ_2 norm. Without loss of generality, suppose that $\|y'_1 - \theta^T x'_1\| \leqslant \|y'_2 - \theta^T x'_2\| \leqslant \ldots \leqslant \|y'_n - \theta^T x'_n\|$, we have the inner minimization of (4) as

$$\max\left(\frac{1}{n}\sum_{i=1}^{k(\theta)-1}\|y'_i - \theta^T x'_i\|^r + \frac{1}{n}\left(1 - \frac{n\delta' - \sum_{i=k(\theta)+1}^{n}\|y'_i - \theta^T x'_i\|^r}{\|y'_{k(\theta)} - \theta^T x'_{k(\theta)}\|^r}\right)\|y'_{k(\theta)} - \theta^T x'_{k(\theta)}\|, 0\right),$$

where $k(\theta) := \max_{k\in[n]}\left\{k : \frac{1}{n}\sum_{i=k}^{n}\|y'_k - \theta^T x'_k\|^r \geqslant \delta'\right\}$ and $\delta' = \delta\|(\theta, -1)\|^r$.

We give the proof of Theorem 4 in Appendix C.1. The structure of the optimal rectified distribution also resembles that of Proposition 3: the rectified data points are shifted towards the hyperplane $y = \theta^T x$ that best fits the clean data points. For a more detailed explanation of the long haul structure and extensions to more general cases, interested readers are encouraged to refer to Appendix B.1.

4.3 Computational Procedure

Our procedure is motivated by the empirical efficacy of subgradient descent for training deep neural networks, as described in e.g. Li et al. [2020]. In each iteration, given the current estimate θ, we compute the optimal rectified distribution. For simple applications such as mean estimation and least absolute deviation regression, we can solve this optimization problem efficiently using the quick-select algorithm or by utilizing an existing solver for linear programs to obtain the optimal solution for the dual variable λ, which we denote λ^*. Then, we employ the subgradient method on the rectified data, iterating until convergence. This optimization process automatically rectifies outliers using a fixed budget δ across all iterations. As the value of θ changes, the same budget δ for all iterations results in varying the quantile used for selecting outliers. Thus, our estimator can be regarded as an iteratively adaptive quantile estimation approach.

We now concretely detail the procedure for LAD regression. The procedure for mean estimation is analogous and results from a simple change in the loss function. We must start by initially addressing the computation of the optimal dual variable $\lambda^\star$ for the inner minimization over the probability space. Without loss of generality, we recall that this problem is

$$\max_{\lambda\geqslant 0}\frac{1}{n}\sum_{i=1}^{n}\min\left\{\|\tilde{\theta}^T z'_i\|, \frac{\lambda\|\tilde{\theta}^T z'_i\|^r}{\|\tilde{\theta}\|_*^r}\right\} - \lambda\delta. \tag{8}$$

We show in the Appendix in Section C.1 by Lemma 6 that this problem can be solved by applying the quick-select algorithm to find the optimal $\lambda^\star$ and the knot point $k(\theta_t)$ which demarcates estimated outliers. Applying this lemma yields our estimation approach, which is described in Algorithm 1.

Algorithm 1: Statistically Robust Estimator

Data: Observed data $\{z'_i\}_{i=1}^n$, initial point θ_0, stepsizes $\alpha_t > 0$;

for $t = 0, \ldots, T$ **do**

1. **Sort** the observed data $\{z'_i\}_{i=1}^n$ via the value $\|y'_i - \theta_t^T x'_i\|$.
2. **Quick-Select** algorithm to get the knot point $k(\theta_t)$ and the optimal $\lambda^\star$;
3. **Subgradient step** on the detected clean data:

$$\theta_{t+1} = \theta_t - \frac{\alpha_t}{n}\sum_{i=1}^{k(\theta_t)}\operatorname{sgn}(\theta_t^T x'_i - y'_i)\cdot x'_i$$

end

Additional details on an alternative linear program approach for estimating $k(\theta_t)$ and λ^* are given in Appendix D.1. We also propose a procedure for estimating more general regression models in Appendix D.2. Further details on the optimization problem are given in Appendix D.3.

Limitations. Appendices D.2 and D.3 contain explanations of the limitations of our approach, which include the difficulty of solving the associated optimization problems and the nature of the derivation of our optimization algorithm for more general regression models.

https://doi.org/10.52202/079017-1113

5 Experimental Results

In this section, we demonstrate the effectiveness of the proposed statistically robust estimator through various tasks: mean estimation, LAD regression, and two applications to volatility surface modeling.

5.1 Mean Estimation and LAD Regression Experiments

We perform simulation experiments with mean estimation and least absolute deviation linear regression for our estimator to illustrate its efficacy. For mean estimation, we compare our estimator with the mean, median, and trimmed mean estimators. We observe that our estimator outperforms all other methods, despite providing the oracle corruption level ε to the trimmed mean estimator. Our results are displayed in Table 1 below.

Table 1: We compared our estimator with several standard mean estimation methods by evaluating the average loss on clean data points across various corruption levels. In our evaluation, we set the percent level of trimmed mean equal to the unknown ground truth corruption level. The hyperparameters for our estimator, namely $\delta = 0.5$ and $r = 0.5$, remained constant across all corrupted levels. The last row in the comparison table represents the percentage of all points rectified by our method. Error bars represent two standard deviation confidence intervals assuming normal errors.

Corruption Levels	20%	30%	40%	45%	49%
Mean	5.009 ± 0.056	7.509 ± 0.080	10.007 ± 0.098	11.248 ± 0.098	12.257 ± 0.114
Median	1.680 ± 0.114	1.831 ± 0.118	2.286 ± 0.192	2.843 ± 0.208	4.203 ± 0.344
Trimmed Mean	1.739 ± 0.108	1.947 ± 0.104	2.456 ± 0.212	3.047 ± 0.174	4.399 ± 0.360
Ours	$\mathbf{1.620 \pm 0.106}$	$\mathbf{1.700 \pm 0.106}$	$\mathbf{1.957 \pm 0.160}$	$\mathbf{2.251 \pm 0.150}$	$\mathbf{2.899 \pm 0.240}$
Ours, %Rectified	10.03%	10.12%	10.24%	10.35%	10.55%

For LAD regression, we compare our estimator with the OLS, LAD, and Huber regression estimators. We find that, again, our estimator outperforms all other methods. A complete overview of these two experiments can be found in Appendix E.

5.2 Options Volatility Experiments

In this section, we conduct empirical studies on real-world applications in fitting and predicting option implied volatility surfaces. Options are financial instruments allowing buyers the right to buy (call options) or sell (put options) an asset at a predetermined price (strike price) by a specified expiration date. In this study, we focus on European-style options, which can only be exercised at expiration. Option prices are influenced by the volatility of the underlying asset's price. Implied volatility (IV), derived from an option's market price, indicates the market's volatility expectations. The implied volatility surface (IVS) represents the variation of IV across different strike prices and times to maturity. Accurate IVS modeling is crucial for risk assessment, hedging, pricing, and trading decisions. However, outliers in IV can distort the IVS, necessitating robust estimation methods. Our approach addresses this by estimating the IVS in the presence of outliers. We conduct two experiments to demonstrate the effectiveness of our statistically robust estimator: (a) using a kernelized IVS estimator and (b) using a state-of-the-art deep learning IVS estimator.

5.2.1 Kernelized Volatility Surface Estimation

Data. Our data set comprises nearly 2,000 option chains containing daily US stock option prices from 2019–2021. The options data is sourced from WRDS, a financial research service available to universities. The option chains were identified as containing significant outliers by our industry partner, a firm providing global financial data and analysis. We were blinded to this choice. We randomly draw a training set and test set from each chain and estimate surfaces. We assess the surfaces' out-of-sample performance using mean absolute percentage error (MAPE) and the discretized surface gradient (defined as $\nabla \hat{S}$). MAPE evaluates the error of the surface versus observed implied volatilities. $\nabla \hat{S}$ evaluates the smoothness of the surface. Further details can be found in Appendix G.2.

Benchmarks. We compare our estimator developed in Theorem 4 to two benchmarks. The first is kernel smoothing (denoted "KS"), a well-established method for estimating the IVS [Aït-Sahalia and Lo, 1998]. In KS, each point on the IVS is constructed by a weighted local average of implied volatilities across nearby expiration dates, strike prices, and call/put type. The weights are given by a kernel matrix which depends on these three features of an option (see next paragraph). The second

benchmark is a two-step "remove and fit" kernel smoothing method (denoted "2SKS") which first attempts to remove outliers via Tukey fences [Hoaglin et al., 1986] before applying the KS method.

Kernelized Regression Problem. We now describe the kernel and our estimator. Following the convention in Aït-Sahalia and Lo [1998], the ith option in a chain is featurized as the vector $\mathbb{R}^3 \ni x_i' = (\log \tau_i, u(\Delta_i), \mathbb{I}_{\text{call}(i)})$, where τ_i is the number of days to the option's expiration date, $\Delta_i \in [-1, 1]$ is a relative measure of price termed the Black-Scholes delta, $u(x) = x\mathbb{I}_{x \geqslant 0} + (1 + x)\mathbb{I}_{x<0}$, and $\mathbb{I}_{\text{call}(i)}$ is 1 if the option is a call option and 0 if the option is a put option. The goal is to fit the pair (y_i': the implied volatility of option i, x_i': the features) via a kernelized regression model. We choose a Gaussian-like kernel $K_h(x, x')$ to measure the distance between options x and x' as $K_h(x, x') = \exp\left(-\|(x - x')/2h\|^2\right)$ where division of $x - x'$ by h is element-wise for a vector of bandwidths $h \in \mathbb{R}^3_+$. When the budget $\delta = 0$, we want to solve $\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^n v_i |y_i' - \theta^T K_h^i|$, where v_i is the i-th entry of a vector of vegas for $\{x_i'\}_{i=1}^n$ and K_h^i is the i-th row of the kernel matrix. The implied volatilities are weighted by vega v_i to improve surface stability as in Mayhew [1995]. We conducted a comparison between our estimator and the benchmarks. We note that the KS method can be regarded as a standard kernelized least square approach [Hansen, 2022].

Results. Our approach improves upon the benchmark approach in both MAPE and $\nabla \hat{S}$ on the out-of-sample test sets, showcasing the importance of jointly estimating the rectified distribution and model parameters with our estimator. The results of our experiment are displayed in Table 5.2.1. We compare the KS and 2SKS benchmarks against our estimator with both fixed $\delta = 0.01$ (chosen to correspond to a 1% change in volatility) and δ chosen by cross-validation. More details can be found in Appendix G.2. Our estimator outperforms others with a mean MAPE of 0.225, compared to 0.294 for KS and 0.236 for 2SKS, showing a 23% and 5% improvement respectively. For surface smoothness, our estimator achieves a mean DSG of 6.5, significantly better than 20.2 for KS and 7.5 for 2SKS, indicating 67% and 13% improvements. Notably, our industry partner applied our estimator to an unseen test set collected after this paper was written and was able to release to production 25% more surfaces than they had before when using their existing proprietary method.

Table 2: Results of our experiment with kernelized IVS estimation.

Model MAPE	0.5% Quantile	5% Quantile	Median	Mean	95% Quantile	99.5% Quantile
KS	0.026	0.068	0.232	0.294	0.677	1.438
2SKS	0.026	0.056	0.172	0.236	0.602	1.389
Ours ($\delta = 10^{-2}$)	0.028	0.057	0.170	0.225	0.535	1.207
Ours (CV)	0.028	0.057	0.169	0.224	0.534	1.240
Model $\nabla \hat{S}$	**0.5% Quantile**	**5% Quantile**	**Median**	**Mean**	**95% Quantile**	**99.5% Quantile**
KS	0.313	1.606	14.434	20.188	57.797	108.901
2SKS	0.050	0.124	2.229	7.491	33.276	78.144
Ours ($\delta = 10^{-2}$)	0.048	0.121	1.898	6.502	28.064	66.410
Ours (CV)	0.043	0.109	1.513	5.079	21.096	55.005

Remark 5.1. To contextualize these results, we note that the difference in MAPE of a surface and an option chain containing outliers and the MAPE of a surface and an option chain containing no outliers will not be large if the set of outliers is small. Consider the MAPE of the same surface for two different option chains of size n, O_1 and O_2, where a small fraction k/n of the options of O_2 are outliers. Supposing the modal APE is 0.3 and the outlier APE is a considerable 1.0, for an options chain with $n = 50$ and just $k = 5$, the MAPE difference will be only $\frac{k+0.3(n-k)}{n} - 0.3 = 0.07$.

We perform an additional experiment with the same dataset in Appendix G.1 which demonstrates the usefulness of our method for estimating an IVS for use on the trading day after the initial contaminated surface is observed. We find similar outperformance of our method versus the benchmark methods.

5.2.2 Deep Learning Volatility Surface Estimation

In this section, we apply our statistically robust estimator to state-of-the-art deep learning prediction approaches for modeling volatility developed in Chataigner et al. [2020]. In this work, deep networks are used to estimate local volatility surfaces. The *local volatility* or *implied volatility function* model introduced by Rubinstein [1994], Dupire et al. [1994], and Derman et al. [1996] is a surface which allows for as close of a fit as possible to market-observed implied volatilities without arbitrage, which is of interest to many market participants (smoothing methods, in contrast, do not make this guarantee). Further background and details for this section are available in Appendix G.3.

https://doi.org/10.52202/079017-1113

We select the data set from Chataigner et al. [2020] consisting of (options chain, surface) pairs from the German DAX index. We first contaminate the chains by replacing an ε fraction of each price p with $10p$. We then test our estimated surface against the true surface. To estimate price and volatility surfaces under data corruption, we applied our statistically robust estimator to the benchmark approach. We use the Dupire neural network of Chataigner et al. [2020] as a benchmark, which estimates the surface under local volatility model assumptions and no-arbitrage constraints. This method enforces these conditions on the surface using hard, soft, and hybrid hard/soft constraints.

We evaluate our robust estimator using the same metrics as Chataigner et al. [2020], RMSE and MAPE, repeated over three trials with different random seeds. Our approach outperforms the baseline approach across all averages, and does so more clearly as the corruption level increases, despite the strong regularizing effect of the no-arbitrage constraints and enforcement of Dupire's formula. Our improvement is most impactful for the most accurate model utilizing soft arbitrage constraints. For this model, the test set RMSE and MAPE are reduced by 33% and 34%. This experiment displays the efficacy of our estimator in a state-of-the-art deep learning approach to volatility modeling.

Table 3: Results of our experiment with deep learning surface estimation.

Panel A: Results by Constraint Type			
Model	Hard Constraints	Hybrid Constraints	Soft Constraints
Dupire NN RMSE	0.125	0.140	0.044
Our RMSE	0.110	0.131	0.029
Dupire NN MAPE	0.343	0.546	0.111
Our MAPE	0.310	0.508	0.074
Panel B: Results by Corruption Level			
Corruption Level	$\varepsilon = 20\%$	$\varepsilon = 30\%$	$\varepsilon = 40\%$
Dupire NN RMSE	0.077	0.091	0.168
Our RMSE	0.075	0.084	0.141
Dupire NN MAPE	0.061	0.070	0.115
Our MAPE	0.060	0.066	0.097

6 Conclusion

In conclusion, we propose an automatic outlier rectification mechanism that integrates outlier correction and estimation within a unified optimization framework. Our novel approach leverages the optimal transport distance with a concave cost function to construct a rectification set within the realm of probability distributions. Within this set, we identify the optimal distribution for conducting the estimation task. Notably, the concave cost function's "long hauls" attribute facilitates moving only a fraction of the data to distant positions while preserving the remaining dataset, enabling efficient outlier correction during the optimization process. Through comprehensive simulation and empirical analyses involving mean estimation, least absolute regression, and fitting option implied volatility surfaces, we substantiate the effectiveness and superiority of our method over conventional approaches. This demonstrates the potential of our framework to significantly enhance outlier detection integrated within the estimation process across diverse analytical scenarios.

Acknowledgments and Disclosure of Funding

J. Blanchet, M. Pelger, and G. Zanotti acknowledge support from Morgan Stanley Capital International. Support was provided to J. Li by the Air Force Office of Scientific Research under award number FA9550-20-1-0397.

References

Yacine Aït-Sahalia and Andrew W Lo. Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance*, 53(2):499–547, 1998.

Harbir Antil, Sean P Carney, Hugo Díaz, and Johannes O Royset. Rockafellian relaxation for pde-constrained optimization with distributional uncertainty. *arXiv preprint arXiv:2405.00176*, 2024.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pages 1–19. INFORMS, 2015.

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.

Jose Blanchet and Yang Kang. Distributionally robust groupwise regularization estimator. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2017.

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

Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024.

George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

Marc Chataigner, Stéphane Crépey, and Matthew Dixon. Deep local volatility. *Risks*, 8(3):82, 2020.

Louis L Chen and Johannes O Royset. Rockafellian relaxation in optimization under uncertainty: Asymptotically exact formulations. *arXiv preprint arXiv:2204.04762*, 2022.

Tat-Jun Chin, Jin Yu, and David Suter. Accelerated hypothesis generation for multistructure data via preference analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4): 625–638, 2011.

Stéphane Crépey. Calibration of the local volatility in a trinomial tree using tikhonov regularization. *Inverse Problems*, 19(1):91, 2002.

Stéphane Crépey. Delta-hedging vega risk? *Quantitative Finance*, 4(5):559–579, 2004.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Emanuel Derman, Iraj Kani, and Joseph Z Zou. The local volatility surface: Unlocking the information in index option prices. *Financial Analysts Journal*, 52(4):25–36, 1996.

David L Donoho and Richard C Liu. The automatic robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988a.

David L Donoho and Richard C Liu. Pathologies of some minimum distance estimators. *The Annals of Statistics*, pages 587–608, 1988b.

Bruno Dupire et al. Pricing with a smile. *Risk*, 7(1):18–20, 1994.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 2022.

Joseph Diez Gergonne. Dissertation sur la recherche du milieu le plus probable. *Etc. Annales math. pures appl*, 12(6):181–204, 1821.

Frank R Hampel. *Contributions to the theory of robust estimation*. University of California, Berkeley, 1968.

Frank R Hampel. A general qualitative definition of robustness. *The annals of mathematical statistics*, 42(6):1887–1896, 1971.

 https://doi.org/10.52202/079017-1113

Bruce Hansen. *Econometrics*, chapter 19: Nonparametric Regression. Princeton University Press, 2022.

Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

Xuming He and Stephen Portnoy. Reweighted ls estimators converge at the same rate as the initial estimator. *The Annals of Statistics*, pages 2161–2167, 1992.

David C Hoaglin, Boris Iglewicz, and John W Tukey. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999, 1986.

Peter J Huber. Robust estimation of a location parameter. *Ann Math Stat*, 35:73–101, 1964.

John Hull and Sankarshan Basu. *27.3: The IVF Model*, page 659–660. Pearson, 11 edition, 2022.

María Jaenada, Pedro Miranda, and Leandro Pardo. Robust test statistics based on restricted minimum rényi's pseudodistance estimators. *Entropy*, 24(5):616, 2022.

Nan Jiang and Weijun Xie. Distributionally favorable optimization: A framework for data-driven decision-making with endogenous outliers. *SIAM Journal on Optimization*, 34(1):419–458, 2024.

Sumin Kang and Manish Bansal. Distributionally risk-receptive and risk-averse network interdiction problems with general ambiguity set. *Networks*, 81(1):3–22, 2023.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

Erich L Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020.

Jiajin Li, Sirui Lin, Jose Blanchet, and Viet Anh Nguyen. Tikhonov regularization is optimal transport robust under martingale constraints. In *Advances in Neural Information Processing Systems*, 2022.

Stewart Mayhew. Implied volatility. *Financial Analysts Journal*, 51(4):8–20, 1995.

P Warwick Millar. Robust estimation via minimum distance methods. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):73–89, 1981.

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.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein DRO. *Advances in Neural Information Processing Systems*, 36, 2023.

Chanseok Park, Ayanendranath Basu, and Srabashi Basu. Robust minimum distance inference based on combined distances. *Communications in Statistics-Simulation and Computation*, 24(3):653–673, 1995.

William C Parr and William R Schucany. Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75(371):616–624, 1980.

Benjamin Peirce. Criterion for the rejection of doubtful observations. *The Astronomical Journal*, 2: 161–163, 1852.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Peter J. Rousseeuw and Annick M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., 1987.

Mark Rubinstein. Implied binomial trees. *The Journal of Finance*, 49(3):771–818, 1994.

Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, 2015.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Robin Thompson. A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):53–55, 1985.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

John W Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. On linear optimization over wasserstein balls. *Mathematical Programming*, 195(1-2):1107–1122, 2022.

Luhao Zhang, Jincheng Yang, and Rui Gao. A simple and general duality proof for wasserstein distributionally robust optimization. *arXiv preprint arXiv:2205.00362*, 2022.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256–2283, 2022.

 https://doi.org/10.52202/079017-1113

Appendix

A Organization of the Appendix

We organize the appendix as follows:

- The organization of the Appendix is given in this section (Section A).
- Additional details on the connection between our estimator and prior work are given in Section B.1.
- Additional explanation of the long haul structure induced by our estimator and extensions to more general forms of problems are given in Appendix B.2.
- The proof of Theorem 2 and Theorem 4 are given in Section C.1.
- The long haul structure for linear regression with a concave cost function is provided in Section C.2.
- The detailed computational procedure for mean estimation, linear absolute regression, and more general cases can be found in Section D.
- The main experiments for mean estimation and LAD regression are described in Appendix E.1 and E.2 respectively. Additional details are given in the following appendices.
- Additional details on the illustrative example in the text is given in Section F.1.
- Additional details on mean estimation simulations are given in Section F.2.
- Additional plots for the mean estimation simulations for the concave cost function are given in Section F.3.
- Additional plots for the mean estimation simulations for the convex cost function are given in Section F.4.
- Additional details on LAD regression simulations are given in Section F.5.
- Additional plots for the LAD regression simulations for the concave cost function are given in Section F.6.
- Additional plots for the LAD regression simulations for the convex cost function are given in Section F.7.
- Additional option volatility surface experiment details are given in Section G.1.
- Additional information on the volatility surface data set and losses are given in Section G.2.
- Background and additional details on the deep learning volatility modeling experiment are given in Section G.3.

B Estimator Remarks

B.1 Connections to Prior Work

Additional details on connections to prior work are given in the following remark:

Remark B.1. (i) When we disregard the outer minimization concerning estimation parameters, the inner minimization problem over probability measures is related to the minimum distance functionals-based estimator, initially proposed by Donoho and Liu [1988a]. The estimator is obtained by first projecting the corrupted distribution $\mathbb{P}'_n$ onto a family of distributions $\mathcal{G}$ under a certain distribution discrepancy measure $\mathbb{D}$. Then, the optimal parameters are selected for the resulting distribution. However, even with additional information about our contamination models, this "two-stage" procedure has two major drawbacks: the difficulty in choosing an appropriate family of distributions $\mathcal{G}$ and the inherent computational challenge to project probability distributions. Moreover, the first stage (projection) is usually sensitive to the choice of family $\mathcal{G}$. In contrast, we propose a novel approach that integrates outlier rectification (i.e., explicit projection) and parameter estimation into a joint optimization framework in (3). (ii) This min-min strategy has also been explored in Jiang and Xie [2024] using an artificially constructed rectification set. Their primary focus is to utilize the min-min formulation to recover well-studied robust statistics estimators. In contrast, our focus is on a new conceptual framework for formulating novel estimators which have not been studied before.

B.2 Long Haul Structure and Extensions

An explanation of the long haul structure and extensions to more general forms are given in this subsection.

The proposed rectification set has the potential for broader applicability across various applications and problem domains. This versatility makes the rectification set a valuable tool for addressing and improving solutions in diverse problem settings in the future.

(i) (**Concave cost function**) From Proposition 1, the optimal rectification admits $z_{\text{best}} \in \arg\min_{z\in\mathcal{Z}} \ell(\theta, z') + \lambda c(z, z')$. Suppose we select a concave cost function c that grows strictly slower than the loss function ℓ. In this scenario, when the budget is small, the rectified data point z_{best} consistently exhibits the long haul structure, ensuring automatic outlier identification properties. This implies that the rectification process effectively identifies outliers, as the influence of the cost function dominates over the loss function for small budgets. To further illustrate the concept, we can consider a linear regression problem with the squared loss function $\ell(\theta, z) = (y - \theta^T x)^2$ or a nonlinear loss using the r-th norm as the cost function. We refer the readers to Appendix C.2.

(ii) (**Other distribution metric/discrepancy**) Based on the two examples discussed earlier, we can infer that the optimal transport distance-based rectification set is particularly suitable for regression or problems with continuous response variables. This is because the budget δ is used to compensate for the loss caused by identified outliers. However, in classification tasks, where the loss function may be less sensitive, the effectiveness of the optimal transport distance-based approach may be limited. To address this limitation and complement the rectification set for classification tasks, an alternative approach is to use ϕ-divergences. This approach has already been proposed in a study by Chen and Royset [2022], Antil et al. [2024] for handling outliers in image classification tasks. The ϕ-divergence-based rectification set operates by adjusting the weight assigned to outliers, effectively minimizing their impact on the overall data set. By reducing the weight placed on the detected outliers, the rectification set aims to remove their influence from the data set. This selective adjustment allows for the identification and removal of outliers, leading to a more refined and reliable data set.

https://doi.org/10.52202/079017-1113

C Proof Details

C.1 Proof of Theorem 2 and Theorem 4.

To start with, we give two crucial lemmas 5 and 6.

Lemma 5. Suppose that $a, b, \lambda > 0$ and $r \in (0, 1)$, we have

$$\min_{x \in [0, a/b]} a - bx + \lambda x^r = \min \left\{ a, \frac{\lambda a^r}{b^r} \right\}.$$

Proof. The argument is easy. The function $g(x) = a - bx + \lambda x^r$ is concave as the second derivative is always negative:

$$\nabla^2 g(x) = \lambda r(r-1)x^{r-2}$$

for all $x \geqslant 0$. □

Lemma 6. Suppose that there is an increasing sequence $0 \leqslant x_1 < x_2 < \cdots < x_n$. The optimal solution of

$$\max_{\lambda \geqslant 0} \sum_{i=1}^{n} \alpha_i \min\{x_i, \lambda x_i^r\} - \lambda\delta$$

is $\lambda^\star = x_k^{1-r}$ where

$$k := \max_{k \in [n]} \left\{ k : \sum_{i=k}^{n} \alpha_i x_k^r \geqslant \delta \right\}.$$

Here $\alpha_i \in [0, 1], \sum_{i=1}^{n} \alpha_i = 1$, $\delta > 0$ and $r \in (0, 1)$. Moreover, the optimal function value is

$$\max(\sum_{i=1}^{k-1} \alpha_i x_i + \left(1 - \frac{(\delta - \sum_{i=k+1}^{n} \alpha_i x_i^r)}{\alpha_k x_k^r}\right) \alpha_k x_k, 0).$$

Proof. Without loss of generality, we assume that $0 < x_1 < x_2 < \cdots < x_n$ as the zero part of the sequence does not affect the result.

First, given a fixed $\lambda \in \mathbb{R}+$ and the inequality $x_{t+1} < \lambda x_{t+1}^r$, our goal is to show that $x_t < \lambda x_t^r$. Let $x_t = \eta x_{t+1}$, where $\eta \in (0, 1]$. Then, we have

$$x_t = \eta x_{t+1} < \eta\lambda x_{t+1}^r = \frac{\lambda}{\eta^r}\eta x_t^r = \lambda\eta^{1-r} x_t^r \leqslant \lambda x_t^r.$$

Considering a fixed $\lambda \in \mathbb{R}_+$, we can express the objective function as:

$$\sum_{i=1}^{n} \alpha_i \min\{x_i, \lambda x_i^r\} - \lambda\delta = \sum_{i=1}^{k-1} \alpha_i x_i + \alpha_k \min\{x_k, \lambda x_k^r\} + \lambda \sum_{i=k+1}^{n} \alpha_i x_i^r - \lambda\delta$$

where $1 \leqslant k \leqslant n$ and α_i are weights associated with each x_i. Two cases arise:

1. For all $1 \leqslant t \leqslant k-1$, $x_t < \lambda x_t^r$, or for all $k+1 \leqslant t \leqslant n$, $x_t > \lambda x_t^r$.
2. At the k-th point, we have $x_k = \lambda x_k^r$.

Since we are dealing with a concave piecewise linear function, the optimal λ will be a knot point (case 2). Otherwise, modifying λ would increase the objective value. Therefore, we can establish the first-order optimality condition as follows:

$$\sum_{i=k+1}^{n} \alpha_i x_i^r + \mu\alpha_k x_k^r = \delta.$$

where k and $\mu \in [0, 1]$ are determined through a search using the quick-select algorithm. Consequently, the optimal solution is obtained as $\lambda^* = x_k^{1-r}$ and

$$k = \max_{k \in [n]} \left\{ \sum_{i=k}^{n} \alpha_i x_i^r \geqslant \delta \right\}.$$

Thus, we further get $\mu = \frac{(\delta - \sum_{i=k+1}^{n} \alpha_i x_i^r)}{\alpha_k x_k^r}$ and the optimal function value admits

$$\begin{aligned}
&\sum_{i=1}^{k-1} \alpha_i x_i + \alpha_k \min\{x_k, \lambda x_k^r\} + \lambda \sum_{i=k+1}^{n} \alpha_i x_i^r - \lambda\delta \\
&= \sum_{i=1}^{k-1} \alpha_i x_i + \alpha_k \min\{x_k, \lambda x_k^r\} - \lambda\mu\alpha_k x_k^r \\
&= \sum_{i=1}^{k-1} \alpha_i x_i + (1-\mu)\alpha_k x_k \\
&= \sum_{i=1}^{k-1} \alpha_i x_i + \left(1 - \frac{(\delta - \sum_{i=k+1}^{n} \alpha_i x_i^r)}{\alpha_k x_k^r}\right) \alpha_k x_k
\end{aligned}$$

Based on our discussion, we exclude the corner case where $k = 1$. Therefore, δ is sufficiently large such that $\sum_{i=1}^{n} \alpha_i x_i^r \leqslant \delta$. It is evident that in this trivial scenario, the optimal function value is zero.

We conclude our proof. □

We now give the proof of Theorem 2:

Proof. Before we prove the theorem, it is worth highlighting that for any fixed θ, we can always sort $\{z_i\}_{i=1}^n$ based on the error $\|\theta - z_i'\|$ to satisfy condition (6).

By the strong duality result in Proposition 1, we have

$$\begin{aligned}
&\min_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}_n')} \mathbb{E}_{\mathbb{Q}}[\|\theta - Z'\|] \\
&= \max_{\lambda \geqslant 0} \mathbb{E}_{\mathbb{P}_n'} \left[\min_{\Delta \in \mathbb{R}} \|\theta - Z' - \Delta\| + \lambda \|\Delta\|^r \right] - \lambda\delta \\
&= \max_{\lambda \geqslant 0} \mathbb{E}_{\mathbb{P}_n'} \left[\min_{0 \leqslant \|\Delta\| \leqslant \|\theta - Z'\|} \|\theta - Z'\| - \|\Delta\| + \lambda \|\Delta\|^r \right] - \lambda\delta \\
&= \max_{\lambda \geqslant 0} \mathbb{E}_{\mathbb{P}_n'} [\min \{\|\theta - Z'\|, \lambda \|\theta - Z'\|^r\}] - \lambda\delta \\
&= \max_{\lambda \geqslant 0} \frac{1}{n} \sum_{i=1}^{n} \min\{\|\theta - z_i'\|, \lambda \|\theta - z_i'\|^r\} - \lambda\delta \\
&= \max \left(\frac{1}{n} \sum_{i=1}^{k(\theta)-1} \|\theta - z_i'\| + \frac{1}{n} \left(1 - \frac{n\delta - \sum_{i=k(\theta)+1}^{n} \|\theta - z_i'\|^r}{\|\theta - z_{k(\theta)}'\|^r}\right) \|\theta - z_{k(\theta)}'\|, 0 \right)
\end{aligned}$$

where the third equality follows from Lemma 5 in Appendix C.1 and the fifth one is due to Lemma 6 in Appendix C.1. □

The proof in the LAD regression case follows:

Proof. The proof follows a similar idea to that of Theorem 2. For any fixed θ, we can always sort $\{z_i\}_{i=1}^n$ based on the error $\|y_i' - \theta^T x_i'\|$ to satisfy the condition. For simplicity, we denote $\tilde{\theta} = (\theta, -1)$

https://doi.org/10.52202/079017-1113

and $z = (x, y)$. By the strong duality result in Proposition 1, we have

$$
\begin{aligned}
&\min_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[\|Y - \theta^T X\|] \\
&= \max_{\lambda\geqslant 0} \mathbb{E}_{\mathbb{P}'_n}\left[\min_{\Delta\in\mathbb{R}^{d+1}} \|\tilde{\theta}^T Z' + \tilde{\theta}^T\Delta\| + \lambda\|\Delta\|^r\right] - \lambda\delta \\
&\overset{(a)}{=} \max_{\lambda\geqslant 0} \mathbb{E}_{\mathbb{P}'_n}\left[\min_{\|\Delta\|\leqslant \frac{\|\tilde{\theta}^T Z'\|}{\|\tilde{\theta}\|_*}} \|\tilde{\theta}^T Z'\| - \|\tilde{\theta}\|_*\|\Delta\| + \lambda\|\Delta\|^r\right] - \lambda\delta \\
&\overset{(b)}{=} \max_{\lambda\geqslant 0} \mathbb{E}_{\mathbb{P}'_n}\left[\min\left\{\|\tilde{\theta}^T Z'\|, \frac{\lambda\|\tilde{\theta}^T Z'\|^r}{\|\tilde{\theta}\|_*^r}\right\}\right] - \lambda\delta \\
&= \max_{\lambda\geqslant 0} \frac{1}{n}\sum_{i=1}^n \min\left\{\|\tilde{\theta}^T z'_i\|, \frac{\lambda\|\tilde{\theta}^T z'_i\|^r}{\|\tilde{\theta}\|_*^r}\right\} - \lambda\delta \\
&\overset{(c)}{=} \max\left(\frac{1}{n}\sum_{i=1}^{k(\theta)-1} \|\tilde{\theta}^T z'_i\| + \frac{1}{n}\left(1 - \frac{n\delta' - \sum_{i=k(\theta)+1}^n \|\tilde{\theta}^T z'_i\|^r}{\|\tilde{\theta}^T z'_{k(\theta)}\|^r}\right)\|\tilde{\theta}^T z'_{k(\theta)}\|, 0\right),
\end{aligned}
$$

where the third equality follows from the Holder inequality and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$; the fourth one follows from Lemma 5 and the sixth one is due to Lemma 6.

This completes our proof.

□

C.2 Linear Regression with Concave Cost Function

In this subsection, we aim to illustrate an example as mentioned in Section 4.2. Suppose we choose a concave cost function c that grows strictly slower than the loss function f. In such a scenario, when the budget is limited, the rectified data point z_{best} consistently demonstrates the long haul structure, thereby ensuring automatic outlier identification properties.

Example C.1. Suppose that $\mathcal{Z} = \mathbb{R}^{d+1}$, $\ell(\theta, z) = \|y - \theta^T x\|^2$ and the cost function is defined as $c(z, z') = \|z - z\|^r$ where $r = \frac{1}{2}$. We want to study the best rectification distribution of

$$\min_{\mathbb{Q}\in\mathcal{R}(\mathbb{P}'_n)} \mathbb{E}_{\mathbb{Q}}[\|y - \theta^T x\|^2].$$

As we shall see when the dual variable λ is relatively small, the best rectification distribution will keep the long haul structure although the inner minimization problem is no longer concave.

For simplicity, we denote $\tilde{\theta} = (\theta, -1)$ and $z = (x, y)$. By the strong duality result in Proposition 1, we only have to focus on the inner minimization problem:

$$
\begin{aligned}
&\min_{\Delta\in\mathbb{R}^{d+1}} \left\{(\tilde{\theta}^T(Z' + \Delta))^2 + \lambda\|\Delta\|^r\right\} \\
&= \min_{\|\Delta\|\leqslant \frac{|\tilde{\theta}^T Z'|}{\|\tilde{\theta}\|_*}} \left\{(|\tilde{\theta}^T Z'| - \|\tilde{\theta}\|_*\|\Delta\|)^2 + \lambda\|\Delta\|^r\right\}.
\end{aligned}
$$

Next, we aim at clarifying the structured information of the following one-dimensional optimization problem:

$$\min_{\|\Delta\|\in\left[0, \frac{|\tilde{\theta}^T Z'|}{\|\tilde{\theta}\|_*}\right]} K(\|\Delta\|) := (|\tilde{\theta}^T Z'| - \|\tilde{\theta}\|_*\|\Delta\|)^2 + \lambda\|\Delta\|^r.$$

In general, unlike the case of absolute loss, the function $K(\cdot)$ will not be concave. However, the optimal solution of the resulting optimization problem will be active at the boundary when δ is small and the optimal value of λ is sufficiently large. Initially, let's overlook the constraint and express the

first-order optimality condition.

$$
\begin{aligned}
&2\|\tilde{\theta}\|_*(\|\tilde{\theta}\|_*\|\Delta\| - |\tilde{\theta}^T Z'|) + \frac{1}{2}\lambda\|\Delta\|^{-\frac{1}{2}} = 0 \\
\Rightarrow &\|\tilde{\theta}\|_*^2\|\Delta\| + \frac{\lambda}{4}\|\Delta\|^{-\frac{1}{2}} = |\tilde{\theta}^T Z'|\|\tilde{\theta}\|_* \\
\Rightarrow &\|\tilde{\theta}\|_*^2\|\Delta\|^{\frac{3}{2}} + \frac{\lambda}{4} = |\tilde{\theta}^T Z'|\|\tilde{\theta}\|_*\|\Delta\|^{\frac{1}{2}}.
\end{aligned}
$$

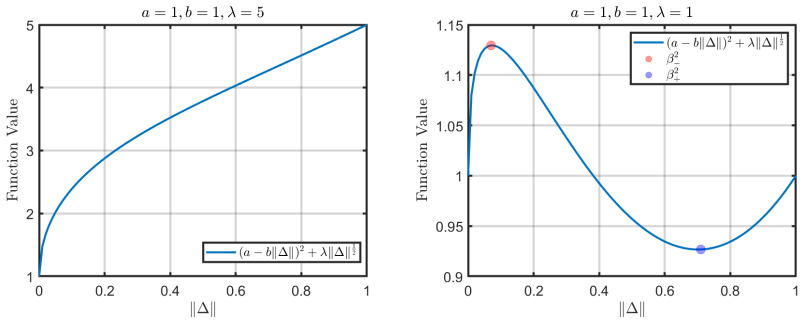
By changing the variable $\beta = \|\Delta\|^{\frac{1}{2}}$, we have

$$
\|\tilde{\theta}\|_*^2\beta^3 + \frac{\lambda}{4} - |\tilde{\theta}^T Z'|\|\tilde{\theta}\|_*\beta = 0.
$$

Now, we observe that $g(\beta) = \|\tilde{\theta}\|_*^2\beta^3 + \frac{\lambda}{4} - |\tilde{\theta}^T Z'|\|\tilde{\theta}\|_*\beta$, and our objective is to find the positive root of this cubic equation in one dimension. Also, the first derivative is $\nabla g(\beta) = 3\|\tilde{\theta}\|_*^2\beta^2 - |\tilde{\theta}^T Z'|\|\tilde{\theta}\|_*$. The stationary point is $\beta^\star = \sqrt{\frac{|\tilde{\theta}^T Z'|}{3\|\tilde{\theta}\|_*}}$ and then the corresponding $\|\Delta^\star\| = \frac{|\tilde{\theta}^T Z'|}{3\|\tilde{\theta}\|_*} \in [0, \frac{|\tilde{\theta}^T Z'|}{\|\tilde{\theta}\|_*}]$.

$$
g(\beta^\star) = \left(\sqrt{\frac{|\tilde{\theta}^T Z'|}{3\|\tilde{\theta}\|_*}}\right)^3 \|\tilde{\theta}\|_*^2 - |\tilde{\theta}^T Z'|\|\tilde{\theta}\|_*\sqrt{\frac{|\tilde{\theta}^T Z'|}{3\|\tilde{\theta}\|_*}} + \frac{\lambda}{4}.
$$

1. When $\lambda \geqslant 4\left(|\tilde{\theta}^T Z'|\|\tilde{\theta}\|_*\sqrt{\frac{|\tilde{\theta}^T Z'|}{3\|\tilde{\theta}\|_*}} - \left(\sqrt{\frac{|\tilde{\theta}^T Z'|}{3\|\tilde{\theta}\|_*}}\right)^3\right)$ (i.e., δ is sufficiently small), we have $g(\beta^\star) \geqslant 0$. As such, the critical point of the unconstrained optimization problem is not in the interval $[0, \frac{|\tilde{\theta}^T Z'|}{\|\tilde{\theta}\|_*}]$. In other words, we can conclude the optimal solution will be 0 and the solution of vanilla least square is already optimal.
2. When $g(\beta^\star) < 0$, we know there are two solutions β_+, β_- for $g(\beta) = 0$ where $\beta_- < \beta_+$. We know $K(\|\Delta\|)$ will be increasing between $[0, \beta_-^2]$ and decreasing between $[\beta_-^2, \beta_+^2]$. Thus, the optimal solution will be either 0 or β_+^2 and ensures the long haul transportation structure. Different from the absolute loss, the cost function $\|z - z'\|^{1/2}$ is not powerful enough to move any points that achieve a perfect fit to the current hyperplane.



(a) Case 1 (b) Case 2

Figure 4: Visualization

https://doi.org/10.52202/079017-1113

D Computational Procedure

D.1 Alternative Estimation Procedure for Inner Problem

We make the following remark on an alternative procedure for estimating $k(\theta_t)$ and λ in the inner problem for mean estimation and LAD regression in Algorithm 1:

Remark D.1 (Alternative procedure for $k(\theta_t)$ and λ^*). It is illuminating to note that steps 1 and 2 in Algorithm 1 can be alternatively replaced with steps that solve a specific linear program for the knot point $k(\theta_t)$ and optimal λ^*. Instead of applying the sorting and quick-select algorithm implied by Lemma 6, we can instead reformulate Equation (8) as a linear programming problem, i.e.,

$$\begin{array}{ll} \max\limits_{\lambda \geqslant 0, t \in \mathbb{R}^n} & \frac{1}{n}\sum_{i=1}^n t_i - \lambda\delta \\ \text{s.t.} & t_i \leqslant \|\tilde{\theta}^T z_i'\|, \forall i \in [n] \\ & \|\tilde{\theta}\|_*^r t_i \leqslant \lambda \|\tilde{\theta}^T z_i'\|^r, \forall i \in [n]. \end{array} \tag{9}$$

The problem in (9) can then be solved instead of performing steps 1 and 2 of Algorithm 1 to find the knot point $k(\theta_t)$ and the optimal λ^*.

D.2 General Computational Procedure

We also propose a procedure for fitting more general regression models with our statistically robust estimator, which is based on the approach for LAD regression above.

Although the procedure can be applied to any model which can be estimated via subgradient methods, we focus on deep learning models. We start by considering a neural network f_θ parameterized by weights θ, and a loss function $\ell(\theta, z) = \ell(y, f_\theta(x))$, where ℓ is a gradient Lipschitz function with a constant L. By Proposition 1, we would like to solve the inner problem

$$\max_{\lambda \geqslant 0} \mathbb{E}_{\mathbb{P}_n'}\left[\min_{z \in \mathcal{Z}} \ell(\theta; z) + \lambda c(z, Z')\right] - \lambda\delta. \tag{10}$$

Although we do not prove a reformulation result for this problem, we solve the problem empirically by employing a computational procedure which is analogous to that of LAD regression. The procedure is heuristic and does not enjoy the same guarantees as Algorithm 1 for LAD regression, but in experiments (see Section 5.2.2), we find that the performance is similarly robust to outliers and significantly outperforms benchmark models trained under empirical risk minimization. Our computational procedure follows in Algorithm 2.

Algorithm 2: Statistically Robust Optimization Procedure

Data: Observed data $\{z_i'\}_{i=1}^n$, initial point $\theta^{(0)}$, stepsizes $\alpha^{(t)} > 0$, sampling distribution $\mathbb{P}_n'$; batch size $m \leqslant n$.

for $t = 0, \ldots, T$ **do**

1. **Sample** m points $\{z_i'\}_{i=1}^m \sim \mathbb{P}_n'$.
2. **Sort** the observed data $\{z_i'\}_{i=1}^m$ via the value $\ell(\theta^{(t)}, z_i')$.
3. **Quick-Select** algorithm to get the knot point $k(\theta^{(t)})$ and the optimal $\lambda^\star$.
4. **Subgradient step** on the detected clean data:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \sum_{i=1}^{k(\theta^{(t)})} v^*(\theta^{(t)}, z_i')$$

where $v^*(\theta^{(t)}, z_i') \in \partial_\theta \ell(\theta^{(t)}, z_i')$.

end

Algorithm 2 is similar to Algorithm 1, but has two principal differences: (1) instead of optimizing over the entire data set, we optimize over mini-batches of size $m \leqslant n$ which are drawn from a data set of size n via a sampling distribution $\mathbb{P}_n'$; and (2) instead of taking the subgradient with respect to θ for the LAD regression problem in calculating the direction of descent, we instead take a local subgradient with respect to θ around each point z_i', each of which is contained within the subdifferetial $\partial_\theta \ell(\theta^{(t)}, z_i')$. The subgradient is in practice computed by the automatic differentiation capability of a software package such as Pytorch, Tensorflow, or JAX. The budget parameter δ, which controls the quantile which identifies outliers, is tuned via cross-validation.

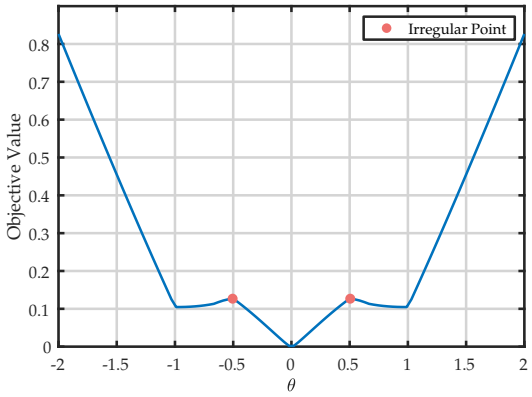


Figure 5: Irregular objective function.

D.3 Optimization Details

Solving the optimization problems in the prior sections can be complicated, as in general, the minimization of a class of convex functions often leads to non-convex problems.

This issue becomes even more critical when dealing with our estimation problem, as these problems may lack weak convexity or subdifferential regularity properties. One significant challenge in such problems is the non-coincidence of different subdifferential concepts, such as Clarke, Fréchet, and limiting subdifferentials. Moreover, some calculus rules (summation and chain rules) do not always hold for these concepts. Thus, computing a first-order oracle, which provides necessary information for optimization algorithms, can be difficult in practice.

To concretely illustrate this challenge, we note that, intuitively, the graph of a subdifferentially regular function cannot exhibit "downward-facing cusps" [Li et al., 2020, Davis et al., 2020]. Here, we give an example of when the objective function (7) is not regular, which may occur in general.

Example D.1 (Irregular Case (Three Point Masses))**.** Let $\mathbb{P}'_n = \frac{1}{3}\delta_{-1} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_1$ and $\delta = 0.7$. Based on the reformulation result given in Theorem 2, we can plot the curve of the loss function from our proposed estimator, see Figure 5.

From Figure 5, it is evident that there are at least two regions with "downward-facing cusps" (marked as irregularities in the figures), which could potentially pose computational challenges. Similar challenges arise in the training of deep neural networks, as demonstrated in Figure 4 of [Li et al., 2020], particularly for two-layer neural networks. However, the "subgradient" method remains effective for addressing these challenges empirically. In our empirical investigation, we have observed a similar phenomenon for our estimator.

https://doi.org/10.52202/079017-1113

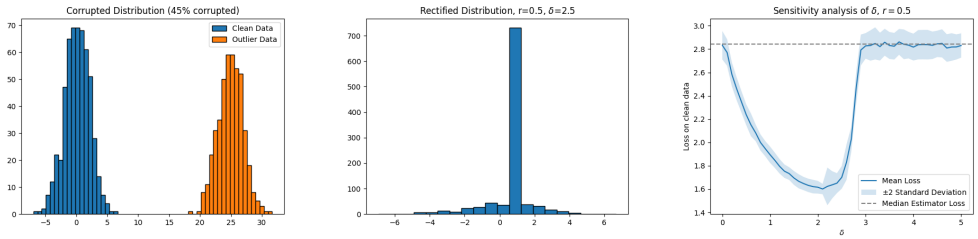
E Mean Estimation and LAD Regression Experiments

In this section, we outline the auxiliary results of the mean estimation experiments and the primary results of the LAD regression experiments. Additional details are deferred to the following appendix section (Appendix F) for the convenience of readers.

E.1 Mean Estimation

We generate the corrupted data by combining two Gaussian distributions: $(1 - \text{corruption level}) \times \mathcal{N}(0, 2)$ and $(\text{corruption level}) \times \mathcal{N}(25, 2)$. For each corruption level in Table 1, 100 random trials are performed for all results presented. We compare our estimator with several widely used baseline methods. Interestingly, we observe that even when setting hyperparameters as constants across all corruption levels, our estimator consistently outperforms the others. Notably, the trimmed mean, which incorporates the ground-truth corruption level, performs even worse than our estimator. Our estimator's outperformance is notable, as in this table we report results specifically using $\delta = 0.5$, which is a nonoptimal choice for δ, as we will show in the sensitivity analysis in Appendix E.1.

We visualize the corrupted distribution and its rectified distribution when the corruption level is 45% in Figure 6. In this case, 55% of the data is drawn from the true distribution $\mathcal{N}(0, 2)$ and 45% is drawn from $\mathcal{N}(25, 2)$. Figure 6(a) displays the original sample of data from the contaminated distribution. The outlier points from the contaminating distribution are clearly visible in orange. Figure 6(b) shows the rectified distribution our estimator produces for $\delta = 2.5$, in which the outlier points have been moved from their original values to their rectified values, which is much closer to the true mean of the clean distribution. Our estimator thus successfully identifies the majority of outliers and relocates them towards the center (our mean estimator), providing further support for our theoretical findings in the previous section. The sensitivity analysis of mean estimation with respect to δ is displayed in Figure 6(c). In this figure, the loss on the clean data is plotted for various values of δ. We see that an approximate minimum occurs at $\delta = 2$, with good performance within the range $\delta \in [0.5, 2.5]$. This illustrates the relative insensitivity of our estimator to different values of δ for a wide and reasonable range. This range can be easily reached by hyperparameter tuning via cross-validation, as the function describing the performance of our estimator is approximately quasiconvex. Moreover, when $\delta = 0$ or when δ is set so large that it can rectify all points, our estimator gracefully degrades to the loss of the median estimator, which is another favorable property. A further sensitivity analysis with respect to r is given in Appendix F.2. In this sensitivity analysis, we find a similarly wide region of good performance for r which improves significantly and almost uniformly on the more typical setting of $r = 1$.



(a) (b) (c)

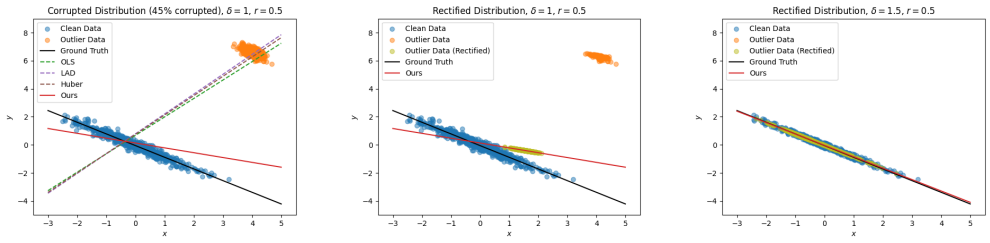
Figure 6: (a): Visualization of contamination model: a mixture of Gaussian $0.55 \times \mathcal{N}(0, 2) + 0.45 \times \mathcal{N}(25, 2)$; (b): The rectified data generated by the proposed statistically robust estimator; (c) The sensitivity analysis of δ. The visualization of (a) and (b) shows that the outlier points are rectified into the bulk of the clean distribution at $\delta = 2.5$. The sensitivity analysis of δ shows that this rectification occurs for a wide range of δ which leads to losses for mean estimation which are lower than that of the median estimator of location. As $\delta \to 0$ and as δ increases, the performance of our estimator gracefully degrades to that the median estimator.

The evolution of the rectified distribution for mean estimation under concave and convex cost functions as δ increases is depicted in Appendices F.3 and F.4. These depictions show that our statistically robust estimator which applies the concave transport cost function acts in a stable and

expected manner as δ increases. In particular, this depiction shows that our estimator moves points in the order of their distance from the estimate of the mean; that is, the budget is "used" on the worst outliers first. This results in an orderly procedure of rectification which is stable and predictable across nearby values of δ, which is a favorable property if δ is being set sequentially or in cross-validation. In contrast, the estimator which applies the convex transport cost function moves all points of the clean and outlier distributions toward each other, which results in poor estimates of the true mean and an improper rectified distribution.

E.2 LAD Regression

As we discussed in the previous result, the theoretical analysis for the LAD estimator resembles that of the mean estimation task. Empirically, we also observe similar performance for the LAD estimator and show that our estimator correctly rectifies most of outliers to the fitted hyperplane under a range of choices for δ. In order to further support the effectiveness of our estimator, we provide additional visualizations: Figure 7 contains a visualization of the lines of best fit produced by various estimations to the LAD regression problem, and displays the effect of different choices of δ on the line of best fit produced by our estimator. In Figure 7(a) our estimator with $\delta = 1$ successfully produces a line of best fit which is closest to the uncontaminated distribution, while the other estimators (OLS, LAD, and Huber regressors) produce lines which are heavily affected by the contaminating distribution, showcasing the robustness of our estimator. Figure 7(b) and 7(c) show the rectified distribution produced by our estimator under $\delta = 1$ (b) and $\delta = 1.5$ (c). The suboptimal value of $\delta = 1$ still rectifies many points from the contaminated distribution and produces a good line of best fit. Setting $\delta = 1.5$ rectifies all of the points from the contaminated distribution and essentially recovers the true line of best fit. Importantly, this simulation shows that even improperly setting δ to the suboptimal value of $\delta = 1$ produces a much better line than any of the other estimators in Figure 7(a). Appendix F.5 contains comparisons over different random trials and a sensitivity analysis with respect to δ and r, along with experimental results.



(a) (b) (c)

Figure 7: (a): Visualization of contamination model: 55% are drawn with $x \sim \mathcal{N}(0, 2)$ and $y = 0.1 - 0.8x + 0.2w$. 45% are drawn with $x \sim \mathcal{N}(4, 2)$ and $y = 10.1 - 0.8x + 0.2w$. In each case $w \sim N(0, 1)$. Fitted models for different baselines are: OLS (ordinary least square), LAD (least absolute deviation regression), Huber (Huber regression with threshold parameter 1.5); (b): The rectified data generated by the proposed statistically robust estimator with a small budget $\delta = 1$; (c): The rectified data with a larger $\delta = 1.5$ is able to rectify all outliers.

https://doi.org/10.52202/079017-1113

F Additional Mean Estimation and LAD Regression Experimental Details

F.1 Illustrative Example Details

In this example, we perform linear regression using our estimator with either a concave or a convex transport cost function. The estimation is performed on a sample of points in $\mathbb{R}^2$ drawn from a contaminated distribution (corresponding to $\mathbb{P}'$) where 45% of the points are distributed according to an outlier distribution and 55% of the points are distributed according to the true uncontaminated distribution (corresponding to $\mathbb{P}_\star$). Our goal is to recover the line of best fit which corresponds to the uncontaminated distribution, which is $y = 0.1 - 0.8x$. In other words, 55% of the points are drawn with $x \sim \mathcal{N}(0, 2)$ and $y = 0.1 - 0.8x + 0.2w$, and 45% of the points are drawn with $x \sim \mathcal{N}(4, 2)$ and $y = 10.1 - 0.8x + 0.2w$. We illustrate the considerable differences which can occur when using each of these cost functions under various choices of δ, the rectification budget parameter. We depict this example in Figure 2 for the concave cost and Figure 3 for the convex cost.

The behavior of the estimator using the concave cost function is depicted in Figures 2(a)-(c). Each of these subfigures shows the points drawn from each distribution, the rectified distribution, the true line of best fit for the uncontaminated distribution, and the line of best fit produced by our estimator, at various choices of δ. In Figure 2(a), $\delta = 0$, in Figure 2(b), $\delta = 0.9$, and in Figure 2(c), $\delta = 1.5$. As can be seen from the figures, as the budget δ increases, we note that the proposed estimator using the concave cost function becomes increasingly adept at mitigating the influence of outlier data. It achieves this by rectifying the outlying points, moving them into the bulk of the uncontaminated distribution. As a result, the line of best fit (depicted in red) aligns much more closely with the true line of best fit (depicted in black). It is able to perform this successful estimation because each outlier point can be moved far with a cheap cost (a "long haul") due to the use of the concave transport cost function.

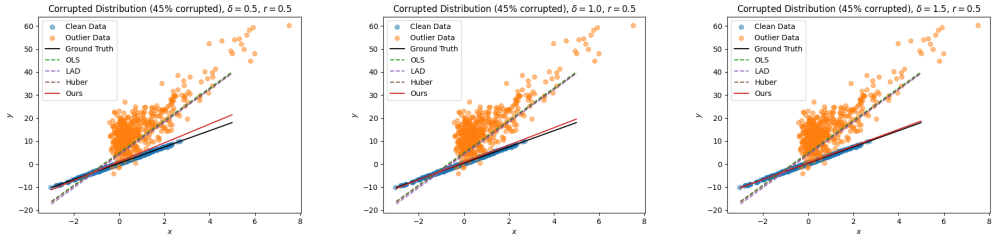
However, the convex cost function may result in suboptimal estimators, which can only move every data point (clean and contaminated) a little bit. Figures 3(a)-(c) display the behavior of the estimator using the convex cost function, which illustrates this effect. Each subfigure displays the rectified distribution as yellow-green points for a different setting of δ (i.e. (a) $\delta = 0$, (b) $\delta = 0.9$, and (c) $\delta = 1.5$). As can be seen from the evolution of the rectified distribution, as δ increases, instead of the outlier points being moved towards the bulk of the distribution, all of the points move towards each other–even the points in the clean distribution. This is not ideal behavior, as the points from the clean distribution should stay in place. This defective behavior causes the line of best fit produced by the estimator using the convex cost function (which is depicted in red) fit to the rectified points to be severely affected by the outliers; instead of being close to the ground truth line of best fit (which is depicted in black), its slope is moved significantly upwards towards the outliers. This occurs because the convex cost function gives lower cost to smaller rectifications of the given data set, which is an inappropriate assumption for data sets containing outliers.

Appendices F.6 and F.7 depict the evolution of the rectified distribution as δ increases for the concave and convex cost functions, respectively. For our estimator which uses the concave cost, the evolution of the rectified distribution as δ increases shows that our estimator increasingly rectifies the points in order of their distance from the line of best fit in a predictable and stable way, which is a favorable property for setting δ via hyperparameter tuning. In contrast, the convex estimator consistently moves all points (clean and corrupted) toward each other as δ changes, which achieves a poor estimate of the true line of best fit for all δ considered.

F.1.1 Non-clustered Outlier Illustrative Example

In this section, we illustrate the performance of an additional experiment under a different model for outliers. Like in the prior section, our goal is to recover the line of best fit which corresponds to the uncontaminated distribution. In this setting, we add additional terms to our data-generating process which cause heteroskedasticity and outlying points in the outlier data. In this setting, points are drawn from the clean distribution and contaminated distribution with the same probabilities (55% and 45%). We generate new random values for the data-generating process: clean points are drawn with $x \sim \mathcal{N}(0, 1)$ and $y = 0.4 + 3.5x + 0.2w$, where $w \sim \mathcal{N}(0, 1)$. Contaminated points are drawn according to $\tilde{x} \sim \mathcal{N}(0, 1)$ and $x = \tilde{x} + \min(\tilde{x}^2, 20) + w$ and $y = 10.4 + 4.5x + 5w$. We illustrate the estimated lines of best fit under various choices of δ, the rectification budget parameter, in Figure 8 for the concave cost, showing that our estimator still performs well against non-clustered, heteroskedastic

outliers. As the value of the budget parameter δ increases, the line of best fit estimated by our method approaches the true line of best fit. Notably, our estimator is not entirely corrupted by the "outliers in the outliers", unlike the other benchmark methods.



(a) (b) (c)

Figure 8: Visualization of the performance of our approach on non-clustered, heteroskedastic outliers, with values of the budget parameter (a) $\delta = 0.5$, (b) $\delta = 1.0$, and (c) $\delta = 1.5$.

F.2 Mean Estimation Details

The experimental details for mean estimation follow. For each (δ, r) point in our experiments, we perform 100 random trials at 100 fixed seeds. In each trial, we run our optimization procedure for with a learning rate of 10^{-2}. We stop when the number of iterations reaches 2000 or the change in the loss function between successive iterations is below a tolerance of 10^{-6}. We initialize θ to the median of the data set.

Below in Figure 9, we visualize the sensitivity analysis of r at $\delta = 1$ when the corruption level is 45%. As shown, our concave transport cost function improves significantly on the linear cost function ($r = 1$) and has a wide region of favorable stable performance.

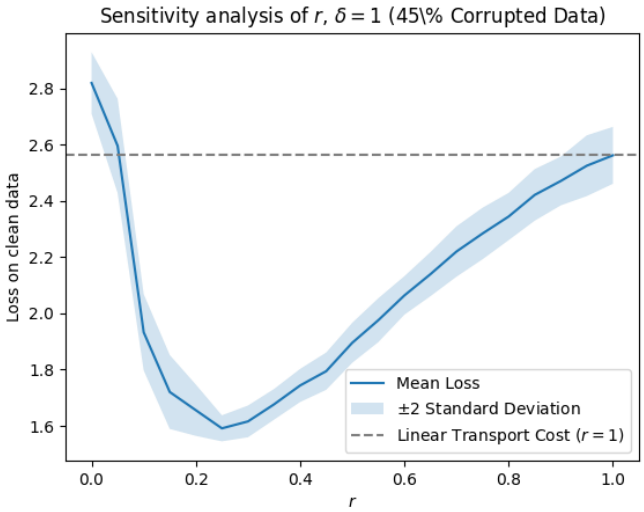


Figure 9: The sensitivity analysis of the loss on clean data with respect to r of our estimator at $\delta = 1$ on data with a 45% corruption level on the mean estimation task.

https://doi.org/10.52202/079017-1113

F.3 Concave Cost Mean Estimation Simulation

In Figure 10 below, we plot the evolution of the rectified distribution produced by our estimator under various values of δ for the **concave** cost function with $r = 0.5$.

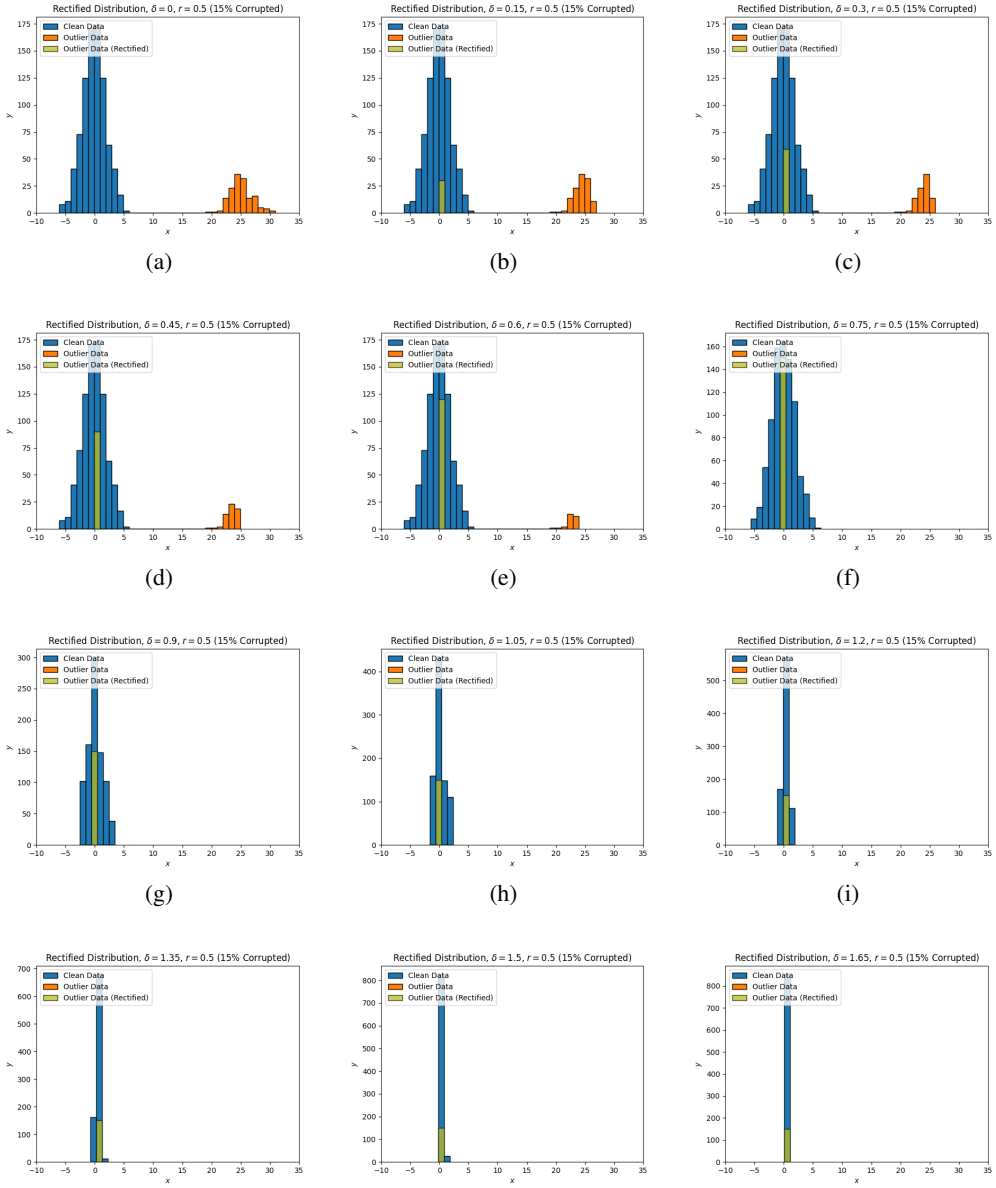
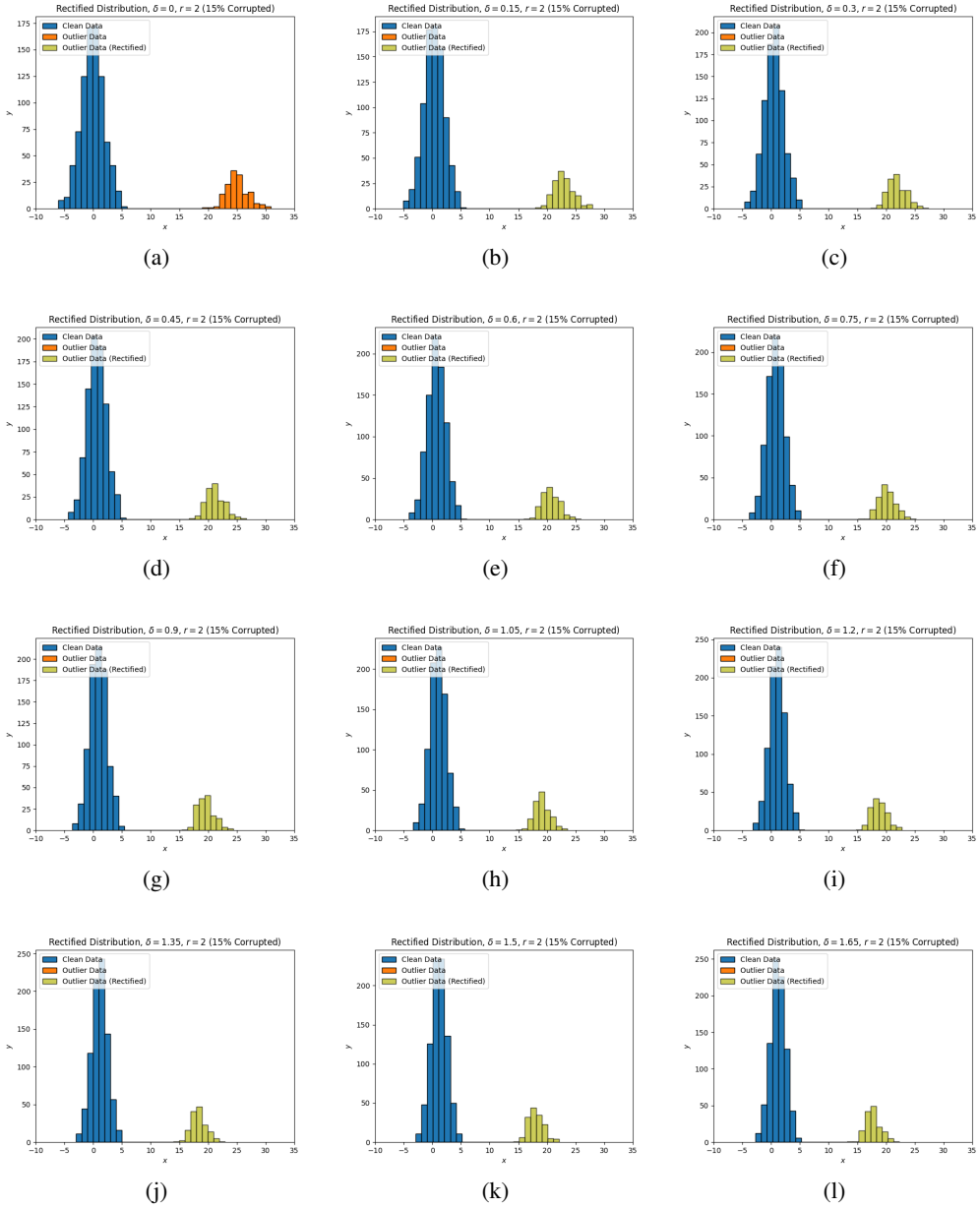


Figure 10: Visualization of the evolution of the rectified distribution.

F.4 Convex Cost Mean Estimation Simulation

In the Figure 11 below, we plot the evolution of the rectified distribution produced by our estimator under various values of δ for the **convex** cost function with $r = 2.0$.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 11: Visualization of the evolution of the rectified distribution.

https://doi.org/10.52202/079017-1113

F.5 Least Absolute Regression Details

We provide additional results on the least absolute deviation simulations below. In Table 4 we report the results of our experiment. We take the same experimental setting as in the regression example from the main text.

For each (δ, r) point in our experiments, we perform 100 random trials at 100 fixed seeds. In each trial, we run our optimization procedure for with a learning rate of 10^{-2} starting from 10 randomly initialized points and take the best loss. In each problem, we stop optimization when the number of iterations reaches 1000 or the change in the loss function between successive iterations is below a tolerance of 10^{-6}. We initialize the slope β according to a $N(0, 1)$ distribution and the bias to zero.

The results are included in Table 4. As can be seen, our estimator outperforms all competing estimators, and attains significant outperformance for most corruption levels.

Table 4: We compared our estimator with several standard regression methods by evaluating the average loss on clean data points across various corruption levels. Mean performance is accompanied by a 95% confidence interval over 100 random trials. In our evaluation, we set the threshold parameter of Huber regression to 1.5. The hyperparameters for our estimator, namely $\delta = 1.5$ and $r = 0.5$, remained constant across all corrupted levels. Error bars represent two standard deviation confidence intervals we have computed manually assuming normal errors.

Corruption Level	20%	30%	40%	45%	49%
OLS	1.569 ± 0.041	1.702 ± 0.043	1.773 ± 0.049	1.803 ± 0.052	1.822 ± 0.054
LAD	1.808 ± 0.131	1.875 ± 0.055	1.892 ± 0.059	1.903 ± 0.061	1.908 ± 0.062
Huber	1.642 ± 0.042	1.776 ± 0.045	1.842 ± 0.053	1.868 ± 0.056	1.882 ± 0.059
Ours	$\mathbf{0.657 \pm 0.638}$	$\mathbf{0.529 \pm 0.430}$	$\mathbf{0.680 \pm 0.467}$	$\mathbf{0.802 \pm 0.597}$	$\mathbf{0.866 \pm 0.619}$

In Figure 12 and Figure 13 below, we plot the sensitivity of our estimator across different values of δ and r. We see that there is a reasonable basin of good performance across both values. Each loss curve approaches the error of benchmark estimator or estimator with a traditional cost function as we let δ and r tend toward values which recover these approaches.

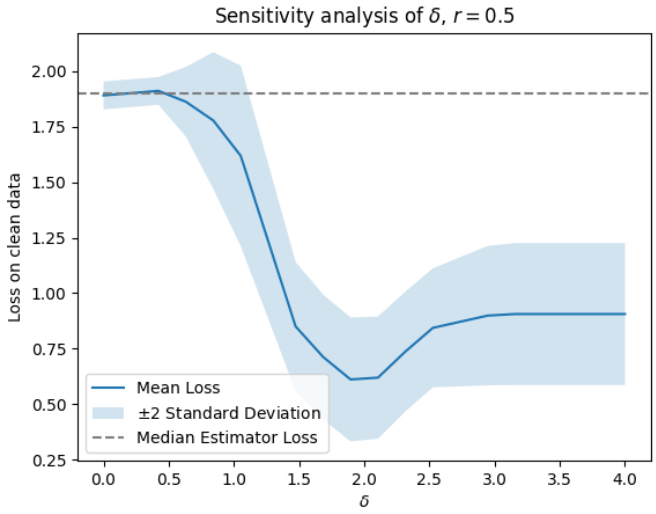


Figure 12: The sensitivity analysis of the loss on clean data with respect to δ of our estimator at $r = 0.5$ on data with a 45% corruption level on the least absolute regression task.

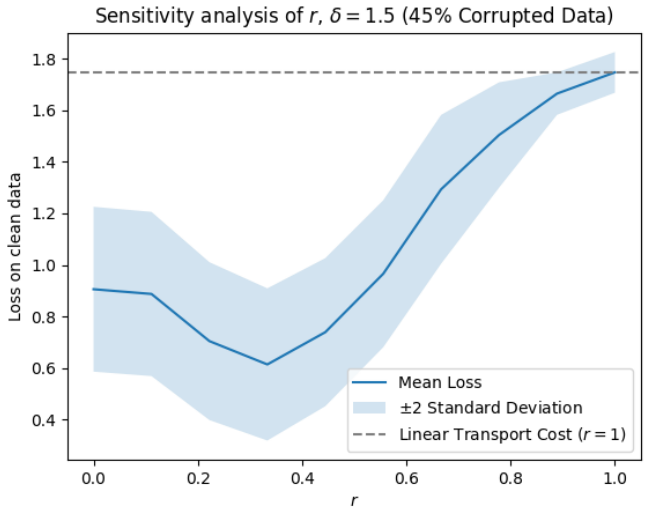


Figure 13: The sensitivity analysis of the loss on clean data with respect to r of our estimator at $\delta = 1.5$ on data with a 45% corruption level on the least absolute regression task.

https://doi.org/10.52202/079017-1113

F.6 Concave Cost Regression Simulation

In the Figure 14, we plot the evolution of the rectified distribution and line of best fit produced by our LAD regression estimator under various values of δ for the **concave** cost function with $r = 0.5$.

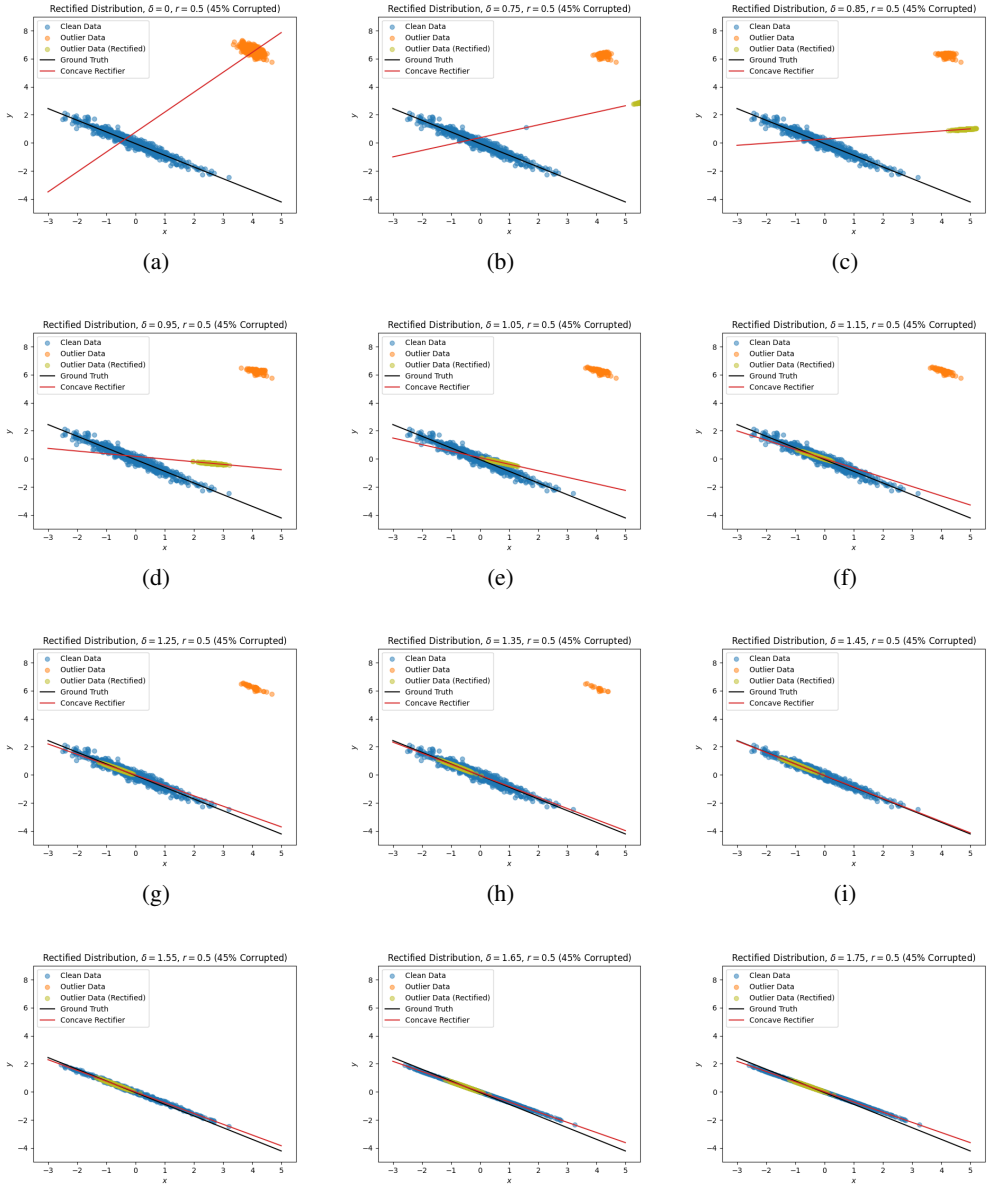


Figure 14: Visualization of the evolution of the rectified distribution.

F.7 Convex Cost Regression Simulation

In the Figure 15, we plot the evolution of the rectified distribution and line of best fit produced by our LAD regression estimator under various values of δ for the **convex** cost function with $r = 2.0$.

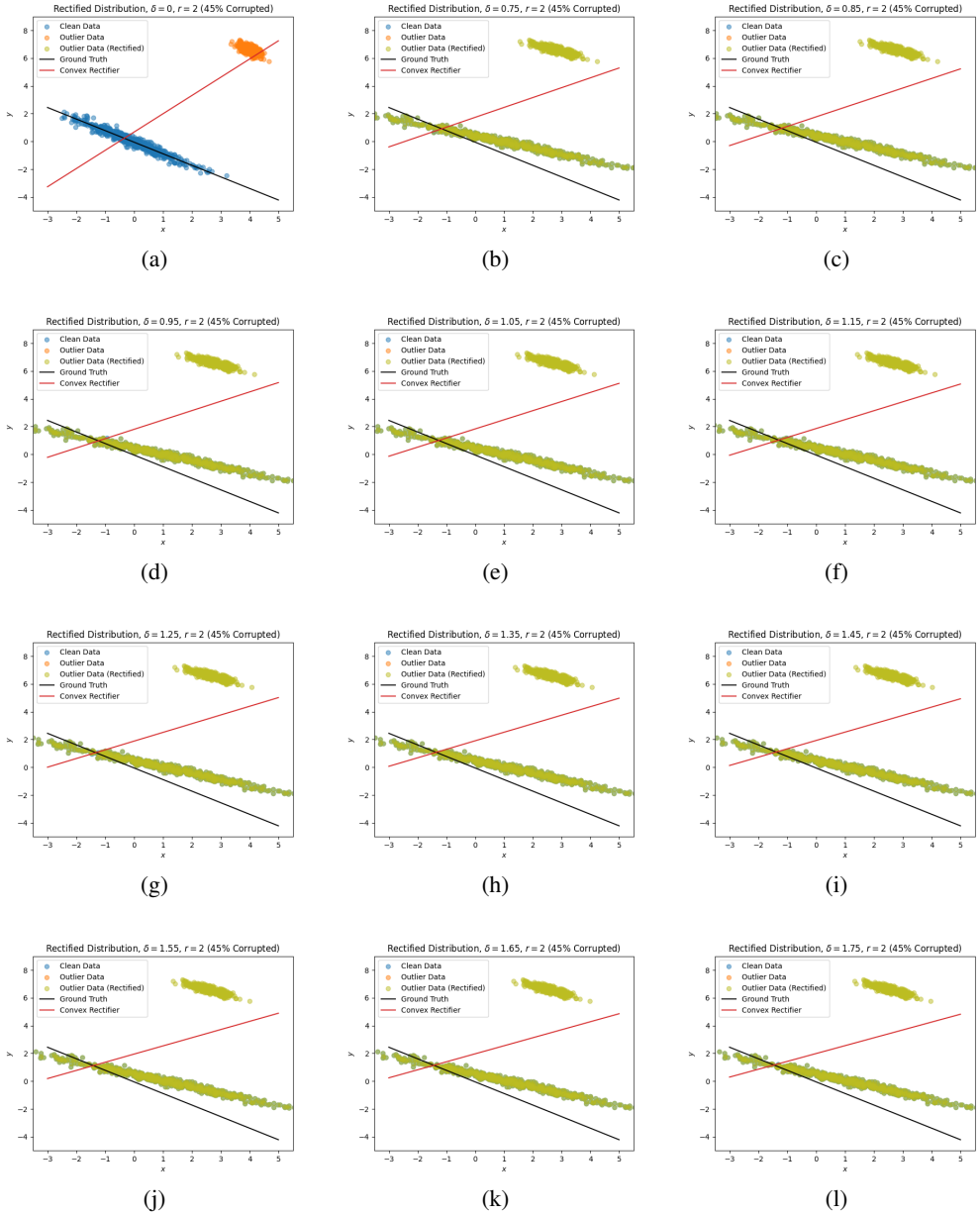


Figure 15: Visualization of the evolution of the rectified distribution.

https://doi.org/10.52202/079017-1113

G Additional Volatility Modeling Details & Experiments

G.1 Additional Volatility Surface Experiment

We have performed an additional experiment to demonstrate the practicality of our method and the selection of the parameter δ with the realistic options outlier data set of our empirical study. This experiment shows that our method leads to better out-of-sample results on complex data and does not require the availability of clean data. The problem is the estimation of the volatility surface in the presence of outliers for the option data.

Our experiment proceeds as follows. In this experiment, the data set is organized as a list of (train day, test day) tuples. Train and test days are consecutive days for the same underlying security. On the train day, we estimate our model with different methods. On the consecutive trading day (that is, the test day) we use the estimated model to obtain out-of-sample evaluation metrics for the surface (MAPE and $\nabla\hat{S}$). Note that this approach follows closely how volatility surfaces are used in practice by traders. This allows us to obtain the out-of-sample MAPE and $\nabla\hat{S}$ as a function of δ. The prior Appendix section provides the details of our empirical setup.

To perform cross-validation, we split the train day into a training and validation sample, which we use to obtain estimates of MAPE and $\nabla\hat{S}$ as a function of δ. We select the delta that optimizes MAPE and $\nabla\hat{S}$ and fit on the entire train day. Then, on the out-of-sample test day, we evaluate our method for the optimally selected δ. Thus, to be clear, for the estimated optimal δ, we estimate the surface on one day, and evaluate it completely out-of-sample on the consecutive trading day.

The results are collected in the table below. The results of our experiment show that our approach outperforms competing approaches and that cross-validation improves upon the fixed δ results reported in the main text. The below tables show the out-of-sample MAPE and $\nabla\hat{S}$ averaged over all out-of-sample test days.

Model MAPE	0.5% Quantile	5% Quantile	Median	Mean	95% Quantile	99.5% Quantile
KS	0.047	0.088	0.236	0.274	0.563	1.033
2SKS	0.036	0.065	0.172	0.216	0.504	0.999
Ours ($\delta = 10^{-2}$)	0.037	0.065	0.170	0.203	0.440	0.841
Ours (CV)	0.036	0.064	0.170	0.203	0.437	0.837
Model $\nabla\hat{S}$	0.5% Quantile	5% Quantile	Median	Mean	95% Quantile	99.5% Quantile
KS	0.323	1.616	14.042	19.653	59.534	105.361
2SKS	0.036	0.121	1.759	6.823	30.020	76.652
Ours ($\delta = 10^{-2}$)	0.035	0.112	1.579	5.880	25.521	70.708
Ours (CV)	0.033	0.104	1.248	4.522	19.645	55.261

Table 5: Median and quantiles of the distribution of MAPE or $\nabla\hat{S}$ across all samples for KS, 2SKS, and our methods. Our estimator achieves significantly lower MAPE and $\nabla\hat{S}$ than the benchmark KS estimator and improves upon the 2SKS estimator. Our CV procedure improves even further.

G.2 Kernelized Regression Experiment Details

Data set. Our data set is derived from European-style US equity options implied volatilities from the years 2019–2021. The data come from the OptionMetrics IvyDB US database accessed through Wharton Research Data Services (WRDS). Implied volatilities created by OptionMetrics are calculated using the Black-Scholes formula using interest rates derived from ICE IBA LIBOR rates and settlement prices of CME Eurodollar futures. Option prices are set to the midpoint of the best bid and offer quoted for the option captured at 3:59 PM ET.

We developed our options surface estimator code in partnership with a major global provider of financial data, indices, and analytical products. This provider had previously identified 1,970 outlier-containing option chains for which they found IVS estimation challenging. These outlier-containing option chains were selected from all listed US equities on days within 2019–2021, and they comprise the data set used in our experiments. Because the code is proprietary, and the data set is a proprietary selection of surfaces, we cannot publicly release them. However, the data can be found in WRDS, and we provide all implementation details in this appendix section.

Implementation. We implement the KS estimator and our estimator in PyTorch. For the 2SKS method, options in an options chain are removed if their implied volatilities fall outside of the region $[q_{0.25} - 1.5 \cdot IQR, q_{0.75} + 1.5 \cdot IQR]$, where $q_{0.25}$ is the 25% quantile of the implied volatilities in the options chain, $q_{0.75}$ is defined similarly, and IQR is the interquartile range of the implied volatilities in the options chain. The surface is then estimated upon the remaining options via the KS method. For our estimator, we estimate the surface by subgradient descent with learning rate $\alpha = 10^{-1}$ and $r = 0.5$, terminating when the relative change in loss reaches 10^{-5}. We denote the test data set's option's implied volatilities y_i' and the estimated surface's implied volatility for option i as $\hat{S}(x_i')$. We begin each iteration by setting $\delta = \ell(x, y)/2\|\theta\|^r$, where ℓ is the loss function of Theorem 4. We initialize θ to those of the benchmark estimator. We perform 5 trials per options chain. In each trial, we randomly select a different 80% train and 20% test set, estimate the surface on the train set, and record the two losses (MAPE and $\nabla\hat{S}$) on the test set. We depict all losses gathered in this way via the histograms of Section 5.2. Experiments are run on a server with a Xeon E5-2398 v3 processor and 756GB of RAM.

Cross-validation. The cross-validation (CV) procedure is standard. We sample 4/5 of the training surface as a CV training set and leave the remaining 1/5 as the CV validation set. Five such splits are made and used to estimate the MAPE of $\delta \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 2, 5, 10\}$. The MAPEs of each δ are averaged across the five CV runs and the δ with the lowest MAPE is then used to fit our estimator on the entire option chain.

Losses. These losses are chosen for their importance in finance and option trading and valuation. MAPE is a preferable error metric for volatility surfaces, as option chains contain options with implied volatilities which can differ in orders of magnitude, and MAPE weighs equally the contribution to error of options with such volatilities. The discretized surface gradient $\nabla\hat{S}$ measures surface smoothness of the estimated volatility surface by computing a discrete gradient across a grid of points in the (time to expiration, strike price) plane. Smooth surfaces are important for several reasons: (1) smoother surfaces allow more stable interpolation or extrapolation of implied volatility for options with market prices which are not directly observable; (2) smoother surfaces produce smaller adjustments for option hedging positions constructed from measures of the IVS, which ultimately lowers hedging transaction costs; (3) smoother surfaces are less likely to have internal arbitrages between options, which are implied by sharp discontinuities in the IVS, putting the surface more in line with established asset pricing theory; and (4) smoother surfaces produce more stable estimates of financial institutions' derivative exposures, which are required to be consistent for financial regulatory and reporting requirements. These measures are standard in options IVS estimation. We define the option implied volatility surface $\hat{S}(x)$ as the function from *option feature vectors* x to *estimated implied volatilities* y. That is, for some given x, we have $y := \hat{S}(x)$. The accuracy measure MAPE is defined as follows for some test set of options $\{(x_i', y_i')\}_{i=1}^n$:

$$l_{\text{MAPE}}(\hat{S}, \{y_i'\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{S}(x_i') - y_i'|}{|y_i'| + c}$$

https://doi.org/10.52202/079017-1113

where c is a small numerical stability factor we set to 0.01. This is a suitable choice, as $\geqslant$99% of implied volatilities are greater than 0.1.

The smoothness measure $\nabla\hat{S}$ is defined as follows:

$$\nabla\hat{S} = \sum_{i=0}^{p-1}\sum_{j=0}^{s-1} \frac{(\hat{S}_{\tau_{i+1},\Delta_j} - \hat{S}_{\tau_i,\Delta_j})^2 + (\hat{S}_{\tau_i,\Delta_{j+1}} - \hat{S}_{\tau_i,\Delta_j})^2}{2}$$

In the expression above, $\hat{S}_{\tau_i,\Delta_j} := \hat{S}(\log\tau_i, u(\Delta_j), z)$ where $z = 1$ if $\Delta_j > 0$ and 0 otherwise. Here, each tuple (τ_i, Δ_j) is a member of the Cartesian product of $\{\tau_j : j \in [p]\}$ and $\{\Delta_k : k \in [s]\}$, where "$[m]$" denotes the set of natural numbers from 0 to m, and $p = 10$ and $s = 39$. By convention, OptionMetrics selects the following 11 discretization points for τ, which we follow:

$$\tau \in \{10, 30, 60, 91, 122, 152, 182, 273, 365, 547, 730\}.$$

For Δ, we use the following 40 points, which are a superset of the set of discretization points which OptionMetrics selects:

$$\Delta \in \{-1.0 + 0.05m : m \in [40]\} \setminus \{0\}.$$

Example implied volatility surfaces. In the following surface figures, we display the significant outlier rectification effect of our estimator on a sample option chain which contains outliers. These examples serves to illustrate the efficacy of the automatic outlier rectification mechanism and to provide intuition for the surface fitting problem.

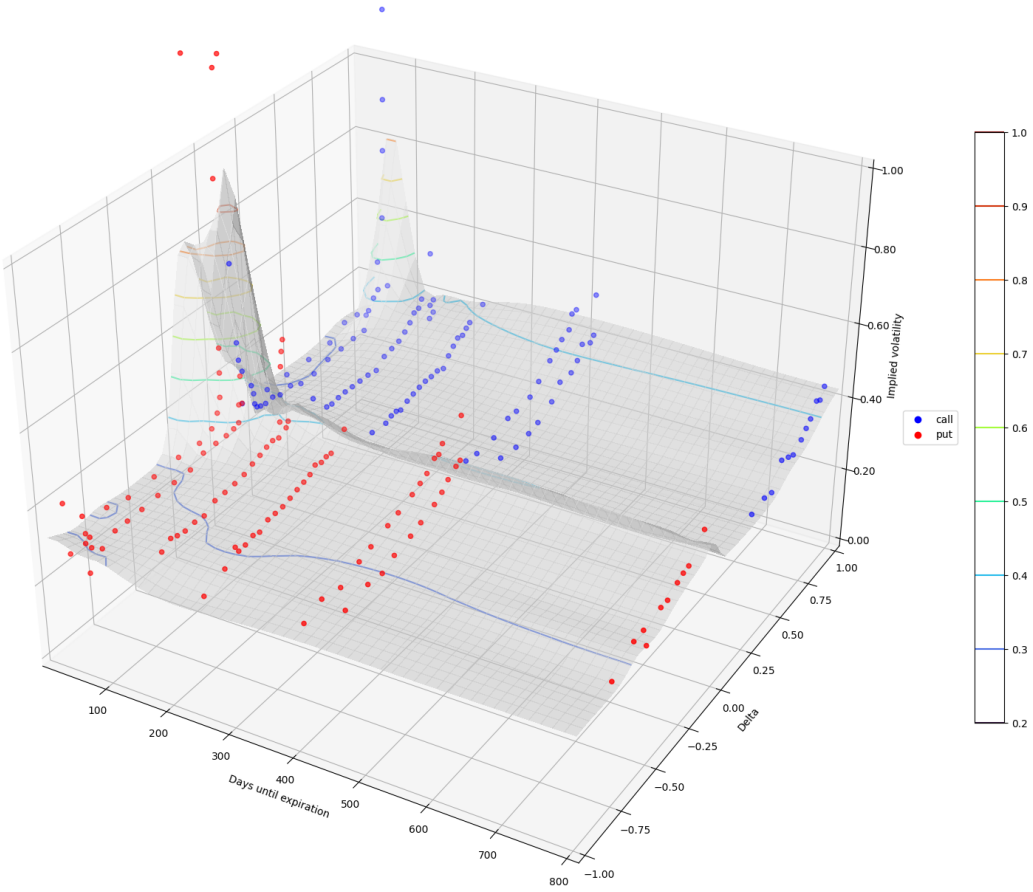


Figure 16: This plot depicts the option implied volatility surface estimated by the benchmark KS method on the options, which are depicted as blue and red dots. Blue denotes call options and red denotes put options. Outliers are present in this option chain near 0-100 days until expiration for deltas around $\Delta = -0.2$ (approximately four put option outliers and 1 call option outlier) and $\Delta = 1.0$ (approximately 3 call option outliers). These outliers heavily corrupt the fitted surface, wildly distorting the values around these deltas and causing a poor fit for surrounding options which are not outliers. The surface reaches values of 90% annualized implied volatility and has a surface gradient $\nabla \hat{S}$ of 1.48.

https://doi.org/10.52202/079017-1113

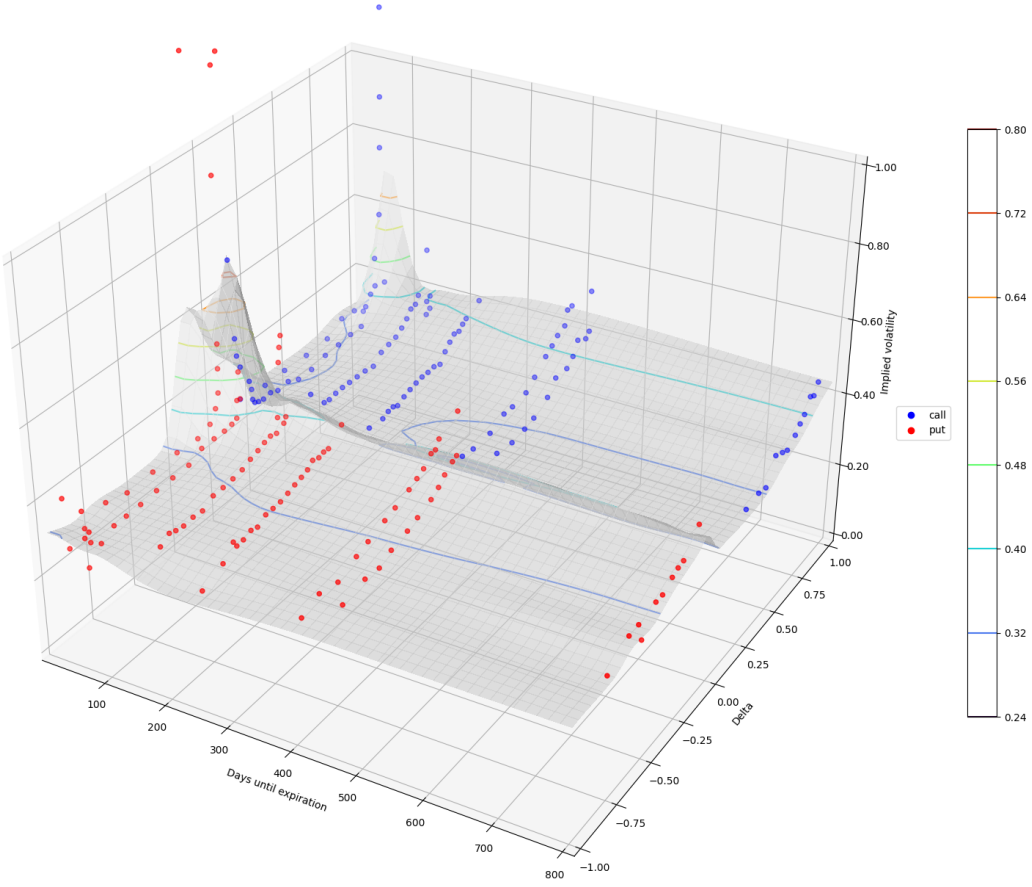


Figure 17: This plot depicts the option implied volatility surface $\hat{S}(x)$ estimated by our estimator with $\delta = 1$ on the displayed options, which are depicted as blue and red dots. Blue denotes call options and red denotes put options. Outliers are present in this option chain near 0-100 days until expiration for deltas around -0.2 and 1.0. However, these outliers do not corrupt the fitted surface, which fits options nearby the outlying options quite well as compared to the surface fit by the benchmark method displayed in Figure G.2. Additionally, in contrast to the surface fit by the benchmark KS method, the surface fit by our estimator only reaches values of $\sim 65\%$ annualized implied volatility and has a surface gradient $\nabla\hat{S}$ of 0.69, a smoothness improvement of over 100%. This example illustrates the capability of our estimator to rectify outliers.

G.3 Deep Learning Experiment Details

The full details and description of the deep learning local volatility surface estimation experiment are given in this section.

Background. IVS estimation provides an approximate surface fit for the volatilities implied by option market prices. However, market participants often require surfaces which fit as closely as possible to the market-implied volatilities.[1] One of the most popular methods used in mathematical finance to estimate surfaces obeying such a constraint under a reasonable stochastic process model for the evolution of asset prices is the *local volatility* or *implied volatility function* model introduced by Rubinstein [1994], Dupire et al. [1994], and Derman et al. [1996]. The local volatility model assumptions imply that a function $\sigma(K, T)$ exists which gives an analytic expression for the volatility σ at a given option strike price K and option time to expiration T. This function is known as Dupire's formula after its originator. Enforcing Dupire's formula in the estimation of a surface results in a local volatility surface, which is of great use to financial market participants. However, estimating such a surface becomes significantly complicated if market data contains corruptions or unrealistic outliers which are unlikely to occur in the future.[2] As human intervention for outlier rectification in a large number of IVS fitting tasks is unrealistic, an automatic method for rectifying outliers in models fitting local volatility surfaces is required.

Data. We utilize the data set of Chataigner et al. [2020], which contains 17 (options chain, IVS) pairs taken from the German DAX index in August 2001. Options chains are captured from daily options market data, and IVSes are estimated for a grid of strikes and maturities via a trinomial tree calibrated by the method of Crépey [2002]. In addition to Chataigner et al. [2020], this data set is also used in Crépey [2002] and Crépey [2004]. In this data set's usage in the literature, the authors train and test using the price surface, which is given by a nonlinear transformation of the IVS. Consequently, we also use the price surface for fitting; however, we report results after transforming back to the IVS. To corrupt the data set, a percentage of the options in each options chain are selected. We refer to this percentage, the corruption level, as ε. The prices of these options x are then corrupted by replacing them with the values $10x$.

Benchmark Method. As a benchmark, we apply the state-of-the-art deep learning approach from Chataigner et al. [2020], the Dupire neural network. In this approach, a neural network estimates the price surface under the local volatility model assumptions encoded in Dupire's formula, and additional no-arbitrage constraints which are useful for empirical applications. These conditions are enforced using different approaches: hard, split, and soft constraints. Hard constraints enforce the conditions by separating the network into subnetworks which have separate input layers for the variables involved in each of the conditions; conditions such as convexity and non-negativity are then enforced by projection steps during optimization and the proper choice of activation functions. Hard constraints thus learn a function which explicitly satisfies the conditions. Soft constraints enforce these conditions instead by penalizing violations of them at each of the training points observed from the market. Hybrid constraints split the network into subnetworks which have separate input layers for each of the constrained variables as in hard constraints, and utilize soft constraints for enforcing the Dupire and no-arbitrage conditions. The function learned only approximately satisfies the conditions in theory, but in practice the approximation is very good (see Chataigner et al. [2020] for details). Because these approaches estimate the price surface, we also estimate the price surface in this application; however, the local volatility surface is easily recovered from the price surface by a nonlinear transformation.

Robust Neural Network Estimation Problem & Implementation. To estimate price and volatility surfaces under the induced corruption, we estimated the benchmark approach using our statistically robust estimator. For each options chain, we tune δ via cross-validation: for each day, we sample 80% of the training set without replacement as a cross-validation (CV) training set, and use the remaining

[1] For example, banks pricing some kinds of exotic options require such surfaces to avoid arbitrage opportunities in surfaces that their internal trading teams may exploit, as outlined in Hull and Basu [2022]. Participants pricing options may moreover simply prefer to accept the assumption that the market prices are usually correct, and thus prefer closely fitting surfaces.

[2] This may occur due to low liquidity, which allows market manipulation or adversarial trading to produce unrealistic prices.

https://doi.org/10.52202/079017-1113

20% of the training set as a CV validation set. We then train our robust neural network estimator on the CV training set with $\delta \in \{10^{-3}, 10^{-2}, ..., 10^{3}\}$. For each δ, we then compute the MAPE on the CV validation set. We then select the δ with the lowest MAPE on CV validation set, and we use this δ to train our estimator on the entire training set. This produces our final estimated model. We base our implementation on the implementation of Chataigner et al. [2020] (see commit e03da98 at this url), which is BSD 3-clause licensed. The experiments were run on a server with a Xeon E5-2398 v3 processor with 756GB of RAM.

Results. To maintain comparability with the benchmark, we test and report the same metrics reported by Chataigner et al. [2020], namely, the RMSE and MAPE for the surface estimation. The metrics for each day and model are repeated for three trials with different random weight initialization seeds. The results are displayed in Table 5.2.2. The first panel displays the RMSE and MAPE for each model using the different constraint techniques, averaged across all trials and corruption levels. The second panel displays the RMSE and MAPE for each corruption level averaged across all models. Our approach outperforms the baseline approach across all averages, and does so more clearly as the corruption level increases, despite the strong regularizing effect of the no-arbitrage constraints and enforcement of Dupire's formula. Our improvement is present across all models, and is especially impactful for the most accurate model utilizing soft arbitrage constraints. For this model, the out-of-sample test error in terms of RMSE and MAPE are reduced by 33% and 34%, respectively. This experiment displays the efficacy of our estimator in a state-of-the-art deep learning model for option price surface modeling.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper claims to give an automatic outlier rectification mechanism that integrates rectification and estimation within a joint optimization framework and demonstrate its effectiveness in experiments. These claims are satisfied in the paper via definitions, theorems, proofs, and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

https://doi.org/10.52202/079017-1113

Answer: [Yes]

Justification: The paper clearly summarizes and refers to the appendix for limitations to our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All important equations are numbered and cross-referenced, and assumptions, lemmas, and proofs are formally given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: All information required for experimental replication is detailed in the experimental appendices.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The work in this paper was conducted with an industry partner, and the code and data are proprietary. However, most of the experiments are reproducible, and experimental details, algorithms, and data sources are clearly stated.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

https://doi.org/10.52202/079017-1113

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We clearly state all experimental details in the appendices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars or quantile information are included in tables, except for the deep learning experiment, which is expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are provided in the experiment appendices.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper concerns work which is purely focused on a mathematical and computational problem.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

https://doi.org/10.52202/079017-1113

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Prior code, data, and models with licenses are clearly cited in the appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper did not conduct crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper has no crowdsourcing or human subject research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1113