On the Adversarial Robustness of Robust Estimators

Lifeng Lai, *Senior Member, IEEE* and Erhan Bayraktar

***Abstract*—Motivated by recent data analytics applications, we study the adversarial robustness of robust estimators. Instead of assuming that only a fraction of the data points are outliers as considered in the classic robust estimation setup, in this paper, we consider an adversarial setup in which an attacker can observe the whole dataset and can modify all data samples in an adversarial manner so as to maximize the estimation error caused by his attack. We characterize the attacker's optimal attack strategy, and further introduce adversarial influence function (AIF) to quantify an estimator's sensitivity to such adversarial attacks. We provide an approach to characterize AIF for any given robust estimator, and then design optimal estimator that minimizes AIF, which implies it is least sensitive to adversarial attacks and hence is most robust against adversarial attacks. From this characterization, we identify a tradeoff between AIF (i.e., robustness against adversarial attack) and influence function, a quantity used in classic robust estimators to measure robustness against outliers, and design estimators that strike a desirable tradeoff between these two quantities.**

***Index Terms*—Robust estimators, adversarial robustness, M-estimator, non-convex optimization.**

I. Introduction

Robust estimation is a classic topic that addresses the outlier or model uncertainty issues. In the existing setup, a certain percentage of the data points are assumed to be outliers. Various concepts such as influence function (IF), breakdown point, and change of variance etc were developed to quantify the robustness of estimators against the presence of outliers, please see [1]–[3] and references therein for details. Furthermore, computationally efficient robust algorithms for high dimensional problems were developed in many recent work [4]–[8].

These concepts are very useful for the classic setup where a *fraction* (up to 50%) of data points are outliers while the remaining data come from the true distribution. In this paper, motivated by recent interest in data analytics, we address the issue of adversarial robustness. In a typical data analytics setup, a dataset is stored in a database. If an attacker has access to the database, he can modify *all* data points (i.e., up to 100%) in an adversarial manner, and hence the existing results on robust statistics are not directly applicable anymore. This scenario also arises in the adversarial example phenomena in deep neural networks that have attracted significant recent research interests [9]–[11]. In the adversarial example in deep neural networks, by making small but carefully chosen changes on the image, the attacker can mislead neural network to make wrong decisions, even though a human will hardly notice changes on the modified image. Certainly, if the attacker can modify all data and no further restrictions on attacker's capability are imposed, then no meaningful estimator can be constructed (this can be viewed as 100% of the data are modified in the classic setup). In this paper, we investigate the scenario that the total amount of change measured by ℓ_p norm is limited, and we will study how these quantities will affect the estimation performance. Towards this goal, we introduce the concept of adversarial influence function (AIF) to quantify how sensitive an estimator is to adversarial attacks. These types of constraints are reasonable and are motivated by real life examples. For example, in generating adversarial examples in images [9], the total distortion should be limited, otherwise human eyes will be able to detect such changes. Our problem formulation could also potentially be useful for investigating the robustness of machine learning algorithms under various constraint on the norm of the attack vector, see for example [12].

We first focus on the scenario with a given data set. For this scenario, we characterize the optimal attack vector that the attacker, who observes the whole data set, can employ to maximize the change of estimation result. Using this characterization, we can then analyze AIF of any given estimator. This analysis enables us to design estimators that are robust to adversarial attacks. In particular, from the estimator's perspective, one would like to design an estimator that minimizes AIF, which implies that such an estimator is least sensitive to adversarial attacks and hence is most robust against adversarial attacks. We derive universal lower bounds on AIF and characterize the conditions under which an estimator can achieve this lower bound (and hence is most robustness against adversarial attacks). We then illustrate these results for two specific models: location estimators and scale estimators.

With the results in the given sample scenario, we then extend our study to the population scenario, in which we investigate the behavior of AIF as the number of samples increases. For this case, we identify a tradeoff between robustness against adversarial attacks vs robustness against outliers. In particular, we first characterize the optimal estimator that minimizes AIF. However, the estimator that minimizes AIF has a poor performance in term of IF [3], [13], a quantity that measures robustness against outliers. Realizing this fact, we then formulate optimization problems to design estimators that strike a desirable tradeoff between AIF (i.e., robustness against adversarial attack) and IF (i.e., robustness against outliers). Using tools from calculus of variations [14], [15], we are able to exploit the unique structure of our problems and obtain

Lifeng Lai is with Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616. Email: lflai@ucdavis.edu. Erhan Bayraktar is with Department of Mathematics, University of Michigan, Ann Arbor, MI 48104. Email: erhan@umich.edu. The work of L. Lai was supported by the National Science Foundation under grants CCF-17-17943, ECCS-17-11468, CNS-18-24553 and CCF-19-08258. The work of E. Bayraktar was supported in part by the National Science Foundation under grant DMS-1613170 and by the Susan M. Smith Professorship.

analytical form of the optimal solution. The obtained solution share similar interpretation as classic robust estimators that it will carefully trim data points that are from the distribution tails. However, the detailed form and thresholds are determined by different criteria.

In the above discussion, we mainly focus on a class of widely used robust estimators: M-estimator. However, the developed tools and analysis can be extended to analyze other types of robust estimators. In this paper, we will use L-estimator as an example to discuss how to extend the analysis to other types of estimators.

Our paper is related to a growing list of recent work on adversarial machine learning. Here we give several examples on data poisoning attack that is related to our work. For example, [16] considers an adversarial principal component analysis (PCA) problem. Different from many interesting work on robust PCA [17], in the model considered in [16], an attacker adds an extra data point in an adversarial manner so as to maximize the error of subspace estimated by PCA. [18] investigates data poisoning attack in regression problems, in which the attacker adds data points to the training dataset with the goal of introducing errors into or guiding the results of regression models. [19] studies an attack that inserts carefully chosen data points to the training set for support vector machine. [20] considers learning problems from untrusted data. In particular, under the assumption that at least α percent of data points are drawn from a distribution of interest, [20] considers two frameworks: 1) list-decodable learning, whose goal is to return a list of answers, with the guarantee that at least one of them is accurate; and 2) semi-verified learning, in which one has a small dataset of trusted data that can be leveraged to enable the accurate extraction of information from a much larger but untrusted dataset. There are also a large number of recent work on robust estimators in high dimensions [4]–[8]. These papers focus on designing computationally efficient algorithms in the high dimension region. A major difference between our work and these interesting work is that the existing work assume that a certain percentage of data points are not compromised, while in our work all data points could be compromised. Our paper is also related to a recent interesting paper [21] that studies the problem of robust linear regression with response variable corruptions. [21] considers an oblivious adversary model, in which the adversary changes a fraction of responses without knowledge of the data, and provides a nearly linear time estimator that is consistent even when the majority of the data is corrupted. In our paper, the modification introduced by the attack can be dependent on the whole dataset.

The remainder of the paper is organized as follows. In Section II, we introduce our problem formulation. In Section III, we introduce the necessary background. In Section IV, we investigate AIF for the given sample scenario. In Section V, we consider the population scenario. We extend the study to L-estimator in Section VI. Numerical examples are given in Section VII. Finally, we offer concluding remarks in Section VIII.

II. Model

We consider an adversarially robust parameter estimation problem in which the adversary has access to the whole dataset. In particular, we have a given data set $\mathbf{x} = \{x_1, \cdots, x_N\}$, in which x_n are i.i.d realizations of random variable $X \in \mathbb{R}$ that has cumulative density function (cdf) $F_\theta(x)$ with unknown parameter $\theta \in \mathbb{R}$. We will use $f_\theta(x)$ to denote the corresponding probability density function (pdf). From this given data set, we would like to estimate the unknown parameter θ. However, as the adversary has access to the whole dataset, it will modify the data to $\mathbf{x}^\Delta = \mathbf{x} + \Delta\mathbf{x} := \{x_1 + \Delta x_1, \cdots, x_N + \Delta x_N\}$, in which $\Delta\mathbf{x} = \{\Delta x_1, \cdots, \Delta x_N\}$ is the attack vector chosen by the adversary after observing $\mathbf{x}$. We will discuss the attacker's optimal attack strategy in choosing $\Delta\mathbf{x}$ in the sequel. In the classic robust estimation setup, it is typically assume that some percentage (up to 50%) of the data points are outliers, that is some entries in $\Delta\mathbf{x}$ are nonzero while the remainders are zero. In this work, we consider the case where the attacker can modify all data points, which is a more suitable setup for recent data analytical applications. However, certain restrictions need to be put on $\Delta\mathbf{x}$, otherwise the estimation problem will not be meaningful. In this paper, we assume that

$$\frac{1}{N}||\Delta\mathbf{x}||_p^p \leq \eta^p, \tag{1}$$

in which $||\cdot||_p$ is the ℓ_p norm. The normalization factor N implies that the per-dimension change (on average) is upper-bound by η^p. As mentioned in the introduction, this type of constraints are reasonable and are motivated by real life examples. The classic setup can be viewed as a special case of our formulation by letting $p \to 0$, i.e., the classic setup has constraint on the total number of data points that the attacker can modify.

Following notation used in robust statistics [2], [3], we will use $T_N(\mathbf{x})$ to denote an estimator. For a given estimator T_N, we would like to characterize how sensitive the estimator is with respect to the adversarial attack. In this paper, we consider a scenario where the goal of the attacker is to maximize the deviation in the estimator's output caused by the attack. In particular, the attacker aims to choose $\Delta\mathbf{x}$ by solving the following optimization problem

$$\begin{aligned} \max_{\Delta\mathbf{x}} \quad & |T_N(\mathbf{x} + \Delta\mathbf{x}) - T_N(\mathbf{x})|, \\ \text{s.t.} \quad & \frac{1}{N}||\Delta\mathbf{x}||_p^p \leq \eta^p. \end{aligned} \tag{2}$$

We use $\Delta T_N(\mathbf{x})$ to denote the optimal value obtained from the optimization problem (2), and define the adversarial influence function (AIF) of estimator T_N at $\mathbf{x}$ under ℓ_p norm constraint as

$$\text{AIF}(T_N, \mathbf{x}, p) = \lim_{\eta \downarrow 0} \frac{\Delta T_N(\mathbf{x})}{\eta}.$$

This quantity, a generalization of the concept of IF used in classic robust estimation (we will briefly review IF in Section III), quantifies the rate at which the attacker can introduce estimation error through its attack.

From the defender's perspective, the smaller AIF is, the more robust the estimator is. In this paper, building on the

characterization of AIF$(T_N, \mathbf{x}, p)$, we will characterize the optimal estimator T_N, among a certain class of estimators $\mathcal{T}$, that minimizes AIF$(T_N, \mathbf{x}, p)$. In particular, we will investigate

$$\min_{T_N \in \mathcal{T}} \text{AIF}(T_N, \mathbf{x}, p).$$

We will show that, for certain class of $\mathcal{T}$, the optimal T_N is independent of $\mathbf{x}$ and p, which is a very desirable property.

Note that AIF$(T_N, \mathbf{x}, p)$ depends on the data realization $\mathbf{x}$. Based on the characterization of AIF for a given data realization $\mathbf{x}$ of length N, we will then study the population version of AIF where each entry of $\mathbf{X} = \{X_1, \cdots, X_N\}$ is i.i.d generated by F_θ. We will examine the behavior of AIF$(T_N, \mathbf{X}, p)$ as N increases. Following the convention in robust statistics, we will assume that there exists a functional T such that

$$T_N(\mathbf{X}) \to T(F_\theta) \tag{3}$$

in probability as $N \to \infty$. We will see that for a large class of estimators AIF$(T_N, \mathbf{X}, p)$ has a well-defined limit as $N \to \infty$. We will use AIF(T, F_θ, p) to denote this limit when it exists.

Similarly, from the defense's perspective, we would like to design an estimator that is least sensitive to the adversarial attack. Again, we will characterize the optimal estimator T, among a certain class of estimators $\mathcal{T}$, that minimizes AIF(T, F_θ, p). That is, for a certain class of estimators $\mathcal{T}$, we will solve

$$\min_{T \in \mathcal{T}} \text{AIF}(T, F_\theta, p). \tag{4}$$

It will be clear in the sequel that the solution to the optimization problem (4), even though is robust against adversarial attacks, has poor performance in guarding against outliers. This motivates us to design estimators that strike a desirable tradeoff between these two robustness measures. In particular, we will solve (4) with an additional constraint on IF. We will need to use tools from calculus of variations for this purpose.

We note that in this paper, we focus on the scalar case (i.e., X and θ are scalers). The problem formulation and analysis can be extended (with additional technical developments) to the more general vector case (including joint location-scale estimation and robust regression etc.). The corresponding results are reported in [22].

III. Background

In this section, we briefly review results from classic robust estimator literature that are closely related to our study.

A. Influence Function (IF)

As mentioned above, in the classic robust estimation setup, it is assumed that a fraction η of data points are outliers, while the remainder of data points are generated from the true distribution F_θ. For a given estimator T, the concept of IF introduced by Hampel [13] is defined

$$\text{IF}(x, T, F_\theta) = \lim_{\eta \downarrow 0} \frac{T((1-\eta)F_\theta + \eta\delta_x) - T(F_\theta)}{\eta}.$$

In this definition, δ_x is a distribution that puts mass 1 at point x, $T(F_\theta)$, introduced in (3), is the obtained estimate when all data points are generated i.i.d from F_θ, and $T((1-\eta)F_\theta + \eta\delta_x)$ is the obtained estimate when $1-\eta$ fraction of data points are generated i.i.d from F_θ while η fraction of the data points are at x. Hence, IF(x, T, F_θ) measures the influence of having outliers at point x as $\eta \downarrow 0$.

To measure the influence of the worst outliers, [13] then further introduced the concept of gross-error sensitivity of T by taking sup over the absolute value of IF(x, T, F_θ):

$$\gamma^*(T, F_\theta) = \sup_x |\text{IF}(x, T, F_\theta)|.$$

Intuitively speaking, $\gamma^*(T, F_\theta)$ can be viewed as the solution of our problem setup for the special case of $p = 0$.

The values of IF(x, T, F_θ) and $\gamma^*(T, F_\theta)$ have been characterized for various class of estimators. Furthermore, under certain conditions, optimal estimator T that minimizes these quantities have been established. Some of these results will be introduced in later sections. More details can be found in [2], [3].

B. M-Estimator

In this paper, we will mainly focus on a class of commonly used estimator in robust statistic: M-estimator [1], in which one obtains an estimate $T_N(\mathbf{x})$ of θ by solving

$$\sum_{n=1}^{N} \psi(x_n, T_N) = 0. \tag{5}$$

Here $\psi(x_n, \theta)$ is a function of data x_n and parameter θ to be estimated. Different choices of ψ lead to different robust estimators. For example, the most likely estimator (MLE) can be obtained by setting $\psi = -f_\theta^{'}/f_\theta$. M-estimator can also be defined as the solution of an optimization problem. The formulation in (5) and the optimization formulation have certain relationship, but they are not always equivalent. Please refer to Chapter 2.3a of [3] for detailed discussion.

As the form of ψ determines T_N, in the remainder of the paper, we will use ψ and T_N interchangeably. For example, we will denote AIF$(T_N, \mathbf{x}, p)$ as AIF$(\psi, \mathbf{x}, p)$. Similarly, we will denote IF(x, T, F_θ) as IF(x, ψ, F_θ).

It is typically assumed that $\psi(x, \theta)$ is continuous and almost everywhere differentiable. This assumption is valid for all ψ's that are commonly used. It is also typically assume the estimator is Fisher consistent [3]:

$$\mathbb{E}_{F_\theta}[\psi(X, \theta)] = 0, \tag{6}$$

in which $\mathbb{E}_{F_\theta}$ means expectation under F_θ. Intuitively speaking, this implies that the true parameter θ is the solution of the M-estimator if there are increasingly more i.i.d. data points generated from F_θ.

For M-estimator, IF(x, ψ, F_θ) was shown to be

$$\text{IF}(x, \psi, F_\theta) = \frac{\psi(x, T(F_\theta))}{-\int \frac{\partial}{\partial\theta}[\psi(y, \theta)]_{\theta = T(F_\theta)} \mathrm{d}F_\theta(y)},$$

see (2.3.5) of [3].

IV. The Fixed Sample Case

In this section, we focus on analyzing AIF$(\psi, \mathbf{x}, p)$ for a given dataset $\mathbf{x}$. We will extend the study to the population case and analyze AIF(ψ, F_θ, p) in Section V.

A. General ψ

We will first characterize AIF$(\psi, \mathbf{x}, p)$ for general ψ, and will then specialize the results to specific problems in later sections. For any given ψ that is continuous and almost everywhere differentiable, we have the following theorem that characterizes AIF$(\psi, \mathbf{x}, p)$.

Theorem 1. *When $p = 1$,*

$$AIF(\psi, \mathbf{x}, 1) = \frac{\left|\frac{\partial}{\partial x}[\psi]_{x=x_{n^*},\theta=T_N}\right|}{\left|\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N}\right|},$$

where

$$n^* = \arg\max_n \left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|. \tag{7}$$

For $p > 1$, we have

$$AIF(\psi, \mathbf{x}, p) = \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N}\right|}.$$

Proof. Please see Appendix A for detailed proof. □

In this theorem, we characterize the result for $p \geq 1$. Ideally, one would like to consider the case with $p < 1$, but this will result in a non-convex optimization, which precludes us from obtaining a closed form solution.

From Theorem 1, we can characterize the form of ψ that leads to the smallest AIF, i.e., the most robust M-estimator against adversarial attacks.

Corollary 1.

$$AIF(\psi, \mathbf{x}, p) \geq \frac{\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|}{\left|\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N}\right|},$$

and the equality holds when

$$\left|\frac{\partial}{\partial x}[\psi]_{x=x_1,\theta=T_N}\right| = \cdots = \left|\frac{\partial}{\partial x}[\psi]_{x=x_N,\theta=T_N}\right|.$$

Proof. For $p > 1$, it is easy to check that $x^{(p-1)/p}$ is a concave function when $x \geq 0$. Hence, using Jensen's inequality, we have

$$\left(\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} \geq \frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|,$$

and the equality holds when $\left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|$ is a constant with respect to n. □

This corollary implies that, from defender's perspective, we should design $\psi(x, \theta)$ such that $\left|\frac{\partial}{\partial x}[\psi]\right|$ is constant in x. It is also interesting that, this result holds for any value of p. And hence we can design an estimator without knowledge about which constraint the attacker is using.

B. Specific Estimators

To illustrate the results obtained above, we specialize results to location estimators and scale estimators.

1) Location Estimator: For location estimator models, $F_\theta(x) = F_0(x - \theta)$, and hence it is natural to use $\psi(x, \theta) = \psi(x - \theta)$, see [2], [3]. For this model, it is easy to check that

$$\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N} = \psi'(x_n - T_N),$$
$$\frac{\partial}{\partial \theta}[\psi]_{x=x_n,\theta=T_N} = -\psi'(x_n - T_N).$$

Plugging these two equations in the AIF expressions in Theorem 1, for the case with $p = 1$, we have

$$\text{AIF}(\psi, \mathbf{x}, 1) = \frac{\left|N\psi'(x_{n^*} - T_N)\right|}{\left|\sum_{n=1}^{N}\psi'(x_n - T_N)\right|} \geq 1,$$

for which the equality holds when $\psi'(x_n - T_N)$ is a constant with respect to n.

For the case with $p > 1$, we have

$$\begin{aligned}\text{AIF}(\psi, \mathbf{x}, p) &= \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\left|\psi'(x_n - T_N)\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N}\psi'(x_n - T_N)\right|} \\ &\overset{(a)}{\geq} \frac{\frac{1}{N}\sum_{n=1}^{N}\left|\psi'(x_n - T_N)\right|}{\left|\frac{1}{N}\sum_{n=1}^{N}\psi'(x_n - T_N)\right|} \geq 1,\end{aligned} \tag{8}$$

in which (a) is due to Jensen's inequality. Both inequalities will hold if $\psi'(x_n - T_N)$ is a constant in n.

Example 1. *Consider an estimator with $\psi(x_n - T_N) = x_n - T_N$. This estimator is simply the empirical sample mean. It is easy to see that $\psi'(x)$ is a constant in n, which implies that this choice of ψ has AIF$(\psi, \mathbf{x}, p) = 1$. It achieves the lower bound established above, regardless of the value of $\mathbf{x}$ and p. Hence, it is the most robust estimator against adversarial attacks. However, as we will discuss in Section V, this choice of ψ is not robust against outliers. In Section V, we will design estimators that strike a desirable balance between robustness against outliers and robustness against adversarial attacks.*

Example 2. *Consider the Huber estimator [1] with*

$$\psi(x_n - T_N) = \min\{b, \max\{x_n - T_N, -b\}\},$$

parameterized by a parameter $0 < b < \infty$. Using (8), it is easy to check that AIF$(\psi, \mathbf{x}, p) = \sqrt{1/\beta}$, in which β is the proportion of points in $\mathbf{x}$ such that $|x_n - T_N| < b$. It is clear that Huber estimator, while being more robust against

outliers [2], is less robust against adversarial attacks than the empirical mean estimator.

2) Scale Estimator: The scale model [2], [3] is given by $F_\theta(x) = F_1(x/\theta)$, and it is typical to consider $\psi(x,\theta) = \psi(x/\theta)$. It is easy to check that for ψ with this form, we have

$$\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N} = \frac{\psi^{'}(x_n/T_N)}{T_N},$$
$$\frac{\partial}{\partial \theta}[\psi]_{x=x_n,\theta=T_N} = \frac{-x_n\psi^{'}(x_n/T_N)}{T_N^2}.$$

Using Theorem 1, for the case with $p=1$, we obtain

$$\begin{aligned}\text{AIF}(\psi,\mathbf{x},1) &= \frac{\left|N\frac{\psi^{'}(x_{n^*}/T_N)}{T_N}\right|}{\left|\sum_{n=1}^{N}\frac{-x_n\psi^{'}(x_n/T_N)}{T_N^2}\right|} \\ &= \frac{\left|\psi^{'}(x_{n^*}/T_N)\right|}{\left|\frac{1}{N}\sum_{n=1}^{N} x_n/T_N\psi^{'}(x_n/T_N)\right|}, \quad (9)\end{aligned}$$

in which n^* is defined in (7).

When $p>1$, we have

$$\begin{aligned}\text{AIF}(\psi,\mathbf{x},p) &= \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\psi^{'}(x_n/T_N)}{T_N}\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N}\frac{-x_n\psi^{'}(x_n/T_N)}{T_N^2}\right|} \\ &= \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\left|\psi^{'}(x_n/T_N)\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N} x_n/T_N\psi^{'}(x_n/T_N)\right|}. \quad (10)\end{aligned}$$

Example 3. *Consider MLE for variance of zero-mean Gaussian random variables, which corresponds to* $\psi(x) = -x(\phi^{'}(x)/\phi(x)) - 1 = x^2 - 1$. *Here,* $\phi(x)$ *is the pdf of zero mean variance one Gaussian random variable. For this choice of* ψ, *we have* $T_N = \frac{1}{N}\sum_{n=1}^{N} x_n^2$ *and* $\psi^{'}(x_n/T_N) = 2x_n/T_N$. *Plugging these values into* (9)*, we obtain*

$$\textit{AIF}(\psi,\mathbf{x},1) = |x_n^*|.$$

Using (10)*, we have*

$$\textit{AIF}(\psi,\mathbf{x},p) = \left(\frac{1}{N}\sum_{n=1}^{N}|x_n|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}},$$

from which we know that when $p=2$, $\textit{AIF}(\psi,\mathbf{x},2) = \sqrt{T_N} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$.

V. Population Case

With the results on the fixed dataset case, we now consider the population version where X_n are i.i.d from F_θ, and analyze the behavior of AIF as $N\to\infty$. Following the convention in classic robust statistics literature, we will focus on the case in which the estimator is Fisher consistent as defined in (6).

It has been shown in Theorem 2.4 of [2] that, under certain mild regularity conditions, $T_N \overset{a.s.}{\to} \theta$. In the following, we will need the following additional regularity conditions:

- $\psi^{'}(x)$ is continuous functions.
- There exist a function $K(x)$ such that $|\psi^{'}(x)| \le K(x)$, $|x\psi_1^{'}(x)| \le K(x)$, and $\mathbb{E}_{F_\theta}[K(X)] < \infty$.

The conditions here are slightly stronger than those conditions needed for the strong law of large numbers, as we will need to use the uniform strong law of large numbers (see Theorem 16 (a) [23]). Under these regularity assumptions, using the uniform strong law of large numbers, Slutsky Theorem (see Chapter 6 of [23]) and the fact that $T_N \overset{a.s.}{\to} \theta$, as $N\to\infty$ we have

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N} \overset{a.s.}{\to} \mathbb{E}_{F_\theta}\left[\frac{\partial}{\partial\theta}[\psi](X,\theta)\right]. \quad (11)$$

Furthermore, using Proposition 3 of [24], as $N\to\infty$ in Theorem 1, we have

$$\left|\frac{\partial}{\partial x}[\psi]_{x=x_{n^*},\theta=T_N}\right| \overset{a.s.}{\to} \max_x\left|\frac{\partial}{\partial x}[\psi](x,\theta)\right|. \quad (12)$$

As the result,

$$\begin{aligned}\text{AIF}(\psi,\mathbf{x},1) &= \frac{\left|\frac{\partial}{\partial x}[\psi]_{x=x_{n^*},\theta=T_N}\right|}{\left|\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N}\right|} \\ &\overset{a.s.}{\to} \frac{\max_x\left|\frac{\partial}{\partial x}[\psi](x,\theta)\right|}{\left|\mathbb{E}_{F_\theta}\left[\frac{\partial}{\partial\theta}[\psi](X,\theta)\right]\right|} \\ &:= \text{AIF}(\psi,F_\theta,1). \quad (13)\end{aligned}$$

For $p>1$, we have

$$\begin{aligned}\text{AIF}(\psi,\mathbf{x},p) &= \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\left|\frac{\partial}{\partial x_n}[\psi]_{x=x_n,\theta=T_N}\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N}\right|} \\ &\overset{a.s.}{\to} \frac{\left(\mathbb{E}_{F_\theta}\left[\left|\frac{\partial}{\partial x}[\psi](X,\theta)\right|^{\frac{p}{p-1}}\right]\right)^{\frac{p-1}{p}}}{\left|\mathbb{E}_{F_\theta}\left[\frac{\partial}{\partial\theta}[\psi](X,\theta)\right]\right|} \\ &:= \text{AIF}(\psi,F_\theta,p). \quad (14)\end{aligned}$$

A. Location Estimator

We now specialize the results to the location model mentioned above. We will first characterize ψ that minimizes $\text{AIF}(\psi,F_\theta,p)$. We will then discuss the tradeoff between the robustness to outliers and robustness to adversarial attacks, and will characterize the optimal ψ that achieves this tradeoff. In the location estimator, we will assume $\psi(x,\theta)$ is monotonic in θ, which will satisfy the regularity conditions established in [2].

1) Minimizing AIF(ψ,F_θ,p)*:* For $p=1$, using (13), we have

$$\text{AIF}(\psi,F_\theta,1) = \frac{\max_x\left|\psi^{'}(x-\theta)\right|}{|\mathbb{E}_{F_\theta}[\psi^{'}(X-\theta)]|}.$$

For $p > 1$, using (14), we obtain

$$\mathrm{AIF}(\psi, F_\theta, p) = \frac{\left(\mathbb{E}_{F_\theta}\left[\left|\psi^{'}(X-\theta)\right|^{\frac{p}{p-1}}\right]\right)^{\frac{p-1}{p}}}{\left|\mathbb{E}_{F_\theta}\left[\psi^{'}(X-\theta)\right]\right|}. \quad (15)$$

In particular, for $p = 2$, we have

$$\mathrm{AIF}(\psi, F_\theta, 2) = \sqrt{\frac{\mathbb{E}_{F_\theta}[\psi^{'}(X-\theta)^2]}{(\mathbb{E}_{F_\theta}[\psi^{'}(X-\theta)])^2}}.$$

From (15) and using Jensen's equality, we have

$$\mathrm{AIF}(\psi, F_\theta, p) \geq 1,$$

for which the equality holds when $\psi^{'}(x-\theta)$ is constant in x.

2) Tradeoff between $AIF(\psi, F_\theta, p)$ and $\gamma^*(\psi, F_\theta)$: From (2.3.12) of [3], we know that the influence function of the location estimator specified by ψ is

$$\mathrm{IF}(x, \psi, F_\theta) = \frac{\psi(x-\theta)}{\mathbb{E}_{F_\theta}[\psi^{'}(X-\theta)]},$$

and hence

$$\gamma^*(\psi, F_\theta) = \sup_x \left|\frac{\psi(x-\theta)}{\mathbb{E}_{F_\theta}[\psi^{'}(X-\theta)]}\right|.$$

As a result, if $\psi^{'}(X-\theta)$ is a constant that minimizes $\mathrm{AIF}(\psi, F_\theta, p)$ as discussed in Section V-A1, then $\gamma^*(\psi, F_\theta)$ might go to ∞, especially for those distributions with unbounded support. To achieve a desirable tradeoff between robustness to outliers (i.e., $\gamma^*(\psi, F_\theta)$ is small) and robustness to adversarial attacks (i.e., $\mathrm{AIF}(\psi, F_\theta, p)$ is small), in the following, we characterize the optimal estimator that minimizes $\mathrm{AIF}(\psi, F_\theta, p)$ subject to a constraint on $\gamma^*(\psi, F_\theta)$.

$$\begin{aligned} \min \quad & \mathrm{AIF}(\psi, F, 2) && (16)\\ \text{s.t.} \quad & \gamma^*(\psi, F_\theta) \leq \xi, && (17)\\ & \mathbb{E}_{F_\theta}[\psi(X-\theta)] = 0, && (18)\\ & \psi^{'}(x) \geq 0, \end{aligned}$$

in which constraint (17) implies that $\gamma^*(\psi, F_\theta)$ is upper-bounded by a positive constant ξ, constraint (18) implies that ψ is Fisher consistent, and the last constraint comes from the condition that ψ is monotonic in θ.

For location estimator, $f_\theta(x) = f_0(x-\theta)$, so all quantities in (16) remain the same by assuming $\theta = 0$ [3]. Hence, in the following, we will solve this optimization problem assuming $\theta = 0$. Once the optimal form of ψ for $\theta = 0$ is characterized, we can obtain the estimate of θ by solving $\sum_{n=1}^{N} \psi(x - T_N) = 0$ for the general case when $\theta \neq 0$.

Theorem 2. *The solution to the optimization problem* (16) *has the following structure:*

- $\psi^{'}(x)$ *satisfies*

$$\psi^{'}(x) = \quad (19)$$
$$\begin{cases} \nu^* - \frac{\vartheta_2^* + (\vartheta_1^* - \vartheta_2^*)F_0(x)}{f_0(x)}, & \nu^* f_0(x) > \vartheta_2^* + (\vartheta_1^* - \vartheta_2^*)F_0(x); \\ 0, & \textit{otherwise,} \end{cases}$$

in which the parameters ν^*, $\vartheta_1^* \geq 0$ and $\vartheta_2^* \geq 0$ are parameters chosen to satisfy the following conditions

$$\mathbb{E}_{F_0}[\psi^{'}(X)] = 1, \quad (20)$$
$$\vartheta_1^* \left(\int \psi^{'}(x) F_0(x) dx - \xi\right) = 0, \quad (21)$$
$$\vartheta_2^* \left(\mathbb{E}_{F_0}\left[\int_{-\infty}^{X} \psi^{'}(t) dt\right] - \xi\right) = 0, \quad (22)$$

along with $\int \psi^{'}(x) F_0(x) dx \leq \xi$ *and* $\mathbb{E}_{F_0}\left[\int_{-\infty}^{X} \psi^{'}(t) dt\right] \leq \xi$.

- $\psi(-\infty)$ *is set as* $-\mathbb{E}_{F_0}\left[\int_{-\infty}^{X} \psi^{'}(t) dt\right]$.

Proof. Please see Appendix B for details. □

The condition $\nu^* f_0(x) > \vartheta_1^* F_0(x) + \vartheta_2^*(1 - F_0(x))$ has a natural interpretation. It will trim data points from the tails. In particular, when x is left tail (i.e. F_0 is small), $1 - F_0$ will be close to 1. On the other hand, when x is in the right tail (i.e. $1 - F_0$ is small), F_0 will be close to 1. In these regions, $\psi^{'} = 0$ if the corresponding $f_0(x)$ is small. Figure 1 illustrates the scenario for estimating the mean of Gaussian variables for the case assuming $\vartheta_1^* > \vartheta_2^*$. It is easy to check that, in this example, if $\nu^* > 2\pi(\vartheta_1^* + \vartheta_2^*)$, there exist a and b such that $\psi^{'}(x) = 0$ when $x < a$ or $x > b$. Correspondingly, $\psi(x)$ is given as

$$\psi(x) = \begin{cases} \xi & x \geq b \\ -\xi + \int_a^x \psi^{'}(t)\mathrm{d}t & a < x < b \\ -\xi & x < a \end{cases}.$$

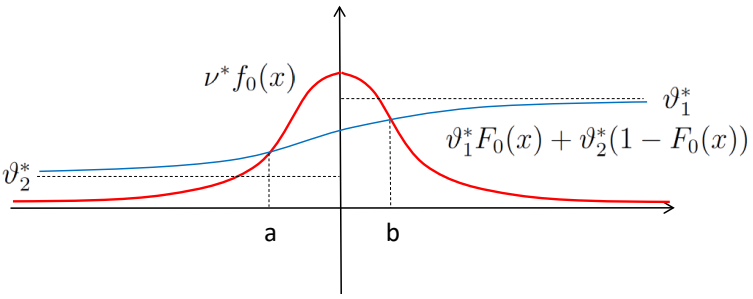

Fig. 1. Gaussian mean example

B. Scale Estimator

We now specialize the results to the scale model where $F_\theta(x) = F_1(x/\theta)$. For this model, it is natural to consider $\psi(x, \theta) = \psi(x/\theta)$ [2], [3]. Similar to the location model, we will first characterize ψ that minimizes $\mathrm{AIF}(\psi, F_\theta, p)$. We will then discuss the tradeoff between the robustness to outliers and robustness to adversarial attacks, and will characterize the optimal ψ that achieves this tradeoff.

For the case with $p=1$, using (13), we obtain

$$\begin{aligned}\text{AIF}(\psi,\mathbf{x},1) &= \frac{\left|NT_N\psi^{'}(x_{n^*}/T_N)\right|}{\left|\sum_{n=1}^{N}x_n\psi^{'}(x_n/T_N)\right|} \xrightarrow{\text{a.s.}} \frac{\max_x\left|\theta\psi^{'}(x/\theta)\right|}{|\mathbb{E}_\theta[X\psi^{'}(X/\theta)]|}\\ &:= \text{AIF}(\psi,F_\theta,1).\end{aligned}$$

For $p>1$, using (13), we have

$$\begin{aligned}\text{AIF}(\psi,\mathbf{x},p) &= \frac{\left(\frac{1}{N}\sum_{n=1}^{N}\left|\psi^{'}(x_n/T_N)\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N}x_n/T_N\psi^{'}(x_n/T_N)\right|}\\ &\xrightarrow{a.s.} \frac{\left(\mathbb{E}_\theta\left[\left|\psi^{'}(X/\theta)\right|^{\frac{p}{p-1}}\right]\right)^{\frac{p-1}{p}}}{|\mathbb{E}_\theta\left[X/\theta\psi^{'}(X/\theta)\right]|}\\ &:= \text{AIF}(\psi,F_\theta,p).\end{aligned}$$

Since in scale model $F_\theta(x)=F_1(x/\theta)$, we have $f_\theta(x)=f_1\left(\frac{x}{\theta}\right)\frac{1}{\theta}$, and hence

$$\text{AIF}(\psi,F_\theta,p)=\frac{\left(\mathbb{E}_{F_1}\left[\left|\psi^{'}(X)\right|^{\frac{p}{p-1}}\right]\right)^{\frac{p-1}{p}}}{|\mathbb{E}_{F_1}\left[X\psi^{'}(X)\right]|}:=\text{AIF}(\psi,F_1,p).$$

For $p=2$, we have

$$\text{AIF}(\psi,F_1,2)=\frac{\left(\mathbb{E}_{F_1}\left[\psi^{'}(X)^2\right]\right)^{\frac{1}{2}}}{|\mathbb{E}_{F_1}\left[X\psi^{'}(X)\right]|}. \tag{23}$$

1) Minimizing $AIF(\psi,F_\theta,p)$: In the following, among Fisher consistent estimators, we aim to design $\psi^{'}$ that minimizes $\text{AIF}(\psi,F_1,2)$.

Theorem 3. *The optimal ψ that minimizes $AIF(\psi,F_1,2)$ has the following structure:*

- *For x in the range of $f_1(x)$, $\psi^{'}$ satisfies*
$$\psi^{'}(x)=\frac{x}{\mathbb{E}_{F_1}[X^2]}.$$
- *$\psi(-\infty)$ is chosen as*
$$\psi(-\infty)=-\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right].$$

With this choice of $\psi(x)$, the minimal value of $AIF(\psi,F_1,2)$ is $1/\sqrt{\mathbb{E}_{F_1}[X^2]}$.

Proof. Please see Appendix C for details. □

We note that for scale estimator $\frac{\partial\psi}{\partial\theta}=-\psi^{'}(x/\theta)x/\theta^2$, hence for this particular choice of $\psi^{'}$ in Theorem 3, $\frac{\partial\psi}{\partial\theta}=-x^2/(\theta^3\mathbb{E}_{F_1}[X^2])$, which means $\psi(x,\theta)$ is monotone in θ. This ensures that the obtained $\psi(x)$ satisfies the regularity conditions [2] mentioned at the beginning of this section.

2) Tradeoff between $AIF(\psi,F_\theta,p)$ and $\gamma^*(\psi,F_\theta)$: Similar to the location estimation case, we can also design ψ to minimize $\text{AIF}(\psi,F_\theta,p)$ with a constraint on $\gamma^*(\psi,F_\theta)$. From (2.3.17) of [3], we know that for scale estimators

$$\text{IF}(x,\psi,F_\theta)=\frac{\psi(x/\theta)\theta}{\mathbb{E}_{F_\theta}[X/\theta\psi^{'}(X/\theta)]}.$$

To facilitate the analysis, we will focus on ψ that is monotonic. Since in scale model, $\psi(x,\theta)=\psi(x/\theta)$, we can simply focus on the case of $\theta=1$. Hence, we will solve the following optimization problem to strike a desirable tradeoff between robustness against outliers and robustness against adversarial attacks.

$$\min \quad \frac{\mathbb{E}_{F_1}\left[\psi^{'}(X)^2\right]}{\left(\mathbb{E}_{F_1}\left[X\psi^{'}(X)\right]\right)^2}, \tag{24}$$

$$\text{s.t.} \quad \gamma^*(\psi,F_1)=\sup_x\left|\frac{\psi(x)}{\mathbb{E}_{F_1}[X\psi^{'}(X)]}\right|\le\xi, \tag{25}$$

$$\mathbb{E}_{F_1}[\psi]=0, \tag{26}$$

$$\psi^{'}(x)\ge 0. \tag{27}$$

Here, constraint (25) is a constraint on the outliers influence, (26) implies that ψ is Fisher consistent.

Theorem 4. *The solution to* (24) *has the following structure:*

- *$\psi^{'}$ has the following form*

$$\psi^{'}(x)= \tag{28}$$
$$\begin{cases}\nu^*x-\frac{\vartheta_2^*+(\vartheta_1^*-\vartheta_2^*)F_1(x)}{f_1(x)}, & \nu^*xf_1(x)>\vartheta_2^*+(\vartheta_1^*-\vartheta_2^*)F_1(x);\\ 0, & \textit{otherwise},\end{cases}$$

in which ν^*, $\vartheta_1^*\ge 0$ and $\vartheta_2^*\ge 0$ are chosen to satisfy

$$\begin{aligned}\mathbb{E}_{F_1}[X\psi^{'}(X)] &= 1,\\ \vartheta_1^*\left(\int_{-\infty}^{\infty}\psi^{'}(x)F_1(x)dx-\xi\right) &= 0,\\ \vartheta_2^*\left(\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right]-\xi\right) &= 0,\end{aligned}$$

along with $\int_{-\infty}^{\infty}\psi^{'}(x)F_1(x)dx\le\xi$ and $\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right]\le\xi$.

- *$\psi(-\infty)$ is set to be $-\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right]$.*

Proof. The proof follows similar strategy as that of the proof of Theorem 2 and 3. Details can be found in Appendix D. □

Similar to the location estimator case, the condition $\nu^*xf_1(x)>\vartheta_1^*F_1(x)+\vartheta_2^*(1-F_1(x))$ will limit the influences of data points at the tails.

VI. Extension: L-estimator

In this section, we briefly discuss how to extend the analysis above to other class of estimators. We will use L-estimator as an example. L-estimator has the following form [2], [3]:

$$T_N(\mathbf{x})=\sum_{n=1}^{N}a_nx_{(n)},$$

where $x_{(1)} \leq \cdots \leq x_{(N)}$ are the ordered sequence of $\mathbf{x}$, and a_n's are coefficients. For example, for location estimator, a natural choice of a_n is

$$a_n = \frac{\int_{(n-1)/N}^{n/N} h(t)\mathrm{d}t}{\int_0^1 h(t)\mathrm{d}t}, \tag{29}$$

for a given function $h(t)$ such that $\int_0^1 h(t)\mathrm{d}t \neq 0$. For example, setting $h(t) = \delta(t - 1/2)$ leads to the median estimator.

We first look at the given sample scenario. Let $\tilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$, and let $\tilde{x}_{(1)} \leq \cdots \leq \tilde{x}_{(N)}$ be the ordered sequence of $\tilde{\mathbf{x}}$. Hence,

$$T_N(\mathbf{x} + \Delta\mathbf{x}) = \sum_{n=1}^{N} a_n \tilde{x}_{(n)}. \tag{30}$$

For general $\Delta\mathbf{x}$, the ordering of $\mathbf{x}+\Delta\mathbf{x}$ may not necessarily be the same as the ordering of $\mathbf{x}$. For example, $\tilde{x}_{(1)}$ might come from $x_{(2)}$, i.e., $\tilde{x}_{(1)} = x_{(2)} + \Delta(x_{(2)})$. This possibility could make the following analysis messy. However, it is easy to see that when η is sufficiently small (more specifically, when $N^{1/p}\eta \leq 1/2 \min_{x_i \neq x_j} |x_i - x_j|$), the ordering of $\mathbf{x} + \Delta\mathbf{x}$ be the same as $\mathbf{x}$ for all $\Delta\mathbf{x}$'s that satisfy the constraint (1). As the result, for the purpose of charactering AIF (which involves making $\eta \downarrow 0$), we can limit (30) to the following form

$$T_N(\mathbf{x} + \Delta\mathbf{x}) = \sum_{n=1}^{N} a_n (x_{(n)} + \Delta(x_{(n)})).$$

Hence

$$T_N(\mathbf{x} + \Delta\mathbf{x}) - T_N(\mathbf{x}) = \sum_{n=1}^{N} a_n \Delta x_{(n)},$$

and (2) becomes

$$\begin{aligned} \min \quad & -\sum_{n=1}^{N} a_n \Delta x_{(n)}, \\ \text{s.t.} \quad & \frac{1}{N} ||\Delta\mathbf{x}||_p^p \leq \eta^p. \end{aligned}$$

Using the exactly same approach as those in the proof of Theorem 1, we have the following characterization. For $p = 1$, let $n^* = \arg\max_n |a_n|$,

$$\Delta x^*_{(n^*)} = \text{sign}\{a_{n^*}\} N\eta,$$

and $\Delta x^*_{(n)} = 0, \forall n \neq n^*$. Hence,

$$\text{AIF}(T_N, \mathbf{x}, 1) = N |a_{n^*}|.$$

For $p > 1$, we have

$$\Delta x^*_{(n)} = \frac{|a_n|^{1/(p-1)} (N)^{1/p}}{(\sum |a_n|^{p/(p-1)})^{1/p}} \text{sign}(a_n)\eta.$$

Hence,

$$\begin{aligned} \text{AIF}(\psi, \mathbf{x}, p) &= \sum a_n \frac{|a_n|^{1/(p-1)} (N)^{1/p}}{(\sum |a_n|^{p/(p-1)})^{1/p}} \text{sign}(a_n) \\ &= \frac{\sum_{n=1}^{N} |a_n|^{p/(p-1)}}{\left(\frac{1}{N} \sum_{n=1}^{N} |a_n|^{p/(p-1)}\right)^{1/p}}. \end{aligned} \tag{31}$$

When $p = 2$, this can be simplified to

$$\text{AIF}(T_N, \mathbf{x}, 2) = \frac{\sqrt{N} \sum_{n=1}^{N} a_n^2}{\sqrt{\sum_{n=1}^{N} a_n^2}} = \sqrt{N \sum_{n=1}^{N} a_n^2}. \tag{32}$$

For example, for α-trimmed estimator [3] defined by

$$T_N^{\alpha}(\mathbf{x}) = \frac{1}{N - 2\lfloor \alpha N \rfloor} \sum_{n=\lfloor \alpha N \rfloor + 1}^{N - \lfloor \alpha N \rfloor} x_{(n)},$$

for a given parameter $0 < \alpha < 1/2$. For this α-trimmed estimator, using (31), we obtain

$$\text{AIF}(T_N^{\alpha}, \mathbf{x}, p) = \frac{N^{1/p}}{(N - 2\lfloor \alpha N \rfloor)^{1/p}}.$$

If a_ns are chosen as (29), then (32) simplifies to

$$\begin{aligned} \text{AIF}(T_N, \mathbf{x}, 2) &= \sqrt{\frac{\frac{1}{N} \sum_{n=1}^{N} \left(\int_{(n-1)/N}^{n/N} h(t)\mathrm{d}t\right)^2}{\left(\frac{1}{N} \int_0^1 h(t)\mathrm{d}t\right)^2}} \\ &\geq \sqrt{\frac{\left(\frac{1}{N} \sum_{n=1}^{N} \int_{(n-1)/N}^{n/N} h(t)\mathrm{d}t\right)^2}{\left(\frac{1}{N} \int_0^1 h(t)\mathrm{d}t\right)^2}} \\ &\geq 1, \end{aligned}$$

in which the first inequality is due to Jensen's inequality, and both inequalities become equality when $a_n = \int_{(n-1)/N}^{n/N} h(t)\mathrm{d}t$ is a constant in n, i.e., $a_n = 1/N$ and the estimator becomes the empirical mean.

VII. Numerical Examples

In this section, we provide numerical examples to illustrate results obtained.

We consider location estimation and illustrate the optimal estimator obtained in Theorem 2 for the case when f_0 is exponential random variable $f_0(x) = e^{-x}, x \geq 0$, hence f_θ is shifted exponential random variable $f_\theta = e^{-(x-\theta)}, x \geq \theta$ and the goal is to estimate θ. As the exponential random variable has a unbounded support, choosing ψ' to be a constant, which minimizes AIF, will lead to an infinite IF. Hence, we use Theorem 2 to characterize the optimal ψ that minimizes AIF while satisfying the condition that IF $\leq \xi$.

For this particular class of distribution, the condition $\nu^* f_0(x) > \vartheta_1^* F_0(x) + \vartheta_2^*(1 - F_0(x))$ becomes $0 \leq x < a$ with the parameter a chosen as

$$e^{-a} = \frac{\vartheta_1^*}{\nu^* + \vartheta_1^* - \vartheta_2^*}. \tag{33}$$

Hence we have

$$\psi'(x) = \begin{cases} \nu^* + \vartheta_1^* - \vartheta_2^* - \vartheta_1^* e^x, & 0 \leq x < a; \\ 0, & \text{otherwise}, \end{cases}$$

for which the parameters $\nu^*, \vartheta_1^*, \vartheta_2^*$ are chosen to satisfy the conditions specified in Theorem 2. After tedious calculation, conditions (20) - (22) can be simplified to

$$\begin{aligned} &(\nu^* + \vartheta_1^* - \vartheta_2^*)(1 - e^{-a}) - \vartheta_1^* a = 1, \\ &\vartheta_1^*((\nu^* + \vartheta_1^* - \vartheta_2^*)(a - 1 + e^{-a}) - \vartheta_1^*(e^a - 1) + a\vartheta_1^* - \xi) = 0, \\ &\vartheta_2^*((\nu^* + \vartheta_1^* - \vartheta_2^*)(1 - e^{-a}) - \vartheta_1^* a - \xi) = 0. \end{aligned}$$

From here, we know that if $\xi > 1$, $\vartheta_2^* = 0$, using this fact along with (33), we have that the conditions are simplified to

$$\begin{aligned} \nu^* - a\vartheta_1^* &= 1, \\ 2a\vartheta_1^* + (a-2)\nu^* &= \xi, \\ \vartheta_2^* &= 0. \end{aligned}$$

Using these, we can express ν^* and ϑ_1^* in terms of a:

$$\begin{aligned} \nu^* &= \frac{\xi + 2}{a}, \\ \vartheta_1^* &= \frac{\xi + 2 - a}{a^2}. \end{aligned}$$

Finally, for any given $\xi > 1$, the value of a can be determined by (33), which is simplified to

$$e^{-a} = \frac{\vartheta_1^*}{\nu^* + \vartheta_1^* - \vartheta_2^*} = \frac{\xi + 2 - a}{(\xi + 1)a + \xi + 2}. \tag{34}$$

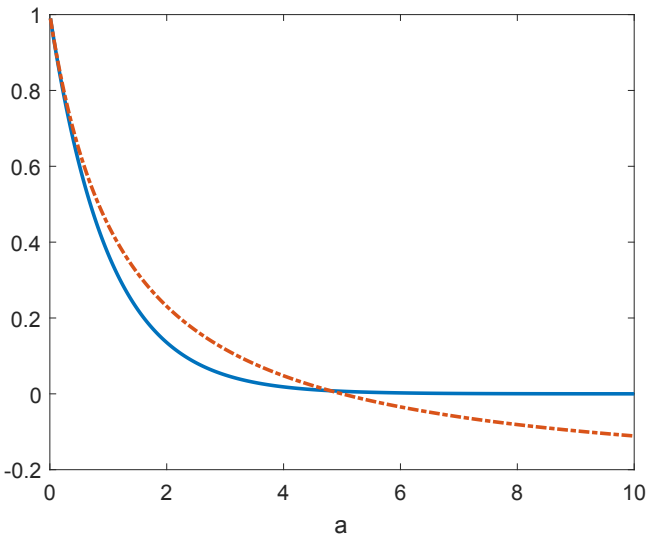

Fig. 2. The solution of a

It is easy to check that, for any given $\xi > 1$, there is always a unique positive solution to (34). For example, Figure 2 illustrates the solution for a when $\xi = 3$. In this figure, the dotted curve is the right side of (34) and the solid curve is the left side of (34). From the figure, we know that these two curves have two intersections $a = 0$ and $a = 4.8$. With these parameters, we know that

$$\psi'(x) = \begin{cases} 1.0417 - 0.0087e^x, & 0 \leq x \leq 4.8; \\ 0, & \text{otherwise}, \end{cases} \tag{35}$$

hence the optimal ψ is

$$\psi'(x) = \begin{cases} \xi, & x \geq 4.8; \\ 1.0417x - 0.0087(e^x - 1) - 1, & 0 \leq x \leq 4.8. \end{cases}$$

Figure 3 illustrates the obtained $\psi(x)$ for the case with $\xi = 3$.

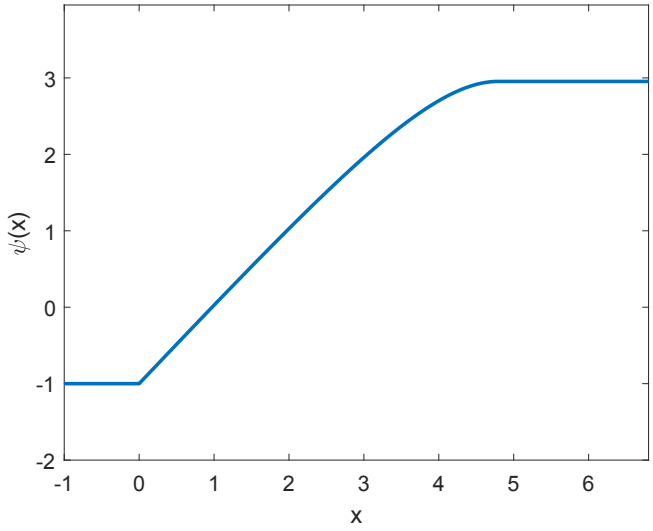

Fig. 3. ψ that minimizes AIF when IF ≤ 3.

Figure 4 illustrates the tradeoff curve between AIF and IF. We obtain this curve by solving (34) and other parameters using different values of ξ. As we can see from the curve, as ξ increases, AIF decreases. Furthermore, the value of AIF converges to 1, the lower bound established in Section V-A1.

VIII. Conclusion

Motivated by recent data analytics applications, we have studied adversarial robustness of robust estimators. We have introduced the concept of AIF to quantify an estimator's sensitivity to such adversarial attacks and have provided an approach to characterize AIF for given robust estimator. We have further designed optimal estimators that minimize AIF. From this characterization, we have identified a tradeoff between AIF and IF, and have designed estimators that strike a desirable tradeoff between these two quantities. We note that AIF only captures the impact of vanishingly small corruptions. It is of interest to investigate the impact of non-vanishing corruptions and its connection with AIF in the future.

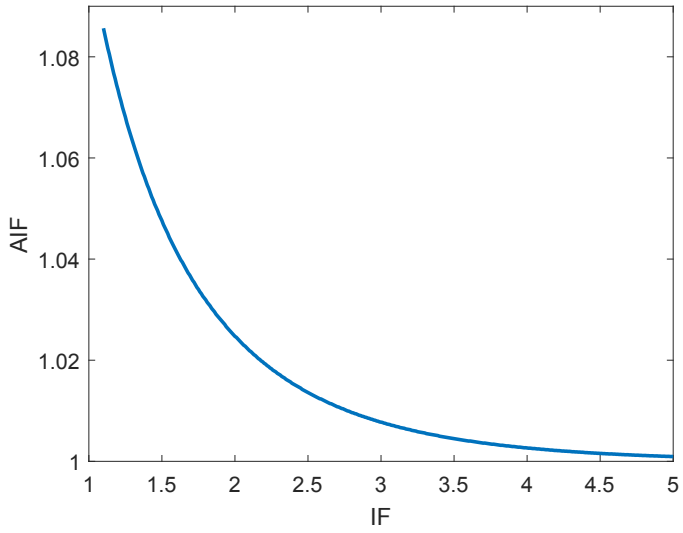

Fig. 4. Tradeoff between AIF and IF of location estimator for exponential random variables.

Appendix A
Proof of Theorem 1

From (5), we know that T_N and $\mathbf{x}$ satisfy

$$\sum_{n=1}^{N} \psi(x_n, T_N) = 0.$$

Hence, we have

$$\frac{\partial}{\partial x_n} T_N = \frac{-\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}}{\sum_{n=1}^{N} \frac{\partial}{\partial \theta}[\psi]_{x=x_n,\theta=T_N}}. \tag{36}$$

Based on Taylor expansion, we have

$$\begin{aligned} &T_N(\mathbf{x}+\Delta\mathbf{x}) - T_N(\mathbf{x}) \\ &\quad = \sum_{n=1}^{N} \Delta x_n \frac{\partial}{\partial x_n} T_N + \text{higher order terms}. \end{aligned}$$

When η is small, the adversary can solve the following problem and obtain an $o(\eta)$ optimal solution

$$\begin{aligned} \min_{\Delta\mathbf{x}} \quad & -\sum_{n=1}^{N} \Delta x_n c_n, \\ \text{s.t.} \quad & ||\Delta\mathbf{x}||_p^p \le N\eta^p, \end{aligned} \tag{37}$$

in which

$$c_n := \frac{\partial}{\partial x_n} T_N.$$

For $p = 1$, this is a linear programing problem, whose solution is simple. In particular, let $n^* = \arg\max_n |\frac{\partial}{\partial x_n} T_N|$, which is the same as $\arg\max_n \left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|$ due to (36), it is easy to check that we have

$$\Delta x_{n^*}^* = \text{sign}\left\{\frac{\partial}{\partial x_{n^*}} T_N\right\} N\eta,$$

and $\Delta x_n^* = 0, \forall n \neq n^*$. Hence,

$$\text{AIF}(\psi, \mathbf{x}, 1) = N\left|\frac{\partial}{\partial x_{n^*}} T_N\right| = \frac{\left|N\frac{\partial}{\partial x}[\psi]_{x=x_{n^*},\theta=T_N}\right|}{\left|\sum_{n=1}^{N} \frac{\partial}{\partial \theta}[\psi]_{x=x_n,\theta=T_N}\right|}.$$

For $\infty > p > 1$, (37) is a convex optimization problem. To solve this, we form Lagrange

$$\mathcal{L}(\Delta\mathbf{x}, \lambda) = -\sum_{n=1}^{N} \Delta x_n c_n + \lambda\left(||\Delta\mathbf{x}||_p^p - N\eta^p\right).$$

The corresponding optimality conditions are:

$$\begin{aligned} -c_n + \lambda^* p \text{sign}(\Delta x_n^*)|\Delta x_n^*|^{p-1} &= 0, \forall n \\ \lambda^* &\ge 0, \\ \lambda^*(||\Delta\mathbf{x}^*||_p^p - N\eta^p) &= 0. \end{aligned} \tag{38}$$

From (38), we know that $\lambda^* \neq 0$, hence

$$||\Delta\mathbf{x}^*||_p^p = N\eta^p, \tag{39}$$

and

$$\text{sign}(\Delta x_n^*)|\Delta x_n^*|^{p-1} = \frac{c_n}{\lambda^* p}. \tag{40}$$

From (40) and the fact that $\lambda^* p$ is positive, we know $\text{sign}(\Delta x_n^*) = \text{sign}(c_n)$, and hence we have

$$|\Delta x_n^*|^{p-1} = \frac{|c_n|}{\lambda^* p},$$

which can be simplified further to

$$\Delta x_n^* = \left(\frac{|c_n|}{\lambda^* p}\right)^{1/(p-1)} \text{sign}(c_n).$$

Combining these with (39), we obtain the value of λ^*:

$$\lambda^* = \frac{1}{p}\left(\frac{\sum_{n=1}^{N} |c_n|^{p/(p-1)}}{N\eta^p}\right)^{(p-1)/p}.$$

As the result, we have

$$\Delta x_n^* = \frac{|c_n|^{1/(p-1)}(N)^{1/p}}{(\sum |c_n|^{p/(p-1)})^{1/p}} \text{sign}(c_n)\eta.$$

Hence,

$$\begin{aligned} \text{AIF}(\psi, \mathbf{x}, p) &= \sum c_n \frac{|c_n|^{1/(p-1)}(N)^{1/p}}{(\sum |c_n|^{p/(p-1)})^{1/p}} \text{sign}(c_n) \\ &= \frac{\sum_{n=1}^{N} |c_n|^{p/(p-1)}}{\left(\frac{1}{N}\sum |c_n|^{p/(p-1)}\right)^{1/p}}. \end{aligned}$$

Using (36), we can further simplify the expression to

$$\text{AIF}(\psi, \mathbf{x}, p) = \frac{\left(\frac{1}{N}\sum_{n=1}^{N} \left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}}}{\left|\frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \theta}[\psi]_{x=x_n,\theta=T_N}\right|}. \tag{41}$$

For $p=\infty$, as $N^{1/p}\overset{p\to\infty}{\to}1$, (37) can be written as

$$\begin{aligned}\min_{\Delta\mathbf{x}} \quad & -\sum_{n=1}^{N}\Delta x_n c_n,\\ \text{s.t.} \quad & ||\Delta\mathbf{x}||_\infty\le\eta.\end{aligned}\tag{42}$$

It is easy to see that the optimal $\Delta x_n^*=\eta\text{sign}\{c_n\}$. Hence,

$$\text{AIF}(\psi,\mathbf{x},p)=\sum_{n=1}^{N}|c_n|=\frac{\sum_{n=1}^{N}\left|\frac{\partial}{\partial x}[\psi]_{x=x_n,\theta=T_N}\right|}{\left|\sum_{n=1}^{N}\frac{\partial}{\partial\theta}[\psi]_{x=x_n,\theta=T_N}\right|},\tag{43}$$

which is the limit of (41) as $p\to\infty$.

Appendix B
Proof of Theorem 2

As $\psi^{'}(x)\ge 0$, we have $\mathbb{E}_{F_0}[\psi^{'}(X)]>0$, and $\sup_x|\psi(x)|$ is either $\psi(\infty)$ or $-\psi(-\infty)$. Hence for $p=2$, the optimization problem (16) is equivalent to

$$\begin{aligned}\min \quad & \frac{\mathbb{E}_{F_0}[\psi^{'}(X)^2]}{(\mathbb{E}_{F_0}[\psi^{'}(X)])^2}\\ \text{s.t.} \quad & \frac{\psi(-\infty)+\int_{-\infty}^{\infty}\psi^{'}(x)\mathrm{d}x}{\mathbb{E}_{F_0}[\psi^{'}(X)]}\le\xi,\\ & \frac{-\psi(-\infty)}{\mathbb{E}_{F_0}[\psi^{'}(X)]}\le\xi,\\ & \psi(-\infty)+\mathbb{E}_{F_0}\left[\int_{-\infty}^{X}\psi^{'}(t)\mathrm{d}t\right]=0,\\ & \psi^{'}\ge 0.\end{aligned}$$

As the objective function does not involve $\psi(-\infty)$, we can first solve

$$\begin{aligned}\min \quad & \frac{\mathbb{E}_{F_0}[\psi^{'}(X)^2]}{(\mathbb{E}_{F_0}[\psi^{'}(X)])^2},\\ \text{s.t.} \quad & \frac{-\mathbb{E}_{F_0}\left[\int_{-\infty}^{X}\psi^{'}(t)\mathrm{d}t\right]+\int_{-\infty}^{\infty}\psi^{'}(x)\mathrm{d}x}{\mathbb{E}_{F_0}[\psi^{'}(X)]}\le\xi,\\ & \frac{\mathbb{E}_{F_0}\left[\int_{-\infty}^{X}\psi^{'}(t)\mathrm{d}t\right]}{\mathbb{E}_{F_0}[\psi^{'}(X)]}\le\xi,\\ & \psi^{'}(x)\ge 0.\end{aligned}$$

After obtaining the solution, we can simply set $\psi(-\infty)=-\mathbb{E}_{F_0}\left[\int_{-\infty}^{X}\psi^{'}(t)\mathrm{d}t\right]$ to make ψ Fisher consistent.

To simplify the notation, in the remainder of the proof, we will use $g(x)$ to denote $\psi^{'}(x)$. We now further simplify the optimization problem. First, we have

$$\begin{aligned}\mathbb{E}_{F_0}\left[\int_{-\infty}^{X}g(t)\mathrm{d}t\right] &= \int_{-\infty}^{\infty}f_0(x)\left[\int_{-\infty}^{x}g(t)\mathrm{d}t\right]\mathrm{d}x\\ &= \int_{-\infty}^{\infty}g(t)\left[\int_{t}^{\infty}f_0(x)\mathrm{d}x\right]\mathrm{d}t\\ &= \int_{-\infty}^{\infty}g(t)\left[1-F_0(t)\right]\mathrm{d}t.\end{aligned}\tag{44}$$

Coupled with the fact that $g(x)\ge 0$ and $f_0(x)\ge 0$, the optimization above is equivalent to

$$\begin{aligned}\min \quad & \frac{\int_{-\infty}^{\infty}g^2(x)f_0(x)\mathrm{d}x}{\left(\int_{-\infty}^{\infty}g(x)f_0(x)\mathrm{d}x\right)^2},\\ \text{s.t.} \quad & \int_{-\infty}^{\infty}g(x)F_0(x)\mathrm{d}x\le\xi\int_{-\infty}^{\infty}g(x)f_0(x)\mathrm{d}x,\\ & \int_{-\infty}^{\infty}g(x)\left[1-F_0(x)\right]\mathrm{d}x\le\xi\int_{-\infty}^{\infty}g(x)f_0(x)\mathrm{d}x,\\ & g(x)\ge 0.\end{aligned}$$

It is clear that the optimization problem is scale invariant in the sense that if $g^*(x)$ is a solution to this problem, then for any positive constant c, $cg^*(x)$ is also a solution to this problem. As a result, without loss of generality, we can assume $\int_{-\infty}^{\infty}g(x)f_0(x)\mathrm{d}x=1$. Using this, we can further simplify the optimization problem to

$$\begin{aligned}\min \quad & \frac{1}{2}\int_{-\infty}^{\infty}g^2(x)f_0(x)\mathrm{d}x,\\ \text{s.t.} \quad & \int_{-\infty}^{\infty}g(x)f_0(x)\mathrm{d}x=1,\\ & \int_{-\infty}^{\infty}g(x)F_0(x)\mathrm{d}x\le\xi,\\ & \int_{-\infty}^{\infty}g(x)\left[1-F_0(x)\right]\mathrm{d}x\le\xi,\\ & g(x)\ge 0.\end{aligned}$$

To solve this convex functional minimization problem, we first form the Lagrangian function

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\int_{-\infty}^{\infty}g^2(x)f_0(x)\mathrm{d}x+\nu\left(-\int_{-\infty}^{\infty}g(x)f_0(x)\mathrm{d}x+1\right)\\ &\quad -\lambda(x)g(x)+\vartheta_1\left(\int_{-\infty}^{\infty}g(x)F_0(x)\mathrm{d}x-\xi\right)\\ &\quad +\vartheta_2\left(\int_{-\infty}^{\infty}g(x)\left[1-F_0(x)\right]\mathrm{d}x-\xi\right).\end{aligned}$$

Let $H=\frac{1}{2}g^2(x)f_0(x)-\nu g(x)f_0(x)+\vartheta_1 g(x)F_0(x)+\vartheta_2 g(x)(1-F_0(x))-\lambda(x)g(x)$. As no derivative $g^{'}(x)$ is involved in H, the optimality condition Euler-Lagrange equation [15]

$$\frac{\partial H}{\partial g}-\frac{d}{dx}\left(\frac{\partial H}{\partial g^{'}}\right)=0$$

simplifies to

$$g^*(x)f_0(x)-\nu^*f_0(x)+\vartheta_2^*+(\vartheta_1^*-\vartheta_2^*)F_0(x)-\lambda^*(x)=0,\tag{45}$$

in which the parameters $\vartheta_1^*\ge 0$, $\vartheta_2^*\ge 0$, $\lambda^*(x)\ge 0$

satisfy [14]

$$\begin{aligned}&\int_{-\infty}^{\infty} g^*(x)f_0(x)\mathrm{d}x = 1,\\ &\vartheta_1^*\left(\int_{-\infty}^{\infty} g^*(x)F_0(x)\mathrm{d}x - \xi\right) = 0,\\ &\vartheta_2^*\left(\int_{-\infty}^{\infty} g^*(x)\left[1-F_0(x)\right]\mathrm{d}x - \xi\right) = 0,\\ &\lambda^*(x)g(x) \geq 0.\end{aligned} \tag{46}$$

From (45), for x in the range of $f_0(x)$, we have

$$g^*(x) = \frac{\lambda^*(x) + \nu^* f_0(x) - \vartheta_2^* - (\vartheta_1^* - \vartheta_2^*)F_0(x)}{f_0(x)}.$$

Combining this with the condition (46), we know that if $\nu^* f_0(x) - \vartheta_2^* - (\vartheta_1^* - \vartheta_2^*)F_0(x) > 0$, then $\lambda^*(x) = 0$. On the other hand, if $\nu^* f_0(x) - \vartheta_2^* - (\vartheta_1^* - \vartheta_2^*)F_0(x) < 0$, then $g^*(x) = 0$. As a result, we have

$$g^*(x) = \begin{cases} \nu^* - \frac{\vartheta_2^* + (\vartheta_1^* - \vartheta_2^*)F_0(x)}{f_0(x)}, & \nu^* f_0(x) > \vartheta_2^* + (\vartheta_1^* - \vartheta_2^*)F_0(x);\\ 0, & \text{otherwise,}\end{cases}$$

which completes the proof.

Appendix C
Proof of Theorem 3

First of all, minimizing (23) is same as solving

$$\min \frac{\mathbb{E}_{F_1}\left[\psi^{'}(X)^2\right]}{\left(\mathbb{E}_{F_1}\left[X\psi^{'}(X)\right]\right)^2}, \tag{47}$$

$$\text{s.t. } \mathbb{E}_{F_1}[\psi(X)] = \psi(-\infty) + \mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)\mathrm{d}t\right] = 0, \tag{48}$$

in which the condition $\mathbb{E}_{F_1}[\psi(X)] = 0$ ensures that the estimator is Fisher consistent.

As $\psi(-\infty)$ does not appear in the objective function, we can solve (47) without the constraint (48) first. After that, we can simply set

$$\psi(-\infty) = -\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)\mathrm{d}t\right]$$

so that the constraint (48) will be satisfied. Furthermore, similar to the proof of Theorem 2, to simplify the notation, we will use $g(x)$ to denote $\psi^{'}(x)$. It is clear from (47) that the cost function is scale-invariant. Hence, without loss of generality, we can assume $\left(\mathbb{E}_{F_1}\left[Xg(X)\right]\right)^2 = 1$, for which we can further focus on $\mathbb{E}_{F_1}\left[Xg(X)\right] = 1$. Combining all these together, the optimization problem can be converted to

$$\begin{aligned}\min \quad & \frac{1}{2}\int_{-\infty}^{\infty} g^2(x)f_1(x)\mathrm{d}x,\\ \text{s.t.} \quad & \int_{-\infty}^{\infty} xg(x)f_1(x)\mathrm{d}x = 1.\end{aligned}$$

For this convex calculus of variations problem, we form Lagrange function

$$\mathcal{L} = \int_{-\infty}^{\infty}\frac{1}{2}g^2(x)f_1(x)\mathrm{d}x + \nu\left(-\int_{-\infty}^{\infty} xg(x)f_1(x)\mathrm{d}x - 1\right).$$

The corresponding Euler-Lagrange equation can be simplified to

$$g^*(x)f_1(x) - \nu^* x f_1(x) = 0, \tag{49}$$

and the optimal value of ν^* is selected to satisfy the condition

$$\int_{-\infty}^{\infty} xg^*(x)f_1(x)\mathrm{d}x = 1. \tag{50}$$

From (49), we know that in the range of X where $f_1(x) > 0$, $g^*(x) = \nu^* x$. Plugging this into (50), we obtain

$$\nu^* = \frac{1}{\int_{-\infty}^{\infty} x^2 f_1(x)\mathrm{d}x}.$$

As the result, for x in the range of $f_1(x)$, the optimal $g^*(x)$ is

$$g^*(x) = \frac{x}{\mathbb{E}_{F_1}[X^2]},$$

and $\psi(-\infty) = -\mathbb{E}_{F_1}\left[\int_{-\infty}^{X} g(t)\mathrm{d}t\right]$.

Appendix D
Proof of Theorem 4

Following the same strategy as those in the proof of Theorem 2, we can first solve the following problem

$$\begin{aligned}\min \quad & \frac{\mathbb{E}_{F_1}\left[\psi^{'}(X)^2\right]}{\left(\mathbb{E}_{F_1}\left[X\psi^{'}(X)\right]\right)^2},\\ \text{s.t.} \quad & -\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right] + \int_{-\infty}^{\infty}\psi^{'}(t)\mathrm{d}t\\ & \qquad \leq \xi\left|\mathbb{E}_{F_1}[X\psi^{'}(X)]\right|,\\ & \mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right] \leq \xi\left|\mathbb{E}_{F_1}[X\psi^{'}(X)]\right|,\\ & \psi^{'}(x) \geq 0,\end{aligned}$$

and then set $\psi(-\infty) = -\mathbb{E}_{F_1}\left[\int_{-\infty}^{X}\psi^{'}(t)dt\right]$ to satisfy the Fisher consistent constraint (26).

Now, we consider two different cases depending on whether $\mathbb{E}_{F_1}[X\psi^{'}(X)]$ is positive or negative. In the following, to simplify notation, we will use $g(x)$ to denote $\psi^{'}(x)$.

We will solve the case with $\mathbb{E}_{F_1}[Xg(X)] > 0$ in detail. The case $\mathbb{E}_{F_1}[Xg(X)] < 0$ can be solved in the similar manner. With $\mathbb{E}_{F_1}[Xg(X)] > 0$, the optimization problem is same as

$$\begin{aligned}\min \quad & \frac{\mathbb{E}_{F_1}\left[g^2(X)\right]}{\left(\mathbb{E}_{F_1}\left[Xg(X)\right]\right)^2},\\ \text{s.t.} \quad & -\mathbb{E}_{F_1}\left[\int_{-\infty}^{X} g(t)dt\right]+\int_{-\infty}^{\infty} g(t)\mathrm{d}t \le \xi\mathbb{E}_{F_1}[Xg(X)],\\ & \mathbb{E}_{F_1}\left[\int_{-\infty}^{X} g(t)dt\right] \le \xi\mathbb{E}_{F_1}[Xg(X)],\\ & g(x)\ge 0.\end{aligned}$$

Similar to the optimization problems in Theorem 2 and 3, the optimization problem is scale-invariant, and hence without loss of generality, we can focus on $\mathbb{E}_{F_1}[Xg(X)] = 1$. Furthermore, similar to (44), we have $\mathbb{E}_{F_1}\left[\int_{-\infty}^{X} g(t)\mathrm{d}t\right] = \int_{-\infty}^{\infty} g(t)\left[1-F_1(t)\right]\mathrm{d}t$. The problem is then converted to

$$\begin{aligned}\min \quad & \int_{-\infty}^{\infty} g^2(x)f_1(x)\mathrm{d}x,\\ \text{s.t.} \quad & \int_{-\infty}^{\infty} xg(x)f_1(x)\mathrm{d}x = 1,\\ & \int_{-\infty}^{\infty} g(x)F_1(x)\mathrm{d}x \le \xi,\\ & \int_{-\infty}^{\infty} g(x)[1-F_1(x)]\mathrm{d}x \le \xi,\\ & g(x)\ge 0.\end{aligned}$$

To solve this convex functional minimization problem, we first form the Lagrangian function

$$\begin{aligned}\mathcal{L} \quad =& \frac{1}{2}\int_{-\infty}^{\infty} g^2(x)f_1(x)\mathrm{d}x + \nu\left(-\int_{-\infty}^{\infty} xg(x)f_1(x)\mathrm{d}x+1\right)\\ & -\lambda(x)g(x)+\vartheta_1\left(\int_{-\infty}^{\infty} g(x)F_1(x)\mathrm{d}x-\xi\right)\\ & +\vartheta_2\left(\int_{-\infty}^{\infty} g(x)\left[1-F_1(x)\right]\mathrm{d}x-\xi\right).\end{aligned}$$

Let $F = \frac{1}{2}g^2(x)f_1(x) - \nu xg(x)f_1(x) + \vartheta_1 g(x)F_1(x) + \vartheta_2 g(x)(1-F_1(x)) - \lambda(x)g(x)$. As no derivative $g'(x)$ is involved in F, the Euler-Lagrange equation

$$\frac{\partial F}{\partial g} - \frac{d}{dx}\left(\frac{\partial F}{\partial g'}\right) = 0$$

simplifies to

$$\begin{aligned}g^*(x)f_1(x) - \nu^* xf_1(x) + \vartheta_2^* + (\vartheta_1^*-\vartheta_2^*)F_1(x) - \lambda^*(x)\\ = 0,\end{aligned} \tag{51}$$

in which the parameters $\vartheta_1^*\ge 0$, $\vartheta_2^*\ge 0$, $\lambda^*(x)\ge 0$ satisfy

$$\begin{aligned}&\int_{-\infty}^{\infty} xg^*(x)f_1(x)\mathrm{d}x = 1,\\ &\vartheta_1^*\left(\int_{-\infty}^{\infty} g^*(x)F_1(x)\mathrm{d}x-\xi\right)=0,\\ &\vartheta_2^*\left(\int_{-\infty}^{\infty} g^*(x)\left[1-F_1(x)\right]\mathrm{d}x-\xi\right)=0,\\ &\lambda^*(x)g^*(x)\ge 0.\end{aligned} \tag{52}$$

From (51), for those x with $f_1(x) > 0$, we have

$$g^*(x) = \frac{\lambda^*(x) + \nu^* xf_1(x) - \vartheta_2^* - (\vartheta_1^*-\vartheta_2^*)F_1(x)}{f_1(x)}.$$

Combining this with the condition (52), we know that if $\nu^* xf_1(x) - \vartheta_2^* - (\vartheta_1^*-\vartheta_2^*)F_1(x) > 0$, then $\lambda^*(x) = 0$. On the other hand, if $\nu^* f_1(x) - \vartheta_2^* - (\vartheta_1^*-\vartheta_2^*)F_1(x) < 0$, then $g^*(x) = 0$. As the result, we have

$$g^*(x) = \begin{cases} \nu^* x - \frac{\vartheta_2^*+(\vartheta_1^*-\vartheta_2^*)F_1(x)}{f_1(x)}, & \nu^* xf_1(x) > \vartheta_2^* + (\vartheta_1^*-\vartheta_2^*)F_1(x);\\ 0, & \text{otherwise.}\end{cases}$$

REFERENCES

[1] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, pp. 73–101, 1964.

[2] P. Huber and E. Ronchetti, *Robust statistics*. Wiley, 2009.

[3] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust statistics: The Approach Based on Influence Functions*. Wiley, 2009.

[4] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," in *Proc. Advances in Neural Information Processing Systems*, (Montreal, Canada), pp. 721–729, 2015.

[5] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, "Robust estimation via robust gradient estimation," *Journal of the Royal Statistical Society*. Submitted.

[6] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *Proc. of IEEE Symposium on Foundations of Computer Science*, (New Brunswick, NJ), Oct. 2016.

[7] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Being robust (in high dimensions) can be practical," in *Proc. International Conference on Machine Learning*, (Sydney, NSW, Australia), pp. 999–1008, 2017.

[8] S. Balakrishnan, S. S. Du, J. Li, and A. Singh, "Computationally efficient robust sparse estimation in high dimensions," in *Proc. Conference on Learning Theory*, Proceedings of Machine Learning Research, (Amsterdam, Netherlands), pp. 169–212, July 2017.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. International Conference on Learning Representations*, (San Diego, CA), May 2015.

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Intl. Symposium on Security and Privacy*, (San Jose, CA), May 2017.

[12] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," in *Proc. AAAI Conference on Artificial Intelligence*, (Austin, TX), Jan. 2015.

[13] F. Hampel, "Contributions to the theory of robust estimation," *Ph.D. thesis, University of California, Berkeley*, 1968.

[14] M. Burger, *Infinite-dimensional Optimization and Optimal Design*. 2003. Lecture notes, available at "ftp://ftp.math.ucla.edu/pub/camreport/cam04-11.pdf".

[15] M. Kot, *A first course in the calculus of variations*. Providence, Rhode Island: American Mathematical Society, 2014.

[16] D. L. Pimentel-Alarcon, A. Biswas, and C. R. Solis-Lemus, "Adversarial principal component analysis," in *Proc. IEEE Intl. Symposium on Inform. Theory*, pp. 2363–2367, July 2017.

[17] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of ACM*, vol. 58, no. 1, pp. 1–37, 2011.

[18] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Intl. Symposium on Security and Privacy*, (San Francisco, CA), May 2018.

[19] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. International Conference on Machine Learning*, (Edinburgh, Scotland), pp. 1467–1474, 2012.

[20] M. Charikar, J. Steinhardt, and G. Valiant, "Learning from untrusted data," in *Proc. Annual ACM SIGACT Symposium on Theory of Computing*, pp. 47–60, 2017.

[21] A. S. Suggala, K. Bhatia, P. Ravikumar, and P. Jain, "Adaptive hard thresholding for near-optimal consistent robust regression," in *Proc. Conference on Learning Theory*, Proceedings of Machine Learning Research, (Phoenix, AZ), June 2019.
[22] E. Bayraktar and L. Lai, "On the adversarial robustness of multivariate robust estimators," *SIAM Journal on Mathematics of Data Science*, Aug. 2019. Submitted.
[23] T. Ferguson, *A course in large sample theory*. Chapman and Hall, London, UK, 1996.
[24] C. Croux, "Limit behavior of the empirical influence function of the median," *Statistics & Probability Letter*, vol. 37, pp. 331–340, 1998.

Lifeng Lai (SM'19) received the B.E. and M.E. degrees from Zhejiang University, Hangzhou, China in 2001 and 2004 respectively, and the Ph.D. from The Ohio State University at Columbus, OH, in 2007. He was a postdoctoral research associate at Princeton University from 2007 to 2009, an assistant professor at University of Arkansas, Little Rock from 2009 to 2012, and an assistant professor at Worcester Polytechnic Institute from 2012 to 2016. Since 2016, he has been an associate professor at University of California, Davis. Dr. Lai's research interests include information theory, stochastic signal processing and their applications in wireless communications, security and other related areas.

Dr. Lai was a Distinguished University Fellow of the Ohio State University from 2004 to 2007. He is a co-recipient of the Best Paper Award from IEEE Global Communications Conference (Globecom) in 2008, the Best Paper Award from IEEE Conference on Communications (ICC) in 2011 and the Best Paper Award from IEEE Smart Grid Communications (SmartGridComm) in 2012. He received the National Science Foundation CAREER Award in 2011, and Northrop Young Researcher Award in 2012. He served as a Guest Editor for IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing Techniques for Wireless Physical Layer Security from 2012 to 2013, and served as an Editor for IEEE Transactions on Wireless Communications from 2013 to 2018. He is currently serving as an Associate Editor for IEEE Transactions on Information Forensics and Security.

Erhan Bayraktar , the holder of the Susan Smith Chair, is a full professor of Mathematics at the University of Michigan, where he has been since 2004 upon getting his Ph.D. from Princeton University. Professor Bayraktar's research is in stochastic analysis, control, applied probability, and mathematical finance. In particular, recently his research focused on mean field games, machine learning and model uncertainty. He has over 130 publications in top journals in these areas.

Professor Bayraktar is recognized as a leader in his areas of research: He is a corresponding editor in the SIAM Journal on Control and Optimization and also serves in the editorial boards of Applied Mathematics and Optimization, Mathematics of Operations Research, Mathematical Finance. His research has been also been continually funded by the National Science Foundation. In particular, he received a CAREER grant. He has been a plenary speaker in numerous conferences and workshops. Professor Bayraktar has had 23 post-docs and 13 Ph.D. students, 9 of whom graduated. They hold prestigious positions in academia and industry.

Professor Bayraktar has been the director of the Risk Management and Quantitative Finance Masters program at the University of Michigan since its inception in 2015.