The Annals of Statistics
2022, Vol. 50, No. 4, 2256–2283
https://doi.org/10.1214/22-AOS2186

GENERALIZED RESILIENCE AND ROBUST STATISTICS

BY BANGHUA ZHU[1,a], JIANTAO JIAO[1,b] AND JACOB STEINHARDT[2,c]

[1]*Department of EECS, University of California, Berkeley,* [a]*banghua@berkeley.edu,* [b]*jiantao@berkeley.edu*
[2]*Department of Statistics, University of California, Berkeley,* [c]*jsteinhardt@berkeley.edu*

Robust statistics traditionally focuses on outliers, or perturbations in total variation distance. However, a dataset could be maliciously corrupted in many other ways, such as systematic measurement errors and missing covariates. We consider corruption in either TV or Wasserstein distance, and show that robust estimation is possible whenever the true population distribution satisfies a property called *generalized resilience*, which holds under moment or hypercontractive conditions. For TV corruption model, our finite-sample analysis improves over previous results for mean estimation with bounded kth moment, linear regression, and joint mean and covariance estimation. For W_1 corruption, we provide the first finite-sample guarantees for second moment estimation and linear regression.

Technically, our robust estimators are a generalization of minimum distance (MD) functionals, which project the corrupted distribution onto a given set of well-behaved distributions. The error of these MD functionals is bounded by a certain modulus of continuity, and we provide a systematic method for upper bounding this modulus for the class of generalized resilient distributions, which usually gives sharp population-level results and good finite-sample guarantees.

1. Introduction. We study the problem of robust estimation from high-dimensional corrupted data. Corruptions can occur in many forms, such as *process error* that affects the outputs, *measurement error* that affects the covariates, or some fraction of arbitrary *outliers*. We will provide a framework for analyzing these and other types of corruptions, study minimal assumptions needed to enable robust estimation at the population level, corruptions, and construct estimators with provably good performance in finite samples.

We model corruptions in terms of a perturbation distance $D(p, q)$. Specifically, we posit a true population distribution p^* that lies in some family of distribution $\mathcal{G}$, but observe samples $X_1, \ldots, X_n$ from a corrupted distribution p such that $D(p^*, p) \le \epsilon$. Our goal is to output an estimate $\hat{\theta}(X_1, \ldots, X_n)$ such that some cost $L(p^*, \hat{\theta})$ is small. Note in particular that our goal is to estimate parameters of the original, uncorrupted distribution p^*. As a result, even as $n \to \infty$ we typically incur some nonvanishing error that depends on ϵ. We also consider a more powerful *adaptive* model where $X_1, \ldots, X_n$ are first sampled from true distribution, and then perturbed by adversary, which is formally defined in Section 2.2.

Throughout the paper, we focus on the case of corruption distance $D = \mathsf{TV}$ or W_1, though many of our results extend to Wasserstein distance over an arbitrary metric space. High dimensional robust statistics for $D = \mathsf{TV}$ has a long history. The majority of the classical statistics papers focus on the minimum distance functional when the true distribution is Gaussian or elliptical (Adrover and Yohai (2002), Chen and Tyler (2002), Donoho and Liu (1988a), Gao (2020), Gao, Yao and Zhu (2020), Huber (1973), Huber (2011)), while recent computationally efficient algorithms are instead based on assumed tail bounds via, for example, moments or sub-Gaussianity (Diakonikolas et al. (2017, 2018a), Bateni and Dalalyan (2019),

Received November 2020; revised February 2022.
MSC2020 subject classifications. Primary 62F35; secondary 62G35.
Key words and phrases. Robust statistics, minimum distance functional, total variation distance perturbation, Wasserstein distance perturbation.

Chen, Gao and Ren (2018), Depersin and Lecué (2022), Diakonikolas, Kane and Stewart (2018), Liu et al. (2018), Steinhardt, Charikar and Valiant (2017), Steinhardt, Koh and Liang (2017)). In this paper, we propose a different assumption called *generalized resilience*, which is more general than either the widely used Gaussian or tail bound assumptions. Generalized resilience enables the systematic design of polynomial sample complexity algorithms, which can also be efficiently computed in some cases. Furthermore, it gives near-optimal statistical rates even in the special case of Gaussian or tail-bounded distributions.

Corruptions under TV only allow an ϵ-fraction of outliers or deletions. In many applications, we might instead believe that all of the data have been slightly corrupted. We can model this as perturbation under the (standard) Wasserstein distance between p and q, defined as the minimum cost in ℓ_2-norm needed to move the points in p to the points in q. For example, measurement bias in sensors usually leads to small perturbations on all the samples, which can be characterized by perturbation in W_1 or W_2 distances (Kowalski (2020)). Stuck pixels on camera introduce coordinate-wise corruption on several coordinates of the output image (Leung et al. (2007)), which can be characterized by perturbation in W_0 distances. For W_1 corruptions, mean estimation is trivial since the adversary can shift the mean by at most ϵ. However, estimation of higher moments, as well as least squares estimation, are nontrivial and we focus on these in the W_1 case.

In this paper, we connect ideas from both the classical and modern approaches to handling TV perturbations, and extend these ideas to Wasserstein perturbations. We summarize our main contributions as follows:

- Motivated by the minimum-distance (MD) functional (Donoho and Liu (1988a)) and the recent progress in efficient algorithms, we construct explicit nonparametric assumptions, *generalized resilience*, under which MD functionals automatically give tight worst-case population error for both TV and W_1 perturbations, matching or improving previous bounds obtained under much stronger assumptions.
- We design finite-sample algorithms based on MD functionals for both TV corruption and W_1 corruption. We propose two different approaches, *weakening the distance* and *expanding the set*, that guarantee polynomial (and sometimes optimal) sample complexity and pave the way for designing computationally efficient algorithms.

For TV corruption, our results improve the best existing bounds for tasks including mean estimation, linear regression and covariance estimation. For W_1 corruption, we are the first to provide any good robustness guarantee under natural assumptions.

1.1. *Overview of structure.* We first summarize the structure of the paper and all the settings and tasks we consider in the paper. For perturbation distance, we consider $D = \mathsf{TV}$ in Section 3 and $D = W_1$ in Section 4. Within each section, we begin with introducing our population assumption, i.e. the design of generalized resilience set that the true distribution lies in, and provide tight bound for its population limit (Sections 3.1 and 4.1).

After the population result in each section, we proceed with finite-sample algorithms. For both TV corruption and W_1 corruption, we propose two different approaches: weaken the distance (Sections 3.2.1 and Section 4.2.1) and expand the set (Sections 3.2.2 and Section 4.2.2). For weaken the distance approach under TV corruption, we study the task of mean estimation, linear regression and joint mean and covariance estimation. For expanding the set approach under TV corruption, we mainly focus on mean estimation under different assumptions. For weaken the distance approach under W_1 corruption, we study second moment estimation and linear regression. For expanding the set approach under W_1 corruption, we illustrate the idea via the example of second moment estimation.

Throughout the paper, we consider two different corruption models, oblivious corruption and adaptive corruption, which are formally defined in Section 2.2. Intuitively, oblivious corruption assumes the adversary corrupts the distribution before sampling, while adaptive corruption allows the adversary to corrupt the data after seeing all the samples. For the convenience of analysis, we focus on oblivious corruption when we consider weaken the distance approach, and focus on adaptive corruption when we consider expand the set approach. However, both methods work under the two different corruption models. We give further discussion in Appendix A of the Supplementary Material (Zhu, Jiao and Steinhardt (2022)).

1.2. *Main results for* TV *corruption* (*Section 3*). Throughout the paper, we design algorithms based on the *minimum distance (MD) functional* estimator (Donoho and Liu (1988a)), which projects the corrupted empirical distribution $\hat{p}_n$ onto some set of distributions $\mathcal{M}$ under a discrepancy measure $\tilde{D}$:

$$\hat{\theta}(\hat{p}_n) = \theta^*(q) \quad \text{where } q = \arg\min_{q \in \mathcal{M}} \tilde{D}(q, \hat{p}_n), \theta^*(q) = \arg\min_{\theta} L(q, \theta) \tag{1}$$

Here $\mathcal{M}$ and $\tilde{D}$ are design parameters to be specified (think of them as relaxations of $\mathcal{G}$ and D). In other words, this estimator projects the observed distribution p onto the distribution set $\mathcal{M}$ to get q, then outputs the optimal parameters for q.

1.2.1. *Design of set*: *Generalized resilience* (*Section 3.1*). We begin with the main results for TV corruption. In the infinite sample case, if the true distribution p^* lies in some family $\mathcal{G}$ and we observe the population corrupted distribution p, the performance of the MD functional estimator $q = \arg\min_{q \in \mathcal{G}} \mathsf{TV}(q, p)$ is upper bounded by the modulus of continuity (Lemma 2.1, (Donoho and Liu (1988a))), defined as $\mathfrak{m}(\mathcal{G}, 2\epsilon, \mathsf{TV}, L) = \sup_{p_1, p_2 \in \mathcal{G}: \mathsf{TV}(p_1, p_2) \le 2\epsilon} L(p_1, \theta^*(p_2))$. While the adversary can choose distributions outside of $\mathcal{G}$, the modulus $\mathfrak{m}$ only involves pairs of distributions that lie within $\mathcal{G}$, making it amenable to analysis. In mean estimation, when the set $\mathcal{G}$ is taken as the set of resilient distributions (Steinhardt, Charikar and Valiant (2018)), defined as

$$\mathcal{G}_{\mathsf{mean}}(\rho, \epsilon) = \left\{ p \mid \|\mu_r - \mu_p\| \le \rho, \forall r \le \frac{p}{1-\epsilon} \right\}, \tag{2}$$

the modulus of continuity can be proved to be upper bounded by 2ρ. The inequality $r \le \frac{p}{1-\eta}$ can be formally understood as $\frac{dr}{dp} \le \frac{1}{1-\eta}$, where $\frac{dr}{dp}$ is the Radon–Nikodym derivative, which can also be understood as $r(A) \le \frac{p(A)}{1-\epsilon}$ for any set A; an equivalent characterization is that r can be obtained from p by conditioning on an event E of probability $1-\epsilon$. Thus, r can be thought of as an "η-deletion" of p, and equation (2) specifies the set of distributions whose mean is stable under deleting an ϵ fraction of points.

The reason for the bounded modulus is a *mid-point property* of TV distance: if $\mathsf{TV}(p_1, p_2) \le \epsilon$ then there is a midpoint r that can be obtained from either of the p_i by conditioning on an event of probability $1-\epsilon$. Thus, μ_r is close to both μ_{p_1} and μ_{p_2} by resilience, and so μ_{p_1} and μ_{p_2} are close by the triangle inequality. The argument appears implicitly in Steinhardt, Charikar and Valiant (2018) and Diakonikolas et al. (2017). Here we make it explicit in Lemma 3.1.

Next, suppose that the loss L is arbitrary. We generalize Steinhardt, Charikar and Valiant's definition of resilience to yield a family with bounded modulus for any given loss L. For loss L, we will need two conditions: the first condition asks that the optimal parameters for p do well on any $r \le \frac{p}{1-\epsilon}$, while the second asks that if a parameter does well on r then it also