Robust Generalized Method of Moments: A Finite Sample Viewpoint

Dhruv Rohatgi[*]
MIT

Vasilis Syrgkanis[†]
Stanford University

Abstract

For many inference problems in statistics and econometrics, the unknown parameter is identified by a set of moment conditions. A generic method of solving moment conditions is the Generalized Method of Moments (GMM). However, classical GMM estimation is potentially very sensitive to outliers. Robustified GMM estimators have been developed in the past, but suffer from several drawbacks: computational intractability, poor dimension-dependence, and no quantitative recovery guarantees in the presence of a constant fraction of outliers. In this work, we develop the first computationally efficient GMM estimator (under intuitive assumptions) that can tolerate a constant ϵ fraction of adversarially corrupted samples, and that has an ℓ_2 recovery guarantee of $O(\sqrt{\epsilon})$. To achieve this, we draw upon and extend a recent line of work on algorithmic robust statistics for related but simpler problems such as mean estimation, linear regression and stochastic optimization. As a special case, we apply our algorithm to instrumental variables linear regression with heterogeneous treatment effects, and experimentally demonstrate that it can tolerate as much as 10 – 15% corruption, significantly improving upon baseline methods.

1 Introduction

Econometric and causal inference methodologies are increasingly being incorporated in automated large scale decision systems. Inevitably these systems need to deal with the plethora of practical issues that arise from automation. One important aspect is being able to deal with corrupted or irregular data, either due to poor data collection, the presence of outliers, or adversarial attacks by malicious agents. Even traditional applications of econometric methods, in social science studies, can greatly benefit from robust inference so as not to draw conclusions solely driven by a handful of samples, as was recently highlighted in [4].

One broad statistical framework, that encompasses the most widely used estimation techniques in econometrics and causal inference, is the framework of estimating models defined via *moment conditions*. In this paper we offer a robust estimation algorithm that extends prior recent work in robust statistics to this more general estimation setting.

For a family of distributions $\{\mathcal{D}_\theta : \theta \in \Theta\}$, identifying the parameter θ is often equivalent to solving

$$\mathbb{E}_{X\sim\mathcal{D}_\theta}[g(X,\theta)] = 0, \tag{1}$$

for an appropriate problem-specific vector-valued function g. This formalism encompasses such problems as linear regression (with covariates X, response Y, and moment $g((X,Y),\theta) = X(Y -$

---

[*] `drohatgi@mit.edu`. This work was partially done while the first author was an intern at Microsoft Research New England.

[†] `vsyrgk@stanford.edu`. This work was partially done while the second author was a Principal Researcher at Microsoft Research New England.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

$X^T\theta$)) and instrumental variables (IV) linear regression (with covariates X, response Y, instruments Z, and moment $g((X,Y,Z),\theta) = Z(Y - X^T\theta)$).

Under simple identifiability assumptions, moment conditions are statistically tractable, and can be solved by the *Generalized Method of Moments* (GMM) [16]. Given independent observations $X_1, \dots, X_n \sim \mathcal{D}_\theta$, the (unweighted) GMM estimator is

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta) \right\|_2^2 .$$

Of course, for general functions g, finding $\hat{\theta}$ (the global minimizer of a potentially non-convex function) may be computationally intractable. Stronger assumptions imply that all approximate *local* minima of the above function are near the true parameter, in which case the GMM estimator is efficiently approximable. For instrumental variables (IV) linear regression, these assumptions follow from standard non-degeneracy assumptions.

Due to its flexibility, the GMM estimator is widely used in practice (along with heuristic variants, in models where it is computationally intractable) [29]. Unfortunately, like most other classical estimators in statistics, the GMM estimator suffers from a lack of robustness: a single outlier in the observations can arbitrarily corrupt the estimate.

Robust statistics Initiated by Tukey and Huber in the 1960s, robust statistics is a broad field studying estimators which have provable guarantees even in the presence of outliers [18]. Outliers can be modelled as samples from a heavy-tailed distribution, or even as adversarially and arbitrarily corrupted data. Classically, robustness of an estimator against arbitrary outliers is measured by breakdown point (the fraction of outliers which can be tolerated without causing the estimator to become unbounded [14]) and influence (the maximum change in the estimator under an infinitesimal fraction of outliers [15]). These metrics have spurred development and study of numerous statistical estimators which are often used in practice to mitigate the effect of outliers (e.g. Huber loss for mean estimation, linear regression, and other problems [17]). Problems such as robust *univariate* mean estimation are by now thoroughly understood [24, 22], and have statistically and computationally efficient estimators.

Unfortunately, in higher dimensions, there has long appeared to be a tradeoff between robustness and computational tractability; as a result, much of the literature on high-dimensional robust statistics has focused on statistical efficiency at the expense of computational feasibility [5, 23, 13]. While there is a rich literature on IV regression and GMM in the context of robust statistics, those works either present computationally intractable estimators [21, 12] or are robust in the sense of bounded influence [1, 27, 20] but not robust against arbitrary outliers. Until the last few years, most high-dimensional statistical problems lacked robust estimators satisfying the following basic properties [7]:

1. Computational tractability (i.e. evading the curse of dimensionality)
2. Robustness to a constant fraction of arbitrary outliers
3. Quantitative error guarantees without dimension dependence.

Recently, a line of work on *algorithmic* robust statistics has blossomed within the theoretical computer science community, with the aim of filling this gap in the high-dimensional statistics literature. Estimators with the above properties have been developed for various fundamental high-dimensional problems, including mean and covariance estimation [7, 9], linear regression [10, 3], and stochastic optimization [26, 8]. However, practitioners in econometrics and applied statistics often employ more sophisticated inference methods such as GMM and IV regression [29, 2]. Such methods are not traditionally under the purview of theoretical computer science and learning theory; perhaps as a result, computationally and statistically efficient robust estimators are still lacking.

Our contribution We address this lack. Methodologically speaking, our main contribution is to introduce GMM to the algorithmic robust statistics literature and vice versa (even aside from robustness, basic algorithmic questions about GMM remain open and surprisingly unstudied). Theoretically speaking, we prove that a simple modification to the SEVER algorithm for robust stochastic optimization [8] (based on using higher-derivative information) yields a computationally efficient and provably robust GMM estimator under intuitive deterministic assumptions about the uncorrupted data. We instantiate this estimator for two important special cases of GMM—instrumental variables linear regression and instrumental variables logistic regression—under distributional assumptions about the covariates, instruments, and responses (and in fact our algorithm also applies to the IV generalized linear model under certain conditions on the link function).

Experimentally, we apply our algorithm to robustly solve IV linear regression. We find that it performs well for a wide range of instrument strengths. In the important setting of heterogeneous treatment effects, our algorithm tolerates as much as 10% corruption. Applied to a seminal dataset previously used to estimate the effect of education on wages [6], we provide evidence for the robustness of the inference, and demonstrate that our algorithm can recover the original inference from corruptions of the dataset, significantly better than baseline approaches.

Technical Overview Our robust GMM algorithm builds upon the SEVER algorithm and framework introduced in [8] for robust stochastic optimization, which itself builds on seminal work on robust multivariate mean estimation via spectral filtering [7, 9]. In this section, we outline the increasing levels of complexity.

First, given samples $v_1, \dots, v_n \in \mathbb{R}^d$ among which ϵn are corrupted, robust mean estimation asks for an estimate of the mean of the uncorrupted samples. The spectral filtering approach due to [9] iteratively does the following, until the sample covariance matrix is bounded: remove outliers in the direction of the largest variance. So long as the uncorrupted samples have bounded covariance, the filtering ensures that at termination, the empirical mean will approximate the uncorrupted mean.

Second, given functions $f_1, \dots, f_n : \mathbb{R}^d \to \mathbb{R}$ among which ϵn are corrupted, robust stochastic optimization asks for an approximate critical point of the mean of the uncorrupted functions. The SEVER algorithm [8] achieves this by alternating between (a) finding a critical point $\hat{w}$ of the current sample set S, and (b) applying one iteration of spectral filtering to the vectors $\{\nabla f_i(\hat{w}) : i \in S\}$, terminating when no samples are removed from S.[3] The termination guarantee of spectral filtering immediately implies that at termination, the average gradient of the uncorrupted samples at $\hat{w}$ is near the average gradient of the final sample set S at $\hat{w}$, which is 0 by part (a). So $\hat{w}$ at termination is an approximate critical point of the mean of the uncorrupted functions.

In our problem, we are given functions $g_1, \dots, g_n : \mathbb{R}^d \to \mathbb{R}^p$ among which ϵn are corrupted, and wish to find an approximate minimizer of $\left\| \frac{1}{|U|} \sum_{i \in U} g_i(w) \right\|_2^2$, where $U \subseteq [n]$ is the set of uncorrupted functions. The obvious approach is to alternate between (a) finding a minimizer $\hat{w}$ of $\left\| \frac{1}{|S|} \sum_{i \in S} g_i(w) \right\|_2^2$, where S is the current sample set, and (b) applying spectral filtering to the vectors $\{g_i(\hat{w}) : i \in S\}$, terminating when no samples are removed from S. The termination guarantee of spectral filtering implies that the final sample average $\frac{1}{|S|} \sum_{i \in S} g_i(\hat{w})$ is near the uncorrupted average $\frac{1}{|U|} \sum_{i \in U} g_i(\hat{w})$. Unfortunately, there is no guarantee that $\frac{1}{|S|} \sum_{i \in S} g_i(\hat{w})$ has small norm: part (a) only implies that $\hat{w}$ is a local minimizer (and hence critical point) of the norm, so

$$\frac{1}{|S|} \sum_{i \in S} (\nabla g_i(\hat{w}))^T \cdot \frac{1}{|S|} \sum_{i \in S} g_i(\hat{w}) = 0.$$

In the above equality, the sample gradient matrix at $\hat{w}$ could be arbitrarily corrupted, so the sample average at $\hat{w}$ could have arbitrarily large norm. In principle, even the *global* minimizer could have large norm. However, this issue can be fixed by using higher-derivative information: specifically, we also apply spectral filtering to (projections of) the matrices $\nabla g_i(\hat{w})$. Under appropriate boundedness and smoothness assumptions, it can then be shown that at termination (when neither filtering step removes samples), $\hat{w}$ is an approximate critical point of the norm of the uncorrupted average $\left\| \frac{1}{|U|} \sum_{i \in U} g_i(w) \right\|_2^2$. By a "strong identifiability" assumption, this implies that $\hat{w}$ is near the minimizer of $\left\| \frac{1}{|U|} \sum_{i \in U} g_i(x) \right\|_2^2$, as desired.

---

[3] A related approach simply applies robust mean estimation to estimate the gradients at each step of gradient descent [26].

2 Preliminaries

For real scalars or vectors $\{\xi_i\}_{i \in S}$ indexed by a set S, we use the notation $\mathbb{E}_S[\xi_i]$ for the sample expectation $\frac{1}{|S|}\sum_{i \in S} \xi_i$. Similarly, if ξ_i are scalars, then we define the sample variance $\operatorname{Var}_S(\xi_i) = \mathbb{E}_S(\xi_i - \mathbb{E}_S \xi_i)^2$. If ξ_i are vectors then we define the sample covariance matrix $\operatorname{Cov}_S(\xi_i) = \mathbb{E}_S(\xi_i - \mathbb{E}_S\xi_i)(\xi_i - \mathbb{E}_S\xi_i)^T$. A random vector X is $(4, 2, \tau)$-hypercontractive if $\mathbb{E}(\langle X, u\rangle)^4 \le \tau(\mathbb{E}(\langle X, u\rangle)^2)^2$ for all vectors u.

Definition 2.1. For a closed set $\mathcal{H}$, a function $f : \mathcal{H} \to \mathbb{R}$, and $\gamma > 0$, a γ-approximate critical point of f (in $\mathcal{H}$) is some $x \in \mathcal{H}$ such that for any vector v with $x + \delta v \in \mathcal{H}$ for arbitrarily small $\delta > 0$, it holds that $v \cdot \nabla f(x) \ge -\gamma \|v\|_2$.

Definition 2.2. For a closed set $\mathcal{H}$, a γ-approximate critical point oracle $\mathcal{L}_{\gamma,\mathcal{H}}$ is an algorithm which, given a differentiable function $f : \mathcal{H} \to \mathbb{R}$ returns a γ-approximate critical point of f.

Definition 2.3. The (unscaled) *logistic function* $G : \mathbb{R} \to \mathbb{R}$ is defined by $G(x) = 1/(1 + e^{-x})$.

Outline In Section 3, we describe the robust GMM problem, and we describe deterministic assumptions on a set of corrupted sample moments, under which we'll be able to efficiently estimate the parameter which makes the uncorrupted moments small. In Section 4, we describe a key subroutine of our robust GMM algorithm, which is commonly known in the literature as *filtering*. In Section 5, we describe the robust GMM algorithm and prove a recovery guarantee under the assumptions from Section 3. In Section 6, we apply this algorithm to instrumental variable linear and logistic regression, proving that under reasonable stochastic assumptions on the uncorrupted data, arbitrarily ϵ-corrupted moments from these models satisfy the desired deterministic assumptions with high probability. Finally, in Section 7, we evaluate the performance of our algorithm on two corrupted datasets.

3 Robust GMM Model

In this section, we formalize the model in which we will provide a robust GMM algorithm. Classically, the goal of GMM estimation is to identify $\theta \in \Theta$ given data $X_1, \dots, X_n \sim \mathcal{D}_\theta$, using the moment condition $\mathbb{E}_{X \sim \mathcal{D}_\theta}[g(X, \theta)] = 0$. We consider the added challenge of the *ϵ-strong contamination model*, in which an adversary is allowed to inspect the data $X_1, \dots, X_n$ and replace ϵn samples with arbitrary data, before the algorithm is allowed to see the data. This corruption model encompasses most reasonable sources of outliers.

For our main theorem, we do not make stochastic assumptions about $\{\mathcal{D}_\theta : \theta \in \Theta\}$. Instead, we make deterministic assumptions about the empirical moments $g_i(\theta) := g(X_i, \theta)$ of the given data, which are *robust to ϵ-strong contamination*. Concretely, we make the following assumption.

Assumption 3.1. Given differentiable moments $g_1, \dots, g_n : \mathbb{R}^d \to \mathbb{R}^p$, a corruption parameter $\epsilon > 0$, well-conditionedness parameters λ and L, a Lipschitzness parameter L_g, and a noise level parameter σ^2, there is a set $I_{\text{good}} \subseteq [n]$ with $|I_{\text{good}}| \ge (1 - \epsilon)n$ (the "uncorrupted samples"), a vector $w^* \in \mathbb{R}^d$ (the "true parameter"), and a radius $R_0 \ge \|w^*\|_2$ with the following properties:

- **Strong identifiability.** $\sigma_{\min}(\mathbb{E}_{I_{\text{good}}} \nabla g(w^*)) \ge \lambda$
- **Bounded-variance gradient.** $\mathbb{E}_{I_{\text{good}}}(u^T \nabla g(w^*) v)^2 \le L^2$ for all unit-vectors $u \in \mathbb{R}^p$, $v \in \mathbb{R}^d$
- **Bounded-variance noise.** $\mathbb{E}_{I_{\text{good}}}(v \cdot g(w^*))^2 \le \sigma^2 L$ for all unit vectors v
- **Well-specification.** $\left\|\mathbb{E}_{I_{\text{good}}} g(w^*)\right\|_2 \le \sigma\sqrt{L\epsilon}$
- **Lipschitz gradient.** $\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w) - \mathbb{E}_{I_{\text{good}}} \nabla g(w^*)\right\|_{\text{op}} \le L_g \|w - w^*\|_2$ for all $w \in B_{2R_0}(0)$
- **Stability of gradient.** $R_0 < \lambda/(9L_g)$.

Intuitively, Assumption 3.1 can be thought of as a condition on the uncorrupted samples, because if they satisfy the assumption with parameter ϵ_0, then after ϵ-strong contamination, the corrupted samples will still satisfy the assumption with parameter $\epsilon_0 + \epsilon$. Strong identifiability is needed for parameter recovery (even without corruption). Bounded-variance gradient is a technical condition which e.g. reduces to a 4th moment bound for IV regression. The third and fourth conditions ensure that the data is approximately well-specified by the moment conditions. The fifth and sixth conditions hold trivially for IV linear regression; for non-linear moment problems, such as our logistic IV regression problem, this condition requires that the ℓ_2-norm of the parameters be sufficiently small, such that the logits do not approach the flat region of the logistic function, a condition that is natural to avoid loss of gradient information and extreme propensities.

4 The FILTER Algorithm

In many robust statistics algorithms, an important subroutine is a *filtering* algorithm for robust mean estimation. In this section we describe a filtering algorithm used in numerous prior works, including e.g. [8, 9]. Given a set of vectors $\{\xi_i : i \in S\}$ and a threshold M, the algorithm returns a subset of S, by thresholding outliers in the direction of largest variance. Formally, see Algorithm 1.

Algorithm 1 FILTER

```
procedure FILTER($\{\xi_i : i \in S\}, M$)
    $\hat{\xi} \leftarrow \mathbb{E}_S[\xi_i]$, $\mathrm{Cov}_S(\xi_i) = \mathbb{E}_S[(\xi_i - \hat{\xi})(\xi_i - \hat{\xi})^T]$
    $v \leftarrow$ largest eigenvector of $\mathrm{Cov}_S(\xi_i)$
    $\tau_i \leftarrow (v \cdot (\xi_i - \hat{\xi}))^2$ for $i \in S$
    if $\frac{1}{|S|}\sum_{i \in S} \tau_i \le 24M$ then
        return $S$
    else
        Sample $T \leftarrow \mathrm{Unif}([0, \max \tau_i])$
        return $S \setminus \{i \in S : \tau_i > T\}$
```

This algorithm has two important properties. First, if it does not filter any samples, then the sample mean is provably stable, i.e. it cannot have been affected much by the corruptions, so long as the uncorrupted samples had bounded variance (proof in Appendix B.1).

Lemma 4.1 (see e.g. [8, 9])**.** *Suppose that* FILTER *does not filter out any samples. Then*

$$\|\mathbb{E}_S\xi - \mathbb{E}_I\xi\|_2 \le 3\sqrt{48}\sqrt{(M + \|\mathrm{Cov}_I(\xi)\|_{op})\epsilon}$$

for any $I \subseteq [n]$ and $\epsilon > 0$ such that $|S|, |I| \ge (1-\epsilon)n$.

Second, if the threshold is chosen appropriately (based on the variance of the uncorrupted samples), then the filtering step always in expectation removes at least as many corrupted samples as uncorrupted samples. Equivalently, the size of the symmetric difference between the current sample set and the uncorrupted samples (i.e. the number of corrupted samples in the current set plus the number of uncorrupted samples which have been filtered out of the current set) always decreases in expectation (proof in Appendix B.1.1).

Lemma 4.2 (see e.g. [8, 9])**.** *Consider an execution of* FILTER *with sample set S of size $|S| \ge 2n/3$, and vectors $\{\xi_i : i \in S\}$, and bound M. Let S' be the sample set after this iteration's filtering. Let $I_{good} \subseteq [n]$ satisfy $|I_{good}| \ge (5/6)n$. Suppose that $\mathrm{Cov}_{I_{good}}(\xi_i) \preceq MI$, then*

$$\mathbb{E}|S' \triangle I_{good}| \le \mathbb{E}|S \triangle I_{good}|,$$

where the expectation is over the random threshold, and Δ denotes symmetric difference.

5 The ITERATED-GMM-SEVER Algorithm

In this section, we describe and analyze an algorithm ITERATED-GMM-SEVER for robustly solving moment conditions under Assumption 3.1. The key subroutine is the algorithm GMM-SEVER, which given an initial estimate w_0 and a radius R such that the true parameter is contained in $B_R(w_0)$, returns a refined estimate w such that (with large probability) the radius bound can be decreased by a constant factor. We assume access to an approximate constrained critical point oracle $\mathcal{L}$ (Definition 2.2), which can be efficiently implemented (for arbitrary smooth bounded functions) by gradient descent.

Algorithm 2 GMM-SEVER

```
1: procedure GMM-SEVER($\mathcal{L}, \{g_1, \dots, g_n\}, w_0, R, \gamma, L, \sigma$)
2:     $S \leftarrow [n]$
3:     repeat
4:         Compute a $\gamma$-approximate critical point $w \leftarrow \mathcal{L}_{\gamma, B_R(w_0)}(\|\mathbb{E}_S(g_i(\cdot))\|_2^2)$
5:         $u \leftarrow \mathbb{E}_S g_i(w)$
6:         $S' \leftarrow$ FILTER($\{\nabla g_i(w) \cdot u : i \in S\}, L^2 \|u\|_2^2$)
7:         if $S' \neq S$ then
8:             Set $S \leftarrow S'$ and return to line 4
9:         $S'' \leftarrow$ FILTER($\{g_i(w) : i \in S\}, \sigma^2 L + 4L^2 R^2$)
10:        if $S'' \neq S$ then
11:            Set $S \leftarrow S''$ and return to line 4
12:    until $S'' = S$
13:    return $(w, S)$
```

Algorithm 3 AMPLIFIED-GMM-SEVER

```
1: procedure AMPLIFIED-GMM-SEVER($\mathcal{L}, \{g_1, \dots, g_n\}, w_0, R, \gamma, \epsilon, L, \sigma, \delta$)
2:     $t \leftarrow 0$
3:     repeat
4:         $w, S \leftarrow$ GMM-SEVER($\mathcal{L}, \{g_1, \dots, g_n\}, w_0, R, \gamma, L, \sigma$)
5:         $t \leftarrow t + 1$
6:     until $|S| \geq (1 - 11\epsilon)n$ or $(1/10)^t \leq \delta$
7:     return $w$
```

Like the algorithm SEVER [8], our algorithm GMM-SEVER alternates (a) finding a critical point of a function associated to the current samples, and (b) filtering out "outlier" samples. Unlike SEVER, the function we optimize is not simply an empirical mean over the samples, but rather the squared-norm of the sample moments. Moreover, we need two filtering steps: the moments as well as directional derivatives of the moments, in a carefully chosen direction. See Algorithm 2 for the complete description.

We will only prove a constant failure probability for GMM-SEVER. However, we will show that it can be amplified to an arbitrarily small failure probability δ. We call the resulting algorithm AMPLIFIED-GMM-SEVER; see Algorithm 3. The algorithm ITERATED-GMM-SEVER then consists of iteratively calling AMPLIFIED-GMM-SEVER to refine the parameter estimate and bound the true parameter within successively smaller balls; see Algorithm 4.

We start by analyzing GMM-SEVER. In the next two lemmas, we show that if the algorithm does not filter out too many samples, then we can bound the distance from the output to w^*. First, we show a first-order criticality condition (in the direction $\hat{w} - w^*$) for the norm of the moments of the "good" samples. If there was no corruption, then we would have an inequality of the form

$$\frac{(\hat{w} - w^*)^T}{\|\hat{w} - w^*\|_2} \mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w})^T \mathbb{E}_{I_{\text{good}}} g(\hat{w}) \leq \gamma.$$

With ϵ-corruption, the algorithm is designed so that we can still show the following inequality, matching the above guarantee up to $O(\sqrt{\epsilon})$ (proof in Appendix C.1):

Lemma 5.1. *Suppose that the input parameters R and w_0 satisfy $B_R(w_0) \subseteq B_{2R_0}(0)$. Under Assumption 3.1, at algorithm termination, if $|S| \geq (1 - 10\epsilon)n$, then the output $\hat{w}$ of* GMM-SEVER *satisfies*

$$\frac{(\hat{w} - w^*)^T}{\|\hat{w} - w^*\|_2} \mathbb{E}_{I_{good}} \nabla g(\hat{w})^T \mathbb{E}_{I_{good}} g(\hat{w}) \leq \gamma + 275\sigma L^{3/2}\sqrt{\epsilon} + 603L^2 R\sqrt{\epsilon}$$

Moreover, we can show that any point satisfying the first-order criticality condition must be close to w^*, using the least singular value bound on the gradient (proof in Appendix C.2).

Lemma 5.2. *Suppose that the input parameters R and w_0 satisfy $B_R(w_0) \subseteq B_{2R_0}(0)$. Under Assumption 3.1, suppose that $w \in B_R(w_0)$ satisfies*

$$(w - w^*)^T \mathbb{E}_{I_{good}} \nabla g(w)^T \mathbb{E}_{I_{good}} g(w) \leq \kappa \|w - w^*\|_2 .$$

Algorithm 4 ITERATED-GMM-SEVER

```
procedure ITERATED-GMM-SEVER($\{g_1, \ldots, g_n\}, R_0, \gamma, \epsilon, \lambda, L, \sigma, \delta$)
    $t \leftarrow 1$, $w_1 \leftarrow 0$, $R_1 \leftarrow R_0$, $\delta' \leftarrow c\delta / \log(R\sqrt{L}/(\sigma\sqrt{\epsilon})$, $\gamma = \sigma L^{3/2}\sqrt{\epsilon}$
    repeat
        $\hat{w}_t :=$ AMPLIFIED-GMM-SEVER($\{g_1, \ldots, g_n\}, w_t, R_t, \epsilon, L, \sigma, \gamma, \delta'$)
        $R_{t+1} \leftarrow 2\gamma/\lambda^2 + C((L^2/\lambda^2)R_t\sqrt{\epsilon} + \sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon})$
        $t \leftarrow t + 1$
    until $R_t > R_{t-1}/2$
    return $\hat{w}_{t-1}$
```

Then $\|w - w^*\|_2 \leq 4(\kappa + \sigma L^{3/2}\sqrt{\epsilon})/\lambda^2$.

Putting the above lemmas together, we immediately get the following bound on $\|\hat{w} - w^*\|_2$.

Lemma 5.3. *Suppose that the input parameters* R *and* w_0 *satisfy* $B_R(w_0) \subseteq B_{2R_0}(0)$. *Under Assumption 3.1, at algorithm termination, if* $|S| \geq (1 - 10\epsilon)n$*, then the output* $\hat{w}$ *of* GMM-SEVER *satisfies*

$$\|\hat{w} - w^*\|_2 \leq \frac{4\gamma}{\lambda^2} + 2412(L^2/\lambda^2)R\sqrt{\epsilon} + 1102\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}.$$

It remains to bound the size of S at termination. We follow the super-martingale argument from [8], which uses Lemma 4.2 (proof in Appendix C.3).

Theorem 5.4. *Suppose that the input parameters* R *and* w_0 *satisfy* $B_R(w_0) \subseteq B_{2R_0}(0)$. *Let* $\hat{w}$ *be the output of* GMM-SEVER*. Then with probability at least* $9/10$*, it holds that*

$$\|\hat{w} - w^*\|_2 \leq \frac{4\gamma}{\lambda^2} + 2412(L^2/\lambda^2)R\sqrt{\epsilon} + 1102\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}.$$

The time complexity of GMM-SEVER *is* $O(\text{poly}(n, d, p, T_\gamma))$ *where* T_γ *is the time complexity of the* γ*-approximate learner* $\mathcal{L}$*. Moreover, for any* $\delta > 0$ *the success probability can be amplified to* $1 - \delta$ *by repeating* GMM-SEVER $O(\log 1/\delta)$ *times, or until* $|S| \geq (1 - 10\epsilon)n$ *at termination. We call this* AMPLIFIED-GMM-SEVER*, and it has time complexity* $O(\text{poly}(n, d, p, T_\gamma) \cdot \log(1/\delta))$.

With the above guarantee for GMM-SEVER and AMPLIFIED-GMM-SEVER, we can now analyze ITERATED-GMM-SEVER (proof in Appendix C.4).

Theorem 5.5. *Suppose that the input to* ITERATED-GMM-SEVER *consists of functions* $g_1, \ldots, g_n : \mathbb{R}^d \to \mathbb{R}^p$*, a corruption parameter* $\epsilon > 0$*, well-conditionedness parameters* λ *and* L*, a Lipschitzness parameter* L_g*, a noise level parameter* σ^2*, a radius bound* R_0*, and an optimization error parameter* γ*, such that Assumption 3.1 is satisfied for some unknown parameter* $w^* \in \mathbb{R}^d$*, and* $(L^2/\lambda^2)\sqrt{\epsilon} \leq 1/9648$. [4] *Suppose that the algorithm is also given a failure probability parameter* $\delta > 0$.

Then the output $\hat{w}$ *of* ITERATED-GMM-SEVER *satisfies*

$$\|\hat{w} - w^*\|_2 \leq O(\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon})$$

with probability at least $1 - \delta$*. Moreover, the algorithm has time complexity* $O(\text{poly}(n, d, p, T_\gamma) \cdot \log(1/\delta) \cdot \log(R\sqrt{L}/(\sigma\sqrt{\epsilon})))$*, where* T_γ *is the time complexity of a* γ*-approximate learner and* $\gamma = \sigma L^{3/2}\sqrt{\epsilon}$.

6 Applications

In this section, we apply ITERATED-GMM-SEVER to solve linear and logistic instrumental variables regression in the strong contamination model.

---

[4] This constant may be improved; we focus in this paper on dependence on the parameters of the problem and do not optimize constants.

Robust IV Linear Regression Let Z be the vector of p real-valued instruments, and let X be the vector of d real-valued covariates. Suppose that Z and X are mean-zero. Suppose that the response can be described as $Y = X^T w^* + \xi$ for some fixed $w^* \in \mathbb{R}^d$. The distributional assumptions we will make about X, Y, and Z are described below.

Assumption 6.1. Given a corruption parameter $\epsilon > 0$, well-conditionedness parameters λ and L, hypercontractivity parameter τ, noise level parameter σ^2, and norm bound R_0, we assume the following: (i) **Valid instruments:** $\mathbb{E}[\xi|Z] = 0$, (ii) **Bounded-variance noise:** $\mathbb{E}[\xi^2|Z] \leq \sigma^2$, (iii) **Strong instruments:** $\sigma_{\min}(\mathbb{E}ZX^T) \geq \lambda$, (iv) **Boundedness:** $\|\mathrm{Cov}([Z;X]\|_{\mathrm{op}} \leq L$, (v) **Hypercontractivity:** $[Z;X]$ is $(4,2,\tau)$-hypercontractive, (vi) **Bounded 8th moments:** $\max_i X_i^8 \leq O(\tau^2 L^4)$ and $\max_i Z_i^8 \leq O(\tau^2 L^4)$ (vii) **Bounded norm parameter:** $\|w^*\|_2 \leq R_0$.

For intuition, conditions (i – iii) are standard for IV regression even in the absence of corruption; (iv – vi) are conditions on the moments of the distribution, and hold for a variety of reasonable distributions including but not limited to any multivariate Gaussian distribution with bounded-spectral-norm covariance. Condition (vii) essentially states that we need an initial estimate of w^* (but the time complexity of our algorithm will depend only logarithmically on the initial estimate error R_0).

Define the random variable

$$g(w) = Z(Y - X^T w)$$

for $w \in \mathbb{R}^d$, and let (X_i, Y_i, Z_i) be n independent samples drawn according to (X, Y, Z). Let $\epsilon > 0$. We prove that under the above assumption, if n is sufficiently large, then with high probability, for any ϵ-contamination $(X_i', Y_i', Z_i')_{i=1}^n$ of $(X_i, Y_i, Z_i)_{i=1}^n$, the functions $g_i(w) = Z_i'(Y_i' - (X_i')^T w)$ satisfy Assumption 3.1. Formally, we prove the following theorem (see Appendix D):

Theorem 6.2. *Let $\epsilon > 0$. Suppose that $\epsilon < c\min(\lambda^2/(\tau L^2), \lambda^4/L^4)$ for a sufficiently small constant $c > 0$, and suppose that $n \geq C(d+p)^5 \tau \log((p+d)/\tau\epsilon)/\epsilon^2$ for a sufficiently large constant C. Then with probability at least* 0.95 *over the samples $(X_i, Y_i, Z_i)_{i=1}^n$, the following holds: for any ϵ-corruption of the samples and any upper bound $R_0 \geq \|w^*\|_2$, Assumption 3.1 is satisfied. In that event, if L, λ, σ, and ϵ are known, then there is a* $\mathrm{poly}(n, d, p, \log(1/\delta), \log(R_0/(\sigma\sqrt{\epsilon})))$*-time algorithm which produces an estimate $\hat{w}$ satisfying $\|\hat{w} - w^*\|_2 \leq O(\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon})$ with probability at least* $1 - \delta$.

Robust IV Logistic Regression Let Z be a vector of p real-valued instruments, and let X be a vector of d real-valued covariates. Suppose that Z and X are mean-zero. Suppose that the response can be described as $Y = G(X^T w^*) + \xi$ for some fixed $w^* \in \mathbb{R}^d$, where G is the (unscaled) logistic function. The proofs only use 1-Lipschitzness of G and G', and that $G'(0)$ is bounded away from 0.

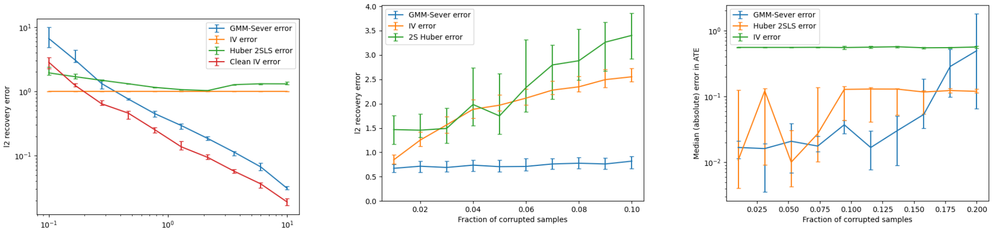
As far as distributional assumptions, we assume in this section that Assumption 6.1 holds, and additionally assume that the norm bound satisfies $R_0 \leq c\min(\lambda^2/L, \lambda/\sqrt{\tau L^3})$ for an appropriate constant c, where λ, L, and τ are as required for the Assumption. We obtain the following algorithmic result (proof in Appendix E):

Theorem 6.3. *Let $\epsilon > 0$. Suppose that $\epsilon < c\min(\lambda^2/(\tau L^2), \lambda^4/L^4)$ for a sufficiently small constant $c > 0$, and suppose that $n \geq C(d+p)^5 \tau \log((p+d)/\tau\epsilon)/\epsilon^2$ for a sufficiently large constant C. Suppose that $\|w^*\|_2 \leq R_0 \leq c\min(\lambda^2/L, \lambda/\sqrt{\tau L^3})$. Then with probability at least* 0.95 *over the samples $(X_i, Y_i, Z_i)_{i=1}^n$, the following holds: for any ϵ-corruption of the samples, Assumption 3.1 is satisfied. In that event, if R_0, L, λ, σ, and ϵ are known, then there is a* $\mathrm{poly}(n, d, p, \log(1/\delta), \log(R_0/(\sigma\sqrt{\epsilon})))$*-time algorithm which produces an estimate $\hat{w}$ satisfying $\|\hat{w} - w^*\|_2 \leq O(\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon})$ with probability at least* $1 - \delta$.

7 Experiments

In this section we corroborate our theory by applying our algorithm ITERATED-GMM-SEVER to several datasets for IV linear regression. See Appendix G for omitted figures and experimental details (e.g. hyperparameter choices and descriptions of the baselines). Error bars are at 25th and 75th percentiles across independent trials.

Varied Instrument Strength. We construct a synthetic dataset with endogenous noise and 1% corruptions, and evaluate our estimator as the instrument strength is varied. Concretely, for dimension



(a) Varied Instrument Strength

(b) Synthetic HE dataset with added corruptions

(c) NLSYM dataset with added corruptions

Figure 1

d and strength α, we draw independent samples $(X_i, Y_i, Z_i)_{i=1}^n$ where for unobserved noise $\eta_i \sim N(0, I_d)$, we define instruments $Z_i \sim N(0, I_d)$ and covariates $X_i = \alpha Z_i + \eta_i$, and response $y_i = \langle X_i, \theta^* \rangle + \langle \eta_i, \mathbb{1} \rangle$. For $k = 0.01n$ of the samples, we introduce corruption by setting $Z_i = -A/(k\sqrt{d})$ and $y_i = \sqrt{d}$ where $A = \sum Z_j y_j$, which zeroes out the IV estimate. We take $n = 10^4$, $d = 20$ and $\theta^* = (1, 0, \dots, 0)$, and vary α from 0.1 to 10. For each α, we do 10 independent trials, comparing median ℓ_2 error of ITERATED-GMM-SEVER with classical IV and two-stage Huber regression. We also compare to the “clean IV” error, i.e. the error of IV on the uncorrupted samples. When α is small, essentially no inference is possible (the clean error is large), but as α increases, our estimator starts to outperform the baselines, and roughly tracks the clean error (Figure 1a). Similar results can be seen for $d = 100$ (Figure 2 in Appendix G.5).

Our next two examples consider *IV linear regression with heterogeneous treatment effects*, a natural setting in which the instruments and covariates are high-dimensional, necessitating dimension-independent robust estimators. Consider a study in which each sample has a vector X of characteristics, a scalar instrument Z, a scalar treatment T, and a response Y. Assuming that the control response and treatment effect are linear in the characteristics, with unknown coefficients β^* and θ^* respectively, and that the response noise is mean-zero conditioned on Z and X (but may correlate with the treatment), we can write the moment conditions

$$\mathbb{E}[XZ(Y - T\langle X, \theta^* \rangle - \langle X, \beta^* \rangle)] = \mathbb{E}[X(Y - T\langle X, \theta^* \rangle - \langle X, \beta^* \rangle)] = 0.$$

This can be interpreted as an IV linear regression with covariates (TX, X) and instruments (ZX, X).

Synthetic HE dataset. For parameters n, d, we generate a unknown d-dimensional parameter vector $\theta^* \sim N(0, I_d)$. We then generate independent samples $(X_i, Y_i, Z_i)_{i=1}^n$ as follows. Draw $X_i \sim N(0, I_d)$ and $Z_i \sim \text{Ber}(1/2)$. The binary treatment is drawn $T_i \sim \text{Ber}(p_i)$ with

$$p_i = \frac{1}{1 + \exp(-Z_i - U_i \bar{X}_i)},$$

where $U_i \sim N(0, 1)$ and $\bar{X}_i = d^{-1/2} \langle X_i, \mathbb{1} \rangle$. Finally, the response is $Y_i = \langle X_i, \theta^* \rangle T_i + \langle X_i, \beta^* \rangle + U_i$ with $\beta^* := 0$.

Ordinary least squares would produce a biased estimate of (θ^*, β^*), since $T\bar{X}$ is correlated with the response noise U. However, U is by construction independent of X and Z. Thus, in the absence of corruption, IV linear regression with covariates (TX, X), response Y, and instrument (ZX, X) should approximately recover the true parameters (θ, β).

For $n = 10^3$ and $d = 20$, the IV estimate still has significant variance, and in this regime, even with no added corruptions, we find that ITERATED-GMM-SEVER has lower recovery error than baselines (Table 1 in Appendix G.5). For $n = 10^4$ and $d = 20$, the IV estimate is more accurate. Hence, we corrupt the first ϵn samples, by setting $X_i := \mathbb{1}$ and $Y_i := 3\sqrt{d}$. Varying ϵ from 0.01 to 0.1, we compute the median ℓ_2 recovery error of ITERATED-GMM-SEVER, classical IV, and two-stage Huber regression, across 50 independent trials (for each ϵ). The results (Figure 1b) demonstrate that our algorithm is resilient to up to 10% corruptions, whereas both baselines rapidly degrade as ϵ increases.

NLSYM dataset. In this experiment, we use the data of [6] from the National Longitudinal Survey of Young Men for estimating the average treatment effect (ATE) of education on wages. The data consists of 3010 samples with years of education as the treatment, log wages as the response, and proximity to a 4-year college as the instrument, along with 22 covariates (e.g. geographic indicator variables). For simplicity, we restrict the model to only two covariates (years and squared years of labor force experience) and bias term. We find that the ATE estimated by ITERATED-GMM-SEVER is close to the positive ATE (≈ 0.277) estimated by classical IV, suggesting that Card's inference may be robust (Figure 3 in Appendix G.5). Next, we corrupt a random ϵ-fraction of the responses, in a way that negates the ATE inferred by classical IV regression (see Appendix G.2 for method).

Varying ϵ from 0.01 to 0.2, we perform 10 independent trials (i.e. resampling the subset of corrupted samples each time). For each trial, we compute the ATE estimate of IV regression, the ATE estimate of two-stage Huber regression, and the median ATE estimate of 50 runs of ITERATED-GMM-SEVER. For each ϵ, we then plot the median absolute error of each algorithm across the 10 trials. We find that our algorithm outperforms both baselines, and has lower variance than two-stage Huber regression, up to $\epsilon \approx 0.15$ (Figure 1c; note that error is on log-scale, so the Huber regression is extremely noisy).

References

[1] Takeshi Amemiya. Two stage least absolute deviations estimators. *Econometrica: Journal of the Econometric Society*, pages 689–711, 1982.

[2] Joshua D Angrist and Alan B Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85, 2001.

[3] Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

[4] Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions?, 2021.

[5] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[6] David Card. Using geographic variation in college proximity to estimate the return to schooling, 1993.

[7] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[8] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019.

[9] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.

[10] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

[11] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[12] Gabriela V Cohen Freue, Hernan Ortiz-Molina, and Ruben H Zamar. A natural robustification of the ordinary instrumental variables estimator. *Biometrics*, 69(3):641–650, 2013.

[13] Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

[14] Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

[15] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[16] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982.

[17] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[18] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[19] Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. *arXiv preprint arXiv:2106.11938*, 2021.

[20] Tae-Hwan Kim and Christophe Muller. Two-stage huber estimation. *Journal of statistical planning and inference*, 137(2):405–418, 2007.

[21] William S Krasker. Two-stage bounded-Influence estimators for simultaneous-equations models. *Journal of Business & Economic Statistics*, 4(4):437–444, 1986.

[22] Jasper CH Lee and Paul Valiant. Optimal sub-gaussian mean estimation in r. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE, 2022.

[23] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019.

[24] Roberto I Oliveira and Paulo Orenstein. The sub-gaussian property of trimmed means estimators. Technical report, Technical report, IMPA, 2019.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[27] Elvezio Ronchetti and Fabio Trojani. Robust inference with gmm estimators. *Journal of econometrics*, 101(1):37–69, 2001.

[28] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[29] Jeffrey M Wooldridge. Applications of generalized method of moments estimation. *Journal of Economic perspectives*, 15(4):87–100, 2001.

Checklist

1. For all authors...
    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Theorems 5.5, 6.2, 6.3, and Section 7
    (b) Did you describe the limitations of your work? [Yes] See Assumptions 3.1 and 6.1
    (c) Did you discuss any potential negative societal impacts of your work? [N/A]
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 3.1 and 6.1
    (b) Did you include complete proofs of all theoretical results? [Yes] See supplemental material

3. If you ran experiments...
    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material
    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 7 and Appendix G in supplemental material
    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 7 and Appendix G in supplemental material
    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    (a) If your work uses existing assets, did you cite the creators? [Yes] National Longitudinal Survey of Young Men
    (b) Did you mention the license of the assets? [N/A]
    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code included in supplemental material
    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Supplementary lemma for Section 3

We state several consequences of Assumption 3.1 which will be used later.

Lemma A.1. *Under Assumption 3.1, the following bounds hold for all $w \in B_{2R_0}(0)$:*

- $\mathbb{E}_{I_{good}}(u^T \nabla g(w) v)^2 \le 2L^2$ *for all unit vectors* $u \in \mathbb{R}^p$ *and* $v \in \mathbb{R}^d$
- $\mathrm{Cov}_{I_{good}}(g(w)) \preceq 2\sigma^2 L + 4L^2 \left\|w - w^*\right\|_2^2 I$
- $\sigma_{min}(\mathbb{E}_{I_{good}} \nabla g(w)) \ge \lambda/2$
- $\left\|\mathbb{E}_{I_{good}} g(w)\right\|_2 \le \sigma\sqrt{L\epsilon} + 2L\left\|w - w^*\right\|_2$
- $\left\|\mathbb{E}_{I_{good}} \nabla g(w)\right\|_{op} \le L\sqrt{2}$

Proof.

First claim. Note that

$$
\begin{aligned}
&|\mathbb{E}_{I_{\text{good}}}(u^T \nabla g(w) v)^2 - \mathbb{E}_{I_{\text{good}}}(u^T \nabla g(w^*) v)^2| \\
&\qquad = |\mathbb{E}_{I_{\text{good}}}(u^T(\nabla g(w) - \nabla g(w^*))v)(u^T(\nabla g(w) + \nabla g(w^*))v)| \\
&\qquad \le \mathbb{E}_{I_{\text{good}}}(u^T(\nabla g(w) - \nabla g(w^*))v)^2 + 2\sqrt{\mathbb{E}_{I_{\text{good}}}(u^T(\nabla g(w) - \nabla g(w^*))v)^2 \mathbb{E}_{I_{\text{good}}}(u^T \nabla g(w^*) v)^2} \\
&\qquad \le L_g^2 \left\|w - w^*\right\|_2^2 + 2L_g L \left\|w - w^*\right\|_2 \\
&\qquad \le \lambda^2/9 + 2\lambda L/3 \\
&\qquad \le L^2
\end{aligned}
$$

where the first inequality expands $\nabla g(w) + \nabla g(w^*)$ as $(\nabla g(w) - \nabla g(w^*)) + 2\nabla g(w^*)$ and applies Cauchy-Schwarz to the resulting second term; the second inequality applies the Lipschitz gradient assumption and bounded-variance gradient assumption at w^*; the third inequality applies the stability of gradient assumption together with the bound $\left\|w - w^*\right\|_2 \le 3R_0$; and the fourth inequality uses that $\lambda \le L$. It follows that

$$
\mathbb{E}_{I_{\text{good}}}(u^T \nabla g(w) v)^2 \le L^2 + \mathbb{E}_{I_{\text{good}}}(u^T \nabla g(w^*) v)^2 \le 2L^2
$$

as claimed.

Second claim. Observe that for any unit vector v,

$$
\mathbb{E}_{I_{\text{good}}}(v \cdot g(w))^2 \le 2\mathbb{E}_{I_{\text{good}}}(v \cdot g(w^*))^2 + 2\mathbb{E}_{I_{\text{good}}}(v \cdot (g(w) - g(w^*)))^2.
$$

The first term is at most $2\sigma^2 L$ by the bounded-variance noise assumption. The second term can be written and bounded as

$$
\begin{aligned}
\mathbb{E}_{I_{\text{good}}}(v \cdot (g(w) - g(w^*)))^2 &= \mathbb{E}_{I_{\text{good}}}\left(\int_0^1 v^T \nabla g(tw + (1-t)w^*)(w - w^*)\, dt\right)^2 \\
&\le \int_0^1 \mathbb{E}_{I_{\text{good}}}(v^T \nabla g(tw + (1-t)w^*)(w - w^*))^2 \\
&\le 2L^2 \left\|w - w^*\right\|_2^2
\end{aligned}
$$

by the first claim. This proves the second claim.

Third claim. We have for any $w \in B_{2R_0}(0)$ that $\left\|w - w^*\right\|_2 \le 4R_0$, so

$$
\begin{aligned}
\sigma_{\min}(\mathbb{E}_{I_{\text{good}}} \nabla g(w)) &\ge \sigma_{\min}(\mathbb{E}_{I_{\text{good}}} \nabla g(w^*)) - \left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w) - \mathbb{E}_{I_{\text{good}}} \nabla g(w^*)\right\|_{\text{op}} \\
&\ge \lambda - L_g \cdot 4R_0 \\
&\ge \lambda/2
\end{aligned}
$$

as claimed, where the second inequality uses the strong identifiability assumption and Lipschitz gradient assumption, and the third inequality uses the stability of gradient assumption.

Fourth claim. We note that

$$\mathbb{E}_{I_{\text{good}}} g(w) - \mathbb{E}_{I_{\text{good}}} g(w^*) = \int_0^1 \mathbb{E}_{I_{\text{good}}} \nabla g(tw + (1-t)w^*)(w - w^*)\, dt.$$

The expectation of the gradient has operator norm at most $L + L_g \|w - w^*\|_2$ by bounded-variance and Lipschitzness of the gradient, and this is at most $2L$ by stability of the gradient and the inequality $\lambda \leq L$. As a result,

$$\left\|\mathbb{E}_{I_{\text{good}}} g(w) - \mathbb{E}_{I_{\text{good}}} g(w^*)\right\|_2 \leq 2L \|w - w^*\|_2,$$

so together with well-specification it follows that $\left\|\mathbb{E}_{I_{\text{good}}} g(w)\right\|_2 \leq \sigma\sqrt{L\epsilon} + 2L\|w - w^*\|_2$ as claimed.

Fifth claim. This follows immediately from the first claim. Indeed, for any $w \in B_{2R_0}(0)$ and unit vectors $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^d$,

$$(u^T \mathbb{E}_{I_{\text{good}}} \nabla g(w) v)^2 = (\mathbb{E}_{I_{\text{good}}} u^T \nabla g(w) v)^2 \leq \mathbb{E}_{I_{\text{good}}} (u^T \nabla g(w) v)^2 \leq 2L^2.$$

Taking the supremum over all u, v we get that $\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)\right\|_{\text{op}} \leq L\sqrt{2}$ as claimed. □

B Omitted proofs from Section 4

B.1 Proof of Lemma 4.1

Proof. If the algorithm does not remove any samples, then it holds that

$$\|\text{Cov}_S(\xi)\|_{\text{op}} = \text{Var}_S(v \cdot \xi) = \frac{1}{|S|} \sum_{i \in S} \tau_i \leq 24M.$$

The claim then follows from application of Lemma F.1 to sets S and I, since the total variation distance between the uniform distribution on S and the uniform distribution on I is at most 2ϵ. □

B.1.1 Proof of Lemma 4.2

Proof. If no elements are filtered out, then the inequality trivially holds. Suppose otherwise. The difference $|S' \triangle I_{\text{good}}| - |S \triangle I_{\text{good}}|$ is precisely the number of good elements (i.e. $i \in I_{\text{good}}$) filtered out in this iteration minus the number of bad elements filtered out in this iteration. Due to the random thresholding, the expectation of the former is $\frac{1}{\max \tau_i} \sum_{i \in S \cap I_{\text{good}}} \tau_i$, and the expectation of the latter is $\frac{1}{\max \tau_i} \sum_{i \in S \setminus I_{\text{good}}} \tau_i$. Thus, we need to show that $\sum_{i \in S \cap I_{\text{good}}} \tau_i \leq \sum_{i \in S \setminus I_{\text{good}}} \tau_i$.

Define $S_{\text{good}} = S \cap I_{\text{good}}$ and $S_{\text{bad}} = S \setminus I_{\text{good}}$. Let v be the largest eigenvector of $\text{Cov}_S(\xi_i)$. We have that

$$\begin{aligned}
\text{Var}_{S_{\text{good}}}(v \cdot \xi_i) &= \mathbb{E}_{S_{\text{good}}}(v \cdot \xi_i - \mathbb{E}_{S_{\text{good}}} v \cdot \xi_i)^2 \\
&\leq \mathbb{E}_{S_{\text{good}}}(v \cdot \xi_i - \mathbb{E}_{I_{\text{good}}} v \cdot \xi_i)^2 \\
&\leq 2\mathbb{E}_{I_{\text{good}}}(v \cdot \xi_i - \mathbb{E}_{I_{\text{good}}} v \cdot \xi_i)^2 \\
&= 2\,\text{Var}_{I_{\text{good}}}(v \cdot \xi_i) \\
&\leq 2M
\end{aligned}$$

where the first inequality uses the fact that variance is the smallest second moment obtainable by shifting; the second inequality uses that $|S_{\text{good}}| \geq (2/3 - 1/6)n \geq |I_{\text{good}}|/2$ and $S_{\text{good}} \subseteq I_{\text{good}}$; and the third inequality is by the lemma's assumption.

On the other hand, since the algorithm doesn't terminate, it holds that

$$\text{Var}_S(v \cdot \xi_i) = \frac{1}{|S|} \sum_{i \in S} \tau_i \geq 24M.$$

Defining $\mu_{\text{good}} = \mathbb{E}_{S_{\text{good}}} v \cdot \xi_i$, $\mu_{\text{bad}} = \mathbb{E}_{S_{\text{bad}}} v \cdot \xi_i$, and $\mu = \mathbb{E}_S v \cdot \xi_i$, it follows that

$$\frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}}} \tau_i = \mathbb{E}_{S_{\text{good}}}(v \cdot \xi_i - \mu)^2 \leq 2M + (\mu - \mu_{\text{good}})^2.$$

There are two cases to consider:

1. If $(\mu - \mu_{\text{good}})^2 \le 8M$. Then
$$\frac{1}{|S_{\text{good}}|}\sum_{i \in S_{\text{good}}} \tau_i \le 12M \le \frac{1}{2}\operatorname{Var}_S(v \cdot \xi_i) = \frac{1}{2|S|}\sum_{i \in S} \tau_i.$$
   Thus,
$$\sum_{i \in S_{\text{good}}} \tau_i \le \frac{1}{2}\sum_{i \in S} \tau_i \le \sum_{i \in S_{\text{bad}}} \tau_i.$$
2. If $(\mu - \mu_{\text{good}})^2 \ge 8M$. By the above calculation,
$$\frac{1}{|S_{\text{good}}|}\sum_{i \in S_{\text{good}}} \tau_i \le 1.5(\mu - \mu_{\text{good}})^2.$$
   On the other hand,
$$\frac{1}{|S_{\text{bad}}|}\sum_{i \in S_{\text{bad}}} \tau_i = \mathbb{E}_{S_{\text{bad}}}(v \cdot \xi_i - \mu)^2 \ge (\mu - \mu_{\text{bad}})^2.$$
   But $|\mu - \mu_{\text{good}}| \cdot |S_{\text{good}}| = |\mu - \mu_{\text{bad}}| \cdot |S_{\text{bad}}|$. As a result,
$$\sum_{i \in S_{\text{good}}} \tau_i \le 1.5|S_{\text{good}}| \cdot (\mu - \mu_{\text{good}})^2 = 1.5\frac{|S_{\text{bad}}|^2}{|S_{\text{good}}|}(\mu - \mu_{\text{bad}})^2 \le 1.5\frac{|S_{\text{bad}}|}{|S_{\text{good}}|}\sum_{i \in S_{\text{bad}}} \tau_i.$$
   But $1.5|S_{\text{bad}}|/|S_{\text{good}}| \le 1.5(n/6)/(2n/3 - n/6) \le 1$.

In either case, the desired claim holds. $\square$

C Omitted proofs from Section 5

C.1 Proof of Lemma 5.1

Proof. By the termination conditions of GMM-SEVER, no samples are filtered out in the last iteration. Thus, by Lemma 4.1 and the bounds $|S|, |I| \ge (1 - 10\epsilon n)$, since no samples are filtered out on Step 6, it holds that (for the vector $u = \mathbb{E}_S g(\hat{w})$ as defined in Step 5)
$$\begin{aligned}\left\|\mathbb{E}_S \nabla g(\hat{w})^T u - \mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w})^T u\right\|_2 &\le 3\sqrt{48}\sqrt{(L^2 \|u\|_2^2 + \left\|\operatorname{Cov}_{I_{\text{good}}}(\nabla g(\hat{w})^T u)\right\|_{\text{op}}) \cdot 10\epsilon} \\ &\le 36\sqrt{10}L\|u\|_2\sqrt{\epsilon}\end{aligned}$$
where the last inequality uses the guarantee of Lemma A.1 that $\mathbb{E}_{I_{\text{good}}}(u^T \nabla g(\hat{w}) v)^2 \le 2L^2$ for unit vectors u, v.

In the second filter operation, since no samples are filtered out, Lemma 4.1 implies that
$$\begin{aligned}\left\|\mathbb{E}_S g(\hat{w}) - \mathbb{E}_{I_{\text{good}}} g(\hat{w})\right\|_2 &\le 3\sqrt{48}\sqrt{(\sigma^2 L + 4L^2R^2 + \left\|\operatorname{Cov}_{I_{\text{good}}}(g(\hat{w}))\right\|_{\text{op}}) \cdot 10\epsilon} \\ &\le 36\sqrt{10}\sigma\sqrt{L\epsilon} + 120\sqrt{6}LR\sqrt{\epsilon}\end{aligned}$$
where the last inequality uses that $\operatorname{Cov}_{I_{\text{good}}}(g_i(\hat{w})) \preceq 2\sigma^2 L + 16L^2R^2$ by Lemma A.1. Next, since $\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w})\right\|_{\text{op}} \le \sqrt{2}L$ by Lemma A.1, it follows that
$$\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w})^T(\mathbb{E}_S g(\hat{w}) - \mathbb{E}_{I_{\text{good}}} g(\hat{w}))\right\|_2 \le 72\sqrt{5}\sigma L^{3/2}\sqrt{\epsilon} + 240\sqrt{3}L^2R\sqrt{\epsilon}.$$
Together with the first inequality, we get that
$$\left\|\mathbb{E}_S \nabla g(\hat{w})^T \mathbb{E}_S g(\hat{w}) - \mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w}) \mathbb{E}_{I_{\text{good}}} g(\hat{w})\right\|_2 \le 72\sqrt{5}\sigma L^{3/2}\sqrt{\epsilon} + 240\sqrt{3}L^2R\sqrt{\epsilon} + 36\sqrt{10}L\|u\|_2\sqrt{\epsilon}.$$
By assumption, $\left\|\mathbb{E}_{I_{\text{good}}} g(\hat{w})\right\|_2 \le \sigma\sqrt{L\epsilon} + L\|\hat{w} - w^*\|_2 \le \sigma\sqrt{L\epsilon} + 2LR$. Therefore $\|u\|_2 = \|\mathbb{E}_S g(\hat{w})\|_2 \le \sigma\sqrt{L} + 3LR$ assuming that $\max(36\sqrt{10}, 120\sqrt{6})\sqrt{\epsilon} \le 1$. Substituting this bound, we get
$$\left\|\mathbb{E}_S \nabla g(\hat{w})^T \mathbb{E}_S g(\hat{w}) - \mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w}) \mathbb{E}_{I_{\text{good}}} g(\hat{w})\right\|_2 \le (72\sqrt{5} + 36\sqrt{10})\sigma L^{3/2}\sqrt{\epsilon} + (240\sqrt{3} + 108\sqrt{3})L^2R\sqrt{\epsilon}.$$

Now recall that $\hat{w}$ is a γ-critical point of $\|\mathbb{E}_S g(w)\|_2^2$ in the region $B_R(w_0)$. Since $w^* \in B_R(w_0)$, the line segment between $\hat{w}$ and w^* is also contained in $B_R(w_0)$, so by definition of a γ-critical point, it holds that

$$(w^* - \hat{w}) \cdot \mathbb{E}_S \nabla g(\hat{w})^T \mathbb{E}_S g(\hat{w}) \geq -\gamma \left\|\hat{w} - w^*\right\|_2.$$

So by the triangle inequality, and rounding up the above constants to integers,

$$(\hat{w} - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(\hat{w})^T \mathbb{E}_{I_{\text{good}}} g(\hat{w}) \leq \gamma \left\|\hat{w} - w^*\right\|_2 + (275\sigma L^{3/2}\sqrt{\epsilon} + 603L^2 R\sqrt{\epsilon}) \left\|\hat{w} - w^*\right\|_2$$

as claimed. ☐

C.2 Proof of Lemma 5.2

Proof. Expanding $\mathbb{E}_{I_{\text{good}}} g(w) - g(w^*)$ as an integral, we have that

$$\begin{aligned}
&(w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \mathbb{E}_{I_{\text{good}}} (g(w) - g(w^*)) \\
&= (w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \int_0^1 \mathbb{E}_{I_{\text{good}}} \nabla g(tw + (1-t)w^*)(w - w^*)\, dt \\
&= (w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \int_0^1 \mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\, dt \\
&\quad + (w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \int_0^1 (\mathbb{E}_{I_{\text{good}}} \nabla g(tw + (1-t)w^*) - \mathbb{E}_{I_{\text{good}}} \nabla g(w))(w - w^*)\, dt.
\end{aligned}$$

Now, the first term is precisely $\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2^2$. We bound the absolute value of the second term by Cauchy-Schwarz and the Lipschitzness of the gradient; it is at most

$$\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2 \cdot L_g \left\|w - w^*\right\|_2^2.$$

As a result,

$$\begin{aligned}
(w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \mathbb{E}_{I_{\text{good}}} (g(w) - g(w^*)) \geq{}& \left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2^2 \\
&- L_g \left\|w - w^*\right\|_2^2 \left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2.
\end{aligned}$$

Suppose that $\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2 \leq 2L_g \left\|w - w^*\right\|_2^2$. Then by assumption that $\sigma_{\min}(\mathbb{E}_{I_{\text{good}}} \nabla g(w)) \geq \lambda$, it follows that $\left\|w - w^*\right\|_2 \leq (2L_g/\lambda) \left\|w - w^*\right\|_2^2$. Thus $\left\|w - w^*\right\|_2 \geq \lambda/(2L_g)$, which contradicts the assumptions that $R_0 < \lambda/(4L_g)$ and $w \in B_R(w_0) \subseteq B_{2R_0}(0)$. We conclude that in fact $\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2 > 2L_g \left\|w - w^*\right\|_2^2$, so that

$$(w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \mathbb{E}_{I_{\text{good}}} (g(w) - g(w^*)) \geq \frac{1}{2} \left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2^2.$$

However, by Assumption 3.1 and Lemma A.1,

$$|(w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \mathbb{E}_{I_{\text{good}}} g(w^*)| \leq \sqrt{2}\sigma L^{3/2} \left\|w - w^*\right\|_2 \sqrt{\epsilon}.$$

So together with the lemma's assumption,

$$(w - w^*)^T \mathbb{E}_{I_{\text{good}}} \nabla g(w)^T \mathbb{E}_{I_{\text{good}}} (g(w) - g(w^*)) \leq (\kappa + \sqrt{2}\sigma L^{3/2}\sqrt{\epsilon}) \left\|w - w^*\right\|_2.$$

As a result,

$$\left\|\mathbb{E}_{I_{\text{good}}} \nabla g(w)(w - w^*)\right\|_2^2 \leq 2(\kappa + \sqrt{2}\sigma L^{3/2}\sqrt{\epsilon}) \left\|w - w^*\right\|_2,$$

so that by the least singular value bound in Lemma A.1, $\left\|w - w^*\right\|_2 \leq 4(\kappa + \sqrt{2}\sigma L^{3/2}\sqrt{\epsilon})/\lambda^2$ as claimed. ☐

C.3 Proof of Theorem 5.4

Proof. For $t \geq 1$ let S_t be the algorithm's sample set at the beginning of the t-th iteration, so that $S_1 = [n]$. Define a "sticky" stochastic process based on $|S_t \triangle I_{\text{good}}|$:

$$X_t = \begin{cases} |S_t \triangle I_{\text{good}}| & \text{if } t = 1 \text{ or } |S_{t-1}| \geq 2n/3 \\ X_{t-1} & \text{otherwise} \end{cases}.$$

By soundness of the filtering algorithm (Lemma 4.2), we know that $(X_t)_{t\geq 1}$ is a super-martingale. By Ville's maximal inequality [11] and since $\mathbb{E}X_1 = \epsilon n$, it holds with probability at least $8/9$ that $\sup_t X_t \leq 9\epsilon n$. In this event, we claim that $|S_t| \geq 2n/3$ for all t. Indeed, this can be proved by induction: suppose that there is some t such that $|S_t| < 2n/3$ and let t^* be the minimal such t. Then $|S_{t^*-1}| \geq 2n/3$, so $|S_{t^*} \triangle I_{\text{good}}| = X_{t^*} \leq 9\epsilon n$. Therefore $|S_{t^*}| \geq |I_{\text{good}}| - 9\epsilon n \geq (1-\epsilon)n - 9\epsilon n \geq 2n/3$ since $10\epsilon n < n/3$. Contradiction, so $|S_t| \geq 2n/3$ for all t.

By definition of the process $(X_t)_t$ and by the preceding bound on $\sup_t X_t$, it follows that $\sup_t |S_t \triangle I_{\text{good}}| \leq 9\epsilon n$, and therefore $\inf_t |S_t| \geq (1-10\epsilon)n$. In particular, $|S| \geq (1-10\epsilon)n$, where S is the terminal sample set. Then by Lemma 5.3, it follows that

$$\|\hat{w} - w^*\|_2 \leq \frac{4\gamma}{\lambda^2} + 2412(L^2/\lambda^2)R\sqrt{\epsilon} + 1102\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}.$$

where $\hat{w}$ is the output of GMM-Sever. Since this bound holds deterministically whenever $|S| \geq (1-10\epsilon)n$, the failure probability can be decreased to δ by repeating GMM-Sever until either $|S| \geq (1-10\epsilon)n$, or $O(\log 1/\delta)$ repetitions have occurred.

The time complexity bound follows from observing that the FILTER algorithm runs in polynomial time, and in each repetition at least one sample is removed from S, so the algorithm terminates after at most n repetitions. □

C.4 Proof of Theorem 5.5

Proof. Formally, ITERATED-GMM-SEVER does the following procedure:

1. Initialize $t = 1$, $w_1 = 0$, $R_1 = R_0$, $\delta' = c\delta/\log(R\sqrt{L}/(\sigma\sqrt{\epsilon}))$, and $\gamma = \sigma L^{3/2}\sqrt{\epsilon}$
2. Compute $\hat{w}_t :=$ AMPLIFIED-GMM-SEVER$(\{g_1, \dots, g_n\}, w_t, R_t, \epsilon, \lambda, L, \sigma, \gamma, \delta')$
3. Set $R'_t := 4\gamma/\lambda^2 + 2412((L^2/\lambda^2)R_t\sqrt{\epsilon} + \sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon})$
4. If $R'_t > R_t/2$, then terminate and return $\hat{w}_t$. Otherwise, set $w_{t+1} := \hat{w}_t$, $R_{t+1} := R'_t$, and return to step (2).

First, note that by induction and the termination condition, R is halved in every iteration, so it holds for all $t \geq 1$ that $R_t \leq R_0/2^{t-1}$.

Runtime. The termination condition is deterministic. In particular, the algorithm will terminate once

$$\frac{4\gamma}{\lambda^2} + 2412((L^2/\lambda^2)R_t\sqrt{\epsilon} + \sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}) > R_t/2.$$

This holds if $R_t < 4824\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}$. Since R_t halves in every iteration and $\lambda \leq L$, the algorithm will therefore terminate after at most $O(\log(R\sqrt{L}/(\sigma\sqrt{\epsilon})))$ iterations. By the runtime bound on AMPLIFIED-GMM-SEVER, it follows that ITERATED-GMM-SEVER has time complexity $O(\text{poly}(n, d, p, T_\gamma) \cdot \log(1/\delta) \cdot \log(R\sqrt{L}/(\sigma\sqrt{\epsilon}))$.

Correctness. Next, we claim by induction that after the t-th call to AMPLIFIED-GMM-SEVER, it holds with probability at least $1 - \delta' t$ that $\|\hat{w}_t - w^*\|_2 \leq R'_t$. For $t = 1$, this follows from Theorem 5.4 and the assumption that $w^* \in B_{R_0}(0)$ (which implies that $w^* \in B_{R_1}(w_1) \subseteq B_{2R_0}(0)$).

Now fix any $t > 1$ for which the algorithm has not yet terminated, and condition on $\|\hat{w}_{t-1} - w^*\|_2 \leq R'_{t-1}$. Then by the triangle inequality,

$$\|w_t\|_2 = \|\hat{w}_{t-1}\|_2 \leq \|w^*\|_2 + R'_{t-1} \leq R_0 + R_0/2^{t-1}.$$

As a result, $B_{R_t}(w_t) \subseteq B_{R_0 + 2R_0/2^{t-1}}(0) \subseteq B_{2R_0}(0)$. In this event, by Theorem 5.4, it holds with probability at least $1 - \delta'$ that $\|\hat{w}_t - w^*\|_2 \leq R'_t$. By the induction hypothesis, the event we conditioned on occurs with probability at least $1 - \delta'(t-1)$, so by a union bound, it holds that $\|\hat{w}_t - w^*\|_2 \leq R'_t$ with probability at least $1 - \delta' t$, completing the induction.

Now consider the final iteration t. Restating the termination condition, we have

$$\frac{4\gamma}{\lambda^2} + 2412((L^2/\lambda^2)R_t\sqrt{\epsilon} + \sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}) > R_t/2.$$

By assumption that $(L^2/\lambda^2)\sqrt{\epsilon} \leq 1/9648$, it follows that

$$R_t \leq \frac{16\gamma}{\lambda^2} + 9648\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}.$$

Thus, with probability at least $1 - \delta' t$, the output $\hat{w}_t$ of ITERATED-GMM-SEVER satisfies

$$\begin{aligned}\|\hat{w}_t - w^*\|_2 &\leq R'_t \\ &\leq \frac{4\gamma}{\lambda^2} + \frac{R_t}{4} + 2412\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon} \\ &\leq \frac{4\gamma}{\lambda^2} + 4824\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon}.\end{aligned}$$

By the choice of γ, this bound is $O(\sigma(L^{3/2}/\lambda^2)\sqrt{\epsilon})$. By the iteration bound and choice of δ', the overall failure probability is at most δ. ☐

D Proof of Theorem 6.2

We need to prove that the contaminated samples $(X'_i, Y'_i, Z'_i)_{i=1}^n$ satisfy Assumption 3.1 with some set I_{good} of size $(1 - O(\epsilon))n$. To this end, it suffices to prove that with high probability over the original samples $(X_i, Y_i, Z_i)_{i=1}^n$, there is a subset I of these original samples, with $|I| \geq (1-\epsilon)n$, such that for any subset $S \subseteq I$ of size at least $(1-2\epsilon)n$, the conditions of Assumption 3.1 are satisfied. In this event, the intersection of I with the uncontaminated samples certifies the assumption.

In the subsequent lemmas, we verify one by one that each condition of Assumption 3.1 is satisfied with high probability for all subsets S of size at least $(1-\epsilon)n$ of a set I of size at least $(1-\epsilon/100)n)$; we then take the intersection of the sets I to yield a set I_{good} witnessing Assumption 3.1.

Lemma D.1. *Let $\epsilon > 0$ be sufficiently small. If $n \geq C(d+p)^3\sqrt{\tau}\log(1/\tau\epsilon)/\epsilon^2$ then with probability at least 0.99, there is a subset $I \subseteq [n]$ of size $|I| \geq (1-\epsilon/100)n$ such that for every subset $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that*

$$\left\| \frac{1}{|S|}\sum_{i \in S} \begin{bmatrix} Z \\ X \end{bmatrix} \begin{bmatrix} Z^T & X^T \end{bmatrix} - \mathbb{E}\begin{bmatrix} Z \\ X \end{bmatrix}\begin{bmatrix} Z^T & X^T \end{bmatrix} \right\|_{op} \leq O(\sqrt{\tau\epsilon}L).$$

As a consequence, if $(L/\lambda)\sqrt{\tau\epsilon}$ is less than a sufficiently small constant, then

$$\sigma_{\mathit{min}}\left(\frac{1}{|S|}\sum_{i \in S} Z_i X_i^T\right) \geq \lambda/2.$$

Proof. The first statement follows from Corollary F.4, τ-hypercontractivity of $[Z; X]$, and the covariance upper bound on $[Z; X]$. Let $\hat{M} = \frac{1}{|S|}\sum_{i \in S} Z_i X_i^T$. It follows from the first statement, that for any u,

$$\left\|\hat{M}u - \mathbb{E}ZX^T u\right\|_2 \leq O(L\sqrt{\tau\epsilon})\|u\|_2 \leq \frac{\lambda}{2}\|u\|_2.$$

By assumption that $\sigma_{\text{min}}(\mathbb{E}ZX^T) \geq \lambda$, it follows that $\left\|\hat{M}u\right\|_2 \geq (\lambda/2)\|u\|_2$. The second statement follows. ☐

Lemma D.2. *Let $\epsilon > 0$ and suppose that $n \geq C(p+d)^5\log((p+d)/\epsilon)/\epsilon^2$ for an appropriate constant C. Then with probability 0.98, there is a set $I \subseteq [n]$ with $|I| \leq (1-\epsilon)n$ such that $\mathbb{E}_I(u^T Z)^2(v^T X)^2 \preceq C\tau L^2 I$ for all unit vectors $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^d$.*

Proof. By hypercontractivity, we have

$$\mathbb{E}\langle X, u\rangle^4 \leq \tau\left(\mathbb{E}\langle X, u\rangle^2\right)^2 \leq \tau L^2\|u\|_2^4$$

for any vector $u \in \mathbb{R}^d$, and similarly for Z. Moreover, we have assumed that the coordinates of X and Z have 8th moments bounded by $O(\tau^2 L^4)$. Thus, we can apply Lemma F.6 to $X/\sqrt[4]{\tau L^2}$ and $Z/\sqrt[4]{\tau L^2}$ to get sets $I_1, I_2 \subseteq [n]$ each of size at least $(1-\epsilon/2)n$, that with probability 0.98 satisfy

$$\frac{1}{|I_1|}\sum_{i \in I_1}\langle X_i, u\rangle^4 \leq C\tau L^2$$

and

$$\frac{1}{|I_2|}\sum_{i\in I_2}\langle Z_i, v\rangle^4 \le C\tau L^2$$

for all unit vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^p$. Let $I = |I_1 \cap I_2|$. Then $|I| \ge (1-\epsilon)n$, and the above bounds hold over I as well up to a constant factor loss. Thus,

$$\mathbb{E}_I(u^T Z)^2(X^T v)^2 \le \sqrt{\mathbb{E}_I\langle Z, u\rangle^4 \mathbb{E}_I\langle X, v\rangle^4} \le C\tau L^2.$$

The lemma follows. □

Lemma D.3. *Let* $\epsilon > 0$ *and suppose that* $n \ge C(p+d)^3/\epsilon^2$. *Then with probability* 0.99*, there is a set* $I \subseteq [n]$ *with* $|I| \ge (1-\epsilon)n$ *such that* $\mathbb{E}_I(v^T Z\xi)^2 \le C\sigma^2 L$ *for every unit vector* $v \in \mathbb{R}^p$.

Proof. Since $\mathbb{E}[\xi^2|Z] \le \sigma^2$, observe that $\mathbb{E}(v^T Z\xi)^2 \le \sigma^2\mathbb{E}(v^T Z)^2 \le \sigma^2 L$ for every unit vector $v \in \mathbb{R}^p$. The claim follows from Corollary F.4. □

Lemma D.4. *Let* $\epsilon > 0$*, and suppose that* $n \ge C(p^{3/2}/\epsilon)\log(p)$. *With probability* 0.99*, there is a subset* $I \subseteq [n]$ *with* $|I| \ge (1-\epsilon/100)n$ *such that for every* $S \subseteq I$ *with* $|S| \ge (1-\epsilon)n$*, it holds that*

$$\|\mathbb{E}_S Z\xi\|_2 \le O(\sigma\sqrt{L\epsilon}).$$

Proof. Observe that $\mathbb{E}Z\xi = 0$ and $\mathbb{E}ZZ^T\xi^2 \preceq \sigma^2 LI$ by assumption. The claim follows from Lemma F.5. □

Corollary D.5. *Let* $\epsilon > 0$. *Suppose that* $\epsilon < c\min(\lambda^2/(\tau L^2), \lambda^4/L^4)$ *for a sufficiently small constant* $c > 0$*, and suppose that* $n \ge C(d+p)^5\tau\log((p+d)/\tau\epsilon)/\epsilon^2$ *for a sufficiently large constant* C. *Then with probability at least* 0.95*, there is a set* $I \subseteq [n]$ *with* $|I| \ge (1-\epsilon/2)n$ *such that for every subset* $S \subseteq I$ *with* $|S| \ge (1-\epsilon)n$*, the following hold:*

- $\sigma_{min}(\mathbb{E}_S\nabla g(w^*)) \ge \Omega(\lambda)$
- $\mathbb{E}_S(u^T\nabla g(w)v)^2 \le O(\tau L^2)$ *for all unit vectors* u, v *and all* w
- $\mathbb{E}_S(v^T g(w^*))^2 \le O(\sigma^2 L)$ *for all unit vectors* v
- $\|\mathbb{E}_S g(w^*)\|_2 \le O(\sigma\sqrt{L\epsilon})$
- $\nabla g(w)$ *is constant in* w.

Proof. Let $I_1, I_2, I_3, I_4 \subseteq [n]$ be the sets guaranteed by Lemma D.1 (with parameter ϵ), Lemma D.2 (with parameter $\epsilon/100$), Lemma D.3 (with parameter $\epsilon/100$), and D.4 (with parameter ϵ), which satisfy the claims of the respective lemmas with probability at least 0.95. Let $I = I_1 \cap I_2 \cap I_3 \cap I_4$. We have that $|I_1|, |I_2|, |I_3|, |I_4| \ge (1-\epsilon/100)n$, so I is a subset of each of I_1, I_2, I_3, I_4 of size at least $(1-\epsilon/2)n$. Let $S \subseteq I$ have $|S| \ge (1-\epsilon)n$. By Lemma D.1 and since $S \subseteq I_1$, it holds that $\sigma_{\min}(\mathbb{E}_S\nabla g(w^*)) \ge \lambda/2$. By Lemma D.2 and since $S \subseteq I_2$, it holds that $\mathbb{E}_S(u^T\nabla g(w)v)^2 \le 2\mathbb{E}_{I_2}(u^T\nabla g(w)v)^2 \le O(\tau L^2)$. By Lemma D.3 we have $\mathbb{E}_S(v^T g(w^*))^2 \le O(\sigma^2 L)$, and by Lemma D.4 we have $\|\mathbb{E}_S g(w^*)\|_2 \le o(\sigma\sqrt{L\epsilon})$. Finally, $\nabla g(w) = ZX^T$ is clearly constant in w. □

The above corollary validates Assumption 3.1 for linear instrumental variables. Since $\nabla g(w)$ is constant in w, the Assumption holds for any bound R_0 on the norm of the true solution w^*. Formally, we can instantiate Theorem 5.5 to get a provably robust estimator for instrumental variables linear regression, as stated in Theorem 6.2.

Remark 1. Although Theorem 6.2 is stated with a constant probability of failure, this is only for simplicity of presentation; in fact, the probabilities of failure all decay exponentially with n, once n exceeds the sample complexity stated in the theorem.

E Proof of Theorem 6.3

Let (X_i, Y_i, Z_i) be n independent samples drawn according to (X, Y, Z). Let $\epsilon > 0$. We prove that under the above assumptions, if n is sufficiently large, then with high probability, for any ϵ-contamination $(X_i', Y_i', Z_i')_{i=1}^n$ of $(X_i, Y_i, Z_i)_{i=1}^n$, the functions $g_i(w) = Z_i'(Y_i' - G((X_i')^T w))$ satisfy Assumption 3.1. The proof is similar to the previous section, with slight complications introduced by the non-linearity of the non-linear function G.

Lemma E.1. *Let $\epsilon > 0$. Suppose that $n \geq C(p+d)^5 \log((p+d)/\epsilon)/\epsilon^4$ for an appropriate constant C. Then with probability at least 0.97, there is a set $I \subseteq [n]$ with $|I| \geq (1-\epsilon/100)n$ such that for every $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that*

$$\left\|\mathbb{E}_S Z X^T G'(X^T w) - \mathbb{E}_S Z X^T G'(X^T w^*)\right\|_{op} \leq O(\sqrt{\tau L^3}\left\|w - w^*\right\|_2).$$

Proof. Let I_1 be the set guaranteed by Lemma D.2 with parameter $\epsilon/200$, and let I_2 be the set guaranteed by applying Corollary F.4 to $X_1, \dots, X_n$ with parameter $\epsilon/200$. Take $I = I_1 \cap I_2$, so that $|I| \geq (1-\epsilon/100)n$. Let $S \subseteq I$ with $|S| \geq (1-\epsilon)n$. By Cauchy-Schwarz, we have that

$$\begin{aligned}\left\|\mathbb{E}_S Z X^T (G'(X^T w) - G'(X^T w^*))\right\|_{\text{op}} &= \sup_{\|u\|=\|v\|=1} \mathbb{E}_S u^T Z X^T v (G'(X^T w) - G'(X^T w^*)) \\ &\leq \sup_{\|u\|=\|v\|=1} \sqrt{\mathbb{E}_S (u^T Z X^T v)^2 \mathbb{E}_S (G'(X^T w) - G'(X^T w^*))^2}.\end{aligned}$$

First, by the guarantee of Lemma D.2, we have

$$\mathbb{E}_S (u^T Z X^T v)^2 \leq O(\tau L^2).$$

Second, by the guarantee of Corollary F.4 and Lipschitzness of G', we have

$$\mathbb{E}_S (G'(X^T w) - G'(X^T w^*))^2 \leq \mathbb{E}_S (X^T (w - w^*))^2 \leq O(L\left\|w - w^*\right\|_2^2).$$

Together,

$$\left\|\mathbb{E}_S Z X^T (G'(X^T w) - G'(X^T w^*))\right\|_{\text{op}} \leq O(\sqrt{\tau L^3}\left\|w - w^*\right\|_2^2)$$

as claimed. □

Lemma E.2. *Let $\epsilon > 0$ and suppose that $n \geq C(p+d)^5 \log((p+d)/\epsilon)/\epsilon^2$ for an appropriate constant C. Then with probability 0.98, there is a set $I \subseteq [n]$ with $|I| \geq (1-\epsilon)n$ such that*

$$\mathbb{E}_I (u^T Z X^T G'(X^T w) v)^2 \leq O(\tau L^2)$$

for all $w \in \mathbb{R}^d$ and unit vectors u, v.

Proof. Let I be the set guaranteed by Lemma D.2. Simply note that since G is 1-Lipschitz,

$$\mathbb{E}_I (u^T Z X^T G'(X^T w) v)^2 \leq \mathbb{E}(u^T Z)^2 (X^T v)^2 \leq O(\tau L^2)$$

for all unit vectors u, v. □

Lemma E.3. *Let $\epsilon > 0$. Suppose that $n \geq C(p+d)^5 \tau \log((p+d)/\tau\epsilon)/\epsilon^2$ for an appropriate constant C. There is a set $I \subseteq [n]$ with $|I| \geq (1-\epsilon/50)n$ such that for every $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that*

$$\sigma_{min}(\mathbb{E}_S Z X^T G'(X^T w^*)) \geq \lambda/16.$$

Proof. Let I_1 be the set guaranteed by Lemma D.1 with parameter ϵ, and let I_2 be the set guaranteed by Lemma E.1 with parameter ϵ. Let $I = I_1 \cap I_2$, so that $|I| \geq (1-\epsilon/50)n$. Pick any $S \subseteq I$ with $|S| \geq (1-\epsilon)n$. Then $\sigma_{\text{min}}(\mathbb{E}_S Z X^T G'(X^T 0)) \geq \lambda/8$ by Lemma D.1 (and since $G'(0) = 1/4$), and $\left\|\mathbb{E}_S Z X^T (G'(0) - G'(X^T w^*))\right\|_{\text{op}} \leq O(\tau L^2 \left\|w^*\right\|_2)$ by Lemma E.1. It follows that

$$\sigma_{\text{min}}(\mathbb{E}_S Z X^T G'(X^T w^*)) \geq \lambda/8 - O(\sqrt{\tau L^3}\left\|w^*\right\|_2) \geq \lambda/16,$$

where the last inequality is by assumption that $\left\|w^*\right\|_2 \leq R_0 \leq O(\lambda/\sqrt{\tau L^3}))$. □

Lemma E.4. *Let $\epsilon > 0$ and suppose that $n \geq C(p+d)^3/\epsilon^2$. Then with probability 0.99, there is a set $I \subseteq [n]$ with $|I| \geq (1-\epsilon)n$ such that*

$$\mathbb{E}_I(v^T Z\xi)^2 \leq O(\sigma^2 L)$$

for all unit vectors v.

Proof. By assumption, $\mathbb{E}(v^T Z\xi)^2 = \mathbb{E}(v^T Z)^2(Y - G(X^T w^*))^2 \leq \sigma^2 L$. So we can apply Corollary F.4 to conclude. $\square$

Lemma E.5. *Let $\epsilon > 0$, and suppose that $n \geq C(p^{3/2}/\epsilon)\log(p)$. With probability 0.99, there is a subset $I \subseteq [n]$ with $|I| \geq (1-\epsilon/100)n$ such that for every $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that*

$$\|\mathbb{E}_S Z\xi\|_2 \leq O(\sigma\sqrt{L\epsilon}).$$

Proof. Observe that $\mathbb{E}Z\xi = 0$ and $\mathbb{E}ZZ^T\xi^2 \preceq \sigma^2 LI$ by assumption. The claim follows from Lemma F.5. $\square$

As a result of the above lemmas, we get the following corollary, just as in the previous section.

Corollary E.6. *Let $\epsilon > 0$. Suppose that $\epsilon < c\min(\lambda^2/(\tau L^2), \lambda^4/L^4)$ for a sufficiently small constant $c > 0$, and suppose that $n \geq C(d+p)^5\tau\log((p+d)/\tau\epsilon)/\epsilon^2$ for a sufficiently large constant C. Suppose that $R_0 \leq c\min(\lambda^2/L, \lambda/(\tau L^2))$. Then with probability at least 0.95, there is a set $I \subseteq [n]$ with $|I| \geq (1-\epsilon/2)n$ such that for every subset $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, the following hold:*

- $\sigma_{\mathit{min}}(\mathbb{E}_S \nabla g(w^*)) \geq \Omega(\lambda)$
- $\mathbb{E}_S(u^T \nabla g(w) v)^2 \leq O(\tau L^2)$ *for all unit vectors u, v and all $w \in B_{R_0}(0)$*
- $\mathbb{E}_S(v^T g(w^*))^2 \leq O(\sigma^2 L)$ *for all unit vectors v*
- $\|\mathbb{E}_S g(w^*)\|_2 \leq O(\sigma\sqrt{L\epsilon})$
- $\|\mathbb{E}_S \nabla g(w) - \mathbb{E}_S \nabla g(w^*)\|_2 \leq O(\sqrt{\tau L^3}\,\|w - w^*\|_2)$ *for all $w \in B_{R_0}(0)$*

This corollary validates Assumption 3.1 for logistic instrumental variables, and proves Theorem 6.3.

F Technical lemmas

In this section we collect technical lemmas that are needed for our proof. Most of these results are standard in the robust statistics literature (see, e.g., [19]).

The following fact is key to the filtering algorithm and various other bounds.

Lemma F.1. *Let P, Q be distributions on $\mathbb{R}^d$. Let $\epsilon \in [0, 1/2)$ and suppose that $\mathrm{TV}(P, Q) = \epsilon$ and $\|\mathrm{Cov}_P\|_{op}, \|\mathrm{Cov}_Q\|_{op} \leq \sigma^2$. Then if $X \sim P$ and $Y \sim Q$, it holds that*

$$\|\mathbb{E}X - \mathbb{E}Y\|_2 \leq C_\epsilon \sigma\sqrt{\epsilon}$$

where $C_\epsilon = \sqrt{6/(1-4\epsilon^2)}$.

Proof. Since $\mathrm{TV}(P, Q) = \epsilon$ there is some coupling under which $\Pr(X \neq Y) = \epsilon$. As a result,

$$\mathbb{E}[X] - \mathbb{E}[Y] = \epsilon(\mathbb{E}[X|X \neq Y] - \mathbb{E}[Y|X \neq Y]).$$

Thus we have that:

$$\begin{aligned} \|\mathbb{E}X - \mathbb{E}Y\|_2^2 &= \epsilon^2 \|\mathbb{E}[X|X \neq Y] - \mathbb{E}[Y|X \neq Y]\|_2^2 \\ &\leq \epsilon^2 \sup_{v \in \mathbb{R}^d : \|v\|_2 = 1} (v \cdot (\mathbb{E}[X|X \neq Y] - \mathbb{E}[Y|X \neq Y]))^2 \end{aligned}$$

Let $v \in \mathbb{R}^d$ be a unit vector. Bounding the means of $X|X \neq Y$ and $Y|X \neq Y$ by second moments around $\mathbb{E}X$, we have that

$$\begin{aligned}(\mathbb{E}[v \cdot X|X \neq Y] - \mathbb{E}[v \cdot Y|X \neq Y])^2 &= (\mathbb{E}[v \cdot (X - \mathbb{E}X)|X \neq Y] - \mathbb{E}[v \cdot (Y - \mathbb{E}X)|X \neq Y])^2 \\ &\leq 2\mathbb{E}[v \cdot (X - \mathbb{E}X)|X \neq Y]^2 + 2\mathbb{E}[v \cdot (Y - \mathbb{E}X)|X \neq Y]^2 \\ &\leq 2\mathbb{E}[(v \cdot (X - \mathbb{E}X))^2|X \neq Y] + 2\mathbb{E}[(v \cdot (Y - \mathbb{E}X))^2|X \neq Y]\end{aligned}$$

By law of total probability,

$$\mathbb{E}[(v \cdot (X - \mathbb{E}X))^2|X \neq Y] \leq \epsilon^{-1}\mathbb{E}[(v \cdot (X - \mathbb{E}X))^2] = \epsilon^{-1}v^T \operatorname{Cov}(X)v \leq \sigma^2/\epsilon.$$

Similarly,

$$\begin{aligned}\mathbb{E}[(v \cdot (Y - \mathbb{E}X))^2|X \neq Y] &\leq 2\mathbb{E}[(v \cdot (Y - \mathbb{E}Y))^2|X \neq Y] + 2(v \cdot (\mathbb{E}Y - \mathbb{E}X))^2 \\ &\leq 2\epsilon^{-1}\mathbb{E}[(v \cdot (Y - \mathbb{E}Y))^2] + 2\left\|\mathbb{E}Y - \mathbb{E}X\right\|_2^2 \\ &\leq 2\sigma^2/\epsilon + 2\left\|\mathbb{E}Y - \mathbb{E}X\right\|_2^2.\end{aligned}$$

As a result, we get that:

$$(\mathbb{E}[v \cdot X|X \neq Y] - \mathbb{E}[v \cdot Y|X \neq Y])^2 \leq 6\sigma^2/\epsilon + 4\left\|\mathbb{E}Y - \mathbb{E}X\right\|_2^2.$$

We conclude that

$$\left\|\mathbb{E}X - \mathbb{E}Y\right\|_2^2 \leq \ 6\sigma^2\epsilon + 4\epsilon^2\left\|\mathbb{E}X - \mathbb{E}Y\right\|_2^2$$

Re-arranging we get the desired inequality. ☐

The above lemma implies that if an adversary is allowed to corrupt an ϵ-fraction of data, and the original distribution has variance no more than σ^2 in any direction, then the corrupted mean must be within $O(\sigma\sqrt{\epsilon})$ of the original mean, unless the corrupted distribution has significantly larger variance.

Lemma F.2. *Let $\epsilon, \delta > 0$. Suppose that $X_1, \ldots, X_n, X$ are independent and identically distributed with $\mathbb{E}XX^T = I_d$. Suppose that $n \geq Cd^3\log(3/\delta)/\epsilon^2$. Then with probability 0.99 there is a subset $I \subseteq [n]$ with $|I| \geq (1-\epsilon)n$ such that*

$$\frac{1}{n}\sum_{i \in I} X_i X_i^T \preceq (1+\delta)I_d$$

and as a consequence

$$\frac{1}{|I|}\sum_{i \in I} X_i X_i^T \preceq \frac{1+\delta}{1-\epsilon}I_d.$$

Proof. Since $\mathbb{E}\left\|X\right\|_2^2 = \operatorname{Tr}(I_d) = d$, we have that $\Pr[\left\|X\right\|_2^2 \geq 2d/\epsilon] \leq \epsilon/2$. Define $I = \{i \in [n] : \left\|X_i\right\|_2^2 \leq 2d/\epsilon\}$. By a Chernoff bound, we have $|I| \geq (1-\epsilon)n$ with probability $1 - \exp(-\Omega(\epsilon n))$. Fix a unit vector $u \in \mathbb{R}^d$ and define

$$A_i = \langle X_i, u\rangle^2 \mathbb{1}[\left\|X_i\right\|_2^2 \leq 2d/\epsilon]$$

for $i \in [n]$. We have that $\mathbb{E}[A_i] \leq \mathbb{E}\langle X_i, u\rangle^2 = 1$, and also $A_1, \ldots, A_n$ are independent and uniformly bounded by $2d/\epsilon$. Thus, Hoeffding's inequality implies that

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} A_i \geq 1+\delta\right] \leq \exp(-2n\delta^2/(2d/\epsilon)^2).$$

Define

$$f(u) = \frac{1}{n}\sum_{i \in I}\langle X_i, u\rangle^2 = \frac{1}{n}\sum_{i=1}^{n} A_i.$$

For any fixed unit vector u we've shown that $f(u) \leq 1+\delta$ with probability $1 - \exp(-\Omega(n\delta^2\epsilon^2/d^2))$. Let $\mathcal{N}$ be a net of the unit ball in $\mathbb{R}^d$ with resolution α and cardinality at most $(3/\alpha)^d$. By a union bound, it holds that $f(u) \le 1+\delta$ for all $u \in \mathcal{N}$ with probability $1-\exp(d\log(3/\alpha) - \Omega(n\delta^2\epsilon^2/d^2))$. But now

$$|f(u) - f(v)| \le \frac{1}{n}\sum_{i\in I} |\langle X_i, u-v\rangle| \cdot |\langle X_i, u+v\rangle| \le \sqrt{f(u-v)f(u+v)|}$$

for any vectors u, v. Define

$$M = \left\| \frac{1}{n}\sum_{i\in I} X_i X_i^T \right\|_{\text{op}} = \max_{\|u\|=1} f(u).$$

Then

$$M \le 1 + \delta + \sqrt{\alpha M \cdot 2M}.$$

Taking $\alpha = \delta^2/2$, we get that $M \le (1+\delta)/(1-\delta) \le 1 + 4\delta$. So long as $n \ge Cd^3\log(3/\delta)/\epsilon^2$ for a large enough constant C, it holds with probablity at least 0.99 that

$$\frac{1}{n}\sum_{i\in I} X_i X_i^T \preceq (1+4\delta) I_d$$

and moreover $|I| \ge (1-\epsilon)n$. By the latter inequality it also follows that

$$\frac{1}{|I|}\sum_{i\in I} X_i X_i^T \preceq \frac{1+4\delta}{1-\epsilon} I_d$$

as claimed. ☐

Lemma F.3. *Let $\epsilon, \tau > 0$. Suppose that $X_1, \dots, X_n, X$ are independent and identically distributed with $\mathbb{E}XX^T = I_d$. Suppose that $\mathbb{E}\langle u, X\rangle^4 \le \tau(\mathbb{E}\langle u, X\rangle^2)^2$ for all $u \in \mathbb{R}^d$. Suppose that $n \ge Cd\sqrt{\tau}\log(1/(\tau\epsilon))/\epsilon^{3/2}$ for an appropriate absolute constant C. Then with probability 0.99, there is a subset $I \subseteq [n]$ with $|I| \ge (1-\epsilon/100)n$ such that for any $S \subseteq I$ with $|S| \ge (1-\epsilon)n$ it holds that*

$$\frac{1}{|S|}\sum_{i\in S} X_i X_i^T \succeq (1 - 7\sqrt{\tau\epsilon}) I_d.$$

Proof. Let $I \subseteq [n]$ be the subset guaranteed by Lemma F.2, with the properties that $|I| \ge (1-\epsilon/100)n$ and $\frac{1}{|I|}\sum_{i\in I} X_i X_i^T \preceq (1+\epsilon) I_d$.

Fix a unit vector $u \in \mathbb{R}^d$. Let q be such that $\Pr(\langle X_i, u\rangle^2 \ge q) = 4\epsilon$. Define $B_i = \mathbb{1}[\langle X_i, u\rangle^2 \ge q]$. By a Chernoff bound, it holds with probability $1 - \exp(-\Omega(\epsilon n))$ that $\sum_{i=1}^n B_i \ge \epsilon n$. Thus the size of the set $Q = \{i \in [n] : \mathbb{1}[\langle X_i, u\rangle^2 < q]\}$ is at most $(1-\epsilon)n$. As a result, any $S \subseteq [n]$ with $|S| \ge (1-\epsilon)n$, must either contain all elements from the set Q or elements from its complement, whose values $\langle X_i, u\rangle^2$ dominate the value of any element in Q. More formally: note that $|S \cap Q| + |Q - S| = |Q| \le |S| = |S \cap Q| + |S \cap Q^c| \implies |Q - S| \le |S \cap Q^c|$. Since every element in $S \cap Q^c$ has value $\langle X_i, u\rangle^2$ larger than any element in $Q - S$, we thus have: $\sum_{i\in S\cap Q^c}\langle X_i, u\rangle^2 \ge \sum_{i\in Q-S}\langle X_i, u\rangle^2$. Thus, it holds that

$$\sum_{i\in S}\langle X_i, u\rangle^2 \ge \sum_{i\in Q}\langle X_i, u\rangle^2 = \sum_{i=1}^n \langle X_i, u\rangle^2 \mathbb{1}[\langle X_i, u\rangle^2 < q].$$

Next, note that $\langle X_i, u\rangle^2\mathbb{1}[\langle X_i, u\rangle^2 < q]$ is bounded by q^2. Since

$$4\epsilon = \Pr(\langle X, u\rangle^2 \ge q) \le q^{-2}\mathbb{E}\langle X, u\rangle^4 \le q^{-2}\tau\left(\mathbb{E}\langle X, u\rangle^2\right)^2 \le \tau/q^2,$$

we have that $q^2 \le \tau/(4\epsilon)$. Therefore by Bernstein's inequality, with probability

$$1-\exp\left(-\Omega(n\frac{\tau\epsilon}{\mathbb{E}[\langle X_i, u\rangle^4] + q^2\sqrt{\tau\epsilon}})\right) = 1-\exp\left(-\Omega(n\frac{\tau\epsilon}{\tau + \frac{\tau}{4\epsilon}\sqrt{\tau\epsilon}})\right) = 1-\exp\left(-\Omega(n\frac{\epsilon^{3/2}}{\sqrt{\tau}})\right)$$

we have that

$$\frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle^2\mathbb{1}[\langle X_i, u\rangle^2 < q] \ge \mathbb{E}\langle X, u\rangle^2\mathbb{1}[\langle X, u\rangle^2 < q] - \sqrt{\tau\epsilon}.$$

But now

$$\begin{aligned}\mathbb{E}\langle X,u\rangle^2 \mathbb{1}[\langle X,u\rangle^2 < q] &= \mathbb{E}\langle X,u\rangle^2 - \mathbb{E}\langle X,u\rangle^2 \mathbb{1}[\langle X,u\rangle^2 \geq q] \\ &\geq 1 - \sqrt{\mathbb{E}\langle X,u\rangle^4 \Pr(\langle X,u\rangle^2 \geq q)} \\ &\geq 1 - \sqrt{4\tau\epsilon}.\end{aligned}$$

Thus, with probability $1-\exp(-\Omega(\epsilon n))-\exp(-\Omega(n\epsilon^{3/2}/\sqrt{\tau}))$, for all $S \subseteq [n]$ with $|S| \geq (1-\epsilon)n$, we have that

$$\sum_{i\in S}\langle X_i,u\rangle^2 \geq (1-3\sqrt{\tau\epsilon})n.$$

Assume moreover that $S \subseteq I$. Define $f(u) = \sum_{i\in S}\langle X_i,u\rangle^2$. Then for any vectors u,v, we have by Cauchy-Schwarz that

$$|f(u)-f(v)| = \sum_{i\in S}\langle X_i,u\rangle^2 - \langle X_i,v\rangle^2 \leq \sqrt{f(u-v)f(u+v)}.$$

Since $S \subseteq I$ we have that $\sum_{i\in S} X_i X_i^T \preceq 2nI_d$. So

$$|f(u)-f(v)| \leq 2n\left\|u-v\right\|_2 \left\|u+v\right\|_2.$$

Fix a net on the unit sphere in $\mathbb{R}^d$, with resolution $\sqrt{\tau\epsilon}$ and cardinality $(O(1)/\sqrt{\tau\epsilon})^d$. Then with probability $1-\exp(O(d\log(1/(\tau\epsilon)))-\Omega(n\epsilon^{3/2}/\sqrt{\tau}))$ the lower bound holds for all u in the net and all $S \subseteq I$ of size $|S| \geq (1-\epsilon)n$. As a result, for any unit vector $v \in \mathbb{R}^d$ and any such S, it holds that

$$\sum_{i\in S}\langle X_i,u\rangle^2 \geq (1-3\sqrt{\tau\epsilon})n - 4n\sqrt{\tau\epsilon}.$$

We conclude that

$$\frac{1}{|S|}\sum_{i\in S} X_i X_i^T \succeq (1-7\sqrt{\tau\epsilon})I_d.$$

So long as $n \geq Cd\sqrt{\tau}\log(1/(\tau\epsilon))/\epsilon^{3/2}$ for a sufficiently large constant C, this holds with probability at least 0.99 as claimed. ☐

Corollary F.4. *Let $\epsilon,\tau > 0$ be sufficiently small. Suppose that $X_1,\dots,X_n$ are independent and identically distributed d-dimensional random vectors, with positive-definite covariance $\mathbb{E}XX^T = \Sigma$. Suppose that $\mathbb{E}\langle u,X\rangle^4 \leq \tau(\mathbb{E}\langle u,X\rangle^2)^2$ for all u. Suppose that $n \geq Cd^3\sqrt{\tau}\log(1/\tau\epsilon)/\epsilon^2$ for a large constant C. Then with probability 0.99 there is a subset $I \subseteq [n]$ with $|I| \geq (1-\epsilon/100)n$ such that for every subset $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that*

$$(1-O(\sqrt{\tau\epsilon}))\Sigma \preceq \frac{1}{|S|}\sum_{i\in S} X_i X_i^T \preceq (1+O(\epsilon))\Sigma.$$

Proof. We apply Lemmas F.2 and F.3 to $\Sigma^{-1/2}X_1,\dots,\Sigma^{-1/2}X_n$. For the upper bound, we observe that if it holds for I then it holds for every large subset S with only an additional factor of $1+O(\epsilon)$. For the lower bound, we note that hypercontractivity is preserved under this linear transformation. ☐

Lemma F.5. *Let $\epsilon,\sigma > 0$. Let $X_1,\dots,X_n,X$ be i.i.d. d-dimensional random vectors with $\mathbb{E}X = 0$ and $\mathbb{E}XX^T \preceq \sigma^2 I$. If $n \geq C(d^{3/2}/\epsilon)\log(d)$ for a sufficiently large constant C, then with probability at least 0.99, there is a subset $I \subseteq [n]$ with $|I| \geq (1-\epsilon/100)n$ such that for every $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that $\left\|\mathbb{E}_S X\right\|_2 \leq O(\sigma\sqrt{\epsilon})$.*

Proof. Since $\mathbb{E}\left\|X\right\|_2^2 \leq \sigma^2 d$, we have that $\Pr[\left\|X\right\|_2^2 \geq 200\sigma^2 d/\epsilon] \leq \epsilon/200$. Define $I = \{i \in [n] : \left\|X_i\right\|_2^2 \leq 200\sigma^2 d/\epsilon\}$. By a Chernoff bound, we have $|I| \geq (1-\epsilon/100)n$ with probability $1-\exp(-\Omega(\epsilon n))$. Now

$$\mathbb{E}XX^T\mathbb{1}[\left\|X\right\|_2^2 \geq 200\sigma^2 d/\epsilon] \preceq \mathbb{E}XX^T \preceq \sigma^2 I,$$

and the random variables $X_i X_i^T \mathbb{1}[\|X_i\|_2^2 \geq 200\sigma^2 d/\epsilon]$ are independent and bounded in operator norm by $200\sigma^2 d/\epsilon$. Thus, we can apply the Matrix Chernoff bound [28] to get

$$\Pr\left[\frac{1}{n}\sum_{i\in I} X_i X_i^T \preceq 2e\sigma^2 I\right] \geq 1 - d\exp(-2e\sigma^2 n(\epsilon/200\sigma^2 d)\log(2)) \geq 0.999 \tag{2}$$

so long as $n \geq C(d/\epsilon)\log(d)$ for a sufficiently large constant C. Moreover, for any unit vector v,

$$\begin{aligned}\mathbb{E}v^T X\mathbb{1}[\|X\|_2^2 \leq 200\sigma^2 d/\epsilon] &= -\mathbb{E}v^T X\mathbb{1}[\|X\|_2^2 > 200\sigma^2 d/\epsilon] \\ &\leq \sqrt{\mathbb{E}(v^T X)^2 \Pr(\|X\|_2^2 > 200\sigma^2 d/\epsilon)} \\ &\leq \sigma\sqrt{\epsilon}.\end{aligned}$$

Since $X\mathbb{1}[\|X\|_2^2 \leq 200\sigma^2 d/\epsilon]$ is bounded in norm by $\sigma\sqrt{200d/\epsilon}$, a Bernstein bound implies that for any unit vector v,

$$\begin{aligned}\Pr\left(\left|v\cdot\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right| > 1.5\sigma\sqrt{\epsilon}\right) &\leq \exp\left(-\Omega\left(\frac{n\sigma^2\epsilon}{\mathbb{E}[(v^T X)^2] + (\sigma\sqrt{d/\epsilon})(\sigma\sqrt{\epsilon})}\right)\right). \\ &\leq \exp\left(-\Omega\left(\frac{n\sigma^2\epsilon}{\sigma^2+\sigma^2\sqrt{d}}\right)\right).\end{aligned}$$

Take a net over unit vectors in $\mathbb{R}^d$ of granularity $1/100$ and cardinality $\exp(O(d))$. Then the above inequality holds for all v in the net, with probability $\exp(O(d) - \Omega(n\epsilon/(\sqrt{d}))$, which is at least 0.999 if $n \geq Cd^{3/2}/\epsilon$ for an appropriate constant C.

Let N denote the aforementioned net of the unit ball in $\mathbb{R}^d$. We have that in the aforementioned event:

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right\|_2 &= \sup_{w\in\mathbb{R}^d:\|w\|_2=1}\left|w\cdot\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right| \\ &\leq \sup_{v\in N}\left|w\cdot\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right| \\ &\quad + \frac{1}{100}\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right\|_2\end{aligned}$$

Re-arranging yields:

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right\|_2 \leq \frac{100}{99}1.5\sigma\sqrt{\epsilon} \leq 2\sigma\sqrt{\epsilon}$$

Therefore

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\mathbb{1}[\|X_i\|_2^2 \leq 200\sigma^2 d/\epsilon]\right\|_2 \leq 2\sigma\sqrt{\epsilon}\right) \geq 0.999 \tag{3}$$

In the intersection of the above events described by Equations 2 and 3, and the event that $|I| \geq (1-\epsilon/100)n$, which together occur with probability at least 0.99, we get that $\mathrm{Cov}_I(X) \preceq 4e\sigma^2 I$ and $\|\mathbb{E}_I X\|_2 \leq 4\sigma\sqrt{\epsilon}$. By Lemma F.1, for any $S \subseteq I$ with $|S| \geq (1-\epsilon)n$, it holds that $\|\mathbb{E}_I X - \mathbb{E}_S X\|_2 \leq O(\sigma\sqrt{\epsilon})$ so in fact $\|\mathbb{E}_S X\|_2 \leq O(\sigma\sqrt{\epsilon})$. □

Lemma F.6. *Let $\epsilon > 0$. Let $X_1, \ldots, X_n, X$ be independent and identically distributed d-dimensional random vectors with $\mathbb{E}\langle X, u\rangle^4 \leq \|u\|_2^4$ for all $u \in \mathbb{R}^d$ and coordinate-wise bounded 8-th moments, i.e. $\max_{i=1}^d \mathbb{E}X_i^8 \leq C_8$. Suppose that $n \geq Cd^5\log(d/\epsilon)/\epsilon^2$ for a sufficiently large constant C. With probability at least 0.99, there is a set $I \subseteq [n]$ with $|I| \geq (1-\epsilon)n$ such that*

$$\frac{1}{|I|}\sum_{i\in I}\langle X_i, u\rangle^4 \leq c\,\|u\|_2^4$$

for all $u \in \mathbb{R}^p$ and an absolute constant c.

Proof. Since $\mathbb{E}\|X\|_2^2 = \mathrm{Tr}(\mathbb{E}XX^T) \le d$ (since $(\mathbb{E}\langle u, X\rangle^2)^2 \le \mathbb{E}\langle X, u\rangle^4 \le 1$ for any unit vector u), we have that

$$\Pr[\|X\|_2^2 \ge 2d/\epsilon] \le \epsilon/2.$$

By a Chernoff bound, we have that $|\{i \in [n] : \|X_i\|_2^2 \ge 2d/\epsilon\}| \le \epsilon n$ with probability at least $1 - \exp(-\Omega(\epsilon n))$. Now fix a unit vector $u \in \mathbb{R}^d$ and define

$$A_i = \langle X_i, u\rangle^4 \mathbb{1}[\|X_i\|_2^2 \le 2d/\epsilon]$$

for $i \in [n]$. We have that

$$\mathbb{E}[A_i] \le \mathbb{E}\langle X_i, u\rangle^4 \le 1$$

and also $A_1, \dots, A_n$ are independent and uniformly bounded by $(2d/\epsilon)^2$. Thus, the Bernstein bound implies that

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n A_i > 2\right] \le \exp\left(-c_1 \frac{n}{\mathbb{E}\langle X, u\rangle^8 + (2d/\epsilon)^2}\right).$$

for some universal constant c_1. Note that:

$$\mathbb{E}\langle X, u\rangle^8 \le \mathbb{E}\|X\|_2^8 = \mathbb{E}\left(\sum_{i=1}^d X_i^2\right)^4 = d^4\mathbb{E}\left(\frac{1}{d}\sum_i X_i^2\right)^4$$
$$\le d^4\mathbb{E}\frac{1}{d}\sum_i X_i^8 \le d^4 \max_{i=1}^d \mathbb{E}X_i^8 \le d^4 C$$

Thus:

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n A_i > 2\right] \le \exp\left(-c_1 \frac{n}{Cd^4 + (2d/\epsilon)^2}\right).$$

Take $I = \{i : \|X_i\|_2^2 \le 2d/\epsilon\}$. For any fixed unit vector $u \in \mathbb{R}^d$, it holds that $\frac{1}{n}\sum_{i\in I}\langle X_i, u\rangle^4 \le 2$ with probability $\exp(-\Omega(n/(d^4/\epsilon^2)))$. Take $\delta = \epsilon^2/d^2$. We can union bound over a δ-net of the unit ball in $\mathbb{R}^d$, which has cardinality at most $(3/\delta)^d$, and note that

$$\left|\frac{1}{n}\sum_{i\in I}\langle X_i, u\rangle^4 - \frac{1}{n}\sum_{i\in I}\langle X_i, v\rangle^4\right| \le C(2d/\epsilon)^2\|u - v\|_2,$$

so in fact it holds that

$$\frac{1}{n}\sum_{i\in I}\langle X_i, u\rangle^4 \le 2 + C(2d/\epsilon)^2\delta \le C'$$

for all unit vectors $u \in \mathbb{R}^d$, with probability

$$1 - \exp\left(O(d\log(d/\epsilon)) - \Omega\left(\frac{n}{Cd^4 + (d/\epsilon)^2}\right)\right) \ge 0.999$$

since $n \ge C'd^5\log(d/\epsilon)/\epsilon^2$ for a sufficiently large constant C'. Finally, it also holds that $|I| \ge (1-\epsilon)n$ with probability $1 - \exp(-\Omega(\epsilon n))$. It therefore holds with probability at least 0.99 that for all unit vectors $u \in \mathbb{R}^d$,

$$\frac{1}{|I|}\sum_{i\in I}\langle X_i, u\rangle^4 \le C''$$

as claimed. □

G Supplementary experimental details

G.1 Implementation details

Iterated-GMM-Sever. For practical simplicity (e.g. decreasing the number of hyperparameters), we make several modifications in the implementation. First, instead of updating the constraint ball radius R_t based on ϵ, σ, L, λ, and ϵ, we simply halve it in each iteration (i.e. $R_{t+1} \leftarrow R_t/2$), and quit after a fixed number of iterations T. Throughout our experiments, we take $T = 10$. Second, for computational reasons, we omit the amplification step. Third, we took hyperparameters σ and L to be equal; both are only used to set the filtering thresholds, and while tuning them separately could potentially improve performance, we did not attempt to do so. Fourth, since all of our experiments are for IV linear regression where the number of instruments equals the number of covariates, the γ-approximate critical point oracle $\mathcal{L}$ (line 4 of GMM-SEVER) can be implemented exactly: the IV moment condition $\mathbb{E}_S g_i(w) = \mathbb{E}_S Z X^T (y - \langle X, w \rangle) = 0$ has a closed-form solution, which is a global zero (and therefore minimizer) of $\|\mathbb{E}_S g_i(w)\|_2^2$ (this solution $\hat{w}$ may not lie in the constraint set $\|w\|_2 \le R$, but in practice this is not an issue).

As a result of these simplifications, our implementation only depends on two hyperparameters: the initial estimate error R_0, and a threshold parameter L. We pick these hyperparameters ad-hoc without serious tuning attempts. For synthetic experiments we pick R_0 equal to the dimension, since the ground truth parameters have coordinates $O(1)$; for the NYSLM data we take $R_0 = 20$. For the Varied Instrument Strength experiment, we take $L = 0.1$; for the synthetic Heterogeneous Effects experiment we take $L = 0.25$; and for the NYSLM dataset we take $L = 0.01$. Note that in particular we do not need to vary the hyperparameters as the corruption level or instrument strength change.

Two-stage Huber Regression. As a baseline robust IV estimator, we implement two-stage Huber regression [20]. Concretely, the classical IV estimator can be implemented as two-stage least squares: first, regress each covariate against the instruments via Ordinary Least Squares. Second, regress the response against the predicted covariates via Ordinary Least Squares. This can be robustified by replacing Ordinary Least Squares with Huber regression. We implement Huber regression via the function `sklearn.linear_model.HuberRegressor` (with default robustness parameter) in scikit-learn [25].

Classical IV regression. This baseline is simply the estimator which solves the empirical moment condition $\mathbb{E}_S Z X^T (y - \langle X, w \rangle) = 0$ over the whole sample set $S = [n]$.

G.2 Corruption method for NLSYM dataset

We randomly pick ϵn of the datapoints and corrupt the responses of these datapoints so that if the original IV estimate was w^*, then the new IV estimate is roughly $-w^*$. Formally, if $S \subseteq [n]$ is the set of corrupted samples, we set $y_S = q - X_S w^*$ where q solves the linear system $Z_S^T q = -2 Z_{S^c}^T X_{S^c} w^*$. Here, Z and X are matrices of the instruments and samples respectively, and y is the vector of responses. Since w^* approximately solves the moment conditions on the uncorrupted samples, it follows that

$$Z^T y = Z_{S^c}^T y_{S^c} + Z_S^T q \approx Z_{S^c}^T X_{S^c} w^* - 2 Z_{S^c}^T X_{S^c} w^* = -Z_{S^c}^T X_{S^c} w^*,$$

so the corrupted IV estimate is approximately $-w^*$.

G.3 Hyperparameter stability

The primary hyperparameter which governs the performance of our algorithm ITERATED-GMM-SEVER is L, which roughly corresponds to the threshold for filtering outliers. Obviously, if L is chosen too large, then the algorithm will fail to remove outliers. Thus, it's important to understand how to choose L in practice. We chose L ad-hoc without significant tuning. We also experimentally verify that our algorithm is not undueley sensitive to the choice of L, by repeating some of our main experimental results for varied L. We find that there is a sort of "phase transition" in L, beyond which the algorithm fails to identify outliers, but below the transition point, the algorithm is fairly robust to choice of L. See Figure 3, where we estimate the ATE on the uncorrupted NLSYM data as L varies, and get consistent results; Figure 4, where we measure our algorithm's error on the synthetic Heterogeneous Effects dataset with $0.1n$ corruptions as L varies; and Figure 5, where we measure the error in ATE of our algorithm on the corrupted NLSYM data as L varies.

G.4 Computational details

All experiments were done in Python on a Microsoft Surface Laptop, and each plot took at most 12 hours to generate.

G.5 Omitted Figures

Algorithm	**Median ℓ_2 recovery error**	**(25th percentile, 75th percentile)**
Iterated-Gmm-Sever	2.14	(1.71, 2.91)
IV estimator	6.63	(4.29, 13.95)
Two-stage Huber estimator	2.74	(2.11, 4.38)
Zero estimator	4.32	(3.90, 2.79)

Table 1: Median ℓ_2 recovery error of Iterated-Gmm-Sever and three baselines on uncorrupted synthetic dataset for IV regression with Heterogeneous Effects ($n = 10^3$ and $d = 20$, with 100 independent trials)

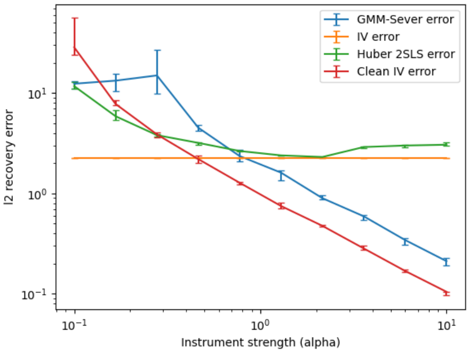


Figure 2: Varied Instrument Strength experiment with $d = 100$. Since the response noise has variance d, to maintain the same signal-to-noise ratio as the $d = 20$ experiment, we set $\theta^* = \sqrt{d/20}(1, 0, \ldots, 0) \in \mathbb{R}^d$. Otherwise, the generative model and corruptions are the same as in the $d = 20$ experiment. When the instruments are weak, all estimators have large error (the corrupted IV error is exactly $\sqrt{100/20}$ by construction: the corruptions were chosen so that the IV estimate is the trivial estimate 0). However, as the instrument strength increases, the clean IV error improves, and our estimator's error improves as well, roughly tracking the clean error up to a constant factor.

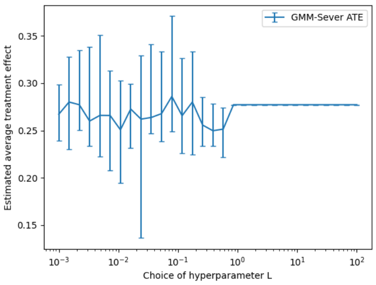


Figure 3: Median estimated ATE of ITERATED-GMM-SEVER on the NLSYM data (with no added corruption) as we vary the hyperparameter L (which controls the algorithm's outlier-removal threshold) from 0.01 to 100. For each choice of L, we took the median over 50 repetitions of our (randomized) algorithm. When L is very large, the algorithm removes no outliers. This plot shows that even when the algorithm aggressively removes outliers (i.e. L is very small), the estimated ATE is quite stable, providing evidence for the robustness of the inference of [6] (that education has a positive effect on wages).

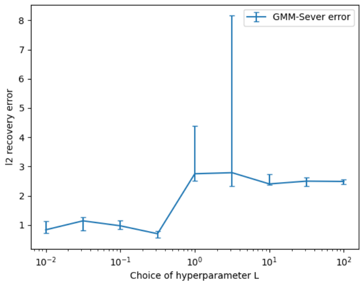


Figure 4: Median ℓ_2 recovery error of ITERATED-GMM-SEVER on synthetic Heterogeneous Effects dataset with $0.1n$ added corruptions, as we vary the hyperparameter L from 0.01 to 100. For each choice of L, we took the median over 10 independent trials (i.e. each trial resamples which subset of $0.1n$ samples to corrupt).

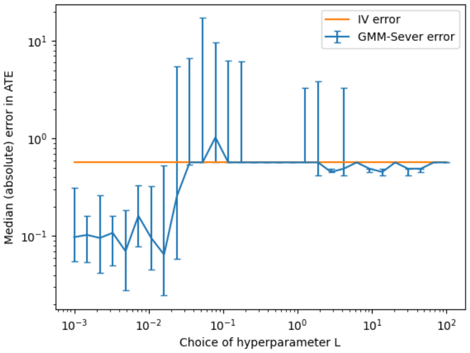


Figure 5: Median ATE error of ITERATED-GMM-SEVER and IV on NLSYM data with $0.1n$ corrupted samples, as we vary the hyperparameter L. For each choice of L, we took the median error (in ATE) over 50 runs of ITERATED-GMM-SEVER. Note that both axes are on log-scale.