Robust Supervised Learning with Coordinate Gradient Descent

Ibrahim Merad* Stéphane Gaïffas†

July 25, 2023

Abstract

This paper considers the problem of supervised learning with linear methods when both features and labels can be corrupted, either in the form of heavy tailed data and/or corrupted rows. We introduce a combination of coordinate gradient descent as a learning algorithm together with robust estimators of the partial derivatives. This leads to robust statistical learning methods that have a numerical complexity *nearly identical* to non-robust ones based on empirical risk minimization. The main idea is simple: while robust learning with gradient descent requires the computational cost of robustly estimating the whole gradient to update all parameters, a parameter can be updated immediately using a robust estimator of a single partial derivative in coordinate gradient descent. We prove upper bounds on the generalization error of the algorithms derived from this idea, that control both the optimization and statistical errors with and without a strong convexity assumption of the risk. Finally, we propose an efficient implementation of this approach in a new `Python` library called `linlearn`, and demonstrate through extensive numerical experiments that our approach introduces a new interesting compromise between robustness, statistical performance and numerical efficiency for this problem.

Keywords. Robust methods; Heavy-tailed data; Outliers; Robust gradient descent; Coordinate gradient descent; Generalization error.

1 Introduction

Outliers and heavy tailed data are a fundamental problem in supervised learning. As explained by [47], an outlier is a sample that differs from the data's "global picture". A rule-of-thumb is that a typical data set may contain between 1% and 10% of outliers [46], or even more than that depending on the considered application. For instance, the inherently complex and random nature of users' web browsing makes web-marketing data sets contain a significant proportion of outliers and have heavy-tailed distributions [44]. Statistical handling of outliers was already considered in the early 50's [32, 43] and motivated in the 70's the development of *robust statistics* [56, 57].

Setting. In this paper, we consider the problem of large-scale supervised learning, where we observe possibly corrupted samples $(X_i, Y_i)_{i=1}^n$ of a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with distribution P, where $\mathcal{X} \subset \mathbb{R}^d$ is the feature space and $\mathcal{Y} \subset \mathbb{R}$ is the set of label values. We focus on linear methods, where the learning task corresponds to finding an approximation of an optimal parameter

$$\theta^\star \in \operatorname*{argmin}_{\theta \in \Theta} R(\theta) \quad \text{where} \quad R(\theta) := \mathbb{E}\big[\ell(X^\top \theta, Y)\big], \tag{1}$$

*LPSM, UMR 8001, Université Paris Diderot, Paris, France

†LPSM, UMR 8001, Université Paris Diderot, Paris, France and DMA, École normale supérieure

where Θ is a convex compact subset of $\mathbb{R}^d$ with diameter Δ containing the origin and $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function satisfying the following. We denote $\ell'(z, y) := \partial \ell(z, y)/\partial z$.

Assumption 1. *The loss $z \mapsto \ell(z, y)$ is convex for any $y \in \mathcal{Y}$, differentiable and γ-smooth in the sense that $|\ell'(z, y) - \ell'(z', y)| \leq \gamma |z - z'|$ for all $z, z' \in \mathbb{R}$ and $y \in \mathcal{Y}$. Moreover, there exists $q \in [1, 2]$, which we will call the asymptotic polynomial degree, and positive constants $C_{\ell,1}, C_{\ell,2}, C'_{\ell,1}$ and $C'_{\ell,2}$ such that*

$$|\ell(z, y)| \leq C_{\ell,1} + C_{\ell,2}|z - y|^q \quad \text{and} \quad |\ell'(z, y)| \leq C'_{\ell,1} + C'_{\ell,2}|z - y|^{q-1}$$

for all $z \in \mathbb{R}$ and $y \in \mathcal{Y}$.

Note that Assumption 1 holds for the majority of loss functions used both for regression and classification, such as the square loss $\ell(z, y) = (z - y)^2/2$ with $q = 2$ or the Huber loss [55] $\ell(z, y) = r_\tau(z - y)$ for $z, y \in \mathbb{R}$ with $\gamma = 1$ and $q = 1$, where $r_\tau(u) = \frac{1}{2}u^2 \mathbf{1}_{|u| \leq \tau} + \tau(|u| - \frac{1}{2}\tau)\mathbf{1}_{|u| > \tau}$ with $\tau > 0$ and the logistic loss $\ell(z, y) = \log(1 + e^{-yz})$ for $z \in \mathbb{R}$ and $y \in \{-1, 1\}$ with $\gamma = 1/4$ and $q = 1$. We will see shortly that a smaller degree q associated to the loss entails looser requirements on the data distribution. If P were known, one could approximate $\theta^\star$ using a first-order optimization algorithm such as *gradient descent* (GD), using iterations of the form

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla R(\theta_t) \quad \text{with} \quad \nabla R(\theta) = \mathbb{E}[\ell'(X^\top \theta, Y) X] \tag{2}$$

for $t = 1, 2, \ldots$ where $\eta > 0$ is a learning rate.

Empirical risk minimization. With P unknown, most supervised learning algorithms rely on *empirical risk minimization* (ERM) [102, 40], which requires (a) the fact that samples are independent and with the same distribution P and (b) that P has sub-Gaussian tails, as explained below. Such assumptions are hardly ever met in practice, and entail implicitly that, for real-world applications, the construction of a training data set requires involved data preparation, such as outlier detection and removal, data normalization and other issues related to feature engineering [110, 65]. An *implicit*[1] ERM estimator of $\theta^\star$ is a minimizer of the empirical risk R_n given by

$$\widehat{\theta}_n^{\texttt{erm}} \in \operatorname*{argmin}_{\theta \in \Theta} R_n(\theta) \quad \text{where} \quad R_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(X_i^\top \theta, Y_i), \tag{3}$$

for which one can prove sub-Gaussian deviation bounds under strong hypotheses such as boundedness of ℓ or sub-Gaussian concentration [80, 69]. In the general case, ERM leads to poor estimations of $\theta^\star$ whenever (a) and/or (b) are not met, corresponding to situations where (a) the data set contains outliers and (b) the data distribution has heavy tails. This fact motivated the theory of robust statistics [55, 57, 45, 46, 99]. The poor performance of ERM stems from the loose deviation bounds of the empirical mean estimator. Indeed, as explained by [17] for the estimation of the expectation of a real random variable, the Chebyshev inequality provably provides the best concentration bound for the empirical mean estimator in the general case, so that the error is $\Omega(1/\sqrt{n\delta})$ for a confidence $1 - \delta$. Gradient Descent (GD) combined with ERM leads to an *explicit* algorithm using iterations (2) with gradients estimated by an average over the samples

$$\widehat{\nabla}^{\texttt{erm}} R(\theta) := \nabla R_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell'(X_i^\top \theta, Y_i) X_i, \tag{4}$$

which is, as explained above, a poor estimator of $\nabla R(\theta)$ beyond (a) and (b).

[1] By *implicit*, we mean defined as the argmin of some functional, as opposed to the *explicit* iterations of an optimization algorithm: an implicit estimator differs from the exact algorithm applied on the data, while an *explicit* algorithm does not.

Robust gradient descent. A growing literature about robust GD estimators [94, 74, 51, 42] suggests to perform GD iterations with $\widehat{\nabla}^{\mathtt{erm}} R(\theta)$ replaced by some robust estimator of $\nabla R(\theta)$. An implicit estimator is considered by [68], based on the minimization of a robust estimate of the risk objective using median-of-means. Robust estimators of $\nabla R(\theta)$ can be built using several approaches including geometric median-of-means [94]; robust coordinate-wise estimators [50] based on a modification of [17]; coordinate-wise median-of-means or trimmed means [74] or robust vector means through projection and truncation [94]. Other works achieve robustness by performing standard training on disjoint subsets of data and aggregating the resulting estimators into a robust one [82, 11]. We discuss such alternative methods in more details in Section 4 below.

These procedures based on GD require to run *costly* subroutines (at the exception of [68, 42]) that induce a considerable computational overhead compared to the non-robust approach based on ERM. The aim of this paper is to introduce *robust* and *explicit* learning algorithms, with performance guarantees under weak assumptions on $(X_i, Y_i)_{i=1}^n$, that have a computational cost *comparable* to the non-robust ERM approach. As explained in Section 2 below, the main idea is to combine *coordinate gradient descent* with robust estimators of the partial derivatives $\partial R(\theta)/\partial \theta_j$, that are scalar (univariate) functionals of the unknown distribution P.

We denote $|A|$ as the cardinality of a finite set A and use the notation $[\![k]\!] = \{1, \dots, k\}$ for any integer $k \in \mathbb{N} \setminus \{0\}$. We denote x^j as the j-th coordinate of a vector x. We will work under the following assumption.

Assumption 2. *The indices of the training samples $[\![n]\!]$ can be divided into two disjoint subsets $[\![n]\!] = \mathcal{I} \cup \mathcal{O}$ of* outliers $\mathcal{O}$ *and* inliers $\mathcal{I}$ *for which we assume the following*: (a) *we have $|\mathcal{I}| > |\mathcal{O}|$*; (b) *the pairs $(X_i, Y_i)_{i\in\mathcal{I}}$ are i.i.d with distribution P and the outliers $(X_i, Y_i)_{i\in\mathcal{O}}$ are arbitrary*; (c) *there is $\alpha \in (0, 1]$ such that*

$$\mathbb{E}\big[|X^j|^{\max(2,q(\alpha+1))}\big] < +\infty, \quad \mathbb{E}\big[|Y^{q-1}X^j|^{1+\alpha}\big] < +\infty \quad \textit{and} \quad \mathbb{E}\big[|Y|^q\big] < +\infty \tag{5}$$

for any $j \in [\![d]\!]$ where $q \in [1, 2]$ is the loss' asymptotic polynomial degree from Assumption 1.

Assumption 2 is purposely vague about $|\mathcal{I}|$ and $|\mathcal{O}|$ and the value of $\alpha \in (0, 1]$. Indeed, conditions on $|\mathcal{O}|$ and α will depend on the considered robust estimator of the partial derivatives, as explained in Section 3 below, including theoretical guarantees with $\alpha < 1$ and cases with $\mathbb{E}[Y^2] = +\infty$ (for the Huber loss for instance). The existence of a second moment for X is indispensable for the objective $R(\theta)$ to be Lipschitz-smooth, see Section 2.2 below.

Square loss. For the square loss we have $q = 2$ and $\mathbb{E}[Y^2] < +\infty$ is required for the risk $R(\theta)$ and its partial derivatives to be well-defined. Note that we have $\mathbb{E}[|\ell'(X^\top\theta, Y)X^j|^{1+\alpha}] = \mathbb{E}[|YX^j|^{1+\alpha}]$ for $\theta = 0 \in \Theta$, which makes (5) somewhat minimal in order to ensure the existence of the moment we need for the loss derivative for all $\theta \in \Theta$.

Huber loss. For the Huber loss, we have $q = 1$ and the only requirement on Y is $\mathbb{E}|Y| < +\infty$ and we have $\max(2, q(\alpha + 1)) = 2$ ensuring that $\mathbb{E}[|X^j|^2] < +\infty$, a requirement for the Lipschitz-smoothness of $R(\theta)$, as detailed in Section 2.2.

Logistic loss. For the logistic loss we have $|Y| \leq 1$ and $q = 1$ so that the only assumption is once again $\mathbb{E}[|X^j|^2] < +\infty$.

Main contributions. This paper introduces a new interesting compromise between robustness, statistical performance and numerical efficiency for supervised learning with linear methods through the following main contributions:

- We introduce a new approach for robust supervised learning with linear methods by combining coordinate gradient descent (CGD) with robust estimators of the partial derivatives

used in its iterations (Section 2). This novel and intuitive idea turns out to be very effective experimentally (see Section 6) and amenable to an in-depth theoretical analysis (see Section 2.2 for guarantees under strong convexity and Section 5 without it).

- We consider state-of-the-art robust estimators of the partial derivatives (Section 3) and provide theoretical guarantees for CGD combined with each of them. For some robust estimators, our analysis requires only weak moments (allowing $\mathbb{E}[Y^2] = +\infty$ in some cases) together with strong corruption (large $|\mathcal{O}|$) which lets our results apply to very general settings with minimal assumptions compared to the relevant literature. We provide guarantees for several variants of CGD namely random uniform sampling, importance sampling and deterministic sampling of the coordinates (Section 2.2).

- We perform extensive numerical experiments, both for regression and classification on several data sets (Section 6). We compare many combinations of gradient descent, coordinate gradient descent and robust estimators of the gradients and partial derivatives. Some of these combinations correspond to state-of-the-art algorithms [67, 50, 94], and we also consider several additional baselines such as Huber regression [89], classification with the modified Huber loss [109], Least Absolute Deviation (LAD) [35] and RANSAC [37]. We carry out an in-depth experimental comparison of state-of-the-art robust methods for supervised linear learning both in terms of statistical precision and numerical complexity. We thereby demonstrate the outstanding performance of our methods on both aspects.

- All the algorithms studied and compared in the paper are made easily accessible in a few lines of code through a new `Python` library called `linlearn`, open-sourced under the BSD-3 License on `GitHub` and available here[2]. This library follows the API conventions of `scikit-learn` [91].

2 Robust coordinate gradient descent

CGD is well-known for its efficiency and fast convergence properties based on both theoretical and practical studies [88, 96, 41, 106] and is the de-facto standard optimization algorithm used in many machine learning libraries. In this paper, we suggest to use CGD with robust estimators $\widehat{g}_j(\theta)$ of the partial derivatives $g_j(\theta) := \partial R(\theta)/\partial \theta_j \in \mathbb{R}$ of the true risk given by Equation (1), several robust estimators $\widehat{g}_j(\theta)$ are described in Section 3 below.

2.1 Iterations

At iteration $t+1$, given the current iterate $\theta^{(t)}$, CGD proceeds as follows. It chooses a coordinate $j_t \in [\![d]\!]$ (several sampling mechanisms are possible, as explained below) and the parameter is updated using

$$\begin{cases} \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} - \beta_j \widehat{g}_j(\theta^{(t)}) & \text{if } j = j_t \\ \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} & \text{otherwise} \end{cases} \tag{6}$$

for all $j \in [\![d]\!]$, where $\beta_j > 0$ is a step-size for coordinate j. A *single* coordinate is updated at each iteration of CGD, and we will designate d iterations of CGD as a *cycle*. The CGD procedure is summarized in Algorithm 1 below, where we denote by $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ the features matrix with rows $X_1^\top, \ldots, X_n^\top$ and where $\boldsymbol{X}_\bullet^j \in \mathbb{R}^n$ stands for its j-th column.

A simple choice for the distribution p is the uniform distribution over $[\![d]\!]$, but improved convergence rates can be achieved using importance sampling, as explained in Theorem 1 below,

[2] `https://github.com/linlearn/linlearn`

Algorithm 1 Robust coordinate gradient descent

1: **Inputs:** Learning rates $\beta_1, \dots, \beta_d > 0$; estimators $(\widehat{g}_j(\cdot))_{j=1}^d$ of the the partial derivatives; initial parameter $\theta^{(0)}$; distribution $p = [p_1 \cdots p_d]$ over $[\![d]\!]$ and number of iterations T.
2: Compute $I^{(0)} \leftarrow \boldsymbol{X}\theta^{(0)}$
3: **for** $t = 0, \dots, T-1$ **do**
4: Sample a coordinate $j_t \in \{1, \dots, d\}$ with distribution p independently of $j_1, \dots, j_{t-1}$
5: Compute $\widehat{g}_{j_t}(\theta^{(t)})$ using $I^{(t)}$ and put $D^{(t)} \leftarrow -\beta_{j_t}\widehat{g}_{j_t}(\theta^{(t)})$
6: Update the inner products using $I^{(t+1)} \leftarrow I^{(t)} + \boldsymbol{X}_\bullet^{j_t} D^{(t)}$
7: Apply the update $\theta_{j_t}^{(t+1)} \leftarrow \theta_{j_t}^{(t)} + D^{(t)}$
8: **end for**
9: **return** The last iterate $\theta^{(T)}$

where the choice of the step-sizes $(\beta_j)_{j=1}^d$ is described as well. The partial derivatives estimators $(\widehat{g}_j(\cdot))_{j=1}^d$ described in Section 3 will determine the statistical error of this explicit learning procedure. Note that line 6 of Algorithm 1 uses the fact that

$$\begin{aligned} I^{(t+1)} = \boldsymbol{X}\theta^{(t+1)} &= \sum_{j \neq j_t} \boldsymbol{X}_\bullet^j \theta_j^{(t+1)} + \boldsymbol{X}_\bullet^{j_t} \theta_{j_t}^{(t+1)} \\ &= \sum_{j \neq j_t} \boldsymbol{X}_\bullet^j \theta_j^{(t)} + \boldsymbol{X}_\bullet^{j_t} \big(\theta_{j_t}^{(t)} + D^{(t)}\big) = I^{(t)} + \boldsymbol{X}_\bullet^{j_t} D^{(t)}. \end{aligned}$$

This computation has complexity $O(n)$, and we will see in Section 3 that the complexity of the considered robust estimators $\widehat{g}_{j_t}(\theta^{(t)})$ at line 5 is also $O(n)$, so that the overall complexity of one iteration of robust CGD is also $O(n)$. This makes the complexity of one cycle of robust CGD $O(nd)$, which corresponds to the complexity of *one iteration of GD using the non-robust estimator* $\widehat{\nabla}^{\mathtt{erm}} R(\theta)$, see Equation (4). A more precise study of these complexities is discussed in Section 3, see in particular Table 1. Moreover, we will see experimentally in Section 6 that our approach is indeed very competitive in terms of the compromise between computational cost and statistical accuracy, compared to all the considered baselines.

Comparison with robust gradient descent. Robust estimators of the expectation of a random vector (such as the geometric median by [82]) require to solve a d-dimensional optimization problem at each iteration step while, in the univariate case, a robust estimator of the expectation can be obtained at a cost comparable to that of an ordinary empirical average. Of course, one can combine such univariate estimators into a full gradient: this is considered for instance by [50, 51, 52, 74, 98], but this approach accumulates errors into the overall estimation of the gradient. This paper introduces an alternative method, where univariate estimators of the partial derivatives are used *immediately* to update the current iterate. We believe that this is the main benefit of using CGD in this context: even if our theoretical analysis hardly explains this, our understanding is that one iteration of CGD is impacted by the estimator error of a *single* partial derivative, that can be corrected straight away in the next iteration, while one iteration of GD is impacted by the accumulated estimation errors of the d partial derivatives, when using d univariate estimators for efficiency, instead of a computationally involved d-dimensional estimator (such as geometric median).

2.2 Theoretical guarantees under strong convexity

In this Section, we provide theoretical guarantees in the form of upper bounds on the risk $R(\theta^{(T)})$ (see Equation (1)) for the output $\theta^{(T)}$ of Algorithm 1. These upper bounds are generic with respect to the considered robust estimators $\big(\widehat{g}_j(\cdot)\big)_{j=1}^d$ and rely on the following definition.

Definition 1. *Let $\delta \in (0,1)$ be a failure probability. We say that a partial derivatives estimator $\widehat{g}$ has an* error vector $\epsilon(\delta) \in \mathbb{R}_+^d$ *if it satisfies*

$$\mathbb{P}\Big[\sup_{\theta\in\Theta} \big|\widehat{g}_j(\theta) - g_j(\theta)\big| \leq \epsilon_j(\delta)\Big] \geq 1-\delta \tag{7}$$

for all $j \in [\![d]\!]$.

In Section 3 below, we specify a value of $\epsilon_j(\delta)$ for each considered robust estimator which will lead to upper bounds on the risk. Recall that $g_j(\theta) = \partial R(\theta)/\partial\theta_j$ and let us denote as e_j the j-th canonical basis vector of $\mathbb{R}^d$. We need the following extra assumptions on the optimization problem itself.

Assumption 3. *There exists $\theta^\star \in \Theta$ satisfying the stationary gradient condition $\nabla R(\theta^\star) = 0$. Moreover, we assume that there are Lipschitz constants $L_j > 0$ such that*

$$\big|g_j(\theta + he_j) - g_j(\theta)\big| \leq L_j|h|$$

for any $j \in [\![d]\!]$, $h \in \mathbb{R}$ and $\theta \in \Theta$ such that $\theta + he_j \in \Theta$. We also consider $L > 0$ such that

$$\big\|g(\theta+h) - g(\theta)\big\| \leq L\|h\|$$

for any $h \in \Theta$ and $\theta \in \Theta$ such that $\theta + h \in \Theta$. We denote $L_{\max} := \max_{j\in[\![d]\!]} L_j$ and $L_{\min} := \min_{j\in[\![d]\!]} L_j$.

Under Assumptions 1 and 2, we know that the Lipschitz constants $(L_j)_{j\in[\![d]\!]}$ and L do exist. Indeed, the Hessian matrix of the risk $R(\theta)$ is given by

$$\nabla^2 R(\theta) = \mathbb{E}\big[\ell''(\theta^\top X, Y) XX^\top\big],$$

where $\ell''(z,y) := \partial^2\ell(z,y)/\partial z^2$, so that

$$L_j = \sup_{\theta\in\Theta} \mathbb{E}\big[\ell''(\theta^\top X, Y)(X^j)^2\big] \quad \text{and} \quad L = \sup_{\theta\in\Theta} \big\|\nabla^2 R(\theta)\big\|_{\mathrm{op}}, \tag{8}$$

where $\big\|H\big\|_{\mathrm{op}}$ stands for the operator norm of a matrix H. Assumption 1 entails $L_j \leq \gamma\mathbb{E}\big[(X^j)^2\big]$, which is finite because of Equation (5) from Assumption 2. In order to derive *linear* convergence rates for CGD, it is standard to require strong convexity [88, 105]. Here, we require strong convexity on the risk $R(\theta)$ itself, as described in the following.

Assumption 4. *We assume that the risk R given by Equation* (1) *is λ-strongly convex, namely that*

$$R(\theta_2) \geq R(\theta_1) + \langle \nabla R(\theta_1), \theta_2 - \theta_1\rangle + \frac{\lambda}{2}\|\theta_2 - \theta_1\|^2 \tag{9}$$

for any $\theta_1, \theta_2 \in \Theta$.

Assumption 4 is satisfied whenever $\lambda_{\min}\big(\nabla^2 R(\theta)\big) \geq \lambda$ for any $\theta \in \Theta$, where $\lambda_{\min}(H)$ stands for the smallest eigenvalue of a symmetric matrix H. For the least-squares loss, this translates into the condition $\lambda_{\min}\big(\mathbb{E}[XX^\top]\big) \geq \lambda$. Note that one can always make the risk λ-strongly convex by considering ridge penalization, namely by replacing $R(\theta)$ by $R(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$, but we provide also guarantees without this Assumption in Section 5 below. The following Theorem provides an upper bound over the risk of Algorithm 1 whenever the estimators $\widehat{g}_j(\cdot)$ have an error vector $\epsilon(\delta)$, as defined in Definition 1. We introduce for short $R^\star = R(\theta^\star) = \min_{\theta\in\Theta} R(\theta)$.

Theorem 1. *Grant Assumptions 1, 3 and 4. Let $\theta^{(T)}$ be the output of Algorithm 1 with step-sizes $\beta_j = 1/L_j$, an initial iterate $\theta^{(0)}$, uniform coordinates sampling $p_j = 1/d$ and estimators of the partial derivatives with error vector $\epsilon(\cdot)$. Then, we have*

$$\mathbb{E}\big[R(\theta^{(T)})\big] - R^\star \leq \big(R(\theta^{(0)}) - R^\star\big)\Big(1 - \frac{\lambda}{L_{\max} d}\Big)^T + \frac{L_{\max}}{2\lambda L_{\min}}\big\|\epsilon(\delta)\big\|_2^2 \tag{10}$$

with probability at least $1-\delta$, where the expectation is w.r.t. the sampling of the coordinates. Now, if Algorithm 1 is run as before, but with an importance sampling *distribution $p_j = L_j / \sum_{k\in[\![d]\!]} L_k$, we have*

$$\mathbb{E}\big[R(\theta^{(T)})\big] - R^\star \leq \big(R(\theta^{(0)}) - R^\star\big)\Big(1 - \frac{\lambda}{\sum_{j\in[\![d]\!]} L_j}\Big)^T + \frac{1}{2\lambda}\big\|\epsilon(\delta)\big\|_2^2 \tag{11}$$

with probability at least $1 - \delta$.

The proof of Theorem 1 is given in Appendix 9. It adapts standard arguments for the analysis of CGD [88, 105] with inexact estimators of the partial derivatives. The statistical error $\|\epsilon(\delta)\|_2^2$ is studied in Section 3 for each considered robust estimator of the partial derivatives. Both (10) and (11) are upper bounds on the excess risk with exponentially vanishing optimization errors (called *linear* rate in optimization) and a constant statistical error. The optimization error term of (11), given by

$$\big(R(\theta^{(0)}) - R^\star\big)\Big(1 - \frac{\lambda}{\sum_{j\in[\![d]\!]} L_j}\Big)^T,$$

goes to 0 exponentially fast as the number of iterations T increases, with a contraction constant better than that of (10) since $\sum_{j\in[\![d]\!]} L_j \leq d L_{\max}$. This can be understood from the fact that importance sampling better exploits the knowledge of the Lipschitz constants L_j. Also, note that T is the number of iterations of CGD, so that $T = Cd$ where C is the number of CGD cycles. Therefore, defining $L' := \frac{1}{d}\sum_{j\in[\![d]\!]} L_j$, we have

$$\Big(1 - \frac{\lambda}{dL'}\Big)^{Cd} \leq \Big(1 - \frac{\lambda}{L'}\Big)^C,$$

for $d \geq 1$, which leads to a linear rate at least similar to the one of GD [12].

Theorem 1 proves an upper bound on the excess risk $R(\theta^{(T)}) - R^\star$ of the iterates of robust CGD directly, without using an intermediate upper bound on $\|\theta^{(T)} - \theta^\star\|_2^2$. This differs from the approaches used by [94, 50] that consider robust GD (while we introduce robust CGD here) to bound the excess risk of the iterates. This allows us to obtain a better contraction factor for the optimization error and a better constant in front of the statistical error. Note that we can derive also an upper bound on $\|\theta^{(T)} - \theta^\star\|_2^2$, see Theorem 4 in Appendix 9.

Note that the iterations considered in Algorithm 1 do not perform a projection in Θ. Indeed, one can show that $\|\theta^{(t)} - \theta^\star\|$ is also subject to a contraction and is therefore decreasing w.r.t. t. Thus, if $\theta^{(0)} = 0$, iterates $\theta^{(t)}$ naturally belong to the ℓ_2 ball of radius $2\|\theta^\star\|$.

Step-sizes. The step-sizes $\beta_j = 1/L_j$ are unknown, since they are functionals of the unknown distribution P. So, we provide, in Appendix 8.1, theoretical guarantees similar to that of Theorem 1 using step-sizes $\widehat{\beta}_j = 1/\widehat{L}_j$, where $\widehat{L}_j$ is a robust estimator of the upper bound $\overline{L}_j := \gamma\mathbb{E}\big[(X^j)^2\big] \geq L_j$ of the Lipschitz constant L_j.

A deterministic result. The previous Theorem 1 provides upper bounds on the expectation of the excess risk with respect to the sampling of the coordinates used in CGD. In Theorem 2 below, we provide an upper bound similar to the one from Theorem 1, but with a fully deterministic variant of CGD, where we replace line 4 of Algorithm 1 with a deterministic cycling through the coordinates.

Theorem 2. *Grant Assumptions 1, 3 and 4. Let $\theta^{(T)}$ be the output of Algorithm 1 with step-sizes $\beta_j = 1/L_j$, an initial iterate $\theta^{(0)}$, deterministic cycling over $[\![d]\!]$ such that*

$$\{j_{td+1}, j_{td+2}, \dots, j_{(t+1)d-1}\} = [\![d]\!]$$

for any t and estimators of the partial derivatives with error vector $\epsilon(\cdot)$. Then, we have

$$R(\theta^{(T)}) - R^\star \leq \big(R(\theta^{(0)}) - R^\star\big)\big(1 - 2\lambda\kappa\big)^T + \frac{3}{8\lambda\kappa L_{\min}}\big\|\epsilon(\delta)\big\|_2^2$$

with probability at least $1-\delta$, where we introduced the constant

$$\kappa = \frac{1}{8L_{\max}(1 + d(L_{\max}/L_{\min}))}.$$

The proof of Theorem 2 is given in Appendix 9 and uses arguments from [7] and [73]. It provides an extra guarantee on the convergence of CGD, for a very general choice of coordinates cycling, at the cost of degraded constants compared to Theorem 1, both for the optimization and statistical error terms.

Note that, our convergence results are set under a Lipschitz-smoothness assumption (see also Theorem 3 for the non strongly convex case), this excludes problems with non-smooth regularization such as Lasso to which CGD has commonly been applied [106, 38, 107]. Although such applications remain beyond the scope of our theory, there is no reason to doubt that plugging robust estimators, such as those given in Section 3 below, into CGD applied to non-smooth problems would lead to improved statistical performance and robustness.

3 Robust estimators of the partial derivatives

We consider three estimators of the partial derivatives

$$g_j(\theta) = \frac{\partial R(\theta)}{\partial \theta_j} = \mathbb{E}\big[\ell'(X^\top\theta, Y)X^j\big]$$

that can be used within Algorithm 1: Median-of-Means in Section 3.1, Trimmed mean in Section 3.2 and an estimator that we will call "Catoni-Holland" in Section 3.3. We provide, for each estimator, a concentration inequality for the estimation of $g_j(\theta)$ for fixed θ under a weak moments assumption (Lemmas 2, 3 and 4). We derive also uniform versions of the bounds in each case (Propositions 1, 2, 3 and 4) which define the error vectors to be plugged into Theorems 1 and 2. We also discuss in details the numerical complexity of each estimator and explain that they all are, in their own way, an interpolation between the empirical mean and the median. We wrap up these results in Table 1 below.

The deviation bound optimality in Table 1 is meant in terms of the dependence, up to a constant, on the sample size n, required confidence $\delta \in (0,1)$ and distribution variance[3]. An estimator's deviation bound is deemed optimal if it fits the lower bounds given by Theorems 1 and 3

[3] or more generally the centered moment of order $1+\alpha$ for $\alpha \in (0,1]$, see below.

	Optimal deviation bound	Robustness to outliers	Numerical complexity	Hyper-parameter
`ERM`	No	None	$O(n)$	None
`MOM`	Yes	Yes for $\lvert\mathcal{O}\rvert < K/2$	$O(n+K)$	$K \in [\![n]\!]$
`CH`	Yes	None	$O(n)$	Scale s
`TM`	Yes	Yes for $\lvert\mathcal{O}\rvert < n/8$	$O(n)$	Proportion $\epsilon \in [0, 1/2)$

Table 1: Properties of some robust estimators, where `ERM` = Empirical Risk Minimizer (ordinary mean), `MOM` = Median-of-Means, `CH` = Catoni-Holland and `TM` = Trimmed Mean. We recall that n = sample size and $|\mathcal{O}|$ = number of outliers. The parameters of each estimators are: the number of blocks K in `MOM`, a scale parameter $s > 0$ in `CH` and a proportion of samples ϵ in `TM`.

in [77]. Let us introduce the centered moment of order $1+\alpha$ of the partial derivatives and its maximum over Θ, given by

$$m_{\alpha,j}(\theta) := \mathbb{E}\Big[\big|\ell'(X^\top\theta, Y)X^j - \mathbb{E}[\ell'(X^\top\theta, Y)X^j]\big|^{1+\alpha}\Big] \qquad \text{and} \qquad M_{\alpha,j} = \sup_{\theta\in\Theta} m_{\alpha,j}(\theta) \quad (12)$$

for $\alpha \in (0, 1]$. Note that $m_{1,j}(\theta) = \mathbb{V}\big[\ell'(X^\top\theta, Y)X^j\big]$ and we know that $m_{\alpha,j}(\theta)$ exists, as explained in the next Lemma.

Lemma 1. *Under Assumptions 1 and 2 the risk $R(\theta)$ is well defined for all $\theta \in \Theta$ and we have*

$$\mathbb{E}\big[\big|\ell'(X^\top\theta, Y)X^j\big|^{1+\alpha}\big] < +\infty$$

for any $j \in [\![d]\!]$ and $\theta \in \Theta$.

The proof of Lemma 1 involves simple algebra and is provided in Section 9 below. Let us introduce

$$g_j^i(\theta) := \ell'(X_i^\top\theta, Y_i)X_i^j, \quad (13)$$

the sample $i \in [\![n]\!]$ partial derivative for coordinate $j \in [\![d]\!]$.

3.1 Median-of-Means

The Median-Of-Means (`MOM`) estimator is the median

$$\widehat{g}_j^{\texttt{MOM}}(\theta) := \text{median}\left(\widehat{g}_j^{(1)}(\theta), \ldots, \widehat{g}_j^{(K)}(\theta)\right) \quad (14)$$

of the block-wise empirical means

$$\widehat{g}_j^{(k)}(\theta) := \frac{1}{|B_k|}\sum_{i\in B_k} g_j^i(\theta) \quad (15)$$

within blocks $B_1, \ldots, B_K$ of roughly equal size that form a partition of $[\![n]\!]$ and that are sampled uniformly at random. This estimator depends on the choice of the number K of blocks used to compute it, which can be understood as an "interpolation" parameter between the ordinary mean ($K = 1$) and the median ($K = n$). It is robust to heavy-tailed data and a limited number of outliers as explained in the following lemma.

Lemma 2. *Grant Assumptions 1 and 2 with $\alpha \in (0,1]$. If $|\mathcal{O}| \leq K/12$, we have*:

$$\mathbb{P}\Big[\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g(\theta)_j\big| > (24 m_{\alpha,j}(\theta))^{1/(1+\alpha)}\Big(\frac{K}{n}\Big)^{\alpha/(1+\alpha)}\Big] \leq e^{-K/18}$$

for any fixed $j \in [\![d]\!]$ and $\theta \in \Theta$. If we fix a confidence level $\delta \in (0,1)$ and choose $K := \lceil 18\log(1/\delta)\rceil$, we have

$$\begin{aligned}\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g(\theta)_j\big| &\leq c_\alpha m_{\alpha,j}(\theta)^{1/(1+\alpha)}\Big(\frac{\log(1/\delta)}{n}\Big)^{\alpha/(1+\alpha)} \\ &\leq c_\alpha M_{\alpha,j}^{1/(1+\alpha)}\Big(\frac{\log(1/\delta)}{n}\Big)^{\alpha/(1+\alpha)} \qquad (16)\end{aligned}$$

with a probability larger than $1-\delta$, where $c_\alpha := 2^{(3+\alpha)/(1+\alpha)}3^{(1+2\alpha)/(1+\alpha)}$.

The proof of Lemma 2 is given in Section 9 below and it adapts simple arguments from [77] and [68]. Compared to [77], it provides additional robustness with respect to $|\mathcal{O}| \geq 1$ outliers and compared to [68] it provides guarantees with weak moments $\alpha < 1$. An inspection of the proof of Lemma 2 shows that it holds also under the assumption $|\mathcal{O}| \leq (1-\varepsilon)K/2$ for any $\varepsilon \in (0,1)$ with an increased constant $c_\alpha = 8 \times 3^{1/(1+\alpha)}/\varepsilon^{(1+2\alpha)/(1+\alpha)}$. This concentration bound is optimal under the $(1+\alpha)$-moment assumption (see Theorems 1 and 3 in [77]) and is sub-Gaussian when $\alpha = 1$ (finite variance). The next proposition provides a *uniform* deviation bound over Θ for $\widehat{g}_j^{\texttt{MOM}}(\theta)$.

Proposition 1. *Grant Assumptions 1 and 2 with $\alpha \in (0,1]$ and $|\mathcal{O}| \leq K/12$. We have*

$$\mathbb{P}\Big[\sup_{\theta\in\Theta}\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big| \leq \epsilon_j^{\texttt{MOM}}(\delta)\Big] \geq 1-\delta$$

for any $j \in [\![d]\!]$, with

$$\begin{aligned}\epsilon_j^{\texttt{MOM}}(\delta) := c_\alpha\Big(M_{j,\alpha} + \frac{m_{L,\alpha}}{n^\alpha}\Big)^{1/(1+\alpha)}\Big(\frac{\log(d/\delta) + d\log(3\Delta n^{\alpha/(1+\alpha)}/2)}{n}\Big)^{\alpha/(1+\alpha)} \\ + (\overline{L} + L_j)\Big(\frac{1}{n}\Big)^{\alpha/(1+\alpha)}\end{aligned}$$

where $\overline{L} = \gamma\mathbb{E}\|X\|^2$, $m_{L,\alpha} = \mathbb{E}|\gamma\|X\|^2 - \overline{L}|^{1+\alpha}$ and $c_\alpha = 2^{(3+2\alpha)/(1+\alpha)}3^{(1+3\alpha)/(1+\alpha)}$.

The proof of Proposition 1 is given in Section 9 and uses methods similar to Lemma 2 with an ε-net argument. This defines the error vector $\epsilon^{\texttt{MOM}}(\delta)$ of the MOM estimator of the partial derivatives in the sense of Definition 1, that can be combined directly with the convergence results from Theorems 1 and 2 from Section 2. Since the optimization error decreases exponentially w.r.t. the number of iterations T in these theorems, while the estimator error $\|\epsilon(\delta)\|_2$ is fixed, one only needs $T = O(\|\epsilon(\delta)\|_2)$ to make both terms of the same order.

About uniform bounds. What is necessary to obtain a control of the excess risk of robust CGD is a control of the noise terms $|\widehat{g}_j(\theta^{(t)}) - g_j(\theta^{(t)})|$, where both iterates $\theta^{(t)}$ and estimators $\widehat{g}_j(\cdot)$ of the partial derivatives depend on the same data. This forbids the direct use of a deviation such as the one from Lemma 2 (and Lemmas 3 and 4 below) where θ must be deterministic. We use in this paper an approach based on uniform deviation bounds (Propositions 1, 3 and 4) in order to bypass this problem, similarly to [52] and many other papers using empirical process theory. This is of course pessimistic, since $\theta^{(t)}$ goes to $\theta^\star$ as t increases. Another approach considered in [94] is to split data into segments of size n/T and to compute the gradient estimator using

a segment independent of the ones used to compute the current iterate. This approach departs strongly from what is actually done in practice, and leads to controls on the excess risk expressed with $\widetilde{\delta} = \delta/T$ and $\widetilde{n} = n/T$ instead of δ and n, hence a deterioration of the control of the excess risk. Our approach based on uniform deviations also suffers from a deterioration, due to the use of an ε-net argument, observed in Proposition 1 through the extra $d^{\alpha/(1+\alpha)}$ factor when compared to Lemma 2. Avoiding such deteriorations is an open difficult problem, either using uniform bounds or data splitting.

In addition to Proposition 1, we propose another uniform deviation bound for $\widehat{g}_j^{\texttt{MOM}}(\theta)$ using the Rademacher complexity, which is a fundamental tool in statistical learning theory and empirical process theory [70, 64, 6]. Let us introduce

$$\mathcal{R}_j(\Theta) = \mathbb{E}\Big[\sup_{\theta\in\Theta} \sum_{i\in\mathcal{I}} \varepsilon_i g_j^i(\theta)\Big]$$

for $j \in [\![d]\!]$, where $(\varepsilon_i)_{i\in\mathcal{I}}$ are i.i.d Rademacher variables and where we recall that $\mathcal{I}$ contains the inliers indices (see Assumption 2).

Proposition 2. *Grant Assumptions 1 and 2 with $\alpha \in (0,1]$. If $|\mathcal{O}| \leq K/12$, we have*

$$\mathbb{P}\bigg[\sup_{\theta\in\Theta} \big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big| \geq \max\Big(\Big(\frac{36M_{\alpha,j}}{(n/K)^\alpha}\Big)^{1/(1+\alpha)}, \frac{64\mathcal{R}_j(\Theta)}{n}\Big)\bigg] \leq e^{-K/18}$$

for any $j \in [\![d]\!]$. If we fix a confidence level $\delta \in (0,1)$ and choose $K := \lceil 18\log(1/\delta)\rceil$, we have

$$\sup_{\theta\in\Theta} \big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g(\theta)_j\big| \leq \max\Big(c_\alpha M_{\alpha,j}^{1/(1+\alpha)}\Big(\frac{\log(d/\delta)}{n}\Big)^{\alpha/(1+\alpha)}, \frac{64\mathcal{R}_j(\Theta)}{n}\Big) \tag{17}$$

with a probability larger than $1-\delta$ for all $j \in [\![d]\!]$, where $c_\alpha := 2^{(2+\alpha)/(1+\alpha)}3^2$. Moreover, if $\mu_{X,j}^{2(1+\alpha)} := \mathbb{E}[(X^j)^{2(1+\alpha)}] < +\infty$ for all $j \in [\![d]\!]$ we have

$$\mathcal{R}_j(\Theta) \leq \gamma\Delta C_\alpha\Big(n\mu_{X,j}^{1+\alpha} \sum_{k\in[\![d]\!]} \mu_{X,k}^{1+\alpha}\Big)^{1/(1+\alpha)} = O((nd)^{1/(1+\alpha)}),$$

where C_α is a constant depending only on α.

The proof of Proposition 2 is given in Section 9 and borrows arguments from [68, 10]. For $\alpha = 1$, the bound (17) has order $O(\sqrt{d/n})$ similarly to Theorem 2 from [68], although we consider here a different quantity (Rademacher complexity of the partial derivatives, towards the study of the *explicit* robust CGD algorithm, while *implicit* algorithms are studied herein). Note also that we do not prove similar uniform bounds using the Rademacher complexity for the `TM` and `CH` algorithms considered below, an interesting open question.

Comparison with [94, 50]. A first distinction of our results compared to [94, 50] is the use and theoretical study of robust CGD instead of robust GD. A second distinction is that we work under $1+\alpha$ moments on the partial derivatives of the risk, while [94, 50] require $\alpha = 1$. Our setting is similar but more general than the one laid out in [50] since the latter does not consider the presence of outliers. Theorem 5 from [50] states linear convergence of the optimization error thanks to strong convexity similarly to our Theorem 1. Their management of the statistical error is quite similar and leads to the same rate. However, our bound involves the sum of the coordinatewise moments of the gradient thanks to Proposition 1, an improvement over the bound from [50] which is only stated in terms of a uniform bound on the coordinate variances. Another reference point is the heavy-tailed setting of [94], which deals with heavy-tails independently from the problem of corruption and requires $\alpha = 1$. More importantly, the approach considered in [94] relies on data-splitting, which departs significantly from what is done in practice, while we do not perform data-spitting but use uniform bounds, as discussed above.

Complexity of $\widehat{g}_j^{\texttt{MOM}}(\theta)$. The computation of $\widehat{g}_j^{\texttt{MOM}}(\theta)$ requires (a) to sample a permutation of $[\![n]\!]$ to sample the blocks $B_1, \dots, B_K$, (b) to compute averages within the blocks and (c) to compute the median of K numbers. Sampling a permutation of $[\![n]\!]$ has complexity $O(n)$ using the Fischer-Yates algorithm [62], and so does the computation of the averages, so that (a) and (b) have complexity $O(n)$. The computation of the median of K numbers can be done using the quickselect algorithm [48] with $O(K)$ average complexity, leading to a complexity $O(n + K) = O(n)$ since $K < n$.

3.2 Trimmed Mean estimator

The idea of the Trimmed Mean (TM) estimator is to exclude a proportion of data in the tails of their distribution to achieve robustness. We are aware of two variants: (1) one in which samples in the tails are removed, the remaining samples being used to compute an empirical mean and (2) another variant in which samples in the tails are clipped but not removed from the empirical mean. Variant (1) is robust to η-corruption[4] whenever the data distribution is sub-exponential [74] or sub-Gaussian [29, 28, 31]. Variant (2), also known as *Winsorized mean*, enjoys a sub-Gaussian deviation [77] for heavy-tailed distributions. Both robustness properties are shown simultaneously (sub-Gaussian deviations under a heavy-tails assumption and η-corruption) in [79] (see Theorem 1 therein). We consider below variant (2), which proceeds as follows.

First, the TM estimator splits $[\![n]\!] = [\![n/2]\!] \cup [\![n/2]\!]^{\texttt{C}}$ where $[\![n/2]\!]^{\texttt{C}} = [\![n]\!] \setminus [\![n/2]\!]$, assuming without loss of generality that n is even, and it computes the sample derivatives $g_j^i(\theta)$ given by (13) for all $i \in [\![n]\!]$. Then, given a proportion $\epsilon \in [0, 1/2)$, it computes the ϵ and $1 - \epsilon$ quantiles of $(g_j^i(\theta))_{i \in [\![n/2]\!]}$ given by

$$q_\epsilon := g_j^{([\epsilon n/2])}(\theta) \quad \text{and} \quad q_{1-\epsilon} := g_j^{([(1-\epsilon)n/2])}(\theta),$$

where $g_j^{(1)}(\theta) \leq \cdots \leq g_j^{(n/2)}(\theta)$ is the order statistics of $(g_j^i(\theta))_{i \in [\![n/2]\!]}$ and where $[x]$ is the lower integer part of $x \in \mathbb{R}$. Finally, the estimator is computed as

$$\widehat{g}_j^{\texttt{TM}}(\theta) = \frac{2}{n} \sum_{i \in [\![n/2]\!]^{\texttt{C}}} q_\epsilon \vee g_j^i(\theta) \wedge q_{1-\epsilon}, \tag{18}$$

where $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$, namely it is the average of the partial derivatives from samples in $[\![n/2]\!]^{\texttt{C}}$ clipped in the interval $[q_\epsilon, q_{1-\epsilon}]$. Note that $\widehat{g}_j^{\texttt{TM}}(\theta)$ is also some form of "interpolation" between the average and the median through ϵ: it is the average of the partial derivatives for $\epsilon = 0$ and their median for $\epsilon = 1/2$. As explained in the next lemma, the TM estimator is robust both to a proportion of corrupted samples and heavy-tailed data.

Lemma 3. *Grant Assumptions 1 and 2 with $\alpha \in (0, 1]$ and assume that $|\mathcal{O}| \leq \eta n$ with $\eta < 1/8$. If we fix a confidence level $\delta \in (0, 1)$ and choose $\epsilon = 8\eta + 12 \log(4/\delta)/n$, we have*

$$\begin{aligned} |\widehat{g}_j^{\texttt{TM}}(\theta) - g_j(\theta)| &\leq 7 m_{\alpha,j}(\theta)^{1/(1+\alpha)} \Big(4\eta + \frac{6\log(4/\delta)}{n}\Big)^{\alpha/(1+\alpha)} \\ &\leq 7 M_{\alpha,j}^{1/(1+\alpha)} \Big(4\eta + \frac{6\log(4/\delta)}{n}\Big)^{\alpha/(1+\alpha)} \end{aligned}$$

with a probability larger than $1 - \delta$.

The proof of Lemma 3 is given in Section 9 and extends Theorem 1 from [79] to $\alpha \in (0, 1]$ instead of $\alpha = 1$ only. It shows that the TM estimator has the remarkable quality of being simultaneously robust to heavy-tailed and a *fraction* of corrupted data, as opposed to MOM which is

[4] We call "η-corruption" the context where the outlier set $\mathcal{O}$ in Assumption 2 satisfies $|\mathcal{O}| = \eta n$ with $\eta \in [0, 1/2)$

only robust to a limited *number* of outliers. Note that for the computation of the TM estimator, the splitting $[\![n]\!] = [\![n/2]\!] \cup [\![n/2]\!]^{\complement}$ is a technical theoretical requirement used to induce independence between $q_\epsilon, q_{1-\epsilon}$ and the sample partial derivatives $(g_j^i(\theta))_{i\in[\![n/2]\!]^{\complement}}$ involved in the average (18). Our implementation does not use this splitting.

Comparison with [94]. A comparison between Lemma 3 and the results from [94] pertaining to the corrupted setting is relevant here. We first point out that corruption in [94] is modeled as receiving data from the "η-contaminated" distribution $(1-\eta)P + \eta Q$ with Q an arbitrary distribution. On the other hand, Lemma 3 considers the more general η-corrupted setting where an η-proportion of the data is replaced by arbitrary outliers *after* sampling. In this case, Lemma 3 results in a statistical error with a dependence of order $\sqrt{\eta d}$ in the corruption (on the vector euclidean norm). On the other hand, Lemma 1 in [94] yields a better dependence of order $\sqrt{\eta \log d}$ in the corresponding case. Keep in mind, however, that Algorithm 2 from [94] which achieves this rate requires recursive SVD decompositions to compute a robust gradient making it computationally heavy and impractical for moderately high dimension. Additionally, the relevant results in [94] require a stronger moment assumption on the gradient and impose additional constraints on the corruption rate η. We also mention Algorithm 5 from [94] which yields an even better dependence on the dimension (see their Lemma 2), although it involves a computationally costly procedure as well. Besides, knowledge of the trace and operator norm of the covariance matrix of the estimated vector is required which makes the algorithm more difficult to use in practice.

Proposition 3. *Grant Assumptions 1 and 2 with $\alpha \in (0,1]$ and $|\mathcal{O}| \leq \eta n$. We have*

$$\mathbb{P}\Big[\sup_{\theta\in\Theta} \big|\widehat{g}_j^{\mathtt{TM}}(\theta) - g_j(\theta)\big| \leq \epsilon_j^{\mathtt{TM}}(\delta)\Big] \geq 1-\delta$$

for any $j \in [\![d]\!]$ with

$$\epsilon_j^{\mathtt{TM}}(\delta) := 28\Big(M_{j,\alpha} + \frac{m_{L,\alpha}}{n^{\alpha(1+\alpha)}}\Big)^{1/(1+\alpha)}\Big(2\eta + 3\frac{\log(4d/\delta) + d\log(3\Delta n^{\alpha/(1+\alpha)}/2)}{n}\Big)^{\alpha/(1+\alpha)} + \frac{\overline{L} + L_j}{n^{\alpha/(1+\alpha)}}$$

where $\overline{L}$ and $m_{L,\alpha}$ are as in Proposition 1.

The proof of Proposition 3 is given in Appendix 9 and uses an ε-net argument to obtain a uniform bound. Similarly to MOM, the resulting statistical error has optimal dependence on the $(1+\alpha)$-moments of the partial derivatives (12).

By plugging, the error vector $\epsilon^{\mathtt{TM}}(\delta)$ into Theorem 1, we obtain the following corollary which summarizes the best learning guarantees we obtain.

Corollary 1. *In the combined settings of Theorem 1 and Proposition 3, let $\widehat{\theta} := \theta^{(T)}$ denote the estimator obtained by running CGD with importance sampling using the TM estimator for T iterations where T will be specified shortly. Then, with probability at least $1-\delta$ we have*

$$\frac{\lambda}{2}\mathbb{E}\big\|\widehat{\theta} - \theta^\star\big\|^2 \leq \mathbb{E}[R(\widehat{\theta})] - R^\star \leq O\Big(\frac{1}{\lambda}\Big(\sum_{j\in[\![d]\!]} M_{j,\alpha}^{2/(1+\alpha)}\Big)\Big(\eta + \frac{\log(d/\delta) + d\log(n)}{n}\Big)^{2\alpha/(1+\alpha)}\Big), \quad (19)$$

where the expectations are w.r.t. the sampling of the coordinates. The above bound holds for a number of iterations T of order

$$T \geq \Omega\left(\log\left[\frac{\lambda(R(\theta^{(0)}) - R^\star)}{\sum_{j\in[\![d]\!]} M_{j,\alpha}^{2/(1+\alpha)}\Big(\eta + \frac{\log(d/\delta)+d\log(n)}{n}\Big)^{2\alpha/(1+\alpha)}}\right] \Big/ \log\Big[\frac{1}{1-\lambda/\sum_{j\in[\![d]\!]} L_j}\Big]\right).$$

The proof of Corollary 1 is given in Appendix 9 and is a straightforward combination of Theorem 1 and Proposition 3 where a big O notation was used to make the bound more legible.

Complexity of $\widehat{g}_j^{\mathtt{TM}}(\theta)$. The most demanding part for the computation of $\widehat{g}_j^{\mathtt{TM}}(\theta)$ is the computation of q_ϵ and $q_{1-\epsilon}$. A naive idea is to sort all n values at an average cost $O(n \log n)$ with quicksort for example [48] and to simply retrieve the desired order statistics afterwards. Of course, better approaches are possible, including the median-of-medians algorithm (not to be confused with MOM), which remarkably manages to keep the cost of finding an order statistic with complexity $O(n)$ even in the worst case (see for instance Chapter 9 of [24]). However, the constant hidden in the previous big-O notations seriously impact performances in real-world implementations: we compared several implementations experimentally and concluded that a variant of the quickselect algorithm [48] was the fastest for this problem.

3.3 Catoni-Holland estimator

This estimator is a variation of the robust mean estimator by Catoni [17] introduced by Holland [50] for robust statistical learning, hence the name "Catoni-Holland", that we will denote $\widehat{g}_j^{\mathtt{CH}}(\theta)$. It is defined as an M-estimator which consists in solving

$$\sum_{i=1}^{n} \psi\Big(\frac{g_j^i(\theta) - \zeta}{\widehat{s}_j(\theta)}\Big) = 0 \tag{20}$$

with respect to ζ, where ψ is an uneven function satisfying $\psi(0) = 0$, $\psi(x) \sim x$ when $x \sim 0$ and $\psi(x) = o(x)$ when $x \to +\infty$ and where $\widehat{s}_j(\theta) > 0$ is a scale estimator. An approximate solution can be found using the fixed-point iterations

$$\zeta_{k+1} = \zeta_k + \frac{\widehat{s}_j(\theta)}{n} \sum_{i=1}^{n} \psi\Big(\frac{g_j^i(\theta) - \zeta_k}{\widehat{s}_j(\theta)}\Big),$$

which can easily be shown to converge to the desired value thanks to the monotonicity and Lipschitz-property of ψ. Following [50], we use the function $\psi(x) = 2\arctan(\exp(x)) - \pi/2$, while functions satisfying $-\log(1 - x + x^2/2) \le \psi(x) \le \log(1 + x + x^2/2)$ are considered in [17]. As explained in [50], the scale estimator is given by

$$\widehat{s}_j(\theta) := \widehat{\sigma}_j(\theta) \sqrt{\frac{n}{2\log(4/\delta)}}, \tag{21}$$

for a confidence level $\delta \in (0, 1)$, where $\widehat{\sigma}_j(\theta)$ is an estimator of the standard deviation of the partial derivative $\sigma_j(\theta) := m_{1,j}(\theta)^{1/2} = \mathbb{V}[\ell'(X^\top\theta, Y)X^j]^{1/2}$, see (12). The estimator $\widehat{\sigma}_j(\theta)$ is defined through another M-estimator solution to

$$\sum_{i=1}^{n} \chi\Big(\frac{g_j^i(\theta) - \bar{g}_j(\theta)}{\sigma}\Big) = 0 \tag{22}$$

with respect to σ, where $\bar{g}_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_j^i(\theta)$ and χ is an even function satisfying $\chi(0) < 0$ and $\chi(x) > 0$ as $x \to +\infty$. We use the same function as in [50] given by $\chi(u) = u^2/(1 + u^2) - c$ where c is such that $\mathbb{E}\chi(Z) = 0$ for Z a standard Gaussian random variable. To compute $\widehat{\sigma}_j(\theta)$ we use also fixed-point iterations

$$\sigma_{k+1} = \sigma_k\Big(1 - \frac{\chi(0)}{n} \sum_{i=1}^{n} \chi\Big(\frac{g_j^i(\theta) - \bar{g}_j(\theta)}{\sigma_k}\Big)\Big). \tag{23}$$

We refer to the supplementary material of [50] for further details on this procedure.

The CH estimator can be understood, once again, as an interpolation between the average and the median of the partial derivatives. Indeed, whenever s is large, the function $\psi(\cdot/s)$ is close to the sign function, which, if used in (20), leads to an M-estimator corresponding to the median [100]. For s small, $\psi(\cdot/s)$ is close to the identity, so that minimizing (20) leads to an ordinary average. As explained in the next lemma, this estimator is robust to heavy-tailed data (with $\alpha = 1$).

Lemma 4. *Grant Assumptions 1 and 2 with $\alpha = 1$ and assume that $\mathcal{O} = \emptyset$ (no outliers). For some failure probability $\delta > 0$, assume that we have, with probability at least $1 - \delta/2$, that $\sigma_j(\theta)/C' \leq \widehat{\sigma}_j(\theta) \leq C'\sigma_j(\theta)$ for some constant $C' > 1$. Then, we have*

$$\big|\widehat{g}_j^{\mathtt{CH}}(\theta) - g_j(\theta)\big| \leq C'\sigma_j(\theta)\sqrt{\frac{8\log(4/\delta)}{n}} \leq C'\Sigma_j\sqrt{\frac{8\log(4/\delta)}{n}}$$

with probability at least $1 - \delta$, where $\Sigma_j = M_{1,j} = \sup_{\theta\in\Theta}\sigma_j(\theta)$.

The proof of Lemma 4 is given in Section 9 and is an almost direct application of the deviation bound from [50]. If $C' \approx 1$, the deviation bound of $\widehat{g}_j^{\mathtt{CH}}(\cdot)$ is better than the ones given in Lemmas 2 and 3 with $\alpha = 1$. This stems from the fact that the analysis of Catoni's estimator [17] results in a deviation with the best possible constant [27]. However, contrary to MOM and TM, an estimator of the scale is necessary: it makes CH computationally much more demanding (see Figure 1 below), since it requires to perform two fixed-point iterations to approximate both $\widehat{\sigma}_j(\theta)$ and $\widehat{g}_j^{\mathtt{CH}}(\theta)$ and it requires Assumption 2 with $\alpha = 1$ so that $\sigma_j(\theta) < +\infty$. Moreover, there is no guaranteed robustness to outliers, a fact confirmed by the numerical experiments performed in Section 6 below.

Proposition 4. *Grant Assumptions 1 and 2 with $\alpha = 1$ and $\mathcal{O} = \emptyset$. Denote $\overline{L} = \mathbb{E}[\gamma\|X\|^2]$, $\sigma_L^2 = \mathbb{V}[\gamma\|X\|^2]$ and assume that for all $\theta, \widetilde{\theta} \in \Theta$ such that $\|\theta - \widetilde{\theta}\| \leq 1/\sqrt{n}$ we have*

$$\frac{1}{2}\sigma_j^2(\widetilde{\theta}) \leq \sigma_j^2(\theta) \leq 2\sigma_j^2(\widetilde{\theta}) \quad \text{and} \quad \frac{\sigma_j(\theta)}{\sigma_L} \geq \frac{1}{\sqrt{n}}.$$

Furthermore, assume that for all $\theta \in \Theta$, the variance estimator $\widehat{\sigma}_j(\theta)$ defined by (22) satisfies $\sigma_j(\theta)/C' \leq \widehat{\sigma}_j(\theta) \leq C'\sigma_j(\theta)$ for some constant $C' > 1$ with probability at least $1 - \delta/2$. Then, we have

$$\mathbb{P}\Big[\sup_{\theta\in\Theta}\big|\widehat{g}_j^{\mathtt{CH}}(\theta) - g_j(\theta)\big| \leq \epsilon_j^{\mathtt{CH}}(\delta)\Big] \geq 1 - \delta$$

for any $j \in [\![d]\!]$ with

$$\epsilon_j^{\mathtt{CH}}(\delta) := 4C'\Big(2\Sigma_j + \frac{\sigma_L}{\sqrt{n}}\Big)\sqrt{\frac{\log(4d/\delta) + d\log(3\Delta\sqrt{n}/2)}{n}} + \frac{\overline{L} + L_j}{\sqrt{n}}$$

where $\overline{L}$ is as in Proposition 1.

The proof of Proposition 4 is given in Section 9. It uses again an ε-net argument combined with a careful control of the variations of $\widehat{g}_j^{\mathtt{CH}}(\theta)$ with respect to θ. Compared with [50], we make a different use of the CH estimator: while it is used therein to estimate the whole gradient $\nabla R(\theta)$ during the robust GD iterations, we use it here to estimate the partial derivatives $g_j(\theta)$ during iterations of robust CGD. The numerical experiments from Section 6 confirm, in particular, that our approach leads to a considerable speedup and improved statistical performances when compared to [50].

The statements of Lemma 4 and Proposition 4 require $\alpha = 1$, while a very recent extension of Catoni's bound [20] is available for $\alpha \in (0, 1)$. However, the necessity to estimate the centered

$(1+\alpha)$-moment subsists (standard-deviation for $\alpha = 1$). Although iteration (23) may be adapted to this case, theoretical guarantees for it do lack. Note that even for $\alpha = 1$, the statements of Lemma 4 and Proposition 4 require assumptions on $\sigma_j^2(\theta)$ and $\widehat{\sigma}_j(\theta)$: an extension to $\alpha \in (0, 1]$ would lead to a set of even more intricate assumptions.

Complexity of $\widehat{g}_j^{\texttt{CH}}(\theta)$. It is not straightforward to analyze the complexity of this estimator, since it involves fixed-point iterations with a number of iterations that can vary from one run to the other. However, each iteration has complexity $O(n)$ and we observe empirically that the number of iterations is of constant order (usually smaller than 10) independently from the required confidence. Therefore, the overall complexity remains in $O(n)$ as demonstrated also by Figure 1 below. The latter also shows that the numerical complexity of `CH` is larger than that of `MOM` and `TM`, which later impacts the overall training time.

3.4 A comparison of the numerical complexities

As explained above, all the considered estimators of the partial derivatives have a numerical complexity $O(n)$. However, they perform different computations and have very different running times in practice. So, in order to compare their actual computational complexities we perform the following experiment. We consider an increasing sample size n between 10^2 and 10^6 on a logarithmic scale and run all the estimators: `MOM`, `TM`, `CH` and `ERM`, which is the average of the per-sample partial derivatives $g_j^i(\theta)$. We fix their parameters so as to obtain deviation bounds with confidence $1 - \delta = 99\%$: this corresponds to 82 blocks for `MOM`, $\epsilon = 72/n$ for `TM` and $\delta = 0.01$ for `CH`, but the conclusion is similar with different combinations of parameters. We use random samples with student $t(2.1)$ distribution (a finite variance distribution but with heavy tails, although run times do not differ by much when using different distributions). This leads to the display proposed in Figure 1, where we display the averaged timings over 100 repetitions (together with standard-deviations).

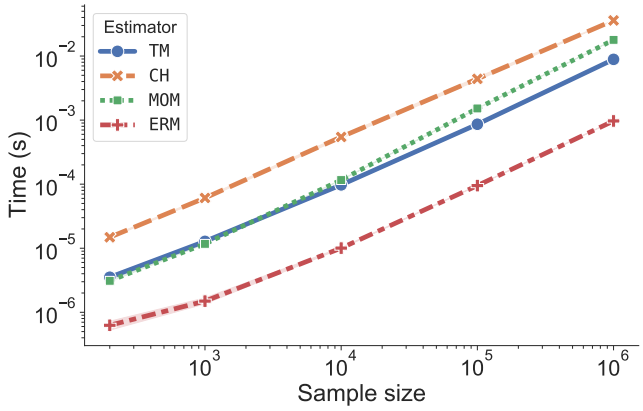

Figure 1: Average running time (y-axis) of all the considered estimators against an increasing sample size (x-axis). The run times increase with a similar slope (on a logarithmic scale), confirming $O(n)$ complexities, but differ significantly: `ERM` is of course the fastest, followed by `TM` and `MOM` (both are close) and finally `CH`, which is the slowest.

We observe that the run times of the estimators increase with a similar slope (on a logarithmic scale) against the sample size, confirming the $O(n)$ complexities. However, their timings differ significantly. `MOM` and `TM` share similar timings (`TM` becomes faster than `MOM` for large samples) and are about 10 times slower than `ERM`. `CH` is the slowest of all and is roughly 50 times slower

than ERM. This is of course related to the fact that CH requires to perform the fixed-point iterations each of which roughly costing $\Theta(n)$. In all cases, the estimators' complexities remain in $O(n)$ so that the complexity of a single iteration of robust CGD (see Algorithm 1) using either of them is $O(n)$, which is identical to the complexity of a non-robust ERM-based CGD. This means that Algorithm 1 achieves robustness at a limited cost, where the computational difference lies only in the constants in front of the big O notations.

4 Related works

The robust statistics field appeared in the 60s with the pioneering works of [99] and [55] and has received longstanding interest since then. Several works pursued the development of robust statistical methods including non-convex M-estimators [57], ℓ_1 tournaments [26, 33] and methods based on depth functions [19, 39, 84], the latter being difficult to use in practice because of their numerical complexity.

Renewed interest has manifested recently, related, on the one hand, to the increasing need for algorithms able to learn from large non-curated data sets and on the other hand, to the development of robust mean estimators with good theoretical guarantees under weak moment assumptions, including Median-of-Means (MOM) [86, 2, 58] and Catoni's estimator [17]. Under adversarial corruption [18], several statistical learning problems are studied in a robust setting, such as parameter estimation [66, 93, 83, 28, 79], regression [60, 75, 22, 8], classification [68, 61, 76], PCA [72, 16, 90] and most recently online learning [101].

In the heavy-tailed setting, a robust learning approach introduced in [11] proposes to optimize a robust estimator of the risk based on Catoni's mean estimator [17] resulting in an implicit estimator for which near-optimal guarantees are shown under weak assumptions on the data. However, the new risk may not be convex (even if the considered loss is), so that its minimization may be expensive and lead to an estimator unrelated to the one theoretically studied, potentially making the associated guarantees inapplicable. More recently, an explicit variant was proposed in [108] which applies Catoni's influence function to each term of the sum defining the empirical risk for linear regression. The associated optimum enjoys a sub-Gaussian bound on the excess risk, albeit with a slow rate since the ℓ_1 loss was used. A follow-up extended this result under weaker distribution assumptions [20]. The main drawback of this approach is that the unconventional use of the influence function introduces a considerable amount of bias which appears in the excess risk bounds.

Another way to obtain a robust estimator was proposed by [82, 54] and consists in computing standard ERMs on disjoint subsets of the data and aggregating them using a multidimensional MOM. This approach recently appeared in [49] as well with various aggregation strategies in order to perform robust distributed learning. Although the previous works use easily implementable aggregation procedures, the associated deviation bounds are sub-optimal (see for instance [77]). Moreover, dividing the data into multiple subsets makes the method impractical for small sample sizes and may introduce bias coming from the choice of such a subdivision.

In the setting where an η-proportion of the data consist of arbitrary outliers, a robust meta-algorithm is introduced in [29], which repeatedly trains a given base learner and filters outliers based on an eccentricity score. The method reaches the target $\sigma\sqrt{\eta}$ error rate with σ the gradient standard deviation, although the requirement of multiple training rounds may be computationally expensive.

More recently, robust solutions to classification problems were proposed in [68] by using MOM to estimate the risk and computing gradients on trustworthy data subsets in order to perform descent. A variant was also proposed by the same authors in [67] where a pair of parameters is alternately optimized for a min-max objective. The resulting algorithm is efficient numerically,

though it requires a vanishing step-size to converge due to the variance coming from gradient estimation. Moreover, the provided theoretical guarantees concern the optimum of the formulated problem but not the optimization algorithm put to use.

Several recent papers [94, 52, 51, 50, 21, 1] use a form of robust gradient descent, where learning is guided by various robust estimators of the true gradient $\nabla R(\theta)$. Two such estimators are proposed in [94]. The first one is a vector analog of MOM where the scalar median is replaced by the geometric median

$$\mathrm{GMed}(g_1, \ldots, g_K) := \operatorname*{argmin}_{g \in \mathbb{R}^d} \sum_{j=1}^{K} \|g - g_j\|_2, \tag{24}$$

which can be computed using the algorithm given in [103]. This vector mean estimator enjoys improved concentration properties over the standard mean as shown in [82] although these remain sub-optimal (see also [77]). A line of works [78, 53, 23, 25, 79, 71, 30] specifically addresses the issue of devising efficient procedures with optimal deviation bounds.

Supervised learning with robustness to heavy-tails and a limited number of outliers is thus achieved but at a possibly high computational cost. The second algorithm called "Huber gradient estimator" is intended for Huber's ϵ-contamination setting. It uses recursive SVD decompositions followed by projections and truncations in order to filter out corruption. The method proves to be robust to data corruption but its computational cost becomes prohibitive as soon as the data has moderately large dimensionality.

Finally, the recent work of [92] proposed to perform robust regression by applying an initial filtering step on the data followed by regression using the robust Huber loss function. Remarkably, the resulting algorithm attains the optimal rates and is simultaneously robust to η-corruption and heavy tails. However, the theoretical guarantees only apply for linear regression and require assumptions which are rarely satisfied in practice such as isotropic covariance of the data.

Table 2 summarizes and compares the characteristics of a number of previously mentioned algorithms with ours. The statistical rate may be understood as the final excess risk or parameter error which are interchangeable up to a constant thanks to strong convexity. We have marked the complexities of some algorithms with a dagger (†) to signal the use of iteratively computed estimators with unpredictable iteration count. This indicates that a big constant is hidden by the big O notation. Note that the rows "GD-Huber gradient" and "GD-Geometric MOM" (drawn from [94]) have statistical rates in terms of $\widetilde{n} = n/T$ and $\widetilde{\delta} = \delta/T$ with T the optimization iterations count. This results from a sample splitting strategy yielding milder dimension dependence. However, one can check that the best choice of T degrades these bounds by a factor $\sim \log n$ roughly[5]. Finally, the statistical rate of "GD-implicit MOM" is marked with a double dagger (‡) because it is derived under the only assumption that the loss function is Lipschitz. Moreover, it only bounds the error on objective value estimation and does not directly apply to the estimate computed by the latter algorithm.

5 Theoretical guarantee without strong convexity

In this section we provide an upper bound similar to that of Theorem 1, but without the strong convexity condition from Assumption 4. As explained in Theorem 3 below, without strong convexity, the optimization error shrinks at a slower sub-linear rate when compared to Theorem 1 (a well-known fact, see [12]). In order to ensure that robust CGD, which uses "noisy" partial derivatives, remains a descent algorithm, we assume that the parameter set can be written as a product

[5]Indeed, considering strong convexity, optimization converges linearly and the final bound is of the form $a\exp(-bT) + cT\log(T/\delta)/n$ for some $a, b, c > 0$ and one can see that $T \sim \log n$ is approximately optimal.

Algorithm	Statistical Performance	Iteration/cycle complexity	Robustness to corruption
CGD-MOM (`MOM`) This paper	$O\Big(\frac{d\log(d/\delta)+d^2\log(n)}{n}\Big)$	$O((n+K)d)$ (cycle)	Yes for $\lvert\mathcal{O}\rvert < K/2$
CGD-TM (`TM`) This paper	$O\Big(d\Big(\eta+\frac{\log(d/\delta)+d\log(n)}{n}\Big)\Big)$	$O(nd)$ (cycle)	Yes for $\lvert\mathcal{O}\rvert < n/8$
CGD-CH (`CH`) This paper	$O\Big(\frac{d\log(d/\delta)+d^2\log(n)}{n}\Big)$	$O(nd)\dagger$ (cycle)	None
GD-Geometric MOM [94] (`GMOM`)	$O\Big(\frac{d\log(1/\widetilde{\delta})}{\widetilde{n}}\Big)$	$O((n+K)d)\dagger$	Yes for $\lvert\mathcal{O}\rvert < K/2$
GD-Huber gradient (`HG`) [94]	$O\Big(\log(d)\Big(\eta+\Big(\frac{d\log(d)\log(\widetilde{n}/(d\widetilde{\delta}))}{\widetilde{n}}\Big)^{3/4}+d\sqrt{\frac{\eta\log(d)\log(d\log(d)/\widetilde{\delta})}{\widetilde{n}}}\Big)\Big)$	$O(nd^2+d^3)\dagger$	Yes for η-*contamination*
GD-implicit MOM [68] (`LLM`)	$O\Big(\sqrt{\frac{d+\log(1/\delta)}{n}}\Big)\ddagger$	$O(nd)$	Yes for $\lvert\mathcal{O}\rvert < K/4$
GD-CH (`CH GD`) [50]	$O\Big(\frac{d\log(d/\delta)+d^2\log(n)}{n}\Big)$	$O(nd)\dagger$	None

Table 2: Summary of the main characteristics of our proposed algorithms (using CGD) and the main competitors in the literature. The notations $\widetilde{n}$ and $\widetilde{\delta}$ stand for n/T and δ/T respectively with T the optimization horizon. All statistical rates are derived under a strong convexity assumption except for "GD-implicit MOM". For each algorithm, the combination of optimization method and gradient estimator is indicated and the associated code name used in the experimental section is given between parentheses. Cycle complexities are given for CGD algorithms for more relevant comparison.

$\Theta = \prod_{j\in[\![d]\!]}\Theta_j$ and replace the iterations (6) (corresponding to Line 5 in Algorithm 1) by

$$\begin{cases}\theta_j^{(t+1)} \leftarrow \mathrm{proj}_{\Theta_j}\left(\theta_j^{(t)} - \beta_j\tau_{\epsilon_j}\big(\widehat{g}_j(\theta^{(t)})\big)\right) & \text{if } j = j_t\\ \theta_j^{(t+1)} \leftarrow \theta_j^{(t)} & \text{otherwise,}\end{cases} \tag{25}$$

where proj_{Θ_j} is the projection onto Θ_j and τ_ϵ is the soft-thresholding operator given by $\tau_\epsilon(x) = \mathrm{sign}(x)(|x|-\epsilon)_+$ with $(x)_+ = \max(x,0)$. In Theorem 3 below we use $\epsilon_j = \epsilon_j(\delta)$, the j-th coordinate of the error vector from Definition 1, which is instantiated for each robust estimator in Section 3. Since it depends on the moment $m_{\alpha,j}$, it is not observable, so we propose in Lemma 6 from Appendix 8.2 an observable upper bound deviation for it based on `MOM`.

This use of soft-thresholding of the partial derivatives can be understood as a form of partial derivatives (or gradient) clipping. However, note that it is rather a theoretical artifact than something to use in practice (we never use τ_ϵ in our numerical experiments from Section 6 below). Indeed, the operator τ_ϵ naturally appears for the following simple reason: consider a convex L-smooth scalar function $f:\mathbb{R}\to\mathbb{R}$ with derivative $g(x) := f'(x)$. An iteration of gradient descent

from x_0 uses an increment δ that minimizes the right-hand side of the following inequality:

$$f(x_0+\delta) \leq Q(\delta, x_0) := f(x_0) + \delta g(x_0) + \frac{L}{2}\delta^2,$$

namely $\operatorname{argmin}_\delta Q(\delta, x_0) = -g(x_0)/L$ leading to the iterate $x_0 - g(x_0)/L$ with ensured improvement of the objective. In our context, $g(x)$ is unknown and we use an estimator $\widehat{g}(x)$ satisfying $|\widehat{g}(x) - g(x)| \leq \epsilon$ with a large probability. Taking this uncertainty into account leads to the upper bound

$$f(x_0+\delta) \leq \widetilde{Q}(\delta, x_0) := f(x_0) + \delta\widehat{g}(x_0) + \frac{L}{2}\delta^2 + \epsilon|\delta|,$$

and, after projection onto the parameter set, to the iteration (25) since $\operatorname{argmin}_\delta \widetilde{Q}(\delta, x_0) = x_0 - \tau_\epsilon(\widehat{g}(x_0))/L$, with guaranteed decrease of the objective.

The clipping of partial derivatives is unnecessary in the strongly convex case since each iteration translates into a contraction of the excess risk, so that the degradations caused by the gradient errors remain controlled (see the proof of Theorem 1). No such contraction can be established without strong convexity, and clipping prevents gradient errors to accumulate uncontrollably.

Theorem 3. *Grant Assumptions 1 and 3 with $\Theta = \prod_{j\in[\![d]\!]} \Theta_j$. Let $\theta^{(T)}$ be the output of Algorithm 1 where we replace iterations* (6) *by* (25) *with step-sizes $\beta_j = 1/L_j$, an initial iterate $\theta^{(0)} \in \Theta$, uniform coordinates sampling $p_j = 1/d$ and estimators of the partial derivatives with error vector $\epsilon(\cdot)$. Then, we have with probability at least $1-\delta$*

$$\mathbb{E}\big[R(\theta^{(T)})\big] - R^\star \leq \frac{d}{T+1}\left(\sum_{j\in[\![d]\!]} \frac{L_j}{2}\big(\theta_j^{(0)} - \theta_j^\star\big)^2 + R(\theta^{(0)})\right) + \frac{2\|\epsilon(\delta)\|_2}{T+1}\sum_{t=0}^{T}\|\theta^{(t)} - \theta^\star\|_2,$$

where the expectation is w.r.t the sampling of the coordinates. Moreover, we have

$$\big\|\theta^{(t)} - \theta^\star\big\|_2 \leq \|\theta^{(t-1)} - \theta^\star\|_2$$

with the same probability, for all $t \in [\![T]\!]$.

The proof of Theorem 3 is given in Appendix 9 and is based on the proof of Theorem 5 from [88] and Theorem 1 from [95] while managing noisy partial derivatives. The optimization error term vanishes at a sublinear $1/T$ rate and is initially of order $R(\theta^{(0)})$ plus the potential $\Phi(\theta) = \sum_{j=1}^d L_j(\theta_j - \theta_j^\star)^2/2$ which is instrumental in the proof. Notice that $\|\epsilon(\delta)\|_2$ appears without the square which translates into "slow" $1/\sqrt{n}$ rates instead of "fast" $1/n$ rates achieved in Section 2. This degradation is an unavoidable consequence of the loss of strong convexity of the risk [97].

6 Numerical Experiments

The theoretical results of Sections 2, 3 and 5 can be applied to multiple supervised linear learning problems, with guaranteed robustness to both heavy-tailed data and outliers. We perform experiments confirming these properties for several tasks (regression, binary classification and multi-class classification) on multiple data sets comparing with a number of baselines including the state-of-the-art.

6.1 Algorithms

We compare our methods with several baselines among the following set of algorithms. For all algorithms, we use, unless specified otherwise, the least-squares loss for regression, and the logistic loss for classification (both for binary and multiclass problems, using the multiclass logistic loss). All considered algorithms can be used easily in a few lines of `Python` code with our library called `linlearn`, open-sourced under the BSD-3 License on `GitHub` and available here: `https://github.com/linlearn/linlearn`. This library follows the API conventions of `scikit-learn` [91].

CGD algorithms: `MOM`, `CH`, `TM` and `CGD ERM`. The `MOM`, `CH` and `TM` algorithms are the variants of robust CGD (Algorithm 1) respectively based on the median-of-means, trimmed mean and Catoni-Holland estimators introduced in Section 3. We also include `CGD ERM` which is CGD using a standard mean as estimator.

GD algorithms: `ERM`, `LLM`, `HG`, `GMOM`, `CH GD` and `Oracle`. These are all GD algorithms using different estimators of the gradient. `ERM` uses a non-robust gradient based on a simple mean. `LLM` corresponds to Algorithm 1 from [68]. It uses a MOM estimation of the risk and performs GD using gradients computed as the mean of the sample gradients from the block corresponding to the median of the risk. `HG` is Algorithm 2 from [94], called Huber Gradient Estimator, which uses recursive SVD decompositions and truncations to compute a robust gradient. `GMOM` is Algorithm 3 from [94], which estimates gradients using a geometric MOM (based on the geometric median). `CH GD` is the robust GD algorithm from [50], which uses gradients computed as coordinate-wise `CH` estimators. We consider also `Oracle`, which is GD performed with "oracle" gradients, namely the gradient of the unobserved true risk (only available for linear regression experiments using simulated data).

Extra algorithms: `RANSAC`, `HUBER` and `LAD`. We also include the following algorithms. For regression, we consider `RANSAC` [37], using the implementation available in the `scikit-learn` library [91]. `HUBER` stands for ERM learning with the modified Huber loss [109] for classification and Huber loss [89] for regression. `LAD` is ERM learning using the least absolute deviation loss [35], namely regression using the mean absolute error instead of least-squares.

6.2 Regression on simulated data

We consider the following simulation setting for linear regression with the square loss. We generate features $X \in \mathbb{R}^d$ with $d = 5$ with a non-isotropic Gaussian distribution with covariance matrix Σ and labels $Y = X^\top \theta^\star + \xi$ for a fixed $\theta^\star \in \mathbb{R}^d$ and simulated noise ξ. Since all distributions are known in this setting, we can compute the true risk and true gradients (used in `Oracle`).

We consider the following settings: (a) ξ is centered Gaussian; (b) ξ is Student with $\nu = 2.1$ degrees of freedom (heavy-tailed noise). In the remaining settings (c), (d), (e) and (f), ξ is as in (b) but 1% of the data is replaced by outliers as follows. For case (c), $X \in \mathbb{R}^5$ is replaced by a constant equal to $\lambda_{\max}(\Sigma)$ (largest eigenvalue of Σ) and labels are replaced by $2y_{\max}$ with $y_{\max} = \max_{i \in \mathcal{I}} |y_i|$; for (d) we do the same as (c) and multiply labels by -1 with probability $1/2$; for (e) we sample $X = 10\lambda_{\max}(\Sigma)v + Z$ where $v \in \mathbb{R}^5$ is a fixed unit vector and Z is a standard Gaussian vector and labels are i.i.d. Bernoulli random variables; finally for (f) we sample $X = 10\lambda_{\max}(\Sigma)V$ where V is uniform on the unit sphere and labels $y = y_{\max} \times (\varepsilon + U)$ where ε is a Rademacher variable and U is uniform in $[-1/5, 1/5]$.

For this experiment, we fix the parameters of the robust partial derivative estimators using the confidence level $\delta = 0.01$ and the number of outliers for `MOM` and `TM`. We report, for all

algorithms and settings (a)-(f), the evolution of the square loss (y-axis) along the iterations (x-axis, corresponding to cycles for CGD and iterations for GD). The results are averaged over 30 repetitions.

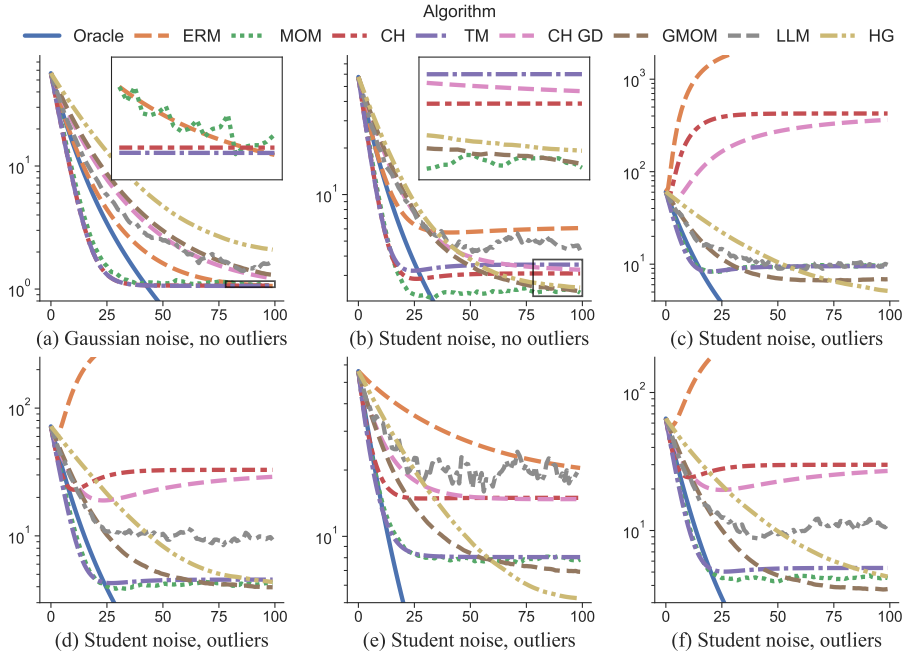

Figure 2: Excess-risk for the square loss (y-axis) against iterations (x-axis) for all the considered algorithms in the simulation settings (a)-(c) (top row) and (d)-(f) (bottom row). We zoom-in the last iterations for settings (a) and (b) to improve readability.

We observe that CGD-based algorithms generally converge faster than GD-based ones, independently of the quality of the optimum found. For setting (a), the final performance of all algorithms is roughly similar to that of `ERM` (as expected since the data are neither heavy-tailed nor corrupted) except for `LLM` and `HG` that converge slowly. For setting (b), the performance of `ERM` degrades visibly. Among robust methods, a slight advantage is observed when a robust vector mean is used as opposed to coordinatewise estimators. Though, `MOM` seems to be an exception to this rule. Different behaviours manifest in settings (c)-(f). We observe that `ERM` and Catoni-Holland estimators (`CH` and `CH GD`) are generally the most sensitive to outliers, especially in setting (c) where corrupted samples are introduced in a single-direction. The best final solutions are often found by `GMOM` and `HG`. This is not surprising since the latter is designed to handle corrupted data and the former is far from its breakdown point with only 1% corruption. Nonetheless, we also observe that `MOM` and `TM` consistently show comparable performance. In particular, for settings (d) and (f), corrupted samples are introduced in multiple directions and the performance gap between `GMOM`/`HG` and `MOM`/`TM` is small. Note that `MOM`/`TM` always converge faster. Finally, while `LLM` seems robust to heavy tails and outliers, its use of a median mini-batch and vanishing steps makes it unstable and often prevents it from converging to a good minimum compared to other algorithms.

6.3 Classification on real data sets

We consider classification tasks (binary and multiclass) on several data sets from the UCI Machine Learning Repository [34]. We use the logistic loss for binary and multiclass classification problems. For k-class problems with $k > 2$, the parameter θ is a $d \times k$ matrix and CGD is performed block-wise along the class axis. In this case, a CGD cycle performs again d iterations, one for each

feature coordinate, each time updating the k associated model weights (a form of block coordinate gradient descent, see [9] for arguments in favor of this approach).

For each data set, we corrupt an increasing random fraction of samples with uninformative outliers or heavy-tailed noise. Each algorithm is hyper-optimized using cross-validation over an appropriate grid of hyper-parameters, see Appendix 8.3 for further details. Subsequently, we train each algorithm with optimal hyper-parameters 10 times over to account for the methods' randomness (most procedures appear to be quite stable across runs) and we finally report in Figure 3 the median accuracy obtained on a 15% test-set (y-axis) for each data set, corruption level (x-axis) and algorithm.

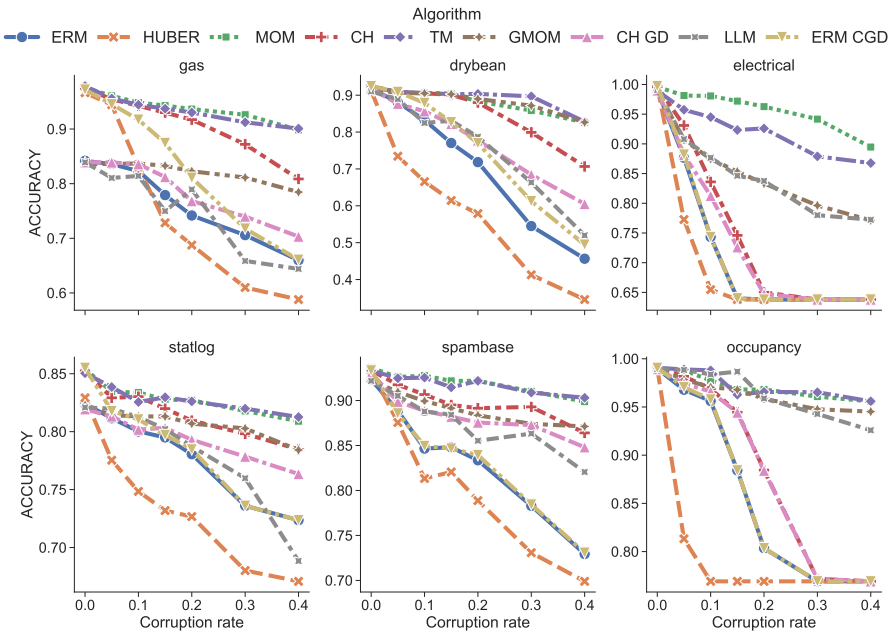

Figure 3: Test accuracy (y-axis) against the proportion of corrupted samples (x-axis) for six data sets and the considered algorithms.

First, we note that better optima can sometimes be found using CGD over GD as can be seen by comparing `ERM` and `ERM CGD` on the *gas* and *statlog* data sets with zero corruption. This is also apparent when corruption is present through the fact that `CH` often outperforms `CH GD`. Unsurprisingly, the accuracy of algorithms deteriorates with increasing corruption. In particular, fast degradations occur for `ERM` and `HUBER` which are not intended to handle corrupted covariates. The best performances are generally achieved by `TM` and `MOM` which only lose a minimal fraction of their accuracy to corruption. Although `CH` has no theoretical guarantees against corruption, we see that it is fairly robust on many data sets, especially at low corruption rates. However, its performance inevitably degrades beyond 20% corruption. The most competitive baseline is `GMOM` which manages to match the performance of `TM` and `MOM` on certain instances but seems to generally lag behind as a GD based algorithm. Finally, `LLM` fails to provide a competitive baseline in most cases and suffers from unsteady performance across data sets.

In order to illustrate the computational performance of each method, we report in Figure 4 the test accuracy (y-axis) against the training time (x-axis) along iterations of each algorithm for two data sets (rows) and 0%, 15% and 30% corruption (resp. first, middle and last column). In all situations, standard methods such as `ERM` and `HUBER` run the fastest due to the absence of computational overhead. However, they only reach poor quality optima when data is corrupted as opposed to robust methods. The results of Figure 4 concur with Figure 1 showing `MOM` to

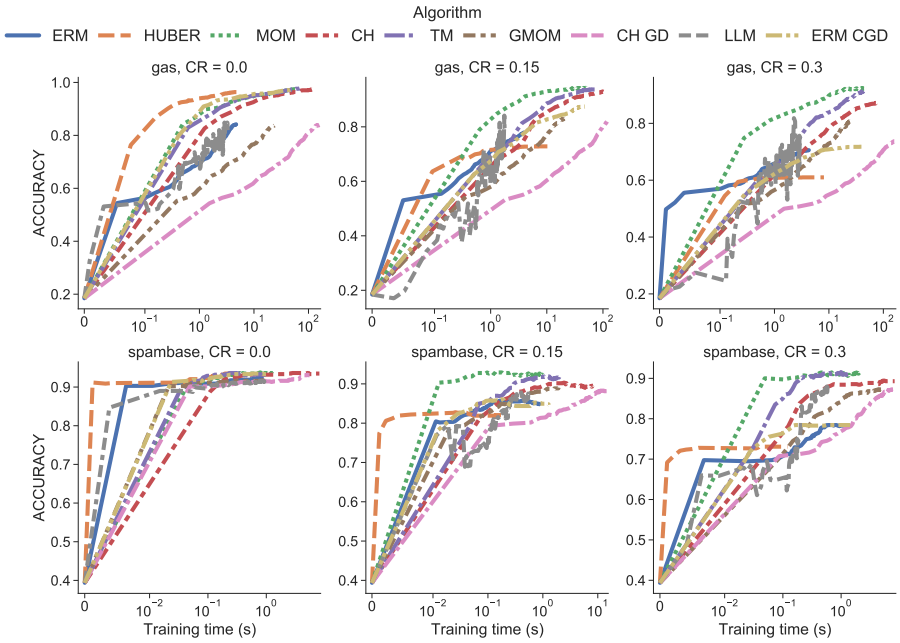

Figure 4: Test accuracy (y-axis) against computation time (x-axis) along training iterations on two data sets (rows) for 0% corruption (first column), 15% corruption (middle column) and 30% corruption (last column). Note the log scale on the x-axis.

be the fastest and `CH` the slowest CGD algorithm. We also observe that `MOM` and `TM` are clear favourites in terms of final performance and convergence speed, especially when corruption is present. Unsurprisingly, we observe that the combination of GD with the Catoni-Holland estimator in `CH GD` results in the slowest method in most cases. In comparison, `GMOM` is a faster alternative whose speed varies between data sets. This may be explained by the varying number of features and distribution of the data sets affecting the vector median computations. Finally, we see that although `LLM` is among the fastest methods (as seen for 0% corruption), its iteration lacks stability and is visibly affected by corruption.

6.4 Regression on real data sets

We consider the same experimental setting (data corruption and hyper-optimization of algorithms) as in Section 6.3 on regression data sets from the UCI Machine Learning Database, see Appendix 8.3 for details. We use the square loss for training and the mean squared error (MSE) as test metric, except for `HUBER`, `RANSAC` and `LAD` which are trained differently. We report the results in Figures 5 and 6. Figure 5 shows the test MSE (y-axis) against the corruption rate (x-axis) for several data sets and algorithms while Figure 6 displays the test MSE against the training time analogously to Figure 4. Note that only final performance and total training time are shown for `RANSAC`, `HUBER` and `LAD` on Figure 6. This is because they were run using `scikit-learn`'s implementation which does not give access to training history. We observe on Figure 5 that `HUBER` and `LAD` often achieve similar performance as they both optimize ℓ_1 objectives. However, `HUBER` finds more precise optima in many cases at low corruption rates. This may be attributed to the quadratic nature of its loss function around zero. Poor performance at low corruption levels is also observed for `RANSAC` in most cases. The latter appears to be somewhat resilient to corruption, suffering limited performance degradation at increasing levels. However, like `LLM`, `RANSAC` displays fluctuating and non competitive results. While `CH` turns out to be especially fragile to corruption

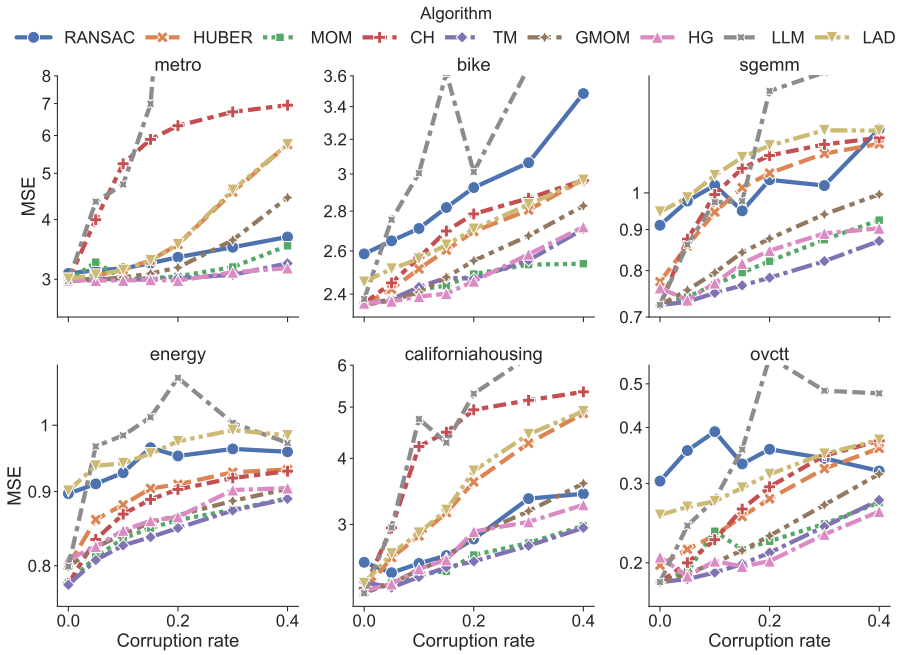

Figure 5: Mean squared error (y-axis) against the proportion of corrupted samples (x-axis) for six data sets and the considered algorithms.

on the regression task, the other CGD algorithms `TM` and `MOM` generally secure the best scores. Close competition and sometimes improved performance is shown by `HG` and `GMOM` which prove to effectively filter out corruption, although `GMOM` seems less reliable at higher levels. Furthermore, Figure 6 shows that the robustness of `HG` and `GMOM` comes at a significantly higher computational price, especially for `HG` whose running time is slower by orders of magnitude on some data sets and outright prohibitive on others.

As for the remaining algorithms, Figure 6 again shows fast convergence for CGD methods with good final performances for `MOM` and `TM`. The iteration of `LLM` is similarly swift but severely destabilized by corruption. Finally, `LAD`, `HUBER` and `RANSAC` sometimes offer short runtimes but lack robustness to corruption.

Our numerical experiments confirm that robust CGD algorithms (`TM` and `MOM`) offer a good compromise between statistical accuracy, robustness and computational cost.

7 Conclusion

In this paper, we introduce new robust algorithms for supervised learning by combining CGD with several robust partial derivative estimators. We derive convergence results for several variants of CGD with noisy partial derivatives and prove deviation bounds for all the considered estimators under minimal moment assumptions, including cases with infinite variance and the presence of arbitrary outliers (except for the `CH` estimator). This leads to very robust learning algorithms, with a numerical cost comparable to that of non-robust approaches based on empirical risk minimization, since it lets us bypass the need of a robust *vector mean* and allows to update model weights immediately using a robust estimator of a *single partial derivative* only. This is substantiated by our numerical experiments which illustrate the good compromise offered by our approach between statistical accuracy, robustness and computational cost. Perspectives include robust learning algorithms in high dimension, achieving sparsity-aware generalization bounds, which is beyond the

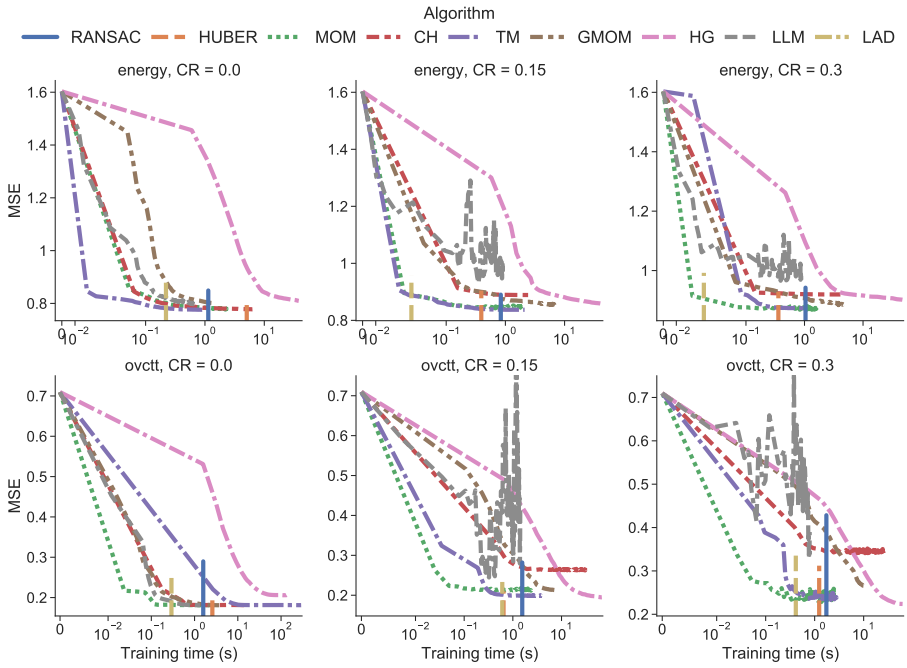

Figure 6: Mean squared error (y-axis) against computation time (x-axis) along training iterations on two data sets (rows) for 0% corruption (first column), 15% corruption (middle column) and 30% corruption (last column).

scope of this paper, since it would require different algorithms based on methods such as mirror descent with an appropriately chosen divergence, see for instance [95, 59].

Acknowledgments

This research is supported by the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program (reference ANR-19-P3IA-0001; PRAIRIE 3IA Institute).

8 Supplementary theoretical results and details on experiments

8.1 The Lipschitz constants L_j are unknown

The step-sizes $(\beta_j)_{j\in[\![d]\!]}$ used in Theorems 1 and 2 are given by $\beta_j = 1/L_j$, where the Lipschitz constants L_j are defined by (8). This makes them non-observable, since they depend on the unknown distribution of the non-corrupted features P_{X_i} for $i \in \mathcal{I}$. We cannot use line-search [3] here, since it requires to evaluate the objective $R(\theta)$, which is unknown as well. In order to provide theoretical guarantees similar to that of Theorem 1 without knowing $(L_j)_{j=1}^d$, we use the following approach. First, we use the upper bound

$$U_j := \gamma \mathbb{E}\big[(X^j)^2\big] \geq L_j, \tag{26}$$

which holds under Assumption 1 and estimate $\mathbb{E}[(X^j)^2]$ to build a robust estimator of U_j. In order to obtain an observable upper bound and to control its deviation with a large probability, we introduce the following condition.

Definition 2. *We say that a real random variable Z satisfies the L^ζ-L^ξ condition with constant $C \geq 1$ whenever it satisfies*

$$\big(\mathbb{E}\big[|Z - \mathbb{E}Z|^\zeta\big]\big)^{1/\zeta} \leq C\big(\mathbb{E}\big[|Z - \mathbb{E}Z|^\xi\big]\big)^{1/\xi}. \tag{27}$$

Using this condition, we can use the `MOM` estimator to obtain a high probability upper bound on $\mathbb{E}[(X^j)^2]$ as stated in the following lemma.

Lemma 5. *Grant Assumption 2 with $\alpha \in (0, 1]$ and suppose that for all $j \in [\![d]\!]$, the variable $(X^j)^2$ satisfies the $L^{(1+\alpha)}$-L^1 condition with a known constant C. For any fixed $j \in [\![d]\!]$, let $\widehat{\sigma}_j^2$ be the `MOM` estimator of $\mathbb{E}[(X^j)^2]$ with K blocks. If $|\mathcal{O}| \leq K/12$, we have*

$$\mathbb{P}\Big[\Big(1 - 12^{1/(1+\alpha)} C\Big(\frac{K}{n}\Big)^{\alpha/(1+\alpha)}\Big)^{-1} \widehat{\sigma}_j^2 \leq \mathbb{E}[(X^j)^2]\Big] \leq \exp(-K/18).$$

If we fix a confidence level $\delta \in (0, 1)$ and choose $K := \lceil 18 \log(1/\delta) \rceil$, we have

$$\Big(1 - 216^{1/(1+\alpha)} C\Big(\frac{\log(1/\delta)}{n}\Big)^{\alpha/(1+\alpha)}\Big)^{-1} \widehat{\sigma}_j^2 > \mathbb{E}[(X^j)^2]$$

with a probability larger than $1 - \delta$.

The proof of Lemma 5 is given in Appendix 9. Denoting $\widehat{U}_j$ the upper bounds it provides on $\mathbb{E}[(X^j)^2]$, we can readily bound the Lipschitz constants as $L_j \leq \gamma \widehat{U}_j$ which leads to the following statement.

Corollary 2. *Grant the same assumptions as in Theorem 1 and Proposition 1. Suppose additionally that for all $j \in [\![d]\!]$, the variable $(X^j)^2$ satisfies the $L^{(1+\alpha)}$-L^1 condition with a known constant C and fix $\delta \in (0, 1)$. Let $\theta^{(T)}$ be the output of Algorithm 1 with step-sizes $\widehat{\beta}_j = 1/\overline{L}_j$ where $\overline{L}_j := \gamma \widehat{U}_j$ and $\widehat{U}_j$ are the upper bounds from Lemma 5 with confidence $\delta/2d$, an initial iterate $\theta^{(0)}$, importance sampling distribution $p_j = \overline{L}_j / \sum_{k\in[\![d]\!]} \overline{L}_k$ and estimators of the partial derivatives with error vector $\epsilon(\cdot)$. Then, we have*

$$\mathbb{E}\big[R(\theta^{(T)})\big] - R^\star \leq (R(\theta^{(0)}) - R^\star)\Big(1 - \frac{\lambda}{\sum_{j\in[\![d]\!]} \overline{L}_j}\Big)^T + \frac{1}{2\lambda}\big\|\epsilon(\delta/2)\big\|_2^2 \tag{28}$$

with probability at least $1 - \delta$.

The proof of Corollary 2 is given in Appendix 9. It is a direct consequence of Theorem 1 and Lemma 5 and shows that an upper bound similar to that of Theorem 1 can be achieved with *observable* step-sizes. One may argue that the $L^{(1+\alpha)}$-L^1 condition simply bypasses the difficulty of deriving an observable upper bound by arbitrarily assuming that a ratio of moments is observed. However, we point out that a hypothesis of this nature is indispensable to obtain bounds such as the one above (alternatively, consider a real random variable with an infinitesimal mass drifting towards infinity). In fact, the $L^{(1+\alpha)}$-L^1 condition is much weaker than the requirement of boundedness (with known range) common to most known empirical bounds [81, 4, 85].

8.2 Observable upper bound for the moment $m_{\alpha,j}$

Since the moment $m_{\alpha,j}$, it is not observable, so we propose in Lemma 6 below an observable upper bound deviation for it based on MOM. Let us introduce now a robust estimator $\widehat{m}^{\texttt{MOM}}_{\alpha,j}(\theta)$ of the unknown moment $m_{\alpha,j}(\theta)$ using the following "two-step" MOM procedure. First, we compute $\widehat{g}^{\texttt{MOM}}_j(\theta)$, the MOM estimator of $g_j(\theta)$ with K blocks given by (14). Then, we compute again a MOM estimator on $|g^i_j(\theta) - \widehat{g}^{\texttt{MOM}}_j(\theta)|^{1+\alpha}$ for $i \in [\![n]\!]$, namely

$$\widehat{m}^{\texttt{MOM}}_{\alpha,j}(\theta) := \operatorname{median}\big(\widehat{m}^{(1)}_{\alpha,j}(\theta), \dots, \widehat{m}^{(K)}_{\alpha,j}(\theta)\big), \tag{29}$$

where

$$\widehat{m}^{(k)}_{\alpha,j}(\theta) := \frac{1}{|B_k|} \sum_{i \in B_k} \big|g^i_j(\theta) - \widehat{g}^{\texttt{MOM}}_j(\theta)\big|^{1+\alpha},$$

using uniformly sampled blocks $B_1, \dots, B_K$ of equal size that form a partition of $[\![n]\!]$.

Lemma 6. *Grant Assumptions 1 and 2 with $\alpha \in (0, 1]$ and suppose that for all $j \in [\![d]\!]$ and $\theta \in \Theta$ the partial derivatives $\ell'(X^\top\theta, Y)X^j$ satisfy the $L^{(1+\alpha)^2}$-$L^{(1+\alpha)}$ condition with known constant C for any $j \in [\![d]\!]$ (see Definition 2). Then, if $|\mathcal{O}| \le K/12$, we have*

$$\mathbb{P}\big[\widehat{m}^{\texttt{MOM}}_{\alpha,j}(\theta) \le (1-\kappa) m_{\alpha,j}(\theta)\big] \le 2\exp(-K/18)$$

where $\kappa = \epsilon + 24(1+\alpha)\big(\frac{(1+\epsilon)K}{n}\big)^{\alpha/(1+\alpha)}$ and $\epsilon = (24(1 + C^{(1+\alpha)^2}))^{1/(1+\alpha)}\big(\frac{K}{n}\big)^{\alpha/(1+\alpha)}$.

The proof of Lemma 6 is given in Appendix 9.

8.3 Experimental details

We provide in this section supplementary information about the numerical experiments conducted in Section 6.

8.3.1 Data sets

The main characteristics of the data sets used from the UCI repository are given in Table 3 and their direct URLs are given in Table 4.

8.3.2 Data corruption

For a given corruption rate η, we obtain a corrupted version of a data set by replacing an η-fraction of its samples with uninformative elements. For a data set of size n we choose $\mathcal{O} \subset [\![n]\!]$ which satisfies $|\mathcal{O}| = \eta n$ up to integer rounding. The corruption is applied prior to any preprocessing except in the regression case where label scaling is applied before. The affected subset is chosen uniformly at random. Since many data sets contain both continuous and categorical data features,

Data set	# Samples	# Features	# Categorical	# Classes
statlog	6,435	36	0	6
spambase	4,601	57	0	2
electrical	10,000	13	0	2
occupancy [14]	20,560	5	0	2
gas [104]	13,910	128	0	6
drybean [63]	13,611	16	0	7
energy [15]	19,735	27	0	-
bike [36]	17,379	10	5	-
metro	48,204	6	1	-
sgemm [5]	241,600	14	0	-
ovctt	68,784	20	2	-
californiahousing	20,640	8	0	-

Table 3: Main characteristics of the data sets used in experiments, including number of samples, number of features, number of categorical features and number of classes.

we distinguish two different corruption mechanisms which we apply depending on their nature. The labels are corrupted as continuous or categorical values when the task is respectively regression or classification. Denote $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n\times(d+1)}$ the data matrix with the vector of labels added to its columns. Let $\widetilde{J} \subset [\![d+1]\!]$ denote the index of continuous columns, we compute $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ their empirical means and standard deviations respectively for $j \in \widetilde{J}$. We also sample a random unit vector u of size $|\widetilde{J}|$.

- For categorical feature columns, for each corrupted index $i \in \mathcal{O}$, we replace $\boldsymbol{X}_{i,j}$ with a uniformly sampled value among $\{\boldsymbol{X}_{\bullet,j}\}$ i.e. among the possible modalities of the categorical feature in question.
- For continuous features, for each corrupted index $i \in \mathcal{O}$, we replace $\boldsymbol{X}_{i,\widetilde{J}}$ with equal probability with one of the following possibilities:
  - a vector ξ sampled coordinatewise according to $\xi_j = r_j + 5\widehat{\sigma}_j\nu$ where r_j is a value randomly picked in the column $\boldsymbol{X}_{\bullet,j}$ and ν is a sample from the Student distribution with 2.1 degrees of freedom.
  - a vector ξ sampled coordinatewise according to $\xi_j = \widehat{\mu}_j + 5\widehat{\sigma}_j u_j + z$ where z is a standard gaussian.
  - a vector ξ sampled according to $\xi = \widehat{\mu} + 5\widehat{\sigma} \otimes w$ where w is a uniformly sampled unit vector.

8.4 Preprocessing

We apply a minimal amount of preprocessing to the data before applying the considered learning algorithms. More precisely, categorical features are one-hot encoded while centering and standard scaling is applied to the continuous features.

8.5 Parameter hyper-optimization

We use the `hyperopt` library to find optimal hyper-parameters for all algorithms. For each data set, the available samples are split into training, validation and test sets with proportions $70\%, 15\%, 15\%$. Whenever corruption is applied, it is restricted to the training set. We run 50 rounds of hyper-parameter optimization which are trained on the training set and evaluated on the validation set. Then, we report results on the test set for all hyper-optimized algorithms. For each

Data set	URL
statlog	https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29
spambase	https://archive.ics.uci.edu/ml/datasets/spambase
electrical	https://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+
occupancy	https://archivehttps://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+
gas	https://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset
drybean	https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset
energy	https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction
bike	https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
metro	https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume
sgemm	https://archive.ics.uci.edu/ml/datasets/SGEMM+GPU+kernel+performance
ovctt	https://archive.ics.uci.edu/ml/datasets/Online+Video+Characteristics+and+Transcoding+Time+Dataset
californiahousing	loaded from scikitlearn.datasets

Table 4: The URLs of all the data sets used in the paper, giving direct download links and supplementary details.

algorithm, the hyper-parameters are tried out using the following sampling mechanism (the one we specify to `hyperopt`):

- `MOM`, `GMOM`, `LLM`: we optimize the number of blocks K used for the median-of-means computations. This is done through a `block_size` $= K/n$ hyper-parameter chosen with log-uniform distribution over $[10^{-5}, 0.2]$
- `CH` and `CH GD`: we optimize the confidence δ used to define the `CH` estimator's scale parameter (see Equation (21)) chosen with log-uniform distribution over $[e^{-10}, 1]$
- `TM`, `HG`: we optimize the percentage used for trimming uniformly in $[10^{-5}, 0.3]$
- `RANSAC`: we optimize the value of the `min_samples` parameter in the scikit-learn implementation, chosen as $4 + m$ with m an integer chosen uniformly in $[\![100]\!]$
- `HUBER`: we optimize the `epsilon` parameter in the scikit-learn implementation chosen uniformly in $[1.0, 2.5]$

9 Proofs

9.1 Proof of Theorem 1

This proof follows, with minor modifications, the proof of Theorem 1 from [105]. Using Definition 1 , we obtain

$$\mathbb{P}[\mathcal{E}] \geq 1 - \delta \quad \text{where} \quad \mathcal{E} := \big\{ \forall j \in [\![d]\!], \quad \forall t \in [T], \quad \big|\widehat{g}_j(\theta^{(t)}) - g_j(\theta^{(t)})\big| \leq \epsilon_j(\delta) \big\}. \tag{30}$$

Let us recall that e_j stands for the j-th canonical basis of $\mathbb{R}^d$ and that, as described in Algorithm 1, we have

$$\theta^{(t+1)} = \theta^{(t)} - \beta_{j_t} \widehat{g}_t e_{j_t},$$

where we use the notations $\widehat{g}_t = \widehat{g}_{j_t}(\theta^{(t)})$ and $g_t = g_{j_t}(\theta^{(t)})$ and where we recall that $j_1, \dots, j_t$ is a i.i.d sequence with distribution p. We introduce also $\epsilon_j := \epsilon_j(\delta)$. Using Assumption 3, we obtain

$$\begin{aligned}
R(\theta^{(t+1)}) &= R\big(\theta^{(t)} - \beta_{j_t} \widehat{g}_t e_{j_t}\big) \\
&\leq R(\theta^{(t)}) - \big\langle g(\theta^{(t)}), \beta_{j_t} \widehat{g}_t e_{j_t} \big\rangle + \frac{L_{j_t}}{2} \beta_{j_t}^2 \widehat{g}_t^2 \\
&= R(\theta^{(t)}) - \beta_{j_t} g_t^2 - \beta_{j_t} g_t (\widehat{g}_t - g_t) + \frac{L_{j_t} \beta_{j_t}^2}{2} \big(g_t^2 + (\widehat{g}_t - g_t)^2 + 2 g_t (\widehat{g}_t - g_t) \big) \\
&= R(\theta^{(t)}) - \beta_{j_t} g_t (1 - L_{j_t} \beta_{j_t}) (\widehat{g}_t - g_t) - \beta_{j_t} \Big(1 - \frac{L_{j_t} \beta_{j_t}}{2} \Big) g_t^2 + \frac{L_{j_t} \beta_{j_t}^2}{2} (\widehat{g}_t - g_t)^2 \\
&= R(\theta^{(t)}) - \frac{1}{2 L_{j_t}} g_t^2 + \frac{1}{2 L_{j_t}} (\widehat{g}_t - g_t)^2 \\
&\leq R(\theta^{(t)}) - \frac{1}{2 L_{j_t}} g_t^2 + \frac{\epsilon_{j_t}^2}{2 L_{j_t}}
\end{aligned} \tag{31}$$

on the event $\mathcal{E}$, where we used the choice $\beta_{j_t} = 1/L_{j_t}$ and the fact that $|\widehat{g}_t - g_t| \leq \epsilon_{j_t}$ on $\mathcal{E}$.

Since $j_1, \dots, j_t$ is a i.i.d sequence with distribution p, we have for any $(j_1, \dots, j_{t-1})$-measurable and integrable function φ that

$$\mathbb{E}_{t-1}\big[\varphi(j_t)\big] = \sum_{j \in [\![d]\!]} \varphi(j) p_j,$$

where we denote for short the conditional expectation $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}_{t-1}[\cdot|j_1, \ldots, j_{t-1}]$. So, taking $\mathbb{E}_{t-1}[\cdot]$ on both sides of (31) leads, whenever $p_j = L_j / \sum_{k=1}^d L_k$, to

$$\mathbb{E}_{t-1}\big[R(\theta^{(t+1)})\big] \leq R(\theta^{(t)}) - \frac{1}{2\sum_k L_k}\big\|g(\theta^{(t)})\big\|^2 + \frac{1}{2\sum_k L_k}\Xi,$$

where we introduced $\Xi := \|\epsilon(\delta)\|_2^2$, while it leads to

$$\mathbb{E}_{t-1}\big[R(\theta^{(t+1)})\big] \leq R(\theta^{(t)}) - \frac{1}{2L_{\max}d}\big\|g(\theta^{(t)})\big\|^2 + \frac{1}{2dL_{\min}}\Xi$$

whenever $p_j = 1/d$, simply using $L_{\min} \leq L_j \leq L_{\max}$. In order to treat both cases simultaneously, consider $\bar{L} = \sum_{k=1} L_k$ and $\bar{\epsilon} = \Xi/(2\sum_k L_k)$ whenever $p_j = L_j / \sum_{k=1}^d L_k$ and $\bar{L} = dL_{\max}$ and $\bar{\epsilon}/(2dL_{\min})$ whenever $p_j = 1/d$ and continue from the inequality

$$\mathbb{E}_{t-1}\big[R(\theta^{(t+1)})\big] \leq R(\theta^{(t)}) - \frac{1}{2\bar{L}}\big\|g(\theta^{(t)})\big\|^2 + \bar{\epsilon}.$$

Introducing $\phi_t := \mathbb{E}\big[R(\theta^{(t)})\big] - R^\star$ and taking the expectation w.r.t. all $j_1, \ldots, j_t$ we obtain

$$\phi_{t+1} \leq \phi_t - \frac{1}{2\bar{L}}\mathbb{E}\big\|g(\theta^{(t)})\big\|^2 + \bar{\epsilon}. \tag{32}$$

Using Inequality (9) with $\theta_1 = \theta^{(t)}$ gives

$$R(\theta_2) \geq R(\theta^{(t)}) + \big\langle \nabla R(\theta^{(t)}), \theta_2 - \theta^{(t)} \big\rangle + \frac{\lambda}{2}\big\|\theta_2 - \theta^{(t)}\big\|^2$$

for any $\theta_2 \in \mathbb{R}^d$, so that by minimizing both sides with respect to θ_2 leads to

$$R^\star \geq R(\theta^{(t)}) - \frac{1}{2\lambda}\big\|g(\theta^{(t)})\big\|^2$$

namely

$$\phi_t \leq \frac{1}{2\lambda}\mathbb{E}\big\|g(\theta^{(t)})\big\|^2,$$

by taking the expectation on both sides. Together with (32) this leads to the following approximate contraction property:

$$\phi_{t+1} \leq \phi_t\Big(1 - \frac{\lambda}{\bar{L}}\Big) + \bar{\epsilon},$$

and by iterating $t = 1, \ldots, T$ to

$$\phi_T \leq \phi_0\Big(1 - \frac{\lambda}{\bar{L}}\Big)^T + \frac{\bar{\epsilon}\bar{L}}{\lambda},$$

which allows to conclude the Proof of Theorem 1. □

9.2 Proof of Theorem 2

This proof reuses ideas from [73] and [7] and adapts them to our context where the gradient coordinates are replaced with high confidence approximations. Without loss of generality, we initially assume that the coordinates are cycled upon in the natural order. We condition on the event (30) which holds with probability $\geq 1-\delta$ as in the proof of Theorem 1 and denote $\epsilon_j = \epsilon_j(\delta)$ and $\epsilon_{Euc} = \|\epsilon(\delta)\|$.

Let the iterations be denoted as $\theta^{(t)}$ for $t = 0, \ldots, T$ and $\theta_{i+1}^{(t)} = \theta_i^{(t)} - \beta_{i+1}\widehat{g}(\theta_i^{(t)})_{i+1}e_{i+1}$ for $i = 0, \ldots, d-1$ with $\beta_i = 1/L_i$, $\theta_0^{(t)} = \theta^{(t)}$ and $\theta_d^{(t)} = \theta^{(t+1)}$. With these notations we have

$$R(\theta^{(t)}) - R(\theta^{(t+1)}) = \sum_{i=0}^{d-1} R(\theta_i^{(t)}) - R(\theta_{i+1}^{(t)}).$$

Similarly to (31) in the proof of Theorem 1 we find:

$$R(\theta_i^{(t)}) - R(\theta_{i+1}^{(t)}) \geq \frac{1}{2L_{i+1}}\big(g(\theta_i^{(t)})_{i+1}^2 - \epsilon_{i+1}^2\big),$$

leading to

$$R(\theta^{(t)}) - R(\theta^{(t+1)}) \geq \sum_{i=0}^{d-1} \frac{1}{2L_{i+1}} g(\theta_i^{(t)})_{i+1}^2 - \frac{1}{2L_{\min}} \sum_{i=0}^{d-1} \epsilon_{i+1}^2. \tag{33}$$

The following aims to find a relationship between $\sum_{i=0}^{d-1} \frac{1}{2L_{i+1}} g(\theta_i^{(t)})_{i+1}^2$ and $\big\|g(\theta^{(t)})\big\|_2^2$ which we do by comparing coordinates. For the first step in a cycle we have $g(\theta^{(t)})_1 = g(\theta_0^{(t)})_1$ because $\theta^{(t)} = \theta_0^{(t)}$. Let $j \in \{1, \ldots, d-1\}$, by the Mean Value Theorem, there exists $\gamma_j^{(t)} \in \mathbb{R}^d$ such that we have:

$$\begin{aligned}
&g(\theta^{(t)})_{j+1} = g(\theta^{(t)})_{j+1} - g(\theta_j^{(t)})_{j+1} + g(\theta_j^{(t)})_{j+1} \\
&= \big(\nabla g_{j+1}(\gamma_j^{(t)})\big)^\top \big(\theta^{(t)} - \theta_j^{(t)}\big) + g(\theta_j^{(t)})_{j+1} \\
&= \left[\frac{\partial R(\gamma_j^{(t)})}{\partial_{j+1}\partial_1}, \ldots, \frac{\partial R(\gamma_j^{(t)})}{\partial_{j+1}\partial_j}, 0, \ldots, 0\right] \big[(\theta^{(t)} - \theta_j^{(t)})_1, \ldots, (\theta^{(t)} - \theta_j^{(t)})_j, 0, \ldots, 0\big]^\top \\
&\quad + g(\theta_j^{(t)})_{j+1} \\
&= [H_{j+1,1}, \ldots, H_{j+1,j}, 0, \ldots, 0]\left[\frac{\widehat{g}_1(\theta_0^{(t)})}{L_1}, \ldots, \frac{\widehat{g}_j(\theta_{j-1}^{(t)})}{L_j}, 0, \ldots, 0\right]^\top + g(\theta_j^{(t)})_{j+1} \\
&= [H_{j+1,1}, \ldots, H_{j+1,j}, 0, \ldots, 0]\left[\frac{g_1(\theta_0^{(t)}) + \delta_1^{(t)}}{L_1}, \ldots, \frac{g_j(\theta_{j-1}^{(t)}) + \delta_j^{(t)}}{L_j}, 0, \ldots, 0\right]^\top \\
&\quad + g(\theta_j^{(t)})_{j+1} \\
&= \underbrace{\left[\frac{H_{j+1,1}}{\sqrt{L_1}}, \ldots, \frac{H_{j+1,j}}{\sqrt{L_j}}, \sqrt{L_{j+1}}, 0, \ldots, 0\right]}_{\widetilde{h}_{j+1}^\top} \underbrace{\left[\frac{g_1(\theta_0^{(t)})}{\sqrt{L_1}}, \ldots, \frac{g_d(\theta_{d-1}^{(t)})}{\sqrt{L_d}}\right]^\top}_{\widetilde{g}_t} \\
&\quad + \underbrace{[H_{j+1,1}, \ldots, H_{j+1,j}, 0, \ldots, 0]}_{h_{j+1}^\top}\left[\frac{\delta_1^{(t)}}{L_1}, \ldots, \frac{\delta_d^{(t)}}{L_d}\right]^\top \\
&= \widetilde{h}_{j+1}\widetilde{g}_t + h_{j+1}A^{-1}\delta^{(t)},
\end{aligned}$$

where we introduced the following quantities: $A \in \mathbb{R}^d$ equal to $A = \operatorname{diag}(L_j)_{j=1}^d$, the vector $\delta^{(t)} \in \mathbb{R}^d$ is such that $\delta_j^{(t)} = \widehat{g}(\theta_{j-1}^{(t)})_j - g(\theta_{j-1}^{(t)})_j$ which satisfies $|\delta_j^{(t)}| \leq \epsilon_j$, the matrix $H = (h_1, \ldots, h_d)^\top$ and $\widetilde{H} = A^{1/2} + HA^{-1/2} = (\widetilde{h}_1, \ldots, \widetilde{h}_d)^\top$. In the case $j = 0$ the vector $h_{j+1} = h_1$

is simply zero. This allows us to obtain the following estimation:

$$\begin{aligned}
\big\|g(\theta^{(t)})\big\|^2 &= \sum_{j=1}^{d} g(\theta^{(t)})_j^2 = \sum_{j=1}^{d} (\widetilde{h}_j^\top \widetilde{g}_t + h_j^\top A^{-1}\delta^{(t)})^2 \\
&\leq \sum_{j=1}^{d} 2(\widetilde{h}_j^\top \widetilde{g}_t)^2 + 2(h_j^\top A^{-1}\delta^{(t)})^2 = 2\big\|\widetilde{H}\widetilde{g}_t\big\|^2 + 2\big\|HA^{-1}\delta^{(t)}\big\|^2 \\
&\leq 2\big\|\widetilde{H}\big\|^2\big\|\widetilde{g}_t\big\|^2 + \frac{2}{L_{\min}^2}\|H\|^2\epsilon_{Euc}^2 \\
&= 2\|\widetilde{H}\|^2 \sum_{i=0}^{d-1} \frac{1}{L_{i+1}} g(\theta_i^{(t)})_{i+1}^2 + \frac{2}{L_{\min}^2}\|H\|^2\epsilon_{Euc}^2. \qquad (34)
\end{aligned}$$

We can bound the spectral norm $\|\widetilde{H}\|$ as follows:

$$\|\widetilde{H}\|^2 = \|A^{1/2} + HA^{-1/2}\|^2 \leq 2\|A^{1/2}{}^2 + 2\|HA^{-1/2}\|^2 \leq 2\Big(L_{\max} + \frac{\|H\|^2}{L_{\min}}\Big).$$

For $\|H\|$, we use the coordinate-wise Lipschitz-smoothness in order to find

$$\|H\|^2 \leq \|H\|_F^2 = \sum_{j=1}^{d} \|h_j\|^2 \leq \sum_{j=1}^{d} \big\|\nabla g_j(\gamma_{j-1}^{(t)})\big\|^2 \leq \sum_{j=1}^{d} L_j^2 \leq dL_{\max}^2.$$

Combining the previous inequality with (33) and (34), we find:

$$\begin{aligned}
&R(\theta^{(t)}) - R(\theta^{(t+1)}) \\
&\geq \frac{1}{8L_{\max}(1 + d\frac{L_{\max}}{L_{\min}})}\big\|g(\theta^{(t)})\big\|^2 - \frac{\epsilon_{Euc}^2}{2}\Big((\frac{1}{L_{\min}} + \frac{d\Big(\frac{L_{\max}}{L_{\min}}\Big)^2}{2L_{\max}(1 + d\frac{L_{\max}}{L_{\min}})}\Big) \\
&\geq \frac{1}{8L_{\max}(1 + d\frac{L_{\max}}{L_{\min}})}\big\|g(\theta^{(t)})\big\|^2 - \frac{\epsilon_{Euc}^2}{2}\Big(\frac{1}{L_{\min}} + \frac{1}{2L_{\min}}\frac{dL_{\max}/L_{\min}}{1 + d\frac{L_{\max}}{L_{\min}}}\Big) \\
&\geq \underbrace{\frac{1}{8L_{\max}(1 + d\frac{L_{\max}}{L_{\min}})}}_{=:\kappa}\big\|g(\theta^{(t)})\big\|^2 - \frac{3}{4L_{\min}}\epsilon_{Euc}^2,
\end{aligned}$$

where the last step uses that $\frac{dL_{\max}/L_{\min}}{1+d\frac{L_{\max}}{L_{\min}}} \leq 1$. Using λ-strong convexity by choosing $\theta_1 = \theta^{(t)}$ in inequality (9) and minimizing both sides w.r.t. θ_2 we obtain:

$$R(\theta^{(t)}) - R^\star \leq \frac{1}{2\lambda}\|g(\theta^{(t)})\|^2,$$

which combined with the previous inequality yields the contraction inequality:

$$R(\theta^{(t+1)}) - R^\star \leq (R(\theta^{(t)}) - R^\star)(1 - 2\lambda\kappa) + \frac{3}{4L_{\min}}\epsilon_{Euc}^2,$$

and after T iterations we have:

$$R(\theta^{(T)}) - R^\star \leq (R(\theta^{(0)}) - R^\star)(1 - 2\lambda\kappa)^T + \frac{3\epsilon_{Euc}^2}{8L_{\min}\lambda\kappa},$$

which concludes the proof of Theorem 2. To see that the proof still holds for any choice of coordinates satisfying the conditions in the main claim, notice that the computations leading up to Inequality (34) work all the same if one were to apply a permutation to the coordinates beforehand.

9.3 Convergence of the parameter error

We state and prove a result about the linear convergence of the parameter under strong convexity.

Theorem 4. *Grant Assumptions 1, 3 and 4. Let $\theta^{(T)}$ be the output of Algorithm 1 with constant step-size $\beta = \frac{2}{\lambda+L}$, an initial iterate $\theta^{(0)}$, uniform coordinates sampling $p_j = 1/d$ and estimators of the partial derivatives with error vector $\epsilon(\cdot)$. Then, we have*

$$\mathbb{E}\big\|\theta^{(T)} - \theta^\star\big\|_2 \leq \big\|\theta^{(0)} - \theta^\star\big\|_2 \Big(1 - \frac{2\beta\lambda L}{d(\lambda + L)}\Big)^T + \frac{\sqrt{d}(\lambda + L)}{\lambda L}\big\|\epsilon(\delta)\big\|_2 \tag{35}$$

with probability at least $1 - \delta$, where the expectation is w.r.t. the sampling of the coordinates.

Proof. As in the proof of Theorem 1, let $(\widehat{g}_j(\theta))_{j=1}^d$ be the estimators used and introduce the notations

$$\widehat{g}_t = \widehat{g}_{j_t}(\theta^{(t)}) \quad \text{and} \quad g_t = g_{j_t}(\theta^{(t)}).$$

We also condition on the event (30) which holds with probability $1 - \delta$ and use the notations $\epsilon_{Euc} = \|\epsilon(\delta)\|_2$ and $\epsilon_j = \epsilon_j(\delta)$. We denote $\|\cdot\|_{L_2}$ the L_2-norm w.r.t. the distribution over j_t i.e. for a random variable ξ we have $\|\xi\|_{L_2} = \sqrt{\mathbb{E}_{j_t}\|\xi\|^2}$. We compute:

$$\big\|\theta^{(t+1)} - \theta^\star\big\|_{L_2} = \big\|\theta^{(t)} - \beta_{j_t}\widehat{g}_t e_{j_t} - \theta^\star\big\|_{L_2} \leq \big\|\theta^{(t)} - \beta_{j_t} g_t e_{j_t} - \theta^\star\big\|_{L_2} + \big\|\beta_{j_t}(\widehat{g}_t - g_t)\big\|_{L_2}. \tag{36}$$

We first treat the first term of (36), in the case of uniform sampling with equal step-sizes $\beta_j = \beta$ we have:

$$\big\|\theta^{(t)} - \beta g_t e_{j_t} - \theta^\star\big\|^2 = \big\|\theta^{(t)} - \theta^\star\big\|^2 + \beta^2 g_t^2 - 2\beta\big\langle g_t e_{j_t}, \theta^{(t)} - \theta^\star\big\rangle.$$

By taking the expectation w.r.t. the random coordinate j_t we find:

$$\begin{aligned}
&\big\|\theta^{(t)} - \beta g_t e_{j_t} - \theta^\star\big\|_{L_2}^2 = \mathbb{E}\big\|\theta^{(t)} - \beta g_t e_{j_t} - \theta^\star\big\|^2 \\
&= \mathbb{E}\big\|\theta^{(t)} - \theta^\star\big\|^2 + \frac{\beta^2}{d}\mathbb{E}\big\|g(\theta^{(t)})\big\|^2 - 2\frac{\beta}{d}\mathbb{E}\big\langle g(\theta^{(t)}), \theta^{(t)} - \theta^\star\big\rangle \\
&= \mathbb{E}\|\theta^{(t)} - \theta^\star\|^2 + \Big(\frac{\beta}{d}\Big)^2\mathbb{E}\|g(\theta^{(t)})\|^2 - 2\frac{\beta}{d}\mathbb{E}\big\langle g(\theta^{(t)}), \theta^{(t)} - \theta^\star\big\rangle + \frac{\beta^2}{d}\mathbb{E}\big\|g(\theta^{(t)})\big\|^2\Big(1 - \frac{1}{d}\Big) \\
&\leq \mathbb{E}\big\|\theta^{(t)} - \theta^\star\big\|^2\Big(1 - \frac{2\beta\lambda L}{d(\lambda + L)}\Big) + \frac{\beta}{d}\Big(\frac{\beta}{d} - \frac{2}{\lambda + L}\Big)\mathbb{E}\big\|g(\theta^{(t)})\big\|^2 + \frac{\beta^2}{d}\mathbb{E}\big\|g(\theta^{(t)})\big\|^2\Big(1 - \frac{1}{d}\Big) \\
&= \mathbb{E}\big\|\theta^{(t)} - \theta^\star\big\|^2\Big(1 - \frac{2\beta\lambda L}{d(\lambda + L)}\Big) + \frac{\beta}{d}\Big(\beta - \frac{2}{\lambda + L}\Big)\mathbb{E}\big\|g(\theta^{(t)})\big\|^2 \\
&\leq \mathbb{E}\big\|\theta^{(t)} - \theta^\star\big\|^2 \underbrace{\Big(1 - \frac{2\beta\lambda L}{d(\lambda + L)}\Big)}_{=:\kappa^2}.
\end{aligned}$$

The first inequality is obtained by applying inequality (2.1.24) from [87] (see also [12] Lemma 3.11) and the second one is due to the choice of β. We can bound the second term as follows:

$$\big\|\widehat{g}_t - g_t\big\|_{L_2}^2 = \mathbb{E}_{j_t}\big|\widehat{g}_t - g_t\big|^2 = \frac{1}{d}\sum_{j=1}^d \big|\widehat{g}_j(\theta^{(t)}) - g_j(\theta^{(t)})\big|^2 \leq \frac{\epsilon_{Euc}^2}{d}.$$

Combining the latter with the former bound, we obtain the approximate contraction:

$$\big\|\theta^{(t+1)} - \theta^\star\big\|_{L_2} \leq \kappa\big\|\theta^{(t)} - \theta^\star\big\|_{L_2} + \frac{\beta\epsilon_{Euc}}{\sqrt{d}}.$$

By iterating this argument on T rounds we find that:

$$\left\|\theta^{(T)} - \theta^\star\right\|_{L_2} \leq \kappa^T \left\|\theta^{(0)} - \theta^\star\right\|_{L_2} + \frac{\beta \epsilon_{Euc}}{\sqrt{d}(1-\kappa)}.$$

Finally, the following inequality yields the result in the case of uniform sampling:

$$\frac{1}{1-\kappa} \leq \frac{1 + \sqrt{1 - \frac{2\beta\lambda L}{d(\lambda+L)}}}{\frac{2\beta\lambda L}{d(\lambda+L)}} \leq \frac{d(\lambda+L)}{\beta\lambda L}.$$

□

9.4 Proof of Lemma 1

Let $\theta \in \Theta$, using Assumption 1 we have:

$$|\ell(\theta^\top X, Y)| \leq C_{\ell,1} + C_{\ell,2}|\theta^\top X - Y|^q \leq C_{\ell,1} + 2^{q-1} C_{\ell,2}(|\theta^\top X|^q + |Y|^q).$$

Taking the expectation and using Assumption 2 shows that the risk $R(\theta)$ is well defined (recall that $q \leq 2$). Next, since $1 \leq q \leq 2$, simple algebra gives

$$\begin{aligned}
&\left|\ell'(\theta^\top X, Y) X_j\right|^{1+\alpha} \\
&\leq \left|\left(C'_{\ell,1} + C'_{\ell,2}|\theta^\top X - Y|^{q-1}\right) X^j\right|^{1+\alpha} \\
&\leq 2^\alpha \left(\left|C'_{\ell,1} X^j\right|^{1+\alpha} + (C'_{\ell,2}(|(\theta^\top X)^{q-1} X^j| + |Y^{q-1} X^j|))^{1+\alpha}\right) \\
&\leq 2^\alpha \left(\left|C'_{\ell,1} X^j\right|^{1+\alpha} + \left(C'_{\ell,2}\left(\sum_{k=1}^d |\theta_k|^{q-1} |(X^k)^{q-1} X^j| + |Y^{q-1} X^j|\right)\right)^{1+\alpha}\right) \\
&\leq 2^\alpha \left(\left|C'_{\ell,1} X^j\right|^{1+\alpha}\right. \\
&\left. + 2^\alpha (C'_{\ell,2})^{1+\alpha} \left(d^\alpha \sum_{k=1}^d |\theta_k|^{(q-1)(1+\alpha)} |(X^k)^{q-1} X^j|^{1+\alpha} + |Y^{q-1} X^j|^{1+\alpha}\right)\right).
\end{aligned}$$

Given Assumption 2, it is straightforward that $\mathbb{E}|X^j|^{1+\alpha} < \infty$ and $\mathbb{E}|Y^{q-1} X^j|^{1+\alpha} < \infty$. Moreover, using a Hölder inequality with exponents $a = \frac{q(1+\alpha)}{(q-1)(1+\alpha)}$ and $b = q$ (the case $q = 1$ is trivial) we find:

$$\mathbb{E}\left|(X^k)^{q-1} X^j\right|^{1+\alpha} \leq \left(\mathbb{E}\left|X^k\right|^{q(1+\alpha)}\right)^{1/a} \left(\mathbb{E}\left|X^j\right|^{q(1+\alpha)}\right)^{1/b},$$

which is finite under Assumption 2. This concludes the proof of Lemma 1.

9.5 Proof of Lemma 2

This proof follows a standard argument from [77, 42] in which we use a Lemma from [13] in order to control the $(1+\alpha)$-moment of the block means instead of their variance. Indeed, we know from Lemma 1 that under Assumptions 1 and 2, the gradient coordinates have finite $(1+\alpha)$-moments, namely $\mathbb{E}[|\ell'(X^\top\theta, Y) X_j|^{1+\alpha}] < +\infty$ for any $j \in [\![d]\!]$. Recall that $(\widehat{g}_j^{(k)}(\theta))_{k \in [\![K]\!]}$ stands for the block-wise empirical mean given by Equation (15) and introduce the set of non-corrupted block indices given by $\mathcal{K} = \{k \in [\![K]\!] \ : \ B_k \cap \mathcal{O} = \emptyset\}$. We will initially assume that the number of outliers satisfies $|\mathcal{O}| \leq (1-\varepsilon)K/2$ for some $0 < \varepsilon < 1$. Note that since samples are i.i.d in B_k for $k \in \mathcal{K}$, we have $\mathbb{E}\left[\widehat{g}_j^{(k)}(\theta)\right] = g_j(\theta)$. We use the following Lemma from [13].

Lemma 7 (Lemma 3 from [13]). *Let $Z, Z_1, \dots, Z_n$ be a i.i.d sequence with $m_\alpha = \mathbb{E}[|Z - \mathbb{E}Z|^{1+\alpha}] < +\infty$ for some $\alpha \in (0,1]$ and put $\bar{Z}_n = \frac{1}{n}\sum_{i\in[\![n]\!]} Z_i$. Then, we have*

$$\bar{Z}_n \leq \mathbb{E}Z + \Big(\frac{3m_\alpha}{\delta n^\alpha}\Big)^{1/(1+\alpha)}$$

for any $\delta \in (0,1)$, with a probability $1-\delta$.

Lemma 7 entails that

$$\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big| \leq \Big(\frac{3m_{j,\alpha}(\theta)}{\delta'(n/K)^\alpha}\Big)^{1/(1+\alpha)} =: \eta_{j,\alpha,\delta'}(\theta)$$

with probability larger than $1-2\delta'$, for each $k \in \mathcal{K}$, since we have n/K samples in block B_k. Now, recalling that $\widehat{g}_j(\theta)$ is the median (see (14)), we can upper bound its failure probability as follows:

$$\begin{aligned}
&\mathbb{P}\Big[\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big| \geq \eta_{j,\alpha,\delta'}(\theta)\Big] \\
&\qquad\qquad \leq \mathbb{P}\Big[\sum_{k\in[\![K]\!]} \mathbf{1}\Big\{\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big| \geq \eta_{j,\alpha,\delta'}(\theta)\Big\} > K/2\Big] \\
&\qquad\qquad \leq \mathbb{P}\Big[\sum_{k\in\mathcal{K}} \mathbf{1}\Big\{\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big| \geq \eta_{j,\alpha,\delta'}(\theta)\Big\} > K/2 - |\mathcal{O}|\Big],
\end{aligned}$$

since at most $|\mathcal{O}|$ blocks contain one outlier. Since the blocks B_k are disjoint and contain i.i.d samples for $k \in \mathcal{K}$, we know that

$$\sum_{k\in\mathcal{K}} \mathbf{1}\Big\{\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big| \geq \eta_{j,\alpha,\delta'}(\theta)\Big\}$$

follows a binomial distribution $\text{Bin}(|\mathcal{K}|, p)$ with $p \leq 2\delta'$. Using the fact that $\text{Bin}(|\mathcal{K}|, p)$ is stochastically dominated by $\text{Bin}(|\mathcal{K}|, 2\delta')$ and that $\mathbb{E}[\text{Bin}(|\mathcal{K}|, 2\delta')] = 2\delta'|\mathcal{K}|$, we obtain, if $S \sim \text{Bin}(|\mathcal{K}|, 2\delta')$, that

$$\begin{aligned}
\mathbb{P}\Big[\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big| \geq \eta_{j,\alpha,\delta'}(\theta)\Big] &\leq \mathbb{P}\big[S > K/2 - |\mathcal{O}|\big] \\
&= \mathbb{P}\big[S - \mathbb{E}S > K/2 - |\mathcal{O}| - 2\delta'|\mathcal{K}|\big] \\
&\leq \mathbb{P}\big[S - \mathbb{E}S > K(\varepsilon - 4\delta')/2\big] \\
&\leq \exp\big(-K(\varepsilon - 4\delta')^2/2\big),
\end{aligned}$$

where we used the fact that $|\mathcal{O}| \leq (1-\varepsilon)K/2$ and $|\mathcal{K}| \leq K$ for the second inequality and the Hoeffding inequality for the last. This concludes the proof of Lemma 2 for the choice $\varepsilon = 5/6$ and $\delta' = 1/8$.

9.6 Proof of Proposition 1

Step 1. First, we fix $\theta \in \Theta$ and try to bound $\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big|$ in terms of quantities only depending on $\widetilde{\theta}$ which is the closest point to θ in an ε-net. Recall that Δ is the diameter of the parameter set Θ and let $\varepsilon > 0$ be a positive number. There exists an ε-net covering Θ with cardinality no more than $(3\Delta/2\varepsilon)^d$ i.e. a set N_ε such that for all $\theta \in \Theta$ there exists $\widetilde{\theta} \in N_\varepsilon$ such that $\|\widetilde{\theta} - \theta\| \leq \varepsilon$. Consider a fixed $\theta \in \Theta$ and $j \in [\![d]\!]$, we wish to bound the quantity

$\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big|$. Using the ε-net N_ε, there exists $\widetilde{\theta}$ such that $\|\widetilde{\theta} - \theta\| \leq \varepsilon$ which we can use as follows:

$$\begin{aligned}\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big| &\leq \big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\widetilde{\theta})\big| + \big|g_j(\widetilde{\theta}) - g_j(\theta)\big| \\ &\leq \big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\widetilde{\theta})\big| + L_j\varepsilon, \end{aligned} \tag{37}$$

where we used the gradient's coordinate Lipschitz constant to bound the second term. We now focus on the second term. Introducing the notation $g_j^i(\theta) = \ell'(\theta^\top X_i, Y_i)X_i^j$, we have

$$g_j^i(\theta) = \ell'(\widetilde{\theta}^\top X_i, Y_i)X_i^j + \underbrace{(\ell'(\theta^\top X_i, Y_i) - \ell'(\widetilde{\theta}^\top X_i, Y_i))X_i^j}_{=:\Delta_i}.$$

Let $(B_k)_{k\in[\![K]\!]}$ be the blocks used to compute the MOM estimator and associated block means $\widehat{g}_j^{(k)}(\theta)$ and $\widehat{g}_j^{(k)}(\widetilde{\theta})$. Notice that the MOM estimator is *monotonous* non decreasing w.r.t. to each of the entries $g_j^i(\theta)$ when the others are fixed. Without loss of generality, assume that $\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\widetilde{\theta}) \geq 0$ then we have:

$$\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\widetilde{\theta})\big| \leq \big|\breve{g}_j^{\texttt{MOM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big|, \tag{38}$$

where $\breve{g}_j^{\texttt{MOM}}(\widetilde{\theta})$ is the MOM estimator obtained using the entries $\ell'\big(\widetilde{\theta}^\top X_i, Y_i\big)X_i^j + \varepsilon\gamma\|X_i\|^2 = g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2$ instead of $g_j^i(\theta)$. Note that $\breve{g}_j^{\texttt{MOM}}(\widetilde{\theta})$ no longer depends on θ except through the fact that $\widetilde{\theta}$ is chosen in N_ε so that $\big\|\widetilde{\theta} - \theta\big\| \leq \varepsilon$. Indeed, using the Lipschitz smoothness of the loss function and a Cauchy-Schwarz inequality we find that:

$$|\Delta_i| \leq \gamma\|\theta - \widetilde{\theta}\| \cdot \|X_i\| \cdot |X_i^j| \leq \varepsilon\gamma\|X_i\|^2.$$

Step 2. We now use the concentration property of MOM to bound the quantity which is in terms of $\widetilde{\theta}$. The samples $(g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)_{i\in[\![n]\!]}$ are independent and distributed according to the random variable $\ell'(\widetilde{\theta}^\top X, Y)X^j + \varepsilon\gamma\|X\|^2$. Denote $\overline{L} = \gamma\mathbb{E}\|X\|^2$ and for $k \in [\![K]\!]$ let $\widehat{g}_j^{(k)}(\widetilde{\theta}) = \frac{K}{n}\sum_{i\in B_k} g_j^i(\widetilde{\theta})$ and $\widehat{L}^{(k)} = \frac{K}{n}\sum_{i\in B_k}\gamma\|X_i\|^2$. We use Lemma 7 for each of these pairs of means to obtain that with probability at least $1 - \delta'/2$:

$$\big|\widehat{g}_j^{(k)}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| \leq \Big(\frac{6m_{j,\alpha}(\widetilde{\theta})}{\delta'(n/K)^\alpha}\Big)^{1/(1+\alpha)} =: \eta_{j,\alpha,\delta'/2}(\widetilde{\theta}),$$

and with probability at least $1 - \delta'/2$

$$\big|\widehat{L}^{(k)} - \overline{L}\big| \leq \Big(\frac{6m_{L,\alpha}}{\delta'(n/K)^\alpha}\Big)^{1/(1+\alpha)} =: \eta_{L,\alpha,\delta'/2},$$

where $m_{L,\alpha} = \mathbb{E}|\gamma\|X\|^2 - \overline{L}|^{1+\alpha}$. Hence for all $k \in [\![K]\!]$

$$\begin{aligned}&\mathbb{P}\big(\big|\widehat{g}_j^{(k)}(\widetilde{\theta}) + \varepsilon\widehat{L}^{(k)} - g_j(\widetilde{\theta})\big| > \eta_{j,\alpha,\delta'/2}(\widetilde{\theta}) + \varepsilon(\overline{L} + \eta_{L,\alpha,\delta'/2})\big) \\ &\quad\leq \mathbb{P}\big(\big|\widehat{g}_j^{(k)}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| > \eta_{j,\alpha,\delta'/2}(\widetilde{\theta})\big) + \mathbb{P}\big(\big|\widehat{L}^{(k)} - \overline{L}\big| > \eta_{L,\alpha,\delta'/2}\big) \\ &\quad\leq \delta'/2 + \delta'/2 = \delta'.\end{aligned}$$

Now defining the Bernoulli variables

$$U_k := \mathbf{1}\Big\{\big|\widehat{g}_j^{(k)}(\widetilde{\theta}) + \delta\widehat{L}^{(k)} - g_j(\widetilde{\theta})\big| > \eta_{j,\alpha,\delta'/2}(\widetilde{\theta}) + \varepsilon\big(\overline{L} + \eta_{L,\alpha,\delta'/2}\big)\Big\},$$

we have just seen they have success probability $\leq \delta'$, moreover

$$\mathbb{P}\Big[|\widetilde{g}_j^{\texttt{MOM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})| \geq \eta_{j,\alpha,\delta'/2}(\widetilde{\theta}) + \varepsilon(\overline{L} + \eta_{L,\alpha,\delta'/2})\Big] \leq \mathbb{P}\Big[\sum_{k\in[\![K]\!]} U_k > K/2\Big]$$
$$\leq \mathbb{P}\Big[\sum_{k\in\mathcal{K}} U_k > K/2 - |\mathcal{O}|\Big],$$

since at most $|\mathcal{O}|$ blocks contain one outlier. Since the blocks B_k are disjoint and contain i.i.d samples for $k \in \mathcal{K}$, we know that $\sum_{k\in\mathcal{K}} U_k$ follows a binomial distribution $\text{Bin}(|\mathcal{K}|, p)$ with $p \leq \delta'$. Using the fact that $\text{Bin}(|\mathcal{K}|, p)$ is stochastically dominated by $\text{Bin}(|\mathcal{K}|, \delta')$ and that $\mathbb{E}[\text{Bin}(|\mathcal{K}|, \delta')] = \delta'|\mathcal{K}|$, we obtain, if $S \sim \text{Bin}(|\mathcal{K}|, \delta')$, that

$$\begin{aligned}\mathbb{P}\Big[|\widetilde{g}_j^{\texttt{MOM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})| \geq \eta_{j,\alpha,\delta'/2}(\widetilde{\theta}) + \varepsilon\big(\overline{L} + \eta_{L,\alpha,\delta'/2}\big)\Big] &\leq \mathbb{P}\big[S > K/2 - |\mathcal{O}|\big] \\ &= \mathbb{P}\big[S - \mathbb{E}S > K/2 - |\mathcal{O}| - \delta'|\mathcal{K}|\big] \\ &\leq \mathbb{P}\big[S - \mathbb{E}S > K(\varepsilon' - 2\delta')/2\big] \\ &\leq \exp\big(-K(\varepsilon' - 2\delta')^2/2\big),\end{aligned}$$

where we used the condition $|\mathcal{O}| \leq (1-\varepsilon')K/2$ and $|\mathcal{K}| \leq K$ for the second inequality and the Hoeffding inequality for the last. To conclude, we choose $\varepsilon' = 5/6$ and $\delta' = 1/4$ and combine (37), (38) and the last inequality in which we take $K = \lceil 18\log(1/\delta)\rceil$ and use a union bound argument to obtain that with probability at least $1-\delta$ for all $j \in [\![d]\!]$

$$|\widetilde{g}_j^{\texttt{MOM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})| \leq \big((24m_{j,\alpha}(\widetilde{\theta}))^{1/(1+\alpha)} + \varepsilon(24m_{L,\alpha})^{1/(1+\alpha)}\big)\Big(\frac{18\log(d/\delta)}{n}\Big)^{\alpha/(1+\alpha)} + \varepsilon\overline{L}. \quad (39)$$

Step 3. We use the ε-net to obtain a uniform bound. For $\theta \in \Theta$ denote $\widetilde{\theta}(\theta) \in N_\varepsilon$ the closest point in N_ε satisfying in particular $\|\widetilde{\theta}(\theta) - \theta\| \leq \varepsilon$, we write, following previous arguments

$$\begin{aligned}\sup_{\theta\in\Theta} \big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big| &\leq \sup_{\theta\in\Theta} \big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\widetilde{\theta}(\theta))\big| + \big|g_j(\widetilde{\theta}(\theta)) - g_j(\theta)\big| \\ &\leq \sup_{\theta\in\Theta} \big|\widetilde{g}_j^{\texttt{MOM}}(\widetilde{\theta}(\theta)) - g_j(\widetilde{\theta}(\theta))\big| + \varepsilon L_j \\ &= \max_{\widetilde{\theta}\in N_\varepsilon} \big|\widetilde{g}_j^{\texttt{MOM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| + \varepsilon L_j.\end{aligned}$$

Here, we make a union bound argument over $\widetilde{\theta} \in N_\varepsilon$ for the inequality (39) and choose $\varepsilon = n^{-\alpha/(1+\alpha)}$ to obtain the final result concluding the proof of Proposition 1.

9.7 Proof of Proposition 2

This proof reuses arguments from the proof of Theorem 2 in [68]. We wish to bound $\big|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)\big|$ with high probability and uniformly on $\theta \in \Theta$. Fix $\theta \in \Theta$ and $j \in [\![d]\!]$, we have $\widehat{g}_j^{\texttt{MOM}}(\theta) = \text{median}\big(\widehat{g}_j^{(1)}(\theta), \ldots, \widehat{g}_j^{(K)}(\theta)\big)$ with $\widehat{g}_j^{(k)}(\theta) = \frac{K}{n}\sum_{i\in B_k} g_j^i(\theta)$ where the blocks $B_1, \ldots, B_K$ constitute a partition of $[\![n]\!]$.

Define the function $\phi(t) = (t-1)\mathbf{1}_{1\leq t\leq 2} + \mathbf{1}_{t>2}$, let $\mathcal{K} = \{k \in [\![K]\!],\ B_k \cap \mathcal{O} = \emptyset\}$ and $\mathcal{J} = \bigcup_{k\in\mathcal{K}} B_k$. Thanks to the inequality $\phi(t) \geq \mathbf{1}_{t\geq 2}$, we have:

$$\begin{aligned}&\sup_{\theta\in\Theta} \sum_{k=1}^K \mathbf{1}\Big\{\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big| > x\Big\} \leq \sup_{\theta\in\Theta} \sum_{k\in\mathcal{K}} \mathbb{E}\big[\phi\big(2\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big|/x\big)\big] + |\mathcal{O}| \\ &+ \sup_{\theta\in\Theta} \sum_{k\in\mathcal{K}} \Big(\phi\big(2\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big|/x\big) - \mathbb{E}\big[\phi\big(2\big|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)\big|/x\big)\big]\Big).\end{aligned}$$

Besides, the inequality $\phi(t) \leq \mathbf{1}_{t\geq 1}$, an application of Markov's inequality and Lemma 7 yield:

$$\mathbb{E}\big[\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big)\big] \leq \mathbb{P}\big(|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)| \geq x/2\big) \leq \frac{3m_{\alpha,j}(\theta)}{(x/2)^{1+\alpha}(n/K)^{\alpha}}.$$

Therefore, recalling that we defined $M_{\alpha,j} := \sup_{\theta\in\Theta} m_{\alpha,j}(\theta)$ we have

$$\sup_{\theta\in\Theta}\sum_{k=1}^{K}\mathbf{1}\Big\{|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)| > x\Big\} \leq K\bigg(\frac{3M_{\alpha,j}}{(x/2)^{1+\alpha}(n/K)^{\alpha}} + \frac{|\mathcal{O}|}{K} + \sup_{\theta\in\Theta}\frac{1}{K}\Big(\sum_{k\in\mathcal{K}}\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big) - \mathbb{E}\big[\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big)\big]\Big)\bigg).$$

Now since for all t we have $0 \leq \phi(t) \leq 1$, McDiarmid's inequality says with probability $\geq 1 - \exp(-2y^2K)$ that:

$$\sup_{\theta\in\Theta}\frac{1}{K}\Big(\sum_{k\in\mathcal{K}}\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big) - \mathbb{E}\big[\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big)\big]\Big) \leq \mathbb{E}\bigg[\sup_{\theta\in\Theta}\frac{1}{K}\Big(\sum_{k\in\mathcal{K}}\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big) - \mathbb{E}\big[\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big)\big]\Big)\bigg] + y.$$

Using a simple symmetrization argument (see for instance Lemma 11.4 in [10]) we find:

$$\mathbb{E}\bigg[\sup_{\theta\in\Theta}\frac{1}{K}\Big(\sum_{k\in\mathcal{K}}\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big) - \mathbb{E}\big[\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big)\big]\Big)\bigg] \leq 2\mathbb{E}\bigg[\sup_{\theta\in\Theta}\frac{1}{K}\sum_{k\in\mathcal{K}}\varepsilon_k\phi\big(2|\widehat{g}^{(k)}(\theta) - g(\theta)|/x\big)\bigg],$$

where the ε_ks are independent Rademacher variables. Since ϕ is 1-Lipschitz and satisfies $\phi(0) = 0$ we can use the contraction principle (see Theorem 11.6 in [10]) followed by another symmetrization step to find

$$2\mathbb{E}\bigg[\sup_{\theta\in\Theta}\frac{1}{K}\sum_{k\in\mathcal{K}}\varepsilon_k\phi\big(2|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\big)\bigg] \leq 4\mathbb{E}\bigg[\sup_{\theta\in\Theta}\frac{1}{K}\sum_{k\in\mathcal{K}}\varepsilon_k|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)|/x\bigg] \leq \frac{8}{xn}\mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i\in\mathcal{J}}\varepsilon_i g_j^i(\theta)\bigg] \leq \frac{8\mathcal{R}_j(\Theta)}{xn}.$$

Taking $|\mathcal{O}| \leq (1-\varepsilon)K/2$, we found that with probability $\geq 1 - \exp(-2y^2K)$

$$\sup_{\theta\in\Theta}\sum_{k=1}^{K}\mathbf{1}\Big\{|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)| > x\Big\} \leq K\bigg(\frac{3M_{\alpha,j}}{(x/2)^{1+\alpha}(n/K)^{\alpha}} + \frac{|\mathcal{O}|}{K} + \frac{8\mathcal{R}_j(\Theta)}{xn}\bigg).$$

Now by choosing $y = 1/4 - |\mathcal{O}|/K$ and $x = \max\Big(\Big(\frac{36M_{\alpha,j}}{(n/K)^{\alpha}}\Big)^{1/(1+\alpha)}, \frac{64\mathcal{R}_j(\Theta)}{n}\Big)$, we obtain the deviation bound:

$$\begin{aligned}\mathbb{P}\bigg(\sup_{\theta\in\Theta}|\widehat{g}_j^{\texttt{MOM}}(\theta) - g_j(\theta)| \geq \max\Big(\Big(\frac{36M_{\alpha,j}}{(n/K)^{\alpha}}\Big)^{1/(1+\alpha)}, \frac{64\mathcal{R}_j(\Theta)}{n}\Big)\bigg) &\leq \mathbb{P}\bigg(\sup_{\theta\in\Theta}\sum_{k=1}^{K}\mathbf{1}\Big\{|\widehat{g}_j^{(k)}(\theta) - g_j(\theta)| > x\Big\} > K/2\bigg)\\ &\leq \exp(-2(\varepsilon - 1/2)^2K/4)\\ &\leq \exp(-K/18),\end{aligned}$$

where the last inequality comes from the choice $\varepsilon = 5/6$. A simple union bound argument lets the previous inequality hold for all $j \in [\![d]\!]$ with high probability.

Finally, assuming that X^j has finite fourth moment for all $j \in [\![d]\!]$, we can control the Rademacher complexity. In this part, we assume without loss of generality that $\mathcal{I} = [\![n]\!]$, we first write

$$\begin{aligned}\mathcal{R}_j(\Theta) &= \mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^n \varepsilon_i \ell'(\theta^\top X_i, Y_i) X_i^j\bigg] \\ &= \mathbb{E}\bigg[\sum_{i=1}^n \varepsilon_i \ell'(0, Y_i) X_i^j + \sup_{\theta\in\Theta}\sum_{i=1}^n \varepsilon_i (\ell'(\theta^\top X_i, Y_i) - \ell'(0, Y_i)) X_i^j\bigg].\end{aligned}$$

Denote $\phi_i(t) = (\ell'(t, Y_i) - \ell'(0, Y_i))X_i^j$ and notice that $\mathbb{E}\big[\sum_{i=1}^n \varepsilon_i \ell'(0, Y_i) X_i^j\big] = 0$. Notice also that $\phi_i(0) = 0$ and ϕ_i is $\gamma|X_i^j|$-Lipschitz for all i. We use a variant of the contraction principle adapted to our case in which functions with different Lipschitz constants appear. We use Lemma 11.7 from [10] and adapt the proof of their Theorem 11.6 to make the following estimations:

$$\begin{aligned}&\mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^n \varepsilon_i \phi_i(\theta^\top X_i)\bigg] \\ &= \mathbb{E}\bigg[\mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^{n-1} \varepsilon_i \phi_i(\theta^\top X_i) + \varepsilon_n \phi_n(\theta^\top X_n)\Big|(\varepsilon_i)_{i=1}^{n-1}, (X_i, Y_i)_{i\in[\![n]\!]}\bigg]\bigg] \\ &\le \mathbb{E}\bigg[\mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^{n-1} \varepsilon_i \phi_i(\theta^\top X_i) + \varepsilon_n \gamma|X_n^j|\theta^\top X_n\Big|(\varepsilon_i)_{i=1}^{n-1}, (X_i, Y_i)_{i\in[\![n]\!]}\bigg]\bigg] \\ &= \mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^{n-1} \varepsilon_i \phi_i(\theta^\top X_i) + \varepsilon_n \gamma|X_n^j|\theta^\top X_n\bigg].\end{aligned}$$

By iterating the previous argument n times we find:

$$\mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^n \varepsilon_i \phi_i(\theta^\top X_i)\bigg] \le \mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^{n-1} \varepsilon_i \gamma|X_i^j|\theta^\top X_i\bigg].$$

Now recalling that the diameter of Θ is Δ, we use Lemma 8 below with $p = 1$ to bound the previous quantity as:

$$\begin{aligned}\mathbb{E}\bigg[\sup_{\theta\in\Theta}\sum_{i=1}^n \varepsilon_i \gamma|X_i^j|\theta^\top X_i\bigg] &= \gamma\mathbb{E}\bigg[\sup_{\theta\in\Theta}\Big\langle\theta, \sum_{i=1}^n \varepsilon_i X_i|X_i^j|\Big\rangle\bigg] \\ &\le \gamma\Delta\mathbb{E}\bigg[\mathbb{E}\bigg[\Big\|\sum_{i=1}^n \varepsilon_i X_i|X_i^j|\Big\|_1\Big|(X_i)_{i\in[\![n]\!]}\bigg]\bigg] \\ &\le \gamma\Delta C_\alpha\mathbb{E}\bigg[\sum_{i=1}^n \|X_i\|^{1+\alpha}|X_i^j|^{1+\alpha}\bigg]^{1/(1+\alpha)} \\ &\le \gamma\Delta C_\alpha\bigg(n\mathbb{E}\big[(X^j)^{2(1+\alpha)}\big]^{1/2}\sum_{k\in[\![d]\!]}\mathbb{E}\big[(X^k)^{2(1+\alpha)}\big]^{1/2}\bigg)^{1/(1+\alpha)},\end{aligned}$$

where we used a Cauchy-Schwarz inequality in the last step, which concludes the proof of Proposition 2. □

Lemma 8 (Khintchine inequality variant)**.** *Let $\alpha \in (0,1]$ and $(x_i)_{i\in[\![n]\!]}$ be real numbers with $n \in \mathbb{N}$ and $p > 0$ and $(\varepsilon_i)_{i\in[\![n]\!]}$ be i.i.d Rademacher random variables then we have the inequality:*

$$\mathbb{E}\Big[\Big|\sum_{i=1}^n \varepsilon_i x_i\Big|^p\Big]^{1/p} \leq B_{p,\alpha}\Big(\sum_{i=1}^n |x_i|^{1+\alpha}\Big)^{1/(1+\alpha)}$$

with the constant $B_{p,\alpha} := 2p\Big(\frac{1+\alpha}{\alpha}\Big)^{\alpha p/(1+\alpha)-1}\Gamma\Big(\frac{\alpha p}{1+\alpha}\Big)$. Moreover, for $p = 1$ the constant $B_{1,\alpha}$ is bounded for any $\alpha \geq 0$.

Proof. This proof is a generalization of Lemma 4.1 from [70] and uses similar methods. For all $\lambda > 0$ we have:

$$\begin{aligned}\mathbb{E}\exp\Big(\lambda\sum_i \varepsilon_i x_i\Big) &= \prod_i \mathbb{E}\exp(\lambda\varepsilon_i x_i) = \prod_i \cosh(\lambda x_i)\\ &\leq \prod_i \exp\Big(\frac{|\lambda x_i|^{1+\alpha}}{1+\alpha}\Big) = \exp\Big(\sum_i \frac{|\lambda x_i|^{1+\alpha}}{1+\alpha}\Big),\end{aligned}$$

where we used the inequality $\cosh(u) \leq \exp\Big(\frac{|u|^{1+\alpha}}{1+\alpha}\Big)$ valid for all $u \in \mathbb{R}$ which can be quickly proven. Since both functions are even, fix $u > 0$ and define $f_u(\alpha) = \exp\Big(\frac{|u|^{1+\alpha}}{1+\alpha}\Big) - \cosh(u)$, we can show that f_u is monotonous on $[0,1]$ separately for $u \in (0,\sqrt{e})$ and $(e,+\infty)$ and notice that $f_u(0)$ and $f_u(1)$ are both non-negative for all $u > 0$ thanks to the famous inequality $\cosh(u) \leq e^{u^2/2}$. Therefore, the inequality holds for $u \in (0,\sqrt{e})$ and $(e,+\infty)$. Finally, for $u \in (\sqrt{e},e)$, the function $f_u(\alpha)$ reaches a minimum at $f_u(1/\log(u)-1) = u^e - \cosh(u)$ and by taking logarithms we have $u^e \geq \cosh(u) \iff \log(1+e^{2u}) \leq u + \log(2) + e\log(u)$ but since the derivatives verify $\frac{2}{1+e^{-2u}} \leq 2 \leq 1 + e/u$ for $u \in (\sqrt{e},e)$ and $e^{e/2} \geq \cosh(\sqrt{e})$ the desired inequality follows by integration.

By homogeneity, we can focus on the case $\big(\sum_{i=1}^n |x_i|^{1+\alpha}\big)^{1/(1+\alpha)} = 1$, we compute:

$$\begin{aligned}\mathbb{E}\Big|\sum_i \varepsilon_i x_i\Big|^p &= \int_0^{+\infty} \mathbb{P}\Big(\Big|\sum_i \varepsilon_i x_i\Big|^p > t\Big)dt\\ &\leq 2\int_0^{+\infty} \exp\Big(\frac{\lambda^{1+\alpha}}{1+\alpha} - \lambda t^{1/p}\Big)dt\\ &= 2\int_0^{+\infty} \exp\Big(-\frac{\alpha}{1+\alpha}u^{(1+\alpha)/\alpha}\Big)du^p\\ &= 2p\Big(\frac{1+\alpha}{\alpha}\Big)^{\alpha p/(1+\alpha)-1}\Gamma\Big(\frac{\alpha p}{1+\alpha}\Big) = B_{p,\alpha}^p,\end{aligned}$$

where we used the previous inequality and chose $\lambda = (t^{1/p})^{1/\alpha}$ in the last step. This proves the main inequality. Finally, it is easy to see that $B_{1,\alpha}$ is bounded for high values of α while for $\alpha \sim 0$ it is consequence of the fact that $\Gamma(x) \sim 1/x$ near 0 and the limit $x^x \to 0$ when $x \to 0^+$. □

9.8 Proof of Lemma 3

As previously, Lemma 1 along with Assumptions 1 and 2 guarantee that the gradient coordinates have finite $(1+\alpha)$-moments. From here, Lemma 3 is a direct application of Lemma 9 stated and proved below. In the following lemma, for any sequence $(z_i)_{i=1}^N$ of real numbers, $(z_i^*)_{i=1}^N$ denotes a non-decreasing reordering of it.

Lemma 9. *Let $\widetilde{X}_1, \dots, \widetilde{X}_N, \widetilde{Y}_1, \dots, \widetilde{Y}_N$ denote an η-corrupted i.i.d sample with rate η from a random variable X with expectation $\mu = \mathbb{E}X$ and with finite $1+\gamma$ centered moment $\mathbb{E}|X-\mu|^{1+\gamma} = M < \infty$ for some $0 < \gamma \leq 1$. Denote $\widehat{\mu}$ the ϵ-trimmed mean estimator computed as $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \phi_{\alpha,\beta}(\widetilde{X}_i)$ with $\phi_{\alpha,\beta}(x) = \max(\alpha, \min(x, \beta))$ and the thresholds $\alpha = \widetilde{Y}^*_{\epsilon N}$ and $\beta = \widetilde{Y}^*_{(1-\epsilon)N}$. Let $1 > \delta \geq e^{-N}/4$, taking $\epsilon = 8\eta + 12\frac{\log(4/\delta)}{n}$, we have*

$$|\widehat{\mu} - \mu| \leq 7M^{\frac{1}{1+\gamma}}(\epsilon/2)^{\frac{\gamma}{1+\gamma}} \tag{40}$$

with probability at least $1-\delta$.

Proof. This proof goes along the lines of the proof of Theorem 1 from [79] with the main difference that only the $(1+\gamma)$-moment is used instead of the variance. Denote X the random variable whose expectation $\mu = \mathbb{E}X$ is to be estimated and $\overline{X} = X - \mu$. Let $X_1, \dots, X_N, Y_1, \dots, Y_N$ the original uncorrupted i.i.d. sample from X and let $\widetilde{X}_1, \dots, \widetilde{X}_N, \widetilde{Y}_1, \dots, \widetilde{Y}_N$ denote the corrupted sample with rate η. We define the following quantity which will intervene in the proof:

$$\overline{\mathcal{E}}(\epsilon, X) := \max\Big\{\mathbb{E}\big[\big|\overline{X} - Q_{\epsilon/2}(\overline{X})\big|\mathbf{1}_{\overline{X} \leq Q_{\epsilon/2}(\overline{X})}\big], \mathbb{E}\big[\big|\overline{X} - Q_{1-\epsilon/2}(\overline{X})\big|\mathbf{1}_{\overline{X} \geq Q_{1-\epsilon/2}(\overline{X})}\big]\Big\}. \tag{41}$$

Step 1. We first derive confidence bounds on the truncation thresholds. Define the random variable $U = \mathbf{1}_{\overline{X} \geq Q_{1-2\epsilon}(\overline{X})}$. Its standard deviation satisfies $\sigma_U \leq \mathbb{P}^{1/2}(\overline{X} \geq Q_{1-2\epsilon}(\overline{X})) = \sqrt{2\epsilon}$. By applying Bernstein's inequality we find with probability $\geq 1 - \exp(-\epsilon N/12)$ that:

$$\big|\big\{i \,:\, Y_i \geq \mu + Q_{1-2\epsilon}(\overline{X})\big\}\big| \geq 3\epsilon N/2,$$

a similar argument with $U = \mathbf{1}_{\overline{X} > Q_{1-\epsilon/2}(\overline{X})}$ yields with probability $\geq 1 - \exp(-\epsilon N/12)$ that:

$$\big|\big\{i \,:\, Y_i \leq \mu + Q_{1-\epsilon/2}(\overline{X})\big\}\big| \geq (1 - (3/4)\epsilon)N,$$

and similarly with probability $\geq 1 - \exp(-\epsilon N/12)$ we have:

$$\big|\big\{i \,:\, Y_i \leq \mu + Q_{2\epsilon}(\overline{X})\big|\big\}\big| \geq 3\epsilon N/2,$$

and with probability $\geq 1 - \exp(-\epsilon N/12)$:

$$\big|\big\{i \,:\, Y_i \geq \mu + Q_{\epsilon/2}(\overline{X})\big\}\big| \geq (1 - (3/4)\epsilon)N,$$

so that with probability $\geq 1 - 4\exp(-\epsilon N/12) \geq 1 - \delta/2$ the four previous inequalities hold simultaneously. We call this event E which only depends on the variables $Y_1, \dots, Y_N$. Since $\eta \leq \epsilon/8$, if $2\eta N$ samples are corrupted we still have:

$$\big|\big\{i \,:\, \widetilde{Y}_i \geq \mu + Q_{1-2\epsilon}(\overline{X})\big|\big\}\big| \geq ((3/2)\epsilon - 2\eta)N \geq \epsilon N$$

and

$$\big|\big\{i \,:\, \widetilde{Y}_i \leq \mu + Q_{1-\epsilon/2}(\overline{X})\big\}\big| \geq (1 - (3/4)\epsilon - 2\eta)N \geq (1-\epsilon)N$$

consequently, the two following bounds hold

$$Q_{1-2\epsilon}(\overline{X}) \leq \widetilde{Y}^*_{(1-\epsilon)N} - \mu \leq Q_{1-\epsilon/2}(\overline{X})$$

and similarly

$$Q_{\epsilon/2}(\overline{X}) \leq \widetilde{Y}^*_{\epsilon N} - \mu \leq Q_{2\epsilon}(\overline{X}).$$

This provides guarantees on the truncation levels used which are $\alpha = \widetilde{Y}^*_{\epsilon N}$ and $\beta = \widetilde{Y}^*_{(1-\epsilon)N}$.

Step 2. We first bound the deviation $\left|\frac{1}{N}\sum_{i=1}^N \phi_{\alpha,\beta}(X_i) - \mu\right|$ in the absence of corruption. W e write:

$$\frac{1}{N}\sum_{i=1}^N \phi_{\alpha,\beta}(X_i) \leq \frac{1}{N}\sum_{i=1}^N \phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X_i) = \mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)\big] + \frac{1}{N}\sum_{i=1}^N \Big(\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X_i) - \mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)\big]\Big). \tag{42}$$

The first term is dominated by:

$$\begin{aligned}
&\mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)\big] = \mathbb{E}\big[\phi_{Q_{2\epsilon}(X),Q_{1-\epsilon/2}(X)}(X)\big] \\
&\quad = \mathbb{E}\big[Q_{2\epsilon}(X)\mathbf{1}_{X\leq Q_{2\epsilon}(X)} + X\mathbf{1}_{Q_{2\epsilon}(X)<X<Q_{1-\epsilon/2}(X)} + Q_{1-\epsilon/2}(X)\mathbf{1}_{X\geq Q_{1-\epsilon/2}(X)}\big] \\
&\quad = \mu + \mathbb{E}\big[(Q_{2\epsilon}(X) - X)\mathbf{1}_{X\leq Q_{2\epsilon}(X)} + (Q_{1-\epsilon/2}(X) - X)\mathbf{1}_{X\geq Q_{1-\epsilon/2}(X)}\big] \\
&\quad \leq \mu + \mathbb{E}\big[(Q_{2\epsilon}(X) - X)\mathbf{1}_{X\leq Q_{2\epsilon}(X)}\big] = \mu + \mathbb{E}\big[(Q_{2\epsilon}(\overline{X}) - \overline{X})\mathbf{1}_{\overline{X}\leq Q_{2\epsilon}(\overline{X})}\big] \\
&\quad \leq \mu + \overline{\mathcal{E}}(4\epsilon, X),
\end{aligned}$$

and lower bounded by:

$$\begin{aligned}
&\mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)\big] \\
&\quad = \mu + \mathbb{E}\big[(Q_{2\epsilon}(X) - X)\mathbf{1}_{X\leq Q_{2\epsilon}(X)} + (Q_{1-\epsilon/2}(X) - X)\mathbf{1}_{X\geq Q_{1-\epsilon/2}(X)}\big] \\
&\quad \geq \mu + \mathbb{E}\big[(Q_{1-\epsilon/2}(X) - X)\mathbf{1}_{X\geq Q_{1-\epsilon/2}(X)}\big] = \mu + \mathbb{E}\big[(Q_{1-\epsilon/2}(\overline{X}) - \overline{X})\mathbf{1}_{\overline{X}\geq Q_{1-\epsilon/2}(\overline{X})}\big] \\
&\quad \geq \mu - \overline{\mathcal{E}}(\epsilon, X).
\end{aligned}$$

The sum in (42) above has terms upper bounded by $Q_{1-\epsilon/2}(\overline{X}) + \overline{\mathcal{E}}(\epsilon, X)$. We need to work with the knowledge that $\mathbb{E}|\overline{X}|^{1+\gamma} = M < \infty$ in order to bound their variance:

$$\begin{aligned}
&\mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X) - \mathbb{E}[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)]\big]^2 \\
&\quad \leq \mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X) - \mu\big]^2 = \mathbb{E}\big[\phi_{Q_{2\epsilon}(\overline{X}),Q_{1-\epsilon/2}(\overline{X})}(\overline{X})^2\big] \\
&\quad = \mathbb{E}\big[Q_{2\epsilon}(\overline{X})\mathbf{1}_{\overline{X}\leq Q_{2\epsilon}(\overline{X})} + \overline{X}\mathbf{1}_{Q_{2\epsilon}(\overline{X})<\overline{X}<Q_{1-\epsilon/2}(\overline{X})} + Q_{1-\epsilon/2}(\overline{X})\mathbf{1}_{\overline{X}\geq Q_{1-\epsilon/2}(\overline{X})}\big]^2 \\
&\quad = \mathbb{E}\big[Q_{2\epsilon}(\overline{X})^2\mathbf{1}_{\overline{X}\leq Q_{2\epsilon}(\overline{X})} + \overline{X}^2\mathbf{1}_{Q_{2\epsilon}(\overline{X})<\overline{X}<Q_{1-\epsilon/2}(\overline{X})} + Q_{1-\epsilon/2}(\overline{X})^2\mathbf{1}_{\overline{X}\geq Q_{1-\epsilon/2}(\overline{X})}\big].
\end{aligned}$$

To control the three terms in the previous expression we mimic the proof of Chebyshev's inequality to obtain that, when $Q_{2\epsilon}(\overline{X}) < 0$:

$$2\epsilon = \mathbb{P}\big(\overline{X} \leq Q_{2\epsilon}(\overline{X})\big) \leq \mathbb{P}\big(|\overline{X}|^{1+\gamma} \geq |Q_{2\epsilon}(\overline{X})|^{1+\gamma}\big) \leq \frac{M}{|Q_{2\epsilon}(\overline{X})|^{1+\gamma}}, \tag{43}$$

analogously, when $Q_{1-\epsilon/2}(\overline{X}) > 0$ we have:

$$\epsilon/2 = \mathbb{P}\big(\overline{X} \geq Q_{1-\epsilon/2}(\overline{X})\big) \leq \mathbb{P}\big(|\overline{X}|^{1+\gamma} \geq |Q_{1-\epsilon/2}(\overline{X})|^{1+\gamma}\big) \leq \frac{M}{|Q_{1-\epsilon/2}(\overline{X})|^{1+\gamma}}, \tag{44}$$

from (43), we deduce that

$$\mathbb{E}\big[Q_{2\epsilon}(\overline{X})^2\mathbf{1}_{\overline{X}\leq Q_{2\epsilon}(\overline{X})}\big] = 2\epsilon Q_{2\epsilon}(\overline{X})^2 \leq 2\epsilon\Big(\frac{M}{2\epsilon}\Big)^{\frac{2}{1+\gamma}} \leq 2\epsilon\Big(\frac{2M}{\epsilon}\Big)^{2/(1+\gamma)},$$

and from (44) we find

$$\mathbb{E}\big[Q_{1-\epsilon/2}(\overline{X})^2\mathbf{1}_{\overline{X}\geq Q_{1-\epsilon/2}(\overline{X})}\big] = Q_{1-\epsilon/2}(\overline{X})^2\epsilon/2 \leq 2\epsilon\Big(\frac{2M}{\epsilon}\Big)^{2/(1+\gamma)}.$$

In the pathological case where we have $Q_{2\epsilon}(\overline{X}) \geq 0$ we use that $Q_{2\epsilon}(\overline{X}) \leq Q_{1-\epsilon/2}(\overline{X})$ (for $\epsilon \leq 2/5$) we deduce $|Q_{2\epsilon}(\overline{X})| \leq |Q_{1-\epsilon/2}(\overline{X})|$ and hence we still have

$$\mathbb{E}\big[Q_{2\epsilon}(\overline{X})^2\mathbf{1}_{\overline{X}\leq Q_{2\epsilon}(\overline{X})}\big] \leq 2\epsilon Q_{1-\epsilon/2}(\overline{X})^2 \leq 2\epsilon\Big(\frac{2M}{\epsilon}\Big)^{2/(1+\gamma)}.$$

The case $Q_{1-\epsilon/2}(\overline{X}) \leq 0$ is similarly handled. Moreover, a simple calculation yields

$$\mathbb{E}\big[\overline{X}^2\mathbf{1}_{Q_{2\epsilon}(\overline{X})\leq\overline{X}\leq Q_{1-\epsilon/2}(\overline{X})}\big] \leq M\max\big\{|Q_{2\epsilon}(\overline{X})|, |Q_{1-\epsilon/2}(\overline{X})|\big\}^{1-\gamma} \leq 2\epsilon\Big(\frac{2M}{\epsilon}\Big)^{2/(1+\gamma)}.$$

All in all, we have shown the inequality:

$$\mathbb{E}\big[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X) - \mathbb{E}[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)]\big]^2 \leq 6\epsilon\Big(\frac{2M}{\epsilon}\Big)^{2/(1+\gamma)},$$

which we now use to apply Bernstein's inequality on the sum in (42) to find, conditionally on $Y_1, \ldots, Y_n$, with probability at least $1 - \delta/4$:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) \\
&\leq \mu + \overline{\mathcal{E}}(4\epsilon, X) + \sqrt{\frac{6\epsilon\log(4/\delta)}{N}}\Big(\frac{2M}{\epsilon}\Big)^{1/(1+\gamma)} + \frac{\log(4/\delta)}{3N}(Q_{1-\epsilon/2}(\overline{X}) + \overline{\mathcal{E}}(\epsilon, X)) \\
&\leq \mu + 2\overline{\mathcal{E}}(4\epsilon, X) + \sqrt{\frac{6\epsilon\log(4/\delta)}{N}}\Big(\frac{2M}{\epsilon}\Big)^{1/(1+\gamma)} + \frac{\log(4/\delta)}{3N}Q_{1-\epsilon/2}(\overline{X}) \\
&\leq \mu + 2\overline{\mathcal{E}}(4\epsilon, X) + (3/2)M^{1/(1+\gamma)}(\epsilon/2)^{\gamma/(1+\gamma)},
\end{aligned}$$

where we used (44), the fact that $\frac{\log(4/\delta)}{N} \leq \epsilon/12$ and the assumption that $\delta \geq e^{-N}/4$. Using the same argument on the lower tail, we obtain, on the event E, that with probability at least $1 - \delta/2$

$$\Big|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \mu\Big| \leq 2\overline{\mathcal{E}}(4\epsilon, X) + (3/2)M^{\frac{1}{1+\gamma}}(\epsilon/2)^{\gamma/(1+\gamma)}.$$

Step 3. Now we show that $\Big|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(\widetilde{X}_i)\Big|$ is of the same order as the previous bounds. There are at most $2\eta N$ indices such that $X_i \neq \widetilde{X}_i$ and for such differences we have the bound:

$$\big|\phi_{\alpha,\beta}(X_i) - \phi_{\alpha,\beta}(\widetilde{X}_i)\big| \leq |Q_{\epsilon/2}(\overline{X})| + |Q_{1-\epsilon/2}(\overline{X})|,$$

and since we have $\eta \leq \epsilon/8$ then

$$\begin{aligned}
\Big|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(\widetilde{X}_i)\Big| &\leq 2\eta\big(|Q_{\epsilon/2}(\overline{X})| + |Q_{1-\epsilon/2}(\overline{X})|\big) \\
&\leq \frac{\epsilon}{2}\max\big\{|Q_{\epsilon/2}(\overline{X})|, |Q_{1-\epsilon/2}(\overline{X})|\big\} \\
&\leq M^{1/(1+\gamma)}(\epsilon/2)^{\gamma/(1+\gamma)},
\end{aligned}$$

where the last step follows from (43) and (44). Finally, using similar arguments along with Hölder's inequality, we show that:

$$\begin{aligned}\mathbb{E}\big[|\overline{X} - Q_{\epsilon/2}(\overline{X})|\mathbf{1}_{\overline{X}\le Q_{\epsilon/2}(\overline{X})}\big] &\le \mathbb{E}\big[|\overline{X}|\mathbf{1}_{\overline{X}\le Q_{\epsilon/2}(\overline{X})}\big] + \mathbb{E}\big[|Q_{\epsilon/2}(\overline{X})|\mathbf{1}_{\overline{X}\le Q_{\epsilon/2}(\overline{X})}\big] \\ &\le M^{1/(1+\gamma)}(\epsilon/2)^{\gamma/(1+\gamma)} + |Q_{\epsilon/2}(\overline{X})|(\epsilon/2) \\ &\le 2M^{1/(1+\gamma)}(\epsilon/2)^{\gamma/(1+\gamma)},\end{aligned}$$

and a similar computation for $\mathbb{E}\big[|\overline{X} - Q_{1-\epsilon/2}(\overline{X})|\mathbf{1}_{\overline{X}\ge Q_{1-\epsilon/2}(\overline{X})}\big]$ leads to

$$\overline{\mathcal{E}}(4\epsilon, X) \le 2M^{1/(1+\gamma)}(2\epsilon)^{\gamma/(1+\gamma)}.$$

This completes the proof of Lemma 9. □

9.9 Proof of Proposition 3

Step 1. Notice that the TM estimator is also a monotonous non decreasing function of each of its entries when the others are fixed. This allows us to replicate Step 1 of the proof of Proposition 1. We define an ε-net N_ε on the set Θ, fix $\theta \in \Theta$ and let $\widetilde{\theta}$ be the closest point in N_ε. We obtain, for all $j \in [\![d]\!]$, the inequalities:

$$\begin{aligned}\big|\widehat{g}_j^{\mathtt{TM}}(\theta) - g_j(\theta)\big| &\le \big|\widehat{g}_j^{\mathtt{TM}}(\theta) - g_j(\widetilde{\theta})\big| + \big|g_j(\widetilde{\theta}) - g_j(\theta)\big| \\ &\le \big|\breve{g}_j^{\mathtt{TM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| + \varepsilon L_j, \end{aligned} \tag{45}$$

where $\breve{g}_j^{\mathtt{TM}}(\widetilde{\theta})$ is the TM estimator obtained for the entries $\ell'\big(\widetilde{\theta}^\top X_i, Y_i\big)X_i^j + \varepsilon\gamma\|X_i\|^2 = g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2$.

Step 2. We use the concentration property of the TM estimator to bound the previous quantity which is in terms of $\widetilde{\theta}$. The terms $\big(g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2\big)_{i\in[\![n]\!]}$ are independent and distributed according to $Z := \ell'\big(\widetilde{\theta}^\top X, Y\big)X^j + \gamma\varepsilon\|X\|^2$. Obviously we have $\mathbb{E}\ell'\big(\theta^\top X, Y\big)X^j = g_j(\theta)$. Furthermore, let $\overline{L} = \mathbb{E}\gamma\|X\|^2$, so that $\mathbb{E}\big[g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2\big] = g_j(\theta) + \varepsilon\overline{L}$. We will apply Lemma 9 for $\breve{g}_j^{\mathtt{TM}}(\widetilde{\theta})$. Before we do so, we need to compute the centered $(1+\alpha)$-moment of Z. Let $m_{j,\alpha}(\widetilde{\theta})$ and $m_{L,\alpha}$ be the centered $(1+\alpha)$-moments of $\ell'(\theta^\top X, Y)X^j$ and $\gamma\|X\|^2$ respectively, we have:

$$\mathbb{E}\big|Z - \mathbb{E}Z\big|^{1+\alpha} \le 2^\alpha\big(m_{j,\alpha}(\theta) + \varepsilon^{1+\alpha}m_{L,\alpha}\big).$$

Now applying Lemma 9 we find with probability no less than $1-\delta$

$$\big|\breve{g}_j^{\mathtt{TM}}(\widetilde{\theta}) - g_j(\widetilde{\theta}) - \varepsilon\overline{L}\big| \le 7\big(m_{j,\alpha}(\widetilde{\theta}) + \varepsilon^{1+\alpha}m_{L,\alpha}\big)^{1/(1+\alpha)}(2\epsilon)^{\alpha/(1+\alpha)},$$

with $\epsilon_\delta = 8\eta + 12\frac{\log(4/\delta)}{n}$. By combining with (45) and using a union bound argument, we deduce that with the same probability, we have for all $j \in [\![d]\!]$

$$\big|\breve{g}_j^{\mathtt{TM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| \le 7\big(m_{j,\alpha}(\widetilde{\theta}) + \varepsilon^{(1+\alpha)^2}m_{L,\alpha}\big)^{1/(1+\alpha)}(4\epsilon_{d\delta})^{\alpha/(1+\alpha)} + \varepsilon\overline{L}. \tag{46}$$

Step 3. We use the ε-net to obtain a uniform bound. We proceed similarly as in the proof of Proposition 1. For $\theta \in \Theta$ denote $\widetilde{\theta}(\theta) \in N_\varepsilon$ the closest point in N_ε satisfying in particular

$\|\widetilde{\theta}(\theta) - \theta\| \leq \varepsilon$, we write, following previous arguments

$$\begin{aligned}\sup_{\theta\in\Theta} \left|\widehat{g}_j^{\mathtt{TM}}(\theta) - g_j(\theta)\right| &\leq \sup_{\theta\in\Theta} \left|\widehat{g}_j^{\mathtt{TM}}(\theta) - g_j(\widetilde{\theta}(\theta))\right| + \left|g_j(\widetilde{\theta}(\theta)) - g_j(\theta)\right| \\ &\leq \sup_{\theta\in\Theta} \left|\widecheck{g}_j^{\mathtt{TM}}(\widetilde{\theta}(\theta)) - g_j(\widetilde{\theta}(\theta))\right| + \varepsilon L_j \\ &= \max_{\widetilde{\theta}\in N_\varepsilon} \left|\widecheck{g}_j^{\mathtt{TM}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\right| + \varepsilon L_j.\end{aligned}$$

Taking union bound over $\widetilde{\theta} \in N_\varepsilon$ for the inequality (46) and choosing $\varepsilon = n^{-\alpha/(1+\alpha)}$ concludes the proof of Proposition 3.

9.10 Proof of Corollary 1

We first write the result of Proposition 3 with a big O notation. This tells us that with probability at least $1-\delta$ for all $\theta \in \Theta$, for all $j \in [\![d]\!]$ we have :

$$\left|\epsilon_j^{\mathtt{TM}}(\delta)\right| \leq O\left(M_{j,\alpha}^{1/(1+\alpha)} \Big(\frac{\log(d/\delta) + d\log(n)}{n}\Big)^{\alpha/(1+\alpha)}\right)$$

It only remains to apply Theorem 1 with importance sampling. The main result corresponds to having the second term (the statistical error) dominate the bound given by Theorem 1. This happens as soon as the number of iterations T is high enough so that

$$\big(R(\theta^{(0)}) - R^\star\big)\Big(1 - \frac{\lambda}{\sum_{j\in[\![d]\!]} L_j}\Big)^T \leq \frac{\left\|\epsilon^{\mathtt{TM}}(\delta)\right\|_2^2}{2\lambda}.$$

From here, it is straightforward to check that the stated number of iterations suffices.

9.11 Proof of Lemma 4

Similarly to the proof of Lemma 2, the assumptions, this time taken with $\alpha = 1$, imply that the gradient has a second moment so that the existence of $\sigma_j^2 = \mathbb{V}(g_j(\theta))$ is guaranteed. We apply Lemma 1 from [50] with $\delta/2$ to obtain:

$$\frac{1}{2}|\widehat{g}_j^{\mathtt{CH}}(\theta) - g_j(\theta)| \leq \frac{C\sigma_j^2}{s} + \frac{s\log(4\delta^{-1})}{n}$$

with probability at least $1-\delta/2$, where C is a constant such that we have:

$$-\log(1 - u + Cu^2) \leq \psi(u) \leq \log(1 + u + Cu^2),$$

and one can easily check that our choice of ψ, the Gudermannian function, satisfies the previous inequality for $C = 1/2$. This, along with the choice of scale s according to (21) and our assumption on $\widehat{\sigma}_j$ yields the announced deviation bound by a simple union bound argument.

9.12 Proof of Proposition 4

In this proof, for a scale $s > 0$ and a set of real numbers $(x_i)_{i\in[\![n]\!]}$, we let $\bar{x} = \frac{1}{n}\sum_{i\in[\![n]\!]} x_i$ be their mean and define the function $\zeta_s\big((x_i)_{i\in[\![n]\!]}\big)$ as the unique x satisfying

$$\sum_{i\in[\![n]\!]} \psi\Big(\frac{x - \bar{x}}{s}\Big) = 0.$$

Since the function ψ is increasing the previous equation has a unique solution. Moreover, for fixed scale s, the function $\zeta_s\big((x_i)_{i\in[\![n]\!]}\big)$ is monotonous non decreasing w.r.t. each x_i when the others are fixed.

Step 1. We proceed similarly as in the proof of Proposition 1 except that we only use the monotonicity of the CH estimator with fixed scale. Let N_ε be an ε-net for Θ with $\varepsilon = 1/\sqrt{n}$. We have $|N_\varepsilon| \leq (3\Delta/2\varepsilon)^d$ with Δ the diameter of Θ. Fix a coordinate $j \in [\![d]\!]$, a point $\theta \in \Theta$ and let $\widetilde{\theta}$ be the closest point to it in the ε-net. We wish to bound the difference

$$\begin{aligned}\big|\widehat{g}_j^{\texttt{CH}}(\theta) - g_j(\theta)\big| &\leq \big|\widehat{g}_j^{\texttt{CH}}(\theta) - g_j(\widetilde{\theta})\big| + \big|g_j(\widetilde{\theta}) - g_j(\theta)\big| \\ &\leq \big|\widehat{g}_j^{\texttt{CH}}(\theta) - g_j(\widetilde{\theta})\big| + \varepsilon L_j,\end{aligned}$$

where we have the CH estimator $\widehat{g}_j^{\texttt{CH}}(\theta) = \zeta_{s(\theta)}\big((g_j^i(\theta))_{i\in[\![n]\!]}\big)$ with scale $s(\theta)$ computed according to (21) and (22). Assume, without loss of generality that $\widehat{g}_j^{\texttt{CH}}(\theta) - g_j(\widetilde{\theta}) \geq 0$. Using the non-decreasing property of the CH estimator at a fixed scale, we find that

$$\begin{aligned}\big|\widehat{g}_j^{\texttt{CH}}(\theta) - g_j(\widetilde{\theta})\big| &= \big|\zeta_{s(\theta)}\big((g_j^i(\theta))_{i\in[\![n]\!]}\big) - g_j(\widetilde{\theta})\big| \\ &\leq \big|\zeta_{s(\theta)}\big((g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)_{i\in[\![n]\!]}\big) - g_j(\widetilde{\theta})\big|.\end{aligned}$$

Indeed, one has

$$\begin{aligned}g_j^i(\theta) &= g_j^i(\widetilde{\theta}) + \big(g_j^i(\theta) - g_j^i(\widetilde{\theta})\big) \\ &\leq g_j^i(\widetilde{\theta}) + \gamma\|\widetilde{\theta} - \theta\| \cdot \|X_i\| \cdot |X_i^j| \\ &\leq g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2.\end{aligned}$$

We introduce the notation $\breve{g}_j^{\texttt{CH}}(\widetilde{\theta}) := \zeta_{s(\theta)}\big((g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)_{i\in[\![n]\!]}\big)$ so that:

$$\big|\widehat{g}_j^{\texttt{CH}}(\theta) - g_j(\widetilde{\theta})\big| \leq \big|\breve{g}_j^{\texttt{CH}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big|.$$

Step 2. We now use the concentration property of CH to bound the previous quantity which is in terms of $\widetilde{\theta}$. We apply Lemma 1 from [50] with $\delta/2$ and scale $s(\theta)$ to the samples $(g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)_{i\in[\![n]\!]}$ which are independent and distributed according to the random variable $\ell'\big(\widetilde{\theta}^\top X, Y\big)X^j + \varepsilon\gamma\|X\|^2$ with expectation $g_j(\widetilde{\theta}) + \varepsilon\overline{L}$. Using our assumptions on $\sigma_L, \sigma_j(\theta), \sigma_j(\widetilde{\theta}), \widehat{\sigma}_j(\theta)$ and the definition of the scale $s(\theta)$ according to (21) we find:

$$\begin{aligned}\frac{1}{2}\big|\breve{g}_j^{\texttt{CH}}(\widetilde{\theta}) - g_j(\widetilde{\theta}) - \varepsilon\overline{L}\big| &= \frac{1}{2}\big|\zeta_{s(\theta)}\big((g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)_{i\in[\![n]\!]}\big) - g_j(\widetilde{\theta}) - \varepsilon\overline{L}\big| \\ &\leq \frac{C\mathbb{V}(g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)}{s(\theta)} + \frac{s(\theta)\log(4/\delta)}{n} \\ &\leq \frac{CC'\mathbb{V}(g_j^i(\widetilde{\theta}) + \varepsilon\gamma\|X_i\|^2)}{\sigma_j(\theta)}\sqrt{\frac{2\log(4/\delta)}{n}} + C'\sigma_j(\theta)\sqrt{\frac{2\log(4/\delta)}{n}} \\ &\leq \frac{CC'2(\sigma_j^2(\widetilde{\theta}) + \varepsilon^2\sigma_L^2)}{\sigma_j(\theta)}\sqrt{\frac{2\log(4/\delta)}{n}} + C'\sigma_j(\theta)\sqrt{\frac{2\log(4/\delta)}{n}} \\ &\leq CC'2\big(\sqrt{2}\sigma_j(\widetilde{\theta}) + \varepsilon\sigma_L\big)\sqrt{\frac{2\log(4/\delta)}{n}} + 2C'\sigma_j(\widetilde{\theta})\sqrt{\frac{\log(4/\delta)}{n}} \\ &\leq 4C'\sigma_j(\widetilde{\theta})\sqrt{\frac{\log(4/\delta)}{n}} + 2C'\varepsilon\sigma_L\sqrt{\frac{\log(4/\delta)}{n}} \\ &\leq 2C'(2\sigma_j(\widetilde{\theta}) + \varepsilon\sigma_L)\sqrt{\frac{\log(4/\delta)}{n}}.\end{aligned}$$

A simple union bound yields that for all $j \in [\![d]\!]$

$$\big|\breve{g}_j^{\texttt{CH}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| \leq 4C'(2\sigma_j(\widetilde{\theta}) + \varepsilon\sigma_L)\sqrt{\frac{\log(4d/\delta)}{n}} + \varepsilon\overline{L}. \tag{47}$$

Step 3. We use the ε-net to obtain a uniform bound. We proceed similarly to the proof of Proposition 1. For $\theta \in \Theta$ denote $\widetilde{\theta}(\theta) \in N_\varepsilon$ the closest point in N_ε satisfying in particular $\|\widetilde{\theta}(\theta) - \theta\| \leq \varepsilon$, we write, following previous arguments

$$\begin{aligned}\sup_{\theta\in\Theta} \big|\widehat{g}_j^{\mathtt{CH}}(\theta) - g_j(\theta)\big| &\leq \sup_{\theta\in\Theta} \big|\widehat{g}_j^{\mathtt{CH}}(\theta) - g_j(\widetilde{\theta}(\theta))\big| + \big|g_j(\widetilde{\theta}(\theta)) - g_j(\theta)\big| \\ &\leq \sup_{\theta\in\Theta} \big|\breve{g}_j^{\mathtt{CH}}(\widetilde{\theta}(\theta)) - g_j(\widetilde{\theta}(\theta))\big| + \varepsilon L_j \\ &= \max_{\widetilde{\theta}\in N_\varepsilon} \big|\breve{g}_j^{\mathtt{CH}}(\widetilde{\theta}) - g_j(\widetilde{\theta})\big| + \varepsilon L_j.\end{aligned}$$

Taking union bound over $\widetilde{\theta} \in N_\varepsilon$ for the inequality (47) and using the choice $\varepsilon = 1/\sqrt{n}$ concludes the proof of Proposition 4.

9.13 Proof of Corollary 2

Under the assumptions made, the constants $(L_j)_{j\in[\![d]\!]}$ are estimated using the `MOM` estimator and we obtain the bounds $(\overline{L}_j)_{j\in[\![d]\!]}$ which hold with probability at least $1 - \delta/2$ by a union bound argument. The rest of the proof is the same as that of Theorem 1 using a failure probability $\delta/2$ instead of δ and replacing the constants $(L_j)_{j\in[\![d]\!]}$ by their upperbounds accordingly. The result then follows after a simple union bound argument.

9.14 Proof of Lemma 5

Let $B_1, \dots, B_K$ be the blocks used for the estimation so that $B_1 \cup \cdots \cup B_K = [\![n]\!]$ and $B_{k_1} \cap B_{k_2} = \emptyset$ for $k_1 \neq k_2$. Let $\mathcal{K}$ denote the uncorrupted block indices $\mathcal{K} = \{k \in [\![K]\!] \text{ such that } B_k \cap \mathcal{O} = \emptyset\}$ and assume $|\mathcal{O}| \leq (1-\varepsilon)K/2$. For $k \in [\![K]\!]$ let $\widehat{\sigma}_k^2 = \frac{K}{n}\sum_{i\in B_k} X_i^2$ be the block means computed by MOM. Denote $N = n/K$, by using (a slight generalization of) Lemma 7 and the $L^{(1+\alpha)}$-L^1 condition satisfied by X^2 with a known constant C, we obtain that with probability at least $1-\delta$ we have

$$|\widehat{\sigma}_k^2 - \sigma^2| \leq \Big(\frac{3\mathbb{E}|X^2-\sigma^2|^{1+\alpha}}{\delta N^\alpha}\Big)^{\frac{1}{1+\alpha}} \leq \Big(\frac{3}{\delta N^\alpha}\Big)^{\frac{1}{1+\alpha}} C\mathbb{E}|X^2-\sigma^2| \leq \Big(\frac{3}{\delta N^\alpha}\Big)^{\frac{1}{1+\alpha}} C\sigma^2,$$

which implies the inequality

$$\sigma^2 \leq \Big(1 - C\Big(\frac{3}{\delta N^\alpha}\Big)^{\frac{1}{1+\alpha}}\Big)^{-1} \widehat{\sigma}_k^2.$$

Define the Bernoulli random variables $U_k = \mathbf{1}\Big\{\sigma^2 > \Big(1 - C\big(\frac{3}{\delta N^\alpha}\big)^{\frac{1}{1+\alpha}}\Big)^{-1}\widehat{\sigma}_k^2\Big\}$ for $k \in [\![K]\!]$ which have success probability $\leq \delta$. Denote $S = \sum_k U_k$, we can bound the failure probability of the estimator as follows:

$$\begin{aligned}\mathbb{P}\Big(\Big(1 - C\Big(\frac{3}{\delta N^\alpha}\Big)^{\frac{1}{1+\alpha}}\Big)^{-1}\widehat{\sigma}^2 < \sigma^2\Big) &\leq \mathbb{P}\big[S > K/2 - |\mathcal{O}|\big] \\ &= \mathbb{P}\big[S - \mathbb{E}S > K/2 - |\mathcal{O}| - \delta|\mathcal{K}|\big] \\ &\leq \mathbb{P}\big[S - \mathbb{E}S > K(\varepsilon - 2\delta)/2\big] \\ &\leq \exp\big(-K(\varepsilon-2\delta)^2/2\big),\end{aligned}$$

where we used the fact that $|\mathcal{O}| \leq (1-\varepsilon)K/2$ and $|\mathcal{K}| \leq K$ for the second inequality and Hoeffding's inequality for the last. The proof is finished by taking $\varepsilon = 5/6$ and $\delta = 1/4$.

9.15 Proof of Lemma 6

Lemma 6 is a direct consequence of the following result.

Lemma 10. *Let $X_1, \dots, X_n$ an i.i.d sample of a random variable X with expectation $\mathbb{E}X = \mu$ and $(1+\alpha)$-moment $\mathbb{E}|X-\mu|^{1+\alpha} = m_\alpha < \infty$. Assume that the variable X satisfies the $L^{(1+\alpha)^2}$-$L^{(1+\alpha)}$ condition with constant $C > 1$. Let $\widehat{\mu}$ be the median-of-means estimate of μ with K blocks and $\widehat{m}_\alpha$ a similarly obtained estimate of m_α from the samples $(|X_i - \widehat{\mu}|^{1+\alpha})_{i\in[\![n]\!]}$. Then, with probability at least $1 - 2\exp(-K/18)$ we have*

$$\widehat{m}_\alpha \geq (1-\kappa)m_\alpha,$$

with $\kappa = \epsilon + 24(1+\alpha)\Big(\frac{1+\epsilon}{n/K}\Big)^{\frac{\alpha}{1+\alpha}}$ and $\epsilon = \Big(\frac{3\times 2^{2+\alpha}(1+C^{(1+\alpha)^2})}{(n/K)^\alpha}\Big)^{\frac{1}{1+\alpha}}$.

Proof. Let $\widehat{\mu}$ be the MOM estimate of μ with K blocks, using Lemma 2, we have with probability at least $1-\exp(-K/18)$,

$$|\mu - \widehat{\mu}| > (24 m_\alpha)^{\frac{1}{1+\alpha}}\Big(\frac{K}{n}\Big)^{\frac{\alpha}{1+\alpha}}. \tag{48}$$

Let $\widehat{m}_\alpha$ be the MOM estimate of m_α obtained from the samples $\big(|X_i - \widehat{\mu}|^{1+\alpha}\big)_{i\in[\![n]\!]}$. Denote $B_1, \dots, B_K$ the blocks we use, we have:

$$\widehat{m}_\alpha = \text{median}\left(\frac{K}{n}\sum_{i\in B_j}|X_i - \widehat{\mu}|^{1+\alpha}\right)_{j\in[\![K]\!]}$$

for any $i \in [\![n]\!]$. Let $N = n/K$, using the convexity of the function $f(x) = |x|^{1+\alpha}$ we find that:

$$\begin{aligned}
\frac{1}{N}\sum_{i\in B_j}\big|X_i - \widehat{\mu}\big|^{1+\alpha} &= \frac{1}{N}\sum_{i\in B_j}\big|(X_i-\mu) + (\mu - \widehat{\mu})\big|^{1+\alpha} \\
&\geq \frac{1}{N}\sum_{i\in B_j}|X_i-\mu|^{1+\alpha} + \frac{1}{N}(1+\alpha)\sum_{i\in B_j}|X_i-\mu|^\alpha \text{sign}(X_i-\mu)(\mu-\widehat{\mu}) \\
&\geq \frac{1}{N}\sum_{i\in B_j}|X_i-\mu|^{1+\alpha} - (1+\alpha)|\mu-\widehat{\mu}|\Big[\frac{1}{N}\sum_{i\in B_j}|X_i-\mu|^\alpha\Big] \\
&\geq \frac{1}{N}\sum_{i\in B_j}|X_i-\mu|^{1+\alpha} - (1+\alpha)|\mu-\widehat{\mu}|\Big[\frac{1}{N}\sum_{i\in B_j}|X_i-\mu|^{1+\alpha}\Big]^{\frac{\alpha}{1+\alpha}},
\end{aligned} \tag{49}$$

where the last step uses Jensen's inequality. Using Lemma 7 we have, for $\delta > 0$, the concentration bound

$$\mathbb{P}\left(\Big|\frac{1}{N}\sum_{i\in B_j}\big|X_i-\mu\big|^{1+\alpha} - m_\alpha\Big| > \Big(\frac{3\mathbb{E}\big||X-\mu|^{1+\alpha} - m_\alpha\big|^{1+\alpha}}{\delta N^\alpha}\Big)^{\frac{1}{1+\alpha}}\right) \leq \delta$$

which, using that X satisfies the $L^{(1+\alpha)^2}$-$L^{(1+\alpha)}$ condition, translates to

$$\begin{aligned}
\mathbb{P}\left(\Big|\frac{1}{N}\sum_{i\in B_j}\big|X_i-\mu\big|^{1+\alpha} - m_\alpha\Big| > \epsilon\right) &\leq \frac{3\mathbb{E}\big||X-\mu|^{1+\alpha} - m_\alpha\big|^{1+\alpha}}{\epsilon^{1+\alpha}N^\alpha} \\
&\leq \frac{3\times 2^\alpha\big(\mathbb{E}|X-\mu|^{(1+\alpha)^2} + m_\alpha^{1+\alpha}\big)}{\epsilon^{1+\alpha}N^\alpha} \\
&\leq \frac{3\times 2^\alpha m_\alpha^{1+\alpha}\big(1+C^{(1+\alpha)^2}\big)}{\epsilon^{1+\alpha}N^\alpha}.
\end{aligned}$$

Replacing ϵ with ϵm_α we find

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i\in B_j}|X_i-\mu|^{1+\alpha}-m_\alpha\Big|>\epsilon m_\alpha\Big)\leq\frac{3\times 2^\alpha\big(1+C^{(1+\alpha)^2}\big)}{N^\alpha\epsilon^{1+\alpha}}.$$

Now conditioning on the event (48) and using the previous bound with $\epsilon=\Big(\frac{3\times 2^\alpha\big(1+C^{(1+\alpha)^2}\big)}{N^\alpha\delta}\Big)^{\frac{1}{1+\alpha}}$ in (49), we obtain that

$$\begin{aligned}&\mathbb{P}\Big(\frac{1}{N}\sum_{i\in B_j}\big|X_i-\widehat{\mu}\big|^{1+\alpha}\leq(1-\epsilon)m_\alpha-(1+\alpha)\Big(\frac{24m_\alpha}{N^\alpha}\Big)^{\frac{1}{1+\alpha}}((1+\epsilon)m_\alpha)^{\frac{\alpha}{1+\alpha}}\Big)\leq\delta\\&\implies\mathbb{P}\Big(\frac{1}{N}\sum_{i\in B_j}\big|X_i-\widehat{\mu}\big|^{1+\alpha}\leq\underbrace{\Big(1-\epsilon-24(1+\alpha)\Big(\frac{1+\epsilon}{N}\Big)^{\frac{\alpha}{1+\alpha}}\Big)}_{=:(1-\kappa)}m_\alpha\Big)\leq\delta.\end{aligned}$$

Now define U_j as the indicator variable of the event in the last probability. We have just seen it has success rate less than δ. We can use the MOM trick, assuming the number of outliers satisfies $|\mathcal{O}|\leq K(1-\varepsilon)/2$ for $\varepsilon\in(0,1)$, we have for $S=\sum_j U_j$

$$\begin{aligned}\mathbb{P}(\widehat{m}_\alpha\leq(1-\kappa)m_\alpha)&\leq\mathbb{P}(S>K/2-|\mathcal{O}|)\\&=\mathbb{P}\big[S-\mathbb{E}S>K/2-|\mathcal{O}|-\delta|\mathcal{K}|\big]\\&\leq\mathbb{P}\big[S-\mathbb{E}S>K(\varepsilon-2\delta)/2\big]\\&\leq\exp\big(-K(\varepsilon-2\delta)^2/2\big).\end{aligned}$$

Taking $\varepsilon=5/6$ and $\delta=1/4$ yields that the previous probability is $\leq\exp(-K/18)$. Finally, recall that we conditioned on the event where the deviation $|\mu-\widehat{\mu}|$ is bounded as previously stated and that this event holds with $\geq 1-\exp(-K/18)$. Taking this conditioning into account and using a union bound argument leads to the fact that the bound

$$\widehat{m}_\alpha\geq(1-\kappa)m_\alpha$$

holds with probability at least $1-2\exp(-K/18)$. □

9.16 Proof of Theorem 3

This proof is inspired from Theorem 5 in [88] and Theorem 1 in [95] while keeping track of the degradations caused by the errors on the gradient coordinates.

We condition on the event (30) and denote $\epsilon_j=\epsilon_j(\delta)$ and $\epsilon_{Euc}=\|\epsilon(\delta)\|_2$. We define for all $\theta\in\Theta$

$$\begin{aligned}u_j(\theta)&=\operatorname*{argmin}_{\vartheta\in\Theta_j}\widehat{g}_j(\theta)(\vartheta-\theta_j)+\frac{L_j}{2}(\vartheta-\theta_j)^2+\epsilon_j|\vartheta-\theta_j|\\&=\mathrm{proj}_{\Theta_j}\big(\theta_j-\beta_j\tau_{\epsilon_j}\big(\widehat{g}_j(\theta)\big)\big)\end{aligned}$$

and denote $\theta^{(t)}$ the optimization iterates for $t=0,\dots,T$ and j_t the random coordinate sampled at step t and let $\widehat{g}_t=\widehat{g}_{j_t}(\theta^{(t)})$ for brevity. We have that $u_{j_t}(\theta^{(t)})$ satisfies the following optimality condition

$$\forall\vartheta\in\Theta_{j_t}\quad\big(\widehat{g}_t+L_{j_t}\big(u_{j_t}(\theta^{(t)})-\theta^{(t)}_{j_t}\big)+\epsilon_{j_t}\rho_t\big)\big(\vartheta-u_{j_t}(\theta^{(t)})\big)\geq 0,$$

where $\rho_t = \operatorname{sign}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)$. Using this condition for $\vartheta = \theta^{(t)}_{j_t}$ and the coordinate-wise Lipschitz smoothness property of R we find

$$\begin{aligned}
R(\theta^{(t+1)}) &\leq R(\theta^{(t)}) + g_{j_t}(\theta^{(t)})\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big) + \frac{L_{j_t}}{2}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2 \\
&\leq R(\theta^{(t)}) + (\widehat{g}_t + \epsilon_{j_t}\rho_t)\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big) + \frac{L_{j_t}}{2}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2 \qquad (50) \\
&\leq R(\theta^{(t)}) - \frac{L_{j_t}}{2}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2. \qquad (51)
\end{aligned}$$

Defining the potential $\Phi(\theta) = \sum_{j=1}^d L_j(\theta_j - \theta^\star_j)^2$, we have:

$$\begin{aligned}
\Phi(\theta^{(t+1)}) &= \Phi(\theta^{(t)}) + 2L_{j_t}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)\big(\theta^{(t)}_{j_t} - \theta^\star_{j_t}\big) + L_{j_t}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2 \\
&= \Phi(\theta^{(t)}) + 2L_{j_t}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)\big(u_{j_t}(\theta^{(t)}) - \theta^\star_{j_t}\big) - L_{j_t}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2 \\
&\leq \Phi(\theta^{(t)}) - 2(\widehat{g}_t + \epsilon_{j_t}\rho_t)\big(u_{j_t}(\theta^{(t)}) - \theta^\star_{j_t}\big) - L_{j_t}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2 \\
&= \Phi(\theta^{(t)}) + 2(\widehat{g}_t + \epsilon_{j_t}\rho_t)\big(\theta^\star_{j_t} - \theta^{(t)}_{j_t}\big) - 2\Big((\widehat{g}_t + \epsilon_{j_t}\rho_t)\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big) \\
&\quad + \frac{L_{j_t}}{2}\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2\Big) \\
&\leq \Phi(\theta^{(t)}) + 2(\widehat{g}_t + \epsilon_{j_t}\rho_t)\big(\theta^\star_{j_t} - \theta^{(t)}_{j_t}\big) + 2\big(R(\theta^{(t)}) - R(\theta^{(t+1)})\big) \\
&\leq \Phi(\theta^{(t)}) + 2g_{j_t}(\theta^{(t)})\big(\theta^\star_{j_t} - \theta^{(t)}_{j_t}\big) + 2\big(R(\theta^{(t)}) - R(\theta^{(t+1)})\big) + 4\epsilon_{j_t}\big|\theta^\star_{j_t} - \theta^{(t)}_{j_t}\big|,
\end{aligned}$$

where the first inequality uses the optimality condition with $\vartheta = \theta^\star_{j_t}$ and the second one uses (50). Now, defining $\Psi(\theta) = \frac{1}{2}\Phi(\theta) + R(\theta)$, taking the expectation w.r.t. j_t and using the convexity of R and a Cauchy-Schwarz inequality, we find

$$\mathbb{E}\big[\Psi(\theta^{(t)}) - \Psi(\theta^{(t+1)})\big] \geq \frac{1}{d}\big(R(\theta^{(t)}) - R(\theta^\star) - 2\epsilon_{Euc}\big\|\theta^{(t)} - \theta^\star\big\|_2\big).$$

Recall that according to (51), we have $R(\theta^{(t+1)}) \leq R(\theta^{(t)})$, summing over $t = 0, \dots, T$ we find:

$$\begin{aligned}
\mathbb{E}\Big[\frac{T+1}{d}\big(R(\theta^{(T)}) - R(\theta^\star)\big)\Big] &\leq \mathbb{E}\Big[\frac{1}{d}\sum_{t=0}^T \big(R(\theta^{(t)}) - R(\theta^\star)\big)\Big] \\
&\leq \sum_{t=0}^T \Big(\mathbb{E}\big[\Psi(\theta^{(t)}) - \Psi(\theta^{(t+1)})\big] + \frac{2\epsilon_{Euc}}{d}\mathbb{E}\big[\big\|\theta^{(t)} - \theta^\star\big\|_2\big]\Big) \\
&= \mathbb{E}\big[\Psi(\theta^{(0)}) - \Psi(\theta^{(t+1)})\big] + \frac{2\epsilon_{Euc}}{d}\sum_{t=0}^T \mathbb{E}\big[\big\|\theta^{(t)} - \theta^\star\big\|_2\big] \\
&\leq \Psi(\theta^{(0)}) + \frac{2\epsilon_{Euc}}{d}\sum_{t=0}^T \mathbb{E}\big[\big\|\theta^{(t)} - \theta^\star\big\|_2\big],
\end{aligned}$$

which yields the result after multiplying by $\frac{d}{T+1}$. To finish, we show that conditionally on any choice of j_t we have $\|\theta^{(t+1)} - \theta^\star\|_2 \leq \|\theta^{(t)} - \theta^\star\|_2$. Indeed a straightforward computation yields

$$\big\|\theta^{(t+1)} - \theta^\star\big\|_2^2 = \big\|\theta^{(t)} - \theta^\star\big\|_2^2 + \big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)^2 + 2\big(u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t}\big)\big(\theta^{(t)}_{j_t} - \theta^\star_{j_t}\big).$$

We need to show that $\delta_t^2 \leq -2\delta_t\big(\theta^{(t)}_{j_t} - \theta^\star_{j_t}\big)$ with $\delta_t = (u_{j_t}(\theta^{(t)}) - \theta^{(t)}_{j_t})$. Notice that δ_t always has the opposite sign of $g_{j_t}(\theta^{(t)})$ (thanks to the thresholding) so by convexity of R along the coordinate

j_t we have $\delta_t\big(\theta^{(t)}_{j_t} - \theta^\star_{j_t}\big) \le 0$ and so it is down to showing $|\delta_t| \le 2\big|\theta^{(t)}_{j_t} - \theta^\star_{j_t}\big|$ which can be seen from

$$|\delta_t| \le \frac{\big|g_{j_t}(\theta^{(t)})\big|}{L_{j_t}} = \frac{\big|g_{j_t}(\theta^{(t)}) - g_{j_t}(\theta^\star)\big|}{L_{j_t}} \le \big|\theta^{(t)}_{j_t} - \theta^\star_{j_t}\big|,$$

which concludes the proof of Theorem 3.

References

[1] Anish Acharya, Abolfazl Hashemi, Prateek Jain, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Robust training in high dimensions via block coordinate geometric median descent. In *International Conference on Artificial Intelligence and Statistics*, pages 11145–11168. PMLR, 2022.

[2] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

[3] Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[4] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. Algorithmic Learning Theory.

[5] Rafael Ballester-Ripoll, Enrique G. Paredes, and Renato Pajarola. Sobol tensor trains for global sensitivity analysis. *Reliability Engineering & System Safety*, 183:311–322, 2019.

[6] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[7] Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[8] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30:2110–2119, 2017.

[9] Mathieu Blondel, Kazuhiro Seki, and Kuniaki Uehara. Block coordinate descent algorithms for large-scale sparse multiclass classification. *Machine learning*, 93(1):31–52, 2013.

[10] Stéphane Boucheron, Gábor Lugosi, Pascal Massart, and Michel Ledoux. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

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

[12] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[14] Luis M Candanedo and Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016.

[15] Luis M. Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

[16] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[17] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

[18] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[19] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

[20] Peng Chen, Xinghu Jin, Xiang Li, and Lihu Xu. A generalized Catoni's M-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15(2):5523–5544, 2021.

[21] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

[22] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I. Jordan, Nicolas Flammarion, and Peter L. Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

[23] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR, 2019.

[24] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT press, 2009.

[25] Jules Depersin and Guillaume Lecué. Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.

[26] Luc Devroye and Laszlo Györfi. *Nonparametric Density Estimation: The L1 View*. Wiley Interscience Series in Discrete Mathematics. Wiley, 1985.

[27] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[28] 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, 2019a.

[29] 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, 2019b.

[30] Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020.

[31] 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.

[32] Will Jim Dixon. Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4):488–506, 1950.

[33] David L. Donoho and Richard C. Liu. The "Automatic" Robustness of Minimum Distance Functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

[34] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[35] Francis Y. Edgeworth. On observations relating to several quantities. *Hermathena*, 6(13):279–285, 1887.

[36] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014.

[37] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

[38] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

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

[40] Sara A. Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[41] Alexander Genkin, David D. Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[42] Chinot Geoffrey, Lecué Guillaume, and Lerasle Matthieu. Robust high dimensional learning for Lipschitz and convex losses. *Journal of Machine Learning Research*, 21, 2020.

[43] Frank E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[44] A. Gupta and S. Kohli. An MCDM approach towards handling outliers in web data: a case study using OWA operators. *Artificial Intelligence Review*, 46(1):59–82, Jun 2016.

[45] Frank R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

[46] Frank R Hampel, Elvezio M Ronchetti, Peter Rousseeuw, and Werner A Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley-Interscience; New York, 1986.

[47] Douglas M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[48] Charles A. R. Hoare. Algorithm 65: Find. *Commun. ACM*, 4(7):321–322, jul 1961.

[49] Matthew Holland. Robustness and scalability under heavy tails, without strong convexity. In *International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2021.

[50] Matthew Holland and Kazushi Ikeda. Better generalization with less data using robust gradient descent. In *International Conference on Machine Learning*, pages 2761–2770. PMLR, 2019.

[51] Matthew J. Holland. Robust descent using smoothed multiplicative noise. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR, 2019.

[52] Matthew J. Holland and Kazushi Ikeda. Efficient learning with robust gradient descent. *Machine Learning*, 108(8):1523–1560, 2019.

[53] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020.

[54] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[55] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[56] Peter J. Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

[57] Peter J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[58] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[59] Anatoli Juditsky, Andrei Kulunchakov, and Hlib Tsyntseus. Sparse recovery by reduced variance stochastic approximation. *Information and Inference: A Journal of the IMA*, 12(2):851–896, 11 2022.

[60] Adam Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

[61] Adam R. Klivans, Philip M. Long, and Rocco A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

[62] Donald E Knuth. Seminumerical algorithms (the art of computer programming 2). *Reading, MA,: AddisonWesley*, pages 124–125, 1969.

[63] Murat Koklu and Ilker Ali Ozkan. Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174:105507, 2020.

[64] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[65] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[66] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[67] Guillaume Lecué, Matthieu Lerasle, et al. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 48(2):906–931, 2020.

[68] Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu. Robust classification via MOM minimization. *Machine Learning*, 109(8):1635–1665, 2020.

[69] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: upper and minimax bounds (2013). *Topics in Learning Theory-Société Mathématique de France,(S. Boucheron and N. Vayatis Eds.)*, 2013.

[70] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Berlin, Heidelberg, 1991.

[71] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-Gaussian rates. In *Conference on Learning Theory*, pages 2598–2612. PMLR, 2020.

[72] Jerry Li. Robust sparse estimation tasks in high dimensions. *arXiv preprint arXiv:1702.05860*, 2017.

[73] Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *The Journal of Machine Learning Research*, 18(1):6741–6764, 2017.

[74] Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust estimation of sparse models via trimmed hard thresholding. *arXiv preprint arXiv:1901.08237*, 2019.

[75] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 411–421. PMLR, 2020.

[76] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

[77] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

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

[79] Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

[80] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

[81] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample-Variance Penalization. In *COLT*, 2009.

[82] Stanislav Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[83] Stanislav Minsker et al. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Annals of Statistics*, 46(6A):2871–2903, 2018.

[84] Ivan Mizera et al. On depth and deep points: a calculus. *The Annals of Statistics*, 30(6):1681–1736, 2002.

[85] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679, 2008.

[86] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[87] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer New York, NY, 2004.

[88] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[89] Art Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443(7):59–72, 2007.

[90] Debolina Paul, Saptarshi Chakraborty, and Swagatam Das. Robust Principal Component Analysis: A Median of Means Approach. *arXiv preprint arXiv:2102.03403*, 2021.

[91] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[92] Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

[93] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A Robust Univariate Mean Estimator is All You Need. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 4034–4044. PMLR, 2020.

[94] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):601–627, 2020.

[95] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic Methods for ℓ_1-Regularized Loss Minimization. *The Journal of Machine Learning Research*, 12:1865–1892, 2011.

[96] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[97] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

[98] Jiyuan Tu, Weidong Liu, Xiaojun Mao, and Xi Chen. Variance Reduced Median-of-Means Estimator for Byzantine-Robust Distributed Inference. *Journal of Machine Learning Research*, 22(84):1–67, 2021.

[99] John W. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485, 1960.

[100] A. W. van der Vaart. *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 1998.

[101] Tim van Erven, Sarah Sachs, Wouter M. Koolen, and Wojciech Kotlowski. Robust Online Convex Optimization in the Presence of Outliers. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 4174–4194. PMLR, 2021.

[102] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, NY, 1999.

[103] Yehuda Vardi and Cun-Hui Zhang. The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

[104] Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A. Ryan, Margie L. Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:320–329, 2012.

[105] Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

[106] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[107] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[108] Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, pages 1084–1094, 2018.

[109] Tong Zhang. Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 116, New York, NY, USA, 2004. Association for Computing Machinery.

[110] Alice Zheng and Amanda Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., 2018.