Robust supervised learning with coordinate gradient descent

Ibrahim Merad¹ · Stéphane Gaïffas¹,²

Received: 2 November 2022 / Accepted: 19 July 2023 / Published online: 1 August 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 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 \texttt{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 Hawkins (1980), 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 (Hampel et al. 1986), 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 (Gupta and Kohli 2016). Statistical handling of outliers was already considered in the early 50’s (Dixon 1950; Grubbs 1969) and motivated in the 70’s the development of robust statistics (Huber 1972, 1981).

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

\[
\theta^* \in \arg\min_{\theta \in \Theta} R(\theta) \quad \text{where} \quad R(\theta) := \mathbb{E}[\ell(X^\top \theta, Y)], \tag{1}
\]

where \(\Theta\) is a convex compact subset of \(\mathbb{R}^d\) with diameter \(\Delta\) containing the origin and \(\ell : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}_+\) 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 \(\gamma\)-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

\[O(1) \text{ Ibrahim Merad}
imera@lpsm.paris
Stéphane Gaïffas
stephanegaiffas@lpsm.paris

¹ LPSM, UMR 8001, Université Paris Diderot, Paris, France
² DMA, École normale supérieure, Paris, France

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

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 (Huber 1964) \( \ell(z, y) = (z - y)^2 / 2 \) with \( q = 2 \) or the Huber loss (Huber 1964) \( \ell(z, y) = r_\gamma(z - y) \) for \( z, y \in \mathbb{R} \) with \( \gamma > 1 \) and \( q = 1 \), where \( r_\gamma(u) = \frac{1}{2} u^2 1_{|u| \leq \gamma} + \tau \left( |u| - \frac{1}{2} \gamma \right) 1_{|u| > \gamma} \) with \( \tau > 0 \) and the logistic loss \( \ell(z, y) = \log(1 + e^{-\gamma z y}) \) 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 \( \theta^* \) using a first-order optimization algorithm such as gradient descent (GD), using iterations of the form

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

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) (Vapnik 1999; Geer and van de Geer 2000), 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 above. 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 (Zheng and Casari 2018; Kuhn and Johnson 2019). An implicit,\(^1\) ERM estimator of \( \theta^* \) is a minimizer of the empirical risk \( \mathcal{L}_n \) given by

\[
\hat{\theta}^\text{erm}_n \in \arg\min_{\theta \in \Theta} \mathcal{L}_n(\theta) \quad \text{where} \quad \mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(X_i^\top \theta, Y_i),
\]

for which one can prove sub-Gaussian deviation bounds under strong hypotheses such as boundedness of \( \ell \) or sub-Gaussian concentration (Massart and Nédélec 2006; Lecué et al. 2013). In the general case, ERM leads to poor estimations of \( \theta^* \) 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 (Huber 1964, 1981; Hampel 1971; Hampel et al. 1986; Tukey 1960). The poor performance of ERM stems from the loose deviation bounds of the empirical mean estimator. Indeed, as explained by Catoni (2012) 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

\[
\hat{\nabla} \mathcal{L}^\text{erm}(\theta) := \nabla \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell'(X_i^\top \theta, Y_i) X_i,
\]

which is, as explained above, a poor estimator of \( \nabla \mathcal{L}(\theta) \) beyond (a) and (b).

**Robust gradient descent**

A growing literature about robust GD estimators (Prasad et al. 2020; Liu et al. 2019; Holland 2019; Geoffrey et al. 2020) suggests to perform GD iterations with \( \hat{\nabla} \mathcal{L}^\text{erm}(\theta) \) replaced by some robust estimator of \( \nabla \mathcal{L}(\theta) \). An implicit estimator is considered by Lecué et al. (2020), based on the minimization of a robust estimate of the risk objective using median-of-means. Robust estimators of \( \nabla \mathcal{L}(\theta) \) can be built using several approaches including geometric median-of-means (Prasad et al. 2020); robust coordinate-wise estimators (Holland et al. 2019) based on a modification of Catoni (2012); coordinate-wise median-of-means or trimmed means (Liu et al. 2019) or robust vector means through projection and truncation (Prasad et al. 2020). Other works achieve robustness by performing standard training on disjoint subsets of data and aggregating the resulting estimators into a robust one (Minsker 2015; Brownlees et al. 2015). We discuss such alternative methods in more details in Sect. 4 below.

These procedures based on GD require to run costly subroutines (at the exception of Lecué et al. (2020); Geoffrey et al. (2020)) 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 Sect. 2 below, the main idea is to combine coordinate gradient descent with robust estimators of the partial derivatives \( \partial \mathcal{L}(\theta) / \partial \theta_j \), 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 \( \|k\| = \{1, \ldots, k\} \) 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 \( \|n\| = \mathcal{O} \cup \mathcal{O}' \) 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
\[ E[|X|^\alpha] \leq \alpha^{q-1} E[|Y|^q] < +\infty \] and $E[|Y|^q] < +\infty$\end{equation}
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 $\alpha$ will depend on the considered robust estimator of the partial derivatives, as explained in Sect. 3 below, including theoretical guarantees with $\alpha < 1$ and cases with $E[Y^2] = +\infty$ (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 Sect. 2.2 below.

\textbf{Square loss.} For the square loss we have $q = 2$ and $E[Y^2] < +\infty$ is required for the risk $R(\theta)$ and its partial derivatives to be well-defined. Note that we have $E[|X|^2] < +\infty$, a requirement for the Lipschitz-smoothness of $R(\theta)$, as detailed in Sect. 2.2.

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

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

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

- We consider state-of-the-art robust estimators of the partial derivatives (Sect. 3) and provide theoretical guarantees for CGD combined with each of them. For some robust estimators, our analysis requires only weak moments (allowing $E[Y^2] = +\infty$ 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 (Sect. 2.2).

- We perform extensive numerical experiments, both for regression and classification on several data sets (Sect. 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 (Lecué and Lerasle 2020; Holland et al. 2019; Prasad et al. 2020), and we also consider several additional baselines such as Huber regression (Owen 2007), classification with the modified Huber loss (Zhang 2004), Least Absolute Deviation (LAD) (Edgeworth 1887) and RANSAC (Fischler and Bolles 1981). 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.\textsuperscript{2} This library follows the API conventions of scikit-learn (Pedregosa et al. 2011).

\section{Robust coordinate gradient descent}

CGD is well-known for its efficiency and fast convergence properties based on both theoretical and practical studies (Nesterov 2012; Shevade and Sathiya Keerthi 2003; Genkin et al. 2007; Wu and Lange 2008) 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 $\hat{g}_j(\theta)$ 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 $\hat{g}_j(\theta)$ are described in Sect. 3 below.

\subsection{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

$$
\begin{cases}
\theta_j^{(t+1)} := \theta_j^{(t)} - \beta \hat{g}_j(\theta^{(t)}) & \text{if } j = j_t \\
\theta_j^{(t+1)} := \theta_j^{(t)} & \text{otherwise}
\end{cases}
$$

\textsuperscript{2} https://github.com/linlearn/linlearn.
for all $j \in \llbracket d \rrbracket$, 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 $X \in \mathbb{R}^{n \times d}$ the features matrix with rows $X_1^T, \ldots, X_n^T$ and where $X^T \in \mathbb{R}^n$ stands for its $j$-th column.

Algorithm 1 Robust coordinate gradient descent

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

A simple choice for the distribution $p$ is the uniform distribution over $\llbracket d \rrbracket$, but improved convergence rates can be achieved using importance sampling, as explained in Theorem 1 below, where the choice of the step-sizes $(\beta_j)_{j=1}^d$ is described as well. The partial derivatives estimators $\tilde{g}_j(\cdot)_{j=1}^d$ described in Sect. 3 will determine the statistical error of this explicit learning procedure. Note that line 6 of Algorithm 1 uses the fact that

$$I^{(t+1)} = X\theta^{(t+1)} = \sum_{j \neq j_t} X_j^T \theta_{j_t}^{(t+1)} + X_{j_t}^T \theta_{j_t}^{(t+1)}$$

$$= \sum_{j \neq j_t} X_j^T \theta_j^{(t)} + X_{j_t}^T (\theta_{j_t}^{(t)} + D^{(t)})$$

$$= I^{(t)} + X_{j_t}^T D^{(t)}.$$ 

This computation has complexity $O(n)$, and we will see in Sect. 3 that the complexity of the considered robust estimators $\tilde{g}_{j_t}(\theta^{(t)})$ 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 $\tilde{\nabla}_j \epsilon(\theta^{(t)})$, see Equation (4). A more precise study of these complexities is discussed in Sect. 3, see in particular Table 1. Moreover, we will see experimentally in Sect. 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 Minsker (2015)] 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 Holland et al. (2019), Holland (2019), Holland and Ikeda (2019), Liu et al. (2019), and Jiyuan et al. (2021), 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 $\mathbb{R}(\theta^{(T)})$ [see Eq. (1)] for the output $\theta^{(T)}$ of Algorithm 1. These upper bounds are generic with respect to the considered robust estimators $(\tilde{g}_j(\cdot))_{j=1}^d$ and rely on the following definition.

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

$$\mathbb{P} \left[ \sup_{\theta \in \Theta} \left| \tilde{g}_j(\theta) - g_j(\theta) \right| \leq \epsilon_j(\delta) \right] \geq 1 - \delta$$

for all $j \in \llbracket d \rrbracket$.

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

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

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

$$\| g(\theta + h) - g(\theta) \| \leq L \| h \|$$
for any $h \in \Theta$ and $\theta \in \Theta$ such that $\theta + h \in \Theta$. We denote $L_{\text{max}} := \max_{j \in [d]} L_j$ and $L_{\text{min}} := \min_{j \in [d]} L_j$.

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

$$\nabla^2 R(\theta) = \mathbb{E}[\ell''(\theta^T X, Y)XX^T],$$

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

$$L_j = \sup_{\theta \in \Theta} \mathbb{E}[\ell''(\theta^T X, Y)(X^j)^2]$$

and

$$L = \sup_{\theta \in \Theta} \|\nabla^2 R(\theta)\|_{\text{op}},$$

where $\|H\|_{\text{op}}$ stands for the operator norm of a matrix $H$.

Assumption 1 entails $L_j \leq \gamma \mathbb{E}[(X^j)^2]$, 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 (Nesterov 2012; Wright 2015). 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 $\lambda$-strongly convex, namely that

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

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

Assumption 4 is satisfied whenever $\lambda_{\text{min}}(\nabla^2 R(\theta)) \geq \lambda$ for any $\theta \in \Theta$, where $\lambda_{\text{min}}(H)$ stands for the smallest eigenvalue of a symmetric matrix $H$. For the least-squares loss, this translates into the condition $\lambda_{\text{min}}(\mathbb{E}[XX^T]) \geq \lambda$. Note that one can always make the risk $\lambda$-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 Sect. 5 below. The following Theorem provides an upper bound over the risk of Algorithm 1 whenever the estimators $\hat{g}_j(\cdot)$ have an error vector $\epsilon(\cdot)$, as defined in Definition 1. We introduce for short $R^* = R(\theta^*) = \min_{\theta \in \Theta} R(\theta)$.

**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

$$\mathbb{E}[R(\theta^{(T)})] - R^* \leq (R(\theta^{(0)}) - R^*) \left(1 - \frac{\lambda}{L_{\text{max}}}\right)^T + \frac{L_{\text{max}}}{2\lambda L_{\text{min}}} \|\epsilon(\delta)\|_2^2$$

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

$$\mathbb{E}[R(\theta^{(T)})] - R^* \leq \left(R(\theta^{(0)}) - R^* \right) \left(1 - \frac{\lambda}{\sum_{j \in [d]} L_j}\right)^T + \frac{1}{2\lambda} \|\epsilon(\delta)\|_2^2$$

with probability at least $1 - \delta$.

The proof of Theorem 1 is given in Appendix B. It adapts standard arguments for the analysis of CGD (Nesterov 2012; Wright 2015) with inexact estimators of the partial derivatives. The statistical error $\|\epsilon(\delta)\|_2^2$ is studied in Sect. 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

$$(R(\theta^{(0)}) - R^*) \left(1 - \frac{\lambda}{\sum_{j \in [d]} L_j}\right)^T,$$

goes to 0 exponentially fast as the number of iterations $T$ increases, with a contraction constant better than that of (10) since $\sum_{j \in [d]} L_j \leq d L_{\text{max}}$. 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

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

for $d \geq 1$, which leads to a linear rate at least similar to the one of GD (Bubeck 2015).

Theorem 1 proves an upper bound on the excess risk $R(\theta^{(T)}) - R^*$ of the iterates of robust CGD directly, without using an intermediate upper bound on $\|\theta^{(T)} - \theta^*\|_2^2$. This differs from the approaches used by Prasad et al. (2020); Holland et al. (2019) 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 $\|\theta^{(T)} - \theta^*\|_2^2$, see Theorem 8 in Appendix B.

Note that the iterations considered in Algorithm 1 do not perform a projection in $\Theta$. Indeed, one can show that $\|\theta^{(T)} - \theta^*\|$ is also subject to a contraction and is therefore decreasing
Given $\theta^{(0)} = 0$, iterates $\theta^{(t)}$ naturally belong to the $\ell_2$ ball of radius $2\|\theta^*\|.$

**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 A.1, theoretical guarantees similar to that of Theorem 1 using step-sizes $\beta_j = 1/L_j$, where $L_j$ is a robust estimator of the upper bound $\hat{L}_j := \gamma\mathbb{E}[(X_j)^2] \geq L_j$ 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

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

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

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

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

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

The proof of Theorem 2 is given in Appendix B and uses arguments from Beck and Tretuashvili (2013) and Li et al. (2017). 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 7 for the non strongly convex case), this excludes problems with non-smooth regularization such as Lasso to which CGD has commonly been applied Wu and Lange (2008); Friedman et al. (2010); Yuan and Lin (2006). 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 Sect. 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

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

that can be used within Algorithm 1: Median-of-Means in Sect. 3.1, Trimmed mean in Sect. 3.2 and an estimator that we will call “Catoni–Holland” in Sect. 3.3. We provide, for each estimator, a concentration inequality for the estimator of $g_j(\theta)$ for fixed $\theta$ under a weak moments assumption (Lemmas 2, 3 and 4). We derive also uniform versions of the bounds in each case (Propositions 3, 4, 5 and 6) 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.

An estimator’s deviation bound is deemed optimal if it fits the lower bounds given by Theorems 1 and 3 in Lugosi and Mendelson (2019a). Let us introduce the centered moment of order $1 + \alpha$ of the partial derivatives and its maximum over $\Theta$, given by

$$
m_{\alpha,j}(\theta) := \mathbb{E}[|\ell'(X^\top \theta, Y)X_j|^1 + \alpha] \\
M_{\alpha,j} := \sup_{\theta \in \Theta} m_{\alpha,j}(\theta)
$$

(12)

for $\alpha \in (0, 1]$.

Note that $m_{1,j}(\theta) = \mathbb{V}[\ell'(X^\top \theta, Y)X_j]$ 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

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

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

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

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

(13)

3 Or more generally the centered moment of order $1 + \alpha$ for $\alpha \in (0, 1]$, see below.
Lemma 2 of the block-wise empirical means

The Median-Of-Means (MOM) estimator is the median
\[ \hat{g}^{(k)}_j(\theta) := \text{median} \left( g^{(1)}_j(\theta), \ldots, g^{(K)}_j(\theta) \right) \]

of the block-wise empirical means
\[ \hat{g}^{(k)}_j(\theta) := \frac{1}{|B_k|} \sum_{i \in B_k} g^j_i(\theta) \]

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:
\[ \text{P}\left[ |\hat{g}^{(k)}_j(\theta) - g(\theta)_j| > (24m_{a,j}(\theta))^{1/(1+\alpha)} \left( \frac{K}{n} \right)^{\alpha/(1+\alpha)} \right] \leq e^{-K/18} \]
for any fixed \( j \in [d] \) and \( \theta \in \Theta \). If we fix a confidence level \( \delta \in (0, 1) \) and choose \( K := \lfloor 18 \log(1/\delta) \rfloor \), we have
\[ |\hat{g}^{(k)}_j(\theta) - g(\theta)_j| \leq c_{\alpha}m_{a,j}(\theta)^{1/(1+\alpha)} \left( \frac{\log(1/\delta)}{n} \right)^{\alpha/(1+\alpha)} \]
\[ \leq c_{\alpha}M_{a,j}^{1/(1+\alpha)} \left( \frac{\log(1/\delta)}{n} \right)^{\alpha/(1+\alpha)} \]
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 Sect.B below and it adapts simple arguments from Lugosi and Mendelson (2019a) and Lecué et al. (2020). Compared to Lugosi and Mendelson (2019a), it provides additional robustness with respect to \( |\mathcal{O}| \geq 1 \) outliers and compared to Lecué et al. (2020) 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 Lugosi and Mendelson (2019a)) and is sub-Gaussian when \( \alpha = 1 \) (finite variance). The next proposition provides a uniform deviation bound over \( \Theta \) for \( \hat{g}^{(k)}_j(\theta) \).

Proposition 3 Grant Assumptions 1 and 2 with \( \alpha \in (0, 1) \) and \( |\mathcal{O}| \leq K/12 \). We have
\[ \text{P}\left[ \sup_{\theta \in \Theta} |\hat{g}^{(k)}_j(\theta) - g(\theta)_j| \leq \epsilon^{\text{MOM}}_j(\delta) \right] \geq 1 - \delta \]
for any \( j \in [d] \), with
\[ \epsilon^{\text{MOM}}_j(\delta) := c_{\alpha} \left( \frac{M_{j,\alpha} \cdot \frac{m_{L,\alpha}}{n^\alpha}}{n} \right)^{1/(1+\alpha)} \left( \frac{\log(d/\delta) + d \log(3\Delta n^\alpha/(1+\alpha)/2)}{n} \right)^{\alpha/(1+\alpha)} + (\overline{L} + L_j) \left( \frac{1}{n} \right)^{\alpha/(1+\alpha)} \]
where \( \overline{L} = \gamma E\|X\|_2^2, M_{L,\alpha} = E\|Y\|X\|_2^2 - \overline{L} \|^1\alpha \) and \( c_{\alpha} = 2^{(3+2\alpha)/(1+\alpha)} 3^{(1+3\alpha)/(1+\alpha)} \).

The proof of Proposition 3 is given in Sect.B and uses methods similar to Lemma 2 with an \( \varepsilon \)-net argument. This defines the error vector \( \epsilon^{\text{MOM}}(\delta) \) 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 Sect.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 \( |\hat{g}^{(k)}_j(\theta^{(t)}) - g^{(k)}_j(\theta^{(t)})| \), where both iterates \( \theta^{(t)} \) and estimators \( \hat{g}^{(k)}_j(\cdot) \) of the

| Properties of some robust estimators, where ERM = Empirical Risk Minimizer (ordinary mean), MOM = Median-of-Means, CH = Catoni–Holland and TM = Trimmed Mean |
|---|---|---|---|---|
| Optimal deviation bound | Robustness to outliers | Numerical complexity | Hyper-parameter |
| ERM | No | None | \( O(n) \) | None |
| MOM | Yes | Yes for \( |\mathcal{O}| < K/2 \) | \( O(n + K) \) | \( K \in [n] \) |
| CH | Yes | None | \( O(s) \) | Scale s |
| TM | Yes | Yes for \( |\mathcal{O}| < n/8 \) | \( O(n) \) | Proportion \( \epsilon \in [0, 1/2) \) |

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 \( \epsilon \) in TM.
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 \( \theta \) must be deterministic. We use in this paper an approach based on uniform deviation bounds (Propositions 3, 5 and 6) in order to bypass this problem, similarly to Holland and Ikeda (2019) and many other papers using empirical process theory. This is of course pessimistic, since \( \theta^{(t)} \) goes to \( \theta^* \) as \( t \) increases. Another approach considered in Prasad et al. (2020) 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 \( \tilde{\delta} = \delta/T \) and \( \tilde{n} = n/T \) instead of \( \delta \) 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 \( \epsilon \)-net argument, observed in Proposition 3 through the extra \( d^{n/(1+\alpha)} \) 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 3, we propose another uniform deviation bound for \( \widehat{g}_{\theta}^{MOM} \) using the Rademacher complexity, which is a fundamental tool in statistical learning theory and empirical process theory Ledoux and Talagrand (1991); Koltchinskii (2006); Bartlett et al. (2005). Let us introduce

\[
\mathcal{R}_j(\theta) = \mathbb{E} \left[ \sup_{\theta \in \Theta} \sum_{i \in \mathcal{I}} e_i g_j^i(\theta) \right]
\]

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

**Proposition 4** Grant Assumptions 1 and 2 with \( \alpha \in (0, 1) \). If \( |\mathcal{O}| \leq K/12 \), we have

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

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

\[
\sup_{\theta \in \Theta} \left| \widehat{g}_{j}^{MOM}(\theta) - g_j(\theta) \right| \\
\leq \max \left( c_\alpha M_{\alpha,j}^{1/(1+\alpha)} \left( \frac{\log(d/\delta)}{n} \right)^{\alpha/(1+\alpha)}, \frac{64 \mathcal{R}_j(\Theta)}{n} \right) \quad (17)
\]

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

\[
\mathcal{R}_j(\theta) \leq \gamma \Delta C_\alpha \left( n \mu_{X,j}^{1+\alpha} \sum_{k \in [d]} \mu_{X,k}^{1+\alpha} \right)^{1/(1+\alpha)} = O(n^{d/(1+\alpha)}),
\]

where \( C_\alpha \) is a constant depending only on \( \alpha \).

The proof of Proposition 4 is given in Sect. B and borrows arguments from Prasad et al. (2020); Boucheron et al. (2013). For \( \alpha = 1 \), the bound (17) has order \( O(\sqrt{d/n}) \) similarly to Theorem 2 from Lecué et al. (2020), 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 \( \mathbb{TM} \) and \( \mathbb{CH} \) algorithms considered below, an interesting open question.

**Comparison with Prasad et al. (2020); Holland et al. (2019)** A first distinction of our results compared to Prasad et al. (2020); Holland et al. (2019) 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 Prasad et al. (2020); Holland et al. (2019) require \( \alpha = 1 \). Our setting is similar but more general than the one laid out in Holland et al. (2019) since the latter does not consider the presence of outliers. Theorem 5 from Holland et al. (2019) 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 3, an improvement over the bound from Holland et al. (2019) which is only stated in terms of a uniform bound on the coordinate variances. Another reference point is the heavy-tailed setting of Prasad et al. (2020), which deals with heavy-tails independently from the problem of corruption and requires \( \alpha = 1 \). More importantly, the approach considered in Prasad et al. (2020) relies on data-splitting, which departs significantly from what is done in practice, while we do not perform data-splitting but use uniform bounds, as discussed above.

**Complexity of \( \widehat{g}_{j}^{MOM}(\theta) \)**

The computation of \( \widehat{g}_{j}^{MOM}(\theta) \) requires (a) to sample a permutation of \( [n] \) to sample the blocks \( B_1, \ldots, B_K \), (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 Knuth (1969), 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.
Hoare (1961) 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 \( \eta \)-corruption\(^4\) whenever the data distribution is sub-exponential Liu et al. (2019) or sub-Gaussian Diakonikolas et al. (2019b,a,c). Variant (2), also known as Winsorized mean, enjoys a sub-Gaussian deviation Lugosi and Mendelson (2021a) for heavy-tailed distributions. Both robustness properties are shown simultaneously (sub-Gaussian deviations under a heavy-tails assumption and \( \eta \)-corruption) in Lugosi and Mendelson (2021a) (see Theorem 1 therein). We consider below variant (2), which proceeds as follows.

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

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

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

\[
\hat{g}_j^{TM}(\theta) = \frac{2}{n} \sum_{i \in [n/2]^c} q_\epsilon \vee g_j'(\theta) \wedge q_{1-\epsilon},
\]

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 \([n/2]^c\) clipped in the interval \([q_\epsilon, q_{1-\epsilon}]\). Note that \( \hat{g}_j^{TM}(\theta) \) is also some form of “interpolation” between the average and the median through \( \epsilon \): 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 \( |O| \leq \eta n \) 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

\[
|\hat{g}_j^{TM}(\theta) - g_j(\theta)| \leq 7M_{\alpha,j}(\theta)^{1/(1+\alpha)} \left(4\eta + \frac{6\log(4/\delta)}{n}\right)^{\alpha/(1+\alpha)}
\]

with a probability larger than \( 1 - \delta \).

The proof of Lemma 3 is given in Sect. B and extends Theorem 1 from Lugosi and Mendelson (2021a) 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 only robust to a limited number of outliers. Note that for the computation of the TM estimator, the splitting \([n] = [n/2] \cup [n/2]^c\) is a technical theoretical requirement used to induce independence between \( q_\epsilon, q_{1-\epsilon} \) and the sample partial derivatives \((g_j'(\theta))_{i \in [n/2]^c}\) involved in the average (18). Our implementation does not use this splitting.

**Comparison with Prasad et al. (2020)**

A comparison between Lemma 3 and the results from Prasad et al. (2020) pertaining to the corrupted setting is relevant here. We first point out that corruption in Prasad et al. (2020) is modeled as receiving data from the “\( \eta \)-contaminated” distribution \((1 - \eta)P + \eta Q\) with \( Q \) an arbitrary distribution. On the other hand, Lemma 3 considers the more general \( \eta \)-corrupted setting where an \( \eta \)-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 \( \sqrt{\eta \log d} \) in the corruption (on the vector euclidean norm). On the other hand, Lemma 1 in Prasad et al. (2020) yields a better dependence of order \( \sqrt{\eta \log d} \) in the corresponding case. Keep in mind, however, that Algorithm 2 from Prasad et al. (2020) 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 Prasad et al. (2020) require a stronger moment assumption on the gradient and impose additional constraints on the corruption rate \( \eta \). We also mention Algorithm 5 from Prasad et al. (2020) 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 5** Grant Assumptions 1 and 2 with \( \alpha \in (0, 1) \) and \( |O| \leq \eta n \). We have

\[
\mathbb{P} \left[ \sup_{\theta \in \Theta} \left| \hat{g}_j^{TM}(\theta) - g_j(\theta) \right| \leq \epsilon_j^{TM}(\delta) \right] \geq 1 - \delta
\]
for any \( j \in [d] \) with

\[
\epsilon_{j}^{\text{T M}}(\delta) := 28 \left( \frac{M_{j,\alpha} + m_{L,\alpha}}{n\alpha(1+\alpha)} \right)^{1/(1+\alpha)} \left( 2\eta + 3\frac{\log(4d/\delta) + d\log(3\Delta n^{\alpha/(1+\alpha)}/2)}{n} \right)^{\alpha/(1+\alpha)} + \frac{\mathcal{T} + L_{j}}{n^{\alpha/(1+\alpha)}},
\]

where \( \mathcal{T} \) and \( m_{L,\alpha} \) are as in Proposition 3.

The proof of Proposition 5 is given in Appendix B and uses an \( \epsilon \)-net argument to obtain a uniform bound. Similarly to M-MOM, the resulting statistical error has optimal dependence on the \((1+\alpha)\)-moments of the partial derivatives (12).

By plugging, the error vector \( \epsilon^{\text{T M}}(\delta) \) 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 5, let \( \hat{\theta} := \theta(T) \) 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

\[
\frac{\lambda}{2} E \left[ \| \hat{\theta} - \theta \|^{2} \right] \leq E \left[ R(\hat{\theta}) \right] - R^{*} \leq O \left( \frac{1}{\lambda} \left( \sum_{j \in [d]} M_{j,\alpha}^{2/(1+\alpha)} \right) \left( \eta + \frac{\log(d/\delta) + d\log(n)}{n} \right)^{2\alpha/(1+\alpha)} \right),
\]

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

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

The proof of Corollary 1 is given in Appendix B and is a straightforward combination of Theorem 1 and Proposition 5 where a big O notation was used to make the bound more legible.

**Complexity of \( \hat{g}_{j}^{\text{T M}}(\theta) \)**

The most demanding part for the computation of \( \hat{g}_{j}^{\text{T M}}(\theta) \) is the computation of \( q_{e} \) 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 Hoare (1961) 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 Cormen et al. (2009)]. 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 Hoare (1961) was the fastest for this problem.

### 3.3 Catoni–Holland estimator

This estimator is a variation of the robust mean estimator by Catoni Catoni (2012) introduced by Holland Holland et al. (2019) for robust statistical learning, hence the name “Catoni–Holland”, that we will denote \( \hat{g}_{j}^{\text{CH}}(\theta) \). It is defined as an M-estimator which consists in solving

\[
\sum_{i=1}^{n} \psi \left( \frac{g_{j}^{(i)}(\theta) - \xi}{\hat{s}_{j}(\theta)} \right) = 0
\]

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

\[
\xi_{k+1} = \xi_{k} + \frac{\hat{s}_{j}(\theta)}{n} \sum_{i=1}^{n} \psi \left( \frac{g_{j}^{(i)}(\theta) - \xi_{k}}{\hat{s}_{j}(\theta)} \right),
\]

which can easily be shown to converge to the desired value thanks to the monotonicity and Lipschitz-property of \( \psi \). Following Holland et al. (2019), we use the function \( \psi(x) = 2 \arctan(\exp(x)) - \pi/2 \), while functions satisfying \( -\log(1 + x + x^{2}/2) \leq \psi(x) \leq \log(1 + x + x^{2}/2) \) are considered in Catoni (2012). As explained in Holland et al. (2019), the scale estimator is given by

\[
\hat{s}_{j}(\theta) := \hat{\sigma}_{j}(\theta) \sqrt{\frac{n}{2 \log(4/\delta)}},
\]

for a confidence level \( \delta \in (0, 1) \), where \( \hat{\sigma}_{j}(\theta) \) is an estimator of the standard deviation of the partial derivative \( \sigma_{j}(\theta) := m_{1,j}(\theta)^{1/2} = \sqrt{\mathbb{E}[\epsilon(X^{\top} \theta, Y)X^{\top}]^{1/2}} \), see (12). The estimator \( \hat{\sigma}_{j}(\theta) \) is defined through another M-estimator solution to

\[
\sum_{i=1}^{n} \chi \left( \frac{g_{j}^{(i)}(\theta) - \hat{g}_{j}(\theta)}{\sigma} \right) = 0
\]
with respect to \( \sigma \), where \( \hat{g}_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_i^j(\theta) \) and \( \chi \) is an even function satisfying \( \chi(0) < 0 \) and \( \chi(x) > 0 \) as \( x \to +\infty \). We use the same function as in Holland et al. (2019) given by \( \chi(u) = u^2/(1+u^2) - c \) where \( c \) is such that \( \mathbb{E}_X(Z) = 0 \) for \( Z \) a standard Gaussian random variable. To compute \( \hat{\sigma}_j(\theta) \) we use also fixed-point iterations

\[
\sigma_{k+1} = \sigma_k \left( 1 - \frac{\chi(0)}{n} \sum_{i=1}^{n} \chi \left( \frac{g_i^j(\theta) - \hat{g}_j(\theta)}{\sigma_k} \right) \right). \tag{23}
\]

We refer to the supplementary material of Holland et al. (2019) for further details on this procedure.

The \( \hat{C}H \) 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 of der V aart (1998). 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 \( \sigma_j(\theta)/C' \leq \hat{\sigma}_j(\theta) \leq C'\sigma_j(\theta) \) for some constant \( C' > 1 \). Then, we have

\[
\left| \hat{g}_j^{CH}(\theta) - g_j(\theta) \right| \leq C'\sigma_j(\theta) \sqrt{\frac{8 \log(4/\delta)}{n}}
\]

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 Sect.B and is an almost direct application of the deviation bound from Holland et al. (2019). If \( C' \approx 1 \), the deviation bound of \( \hat{g}_j^{CH}(\cdot) \) 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 Catoni (2012) results in a deviation with the best possible constant Devroye et al. (2016). However, contrary to \( \text{MOM} \) and \( \text{TM} \), an estimator of the scale is necessary: it makes \( \text{CH} \) computationally much more demanding (see Fig.1 below), since it requires to perform two fixed-point iterations to approximate both \( \hat{\sigma}_j(\theta) \) and \( \hat{g}_j^{CH}(\theta) \) and it requires Assumption 2 with \( \alpha = 1 \) so that \( \sigma_j(\theta) \approx +\infty \). Moreover, there is no guaranteed robustness to outliers, a fact confirmed by the numerical experiments performed in Sect. 6 below.

**Proposition 6** Grant Assumptions 1 and 2 with \( \alpha = 1 \) and \( \mathcal{O} = \emptyset \). Denote \( \bar{L} = \mathbb{E}[\|Y\|^2], \) \( \sigma_L^2 = \mathbb{V}[\|Y\|^2] \) and assume that for all \( \theta, \hat{\theta} \in \Theta \) such that \( \|\theta - \hat{\theta}\| \leq 1/\sqrt{n} \) we have

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

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

\[
\mathbb{P} \left[ \sup_{\theta \in \Theta} \left| \hat{g}_j^{CH}(\theta) - g_j(\theta) \right| \leq \epsilon_j^{CH}(\delta) \right] \geq 1 - \delta
\]

for any \( j \in [d] \) with

\[
\epsilon_j^{CH}(\delta) := 4C' \left( 2\Sigma_j + \frac{\sigma_L}{\sqrt{n}} \right) \sqrt{\frac{\log(4d/\delta) + d \log(3\Delta\sqrt{n}/2)}{n}} + \bar{L} + L_j
\]

where \( \bar{L} \) is as in Proposition 3.

The proof of Proposition 6 is given in Sect. B. It uses again an \( \epsilon \)-net argument combined with a careful control of the variations of \( \hat{g}_j^{CH}(\theta) \) with respect to \( \theta \). Compared with Holland et al. (2019), we make a different use of the \( \text{CH} \) 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 Sect. 6 confirm, in particular, that our approach leads to a considerable speedup and improved statistical performances when compared to Holland et al. (2019).

The statements of Lemma 4 and Proposition 6 require \( \alpha = 1 \), while a very recent extension of Catoni’s bound Chen et al. (2021) 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 6 require assumptions on \( \sigma_j^2(\theta) \) and \( \hat{\sigma}_j(\theta) \): an extension to \( \alpha \in (0, 1] \) would lead to a set of even more intricate assumptions.

**Complexity of \( \hat{g}_j^{CH}(\theta) \)**

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 Fig. 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 $g'_j(\theta)$. 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 Fig. 1, where we display the averaged timings over 100 repetitions (together with standard-deviations).

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 Tukey (1960) and Huber (1964) and has received longstanding interest since then. Several works pursued the development of robust statistical methods including non-convex $M$-estimators (Huber 1981), $\ell_1$ tournaments (Devroye and Györfi 1985; Donoho and Liu 1988) and methods based on depth functions (Chen et al. 2018; Gao 2020; Mizera 2002), 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) (Nemirovskij et al. 1983; Alon et al. 1999; Jerrum et al. 1986) and Catoni’s estimator (Catoni 2012). Under adversarial corruption (Charikar et al. 2017), several statistical learning problems are studied in a robust setting, such as parameter estimation (Lai et al. 2016; Prasad et al. 2020; Minsker 2018; Diakonikolas et al. 2019a; Lugosi and Mendelson 2021), regression (Klivans et al. 2018; Liu et al. 2020; Cherapanamjeri et al. 2020; Bhatia et al. 2017), classification (Lecué et al. 2020; Klivans et al. 2009; Liu and Tao 2015), PCA (Li 2017; Candès et al. 2011; Paul et al. 2021 and most recently online learning (van Erven et al. 2021).

In the heavy-tailed setting, a robust learning approach introduced in Brownlees et al. (2015) proposes to optimize a robust estimator of the risk based on Catoni’s mean estimator (Catoni 2012) 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 Zhang et al. (2018) 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 $\ell_1$ loss was used. A follow-up extended this result under weaker distribution assumptions (Chen et al. 2021). 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 Minsker (2015); Hsu and Sabato (2016) and consists in computing standard ERM s on disjoint subsets of the data and aggregating them using a multidimensional MOM. This approach recently appeared in Holland (2021) 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 Lugosi and Mendelson (2019a)). 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 $\eta$-proportion of the data consist of arbitrary outliers, a robust meta-algorithm is introduced in Diakonikolas et al. (2019b), which repeatedly trains a given base learner and filters outliers based on an eccentricity score. The method reaches the target $\sigma \sqrt{n}$ error rate with $\sigma$ 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 Lecué et al. (2020) 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 Lecué and Lerasle (2020) 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 (Prasad et al. 2020; Holland and Ikeda 2019; Holland 2019; Holland et al. 2019; Chen et al. 2017; Acharya et al. 2022) 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 Prasad et al. (2020). The first one is a vector analog of MOM where the scalar median is replaced by the geometric median

$$GMed(g_1, \ldots, g_K) := \arg\min_{g \in \mathbb{R}^d} \sum_{j=1}^K \| g - g_j \|_2,$$  \hspace{1cm} (24)

which can be computed using the algorithm given in Vardi and Zhang (2000). This vector mean estimator enjoys improved concentration properties over the standard mean as shown in Minsker (2015) although these remain sub-optimal (see also Lugosi and Mendelson (2019a)). A line of works (Lugosi and Mendelson 2019b; Hopkins 2020; Cherapanamjeri et al. 2019; Depersin and Lecué 2022; Lugosi and Mendelson 2021; Lei et al. 2020; Diakonikolas et al. 2020) 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 $\epsilon$-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 Pensia et al. (2020) 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 $\eta$-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 ($\dagger$) 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 Prasad et al. (2020)) have statistical rates in terms of $\tilde{n} = n/T$ and $\tilde{\delta} = \delta/T$ 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. Finally, the statistical rate of “GD-implicit MOM” is marked with a double dagger ($\ddagger$) 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 7 below, without strong convexity, the optimization error shrinks at a slower

\[ \text{Indeed, considering strong convexity, optimization converges linearly and the final bound is of the form } a \exp(-bT) + cT \log(T/\delta)/n \text{ for some } a, b, c > 0 \text{ and one can see that } T \sim \log n \text{ is approximately optimal.} \]
Table 2 Summary of the main characteristics of our proposed algorithms (using CGD) and the main competitors in the literature

| Algorithm                  | Statistical performance | Iteration/cycle complexity | Robustness to corruption |
|----------------------------|-------------------------|-----------------------------|--------------------------|
| CGD-MOM (MC00) This paper | $O\left(\frac{d \log(d) + d^2 \log(n)}{n}\right)$ | $O((n + K)d)$ (cycle)      | Yes for $|\mathcal{O}| < K/2$ |
| CGD-TM (TM) This paper    | $O\left(d \left(\frac{\eta + \log(d/k) + d \log(n)}{n}\right)\right)$ | $O(nd)$ (cycle)            | Yes for $|\mathcal{O}| < n/8$ |
| CGD-CH (CH) This paper    | $O\left(\frac{d \log(d/k) + d^2 \log(n)}{n}\right)$ | $O(nd)^\dagger$ (cycle)    | None                     |
| GD-Geometric MOM (Prasad et al. 2020) (GMOM) | $O\left(d \left(\frac{\log(1/d)}{n}\right)\right)$ | $O((n + K)d)^\dagger$      | Yes for $|\mathcal{O}| < K/2$ |
| GD-Huber gradient (HG) (Prasad et al. 2020) (GMOM) | $O\left(\log(d)\left(\frac{\eta + \log(d/k) + d \log(n)}{n}\right)^{3/4} + d^2\right)$ | $O(nd^2 + d^3)^\dagger$    | Yes for $\eta$-contamination |
| GD-implicit MOM (Lecué et al. 2020) (LLM) | $O\left(\frac{d \log(1/d)}{n}\right)^\dagger$ | $O(nd)$                    | Yes for $|\mathcal{O}| < K/4$ |
| GD-CH (CH GD) Holland et al. (2019) | $O\left(\frac{d \log(d/k) + d^2 \log(n)}{n}\right)$ | $O(nd)^\dagger$            | None                     |

The notations $\tilde{n}$ and $\tilde{d}$ stand for $n/T$ and $d/T$ 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.

Sub-linear rate when compared to Theorem 1 (a well-known fact, see Bubeck (2015)). 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 $\Theta = \prod_{j \in [d]} \Theta_j$ and replace the iterations (6) (corresponding to Line 5 in Algorithm 1) by

$$
\begin{align*}
\theta_{j}^{(t+1)} \leftarrow \text{proj}_{\Theta_j} \left( \theta_{j}^{(t)} - \beta_j \tau_j \left( \tilde{g}_j(\theta_j^{(t)}) \right) \right) & \quad \text{if } j = j_t \\
\theta_{j}^{(t+1)} \leftarrow \theta_j^{(t)} & \quad \text{otherwise,}
\end{align*}
$$

(25)

where $\text{proj}_{\Theta_j}$ is the projection onto $\Theta_j$ and $\tau_j$ is the soft-thresholding operator given by $\tau_j(x) = \text{sign}(x)(|x| - \epsilon)_+$ with $(x)_+ = \max(x, 0)$. In Theorem 7 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 Sect. 3. Since it depends on the moment $m_{\alpha_j}$, it is not observable, so we propose in Lemma 6 from Appendix A.2 an observable upper bound 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 $\tau_j$ in our numerical experiments from Sect. 6 below). Indeed, the operator $\tau_j$ naturally appears for the following simple reason: consider a convex $L$-smooth scalar function $f : \mathbb{R} \rightarrow \mathbb{R}$ with derivative $g(x) := f'(x)$. An iteration of gradient descent from $x_0$ uses an increment $\delta$ 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 $\arg\min_\delta Q(\delta, x_0) = -g(x_0)/L$ 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 $\tilde{g}(x)$ satisfying $|\tilde{g}(x) - g(x)| \leq \epsilon$ with a large probability. Taking this uncertainty into account leads to the upper bound

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

and, after projection onto the parameter set, to the iteration (25) since $\arg\min_\delta \tilde{Q}(\delta, x_0) = x_0 - \tau_\epsilon(\tilde{g}(x_0))/L$, 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 7 Grant Assumptions I 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 \( 1 - \delta \)

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

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

\[
\|\theta(t) - \theta^*\|_2 \leq \|\theta(t-1) - \theta^*\|_2
\]

with the same probability, for all \( t \in \llbracket T \rrbracket \).

The proof of Theorem 7 is given in Appendix B and is based on the proof of Theorem 5 from Nesterov (2012) and Theorem 1 from Shalev-Shwartz and Tewari (2011) 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 \( \Phi(\theta) = \sum_{j=1}^d L_j (\theta_j - \theta_j^*)^2/2 \) 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 Sect. 2. This degradation is an unavoidable consequence of the loss of strong convexity of the risk (Srebro et al. 2010).

### 6 Numerical Experiments

The theoretical results of Sects. 2, 3 and 5 can be applied to multiple supervised linear learning problems, 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 (Pedregosa et al. 2011).

**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 Sect. 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 Lecué et al. (2020). 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 Prasad et al. (2020), called Huber Gradient Estimator, which uses recursive SVD decompositions and truncations to compute a robust gradient. GMOM is Algorithm 3 from Prasad et al. (2020), which estimates gradients using a geometric MOM (based on the geometric median). CH GD is the robust GD algorithm from Holland et al. (2019), 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 (Fischler and Bolles 1981), using the implementation available in the scikit-learn library (Pedregosa et al. 2011). HUBER stands for ERM learning with the modified Huber loss (Zhang 2004) for classification and Huber loss (Owen 2007) for regression. LAD is ERM learning using the least absolute deviation loss (Edgeworth 1887), 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 \( \Sigma \) and labels \( Y = X^\top \theta^* + \xi \) for a fixed \( \theta^* \in \mathbb{R}^d \) and simulated noise \( \xi \). 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) \( \xi \) is centered Gaussian; (b) \( \xi \) is Student with \( v = 2.1 \) degrees of freedom (heavy-tailed noise). In the remaining settings (c), (d), (e) and (f), \( \xi \) 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 \( \Sigma \)) and labels are replaced by \( 2\gamma_{\max} \) with \( \gamma_{\max} = \max_{x \in X} |y| \); 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...
Fig. 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.

unit vector and $Z$ is a standard Gaussian vector and labels are i.i.d. Bernoulli random variables; finally for (f) we sample $X = 10\lambda_{\text{max}}(\Sigma) V$ where $V$ is uniform on the unit sphere and labels $y = y_{\text{max}} \times (\epsilon + U)$ where $\epsilon$ 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 (Fig. 2).

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 (Dua and Graff 2017). We use the logistic loss for binary and multiclass classification problems. For $k$-class problems with $k > 2$, the parameter $\theta$ 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 Blondel et al. (2013) 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 A.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 Fig. 3 the median accuracy obtained on a 15% test-set.
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, 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 Sect. 6.3 on regression data sets from the UCI Machine Learning Database, see Appendix A.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 Figs. 5 and 6. This is because they were run using scikit-learn’s implementation which does not give access to training history.

We observe on Fig. 5 that HUBER and LAD often achieve similar performance as they both optimize $\ell_1$ objectives.
Fig. 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.

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

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 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, Fig. 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.
of magnitude on some data sets and outright prohibitive on others.

As for the remaining algorithms, Fig. 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 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 Shalev-Shwartz and Tewari (2011); Juditsky et al. (2022).

Acknowledgements 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).

Author Contributions Conceptualization: Stéphane Gaïffas; Methodology: Ibrahim Merad, Stéphane Gaïffas; Formal analysis and investigation: Ibrahim Merad; Writing - original draft preparation: Ibrahim Merad; Writing - review and editing: Stéphane Gaïffas; Funding acquisition, Resources and Supervision: Stéphane Gaïffas.

Funding This research is supported by the french Agence Nationale de la Recherche as well as by the PRAIRIE Institute.

Availability of data and materials All data sets used in the experimental sections of this work are publicly available except for the synthetic ones which were randomly generated. Detailed information and sources were given in Appendix A.3.

Code Availability The experiments were carried out using the library called linlearn which was developed as part of this project and open-sourced under the BSD-3 License on GitHub and available here https://github.com/linlearn/linlearn.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.
Ethics approval  The authors approve the ethical rules set by the journal and are committed to abiding by them. Declarations relative to conflicts of interest, research involving Human Participants and/or Animals and informed consent are not applicable for this work.

Consent to participate  Not applicable.

Consent for publication  Not applicable.

Appendix A: Supplementary theoretical results and details on experiments

A.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 $i \in I$. We cannot use line-search (Armijo 1966) here, since it requires to evaluate the objective $K(\theta)$, 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

\[ U_j := \gamma E[(X_j)^2] \geq L_j, \quad (A1) \]

which holds under Assumption 1 and estimate $E[(X_j)^2]$ 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 $L^\xi$ condition with constant $C \geq 1$ whenever it satisfies

\[ (E[|Z - EZ|^\xi])^{1/\xi} \leq C (E[|Z - EZ|])^{1/\xi}. \quad (A2) \]

Using this condition, we can use the MOM estimator to obtain a high probability upper bound on $E[(X_j)^2]$ 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 $L^{(1+\alpha)}$ condition with a known constant $C$. For any fixed $j \in [d]$, let $\tilde{\sigma}_j^2$ be the MOM estimator of $E[(X_j)^2]$ with $K$ blocks. If $|\mathcal{O}| \leq K/12$, we have

\[
\frac{1}{(1 - 12^{1/(1+\alpha)}) C \left( \frac{K}{n} \right)^{\alpha/(1+\alpha)}} - 1 \leq \sigma_j^2 \leq \frac{1}{(1 - 12^{1/(1+\alpha)}) C \left( \frac{K}{n} \right)^{\alpha/(1+\alpha)}} - 1 \leq \tilde{\sigma}_j^2 \leq E[(X_j)^2] \leq \exp(-K/18).
\]

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

\[ \left( 1 - 216^{1/(1+\alpha)} C \left( \frac{\log(1/\delta)}{n} \right)^{\alpha/(1+\alpha)} \right)^{-1} \sigma_j^2 \leq \tilde{\sigma}_j^2 \geq E[(X_j)^2] \]

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

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

Corollary 2  Grant the same assumptions as in Theorem 1 and Proposition 3. Suppose additionally that for all $j \in [d]$, the variable $(X_j)^2$ satisfies the $L^{(1+\alpha)}$ condition with a known constant $C$ and fix $\delta \in (0, 1)$. Let $\theta^{(T)}$ be the output of Algorithm 1 with step-sizes $\beta_j = 1/L_j$ where $L_j := \gamma \hat{U}_j$ and $\hat{U}_j$ are the upper bounds from Lemma 5 with confidence $\delta/2d$, an initial iterate $\theta^{(0)}$, importance sampling distribution $p_j = \hat{U}_j / \sum_{k \in [d]} \hat{U}_k$ and estimators of the partial derivatives with error vector $\epsilon(\cdot)$. Then, we have

\[ E[R(\theta^{(T)})] - R^* \leq (E[R(\theta^{(0)})] - R^*) \left( 1 - \frac{\lambda}{\sum_{j \in [d]} L_j} \right) + \frac{1}{2\pi} \| \epsilon(\delta/2) \|^2_2 \]

(A3)

with probability at least $1 - \delta$.

The proof of Corollary 2 is given in Appendix B. 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 $L^{(1+\alpha)}$-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 $L^{(1+\alpha)}$-condition is much weaker than the requirement of boundedness (with known range) common to most known empirical bounds (Maurer et al. 2009; Audibert et al. 2009; Mnih et al. 2008).

A.2 Observable upper bound for the moment $m_{\alpha,j}$

Since the moment $m_{\alpha,j}$ 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 $\tilde{m}_{\alpha,j}(\theta)$ of the unknown moment $m_{\alpha,j}(\theta)$ using the following “two-step” MOM procedure. First, we compute $\tilde{g}_{\alpha,j}(\theta)$, the MOM estimator of $g_{\alpha,j}(\theta)$ with $K$ blocks.
given by (14). Then, we compute again a MOM estimator on $|g'_j(\theta) - \hat{g}^{\text{MOM}}_j(\theta)|^{1+\alpha}$ for $i \in [n]$, namely

$$m^\text{MOM}_{a,j}(\theta) := \text{median}(\hat{m}^{(1)}_{a,j}(\theta), \ldots, \hat{m}^{(K)}_{a,j}(\theta)), \quad (A4)$$

where

$$\hat{m}^{(k)}_{a,j}(\theta) := \frac{1}{|B_k|} \sum_{i \in B_k} |g'_j(\theta) - \hat{g}^{\text{MOM}}_j(\theta)|^{1+\alpha},$$

using uniformly sampled blocks $B_1, \ldots, B_K$ 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 $L^{(1+\alpha)} - L^{(1+\alpha)}$ condition with known constant $C$ for any $j \in [d]$ (see Definition 2). Then, if $|O| \leq K/12$, we have

$$\Pr[\hat{m}^\text{MOM}_{a,j}(\theta) \leq (1 - \kappa)m_{a,j}(\theta)] \leq 2 \exp(-K/18)$$

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

The proof of Lemma 6 is given in Appendix B.

**A.3 Experimental details**

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

**A.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.

**A.3.2 Data corruption**

For a given corruption rate $\eta$, we obtain a corrupted version of a data set by replacing an $\eta$-fraction of its samples with uninformative elements. For a data set of size $n$ which satisfies $|O| = \eta n$ 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, 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 $\tilde{X} \in \mathbb{R}^{n \times (d+1)}$ the data matrix with the vector of labels added to its columns. Let $\tilde{J} \subset [d + 1]$ denote the index of continuous columns, we compute $\hat{\mu}_j$ and $\hat{\sigma}_j$ their empirical means and standard deviations respectively for $j \in \tilde{J}$. We also sample a random unit vector $u$ of size $|\tilde{J}|$.

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

**A.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.

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

- MOM, GMOM, LLC: 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 $\delta$ used to define the CH estimator’s scale parameter (see Equation (21)) chosen with log-uniform distribution over $[e^{-10}, 1]$
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

| Data set                      | # Samples | # Features | # Categorical | # Classes |
|-------------------------------|-----------|------------|---------------|-----------|
| Statlog                       | 6435      | 36         | 0             | 6         |
| Spambase                      | 4601      | 57         | 0             | 2         |
| Electrical                    | 10,000    | 13         | 0             | 2         |
| Occupancy (Candanedo and Feldheim 2016) | 20,560 | 5         | 0             | 2         |
| Gas (Vergara et al. 2012)    | 13,910    | 128        | 0             | 6         |
| Drybean (Koklu and Ozkan 2020) | 13,611 | 16         | 0             | 7         |
| Energy (Candanedo et al. 2017) | 19,735 | 27         | 0             | –         |
| Bike (Fanace-T and Gama 2014) | 17,379 | 10         | 5             | –         |
| Metro                         | 48,204    | 6          | 1             | –         |
| Sgemm (Ballester-Ripoll et al. 2019) | 241,600 | 14        | 0             | –         |
| Ovctt                         | 68,784    | 20         | 2             | –         |
| Californiahousing             | 20,640    | 8          | 0             | –         |

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

| 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://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                        |

• 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 $[1.0, 2.5]$

\[
\mathbb{P} [E_j] \geq 1 - \delta \quad \text{where} \quad E := \{ \forall j \in [d], \ \forall t \in [T] \}.
\]

\[
\left| \hat{g}_j (\theta(t)) - g_j (\theta(t)) \right| \leq \epsilon_j (\delta), \quad \text{(B5)}
\]

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

\[
\theta^{(t+1)} = \theta^{(t)} - \beta_j \hat{g}_j \epsilon_j,
\]

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

\[
R (\theta^{(t+1)}) = R (\theta^{(t)} - \beta_j \hat{g}_j e_j)
\]

\[
\leq R (\theta^{(t)}) - \{ g (\theta(t)), \beta_j \hat{g}_j e_j \} + \frac{L_j \beta_j^2} {2} \hat{g}_j^2
\]

\[
= R (\theta^{(t)}) - \beta_j \hat{g}_j^2 - \beta_j g_t (\hat{g}_t - g_t) + \frac{L_j \beta_j^2} {2} (\hat{g}_t^2 + (\hat{g}_t - g_t)^2 + 2 g_t (\hat{g}_t - g_t))
\]

\]

Appendix B: Proofs

B.1 Proof of Theorem 1

This proof follows, with minor modifications, the proof of Theorem 1 from Wright (2015). Using Definition 1, we obtain

\[
\mathbb{P} [E] \geq 1 - \delta \quad \text{where} \quad E := \{ \forall j \in [d], \ \forall t \in [T] \}.
\]
\begin{align*}
&= R(\theta^{(t)}) - \beta_j g_t (1 - L_j \beta_j)(g_t - g_t) \\
&\quad - \beta_j \left(1 - \frac{L_j \beta_j}{2}\right) g_t^2 + \frac{L_j \beta_j^2}{2} (g_t - g_t)^2 \\
&= R(\theta^{(t)}) - \frac{1}{2L_j} g_t^2 + \frac{1}{2L_j} (g_t - g_t)^2 \\
&\leq R(\theta^{(t)}) - \frac{1}{2L_j} g_t^2 + \frac{\epsilon_j^2}{2L_j} \tag{B6}
\end{align*}

on the event \(\mathcal{E}\), where we used the choice \(\beta_j = 1/L_j\) and the fact that \(|\hat{g}_t - g_t| \leq \epsilon_j\) on \(\mathcal{E}\).

Since \(j_1, \ldots, j_t\) is a i.i.d sequence with distribution \(\mu\), we have for any \((j_1, \ldots, j_t)\)-measurable and integrable function \(\varphi\) that

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

where we denote for short the conditional expectation 
\(\mathbb{E}_{t-1}[\cdot | \mathcal{E}_{t-1} = \mathcal{E}_{t-1}[j_1, \ldots, j_t-1].\) So, taking \(\mathbb{E}_{t-1}[\cdot]\) on both sides of (B6) leads, whenever \(p_j = L_j/\sum_{k=1}^d L_k\), to

\[
\mathbb{E}_{t-1}[R(\theta^{(t+1)})] \leq R(\theta^{(t)}) - \frac{1}{2} \sum_{k=1}^d L_k g(\theta^{(t)})^2 \\
+ \frac{1}{2} \sum_{k=1}^d L_k \Xi,
\]

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

\[
\mathbb{E}_{t-1}[R(\theta^{(t+1)})] \leq R(\theta^{(t)}) - \frac{1}{2dL_{\min}} \|g(\theta^{(t)})\|_2^2 \\
+ \frac{1}{2dL_{\min}} \Xi
\]

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

\[
\mathbb{E}_{t-1}[R(\theta^{(t+1)})] \leq R(\theta^{(t)}) - \frac{1}{2L} \|g(\theta^{(t)})\|_2^2 + \bar{\epsilon}.
\]

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

\[
\phi_{t+1} \leq \phi_t - \frac{1}{2L} \|g(\theta^{(t)})\|_2^2 + \bar{\epsilon}.
\]

(B7)

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

\[
R(\theta_t) \geq R(\theta^{(t)}) + \langle \nabla R(\theta^{(t)}), \theta_t - \theta^{(t)} \rangle + \frac{\lambda}{2} \|\theta_t - \theta^{(t)}\|^2
\]

for any \(\theta_t \in \mathbb{R}^d\), so that by minimizing both sides with respect to \(\theta_t\) leads to

\[
R^* \geq R(\theta^{(t)}) - \frac{1}{2\lambda} \|g(\theta^{(t)})\|_2^2
\]

namely

\[
\phi_t \leq \frac{1}{2\lambda} \mathbb{E}\|g(\theta^{(t)})\|_2^2,
\]

by taking the expectation on both sides. Together with (B7) this leads to the following approximate contraction property:

\[
\phi_{t+1} \leq \phi_t \left(1 - \frac{\lambda}{T}\right) + \bar{\epsilon},
\]

and by iterating \(t = 1, \ldots, T\) to

\[
\phi_T \leq \phi_0 \left(1 - \frac{\lambda}{T}\right) T + \bar{\epsilon} \bar{L},
\]

which allows to conclude the Proof of Theorem 1. \(\square\)

### B.2 Proof of Theorem 2

This proof reuses ideas from Li et al. (2017) and Beck and Tetruashvili (2013) 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 (B5) 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 \(\theta_{i+1}^{(t)} = \theta_i^{(t)} - \beta_i^g (\theta_i^{(t)} - \beta_i^g (\theta_i^{(t)} - \beta_i^{(t+1)})\) for \(i = 0, \ldots, d - 1\) with \(\beta_i^g = 1/L_i, \beta_i^{(t)} = \theta_i^{(t)}\) and \(\beta_i^{(t+1)} = \theta_i^{(t+1)}\). With these notations we have

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

Similarly to (B6) in the proof of Theorem 1 we find:

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

leading to

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

(B8)
The following aims to find a relationship between $\sum_{j=0}^{d-1} \frac{1}{2L_{j+1}} g(\theta^{(j)}_{i+1})^2$ and $\|g(\theta^{(i)})\|_2^2$, which we do by comparing coordinates. For the first step in a cycle we have $g(\theta^{(0)}_{i+1}) = g(\theta^{(0)}_{i})$ because $\theta^{(0)} = \theta^{(0)}_{0}$. Let $j \in \{1, \ldots, d - 1\}$, by the Mean Value Theorem, there exists $\gamma^{(j)}_j \in \mathbb{R}^d$ such that we have:

$$g(\theta^{(j)}_{j+1}) = g(\theta^{(j)}_{j+1}) - g(\theta^{(j)}_{j}) + g(\theta^{(j)}_{j})$$

$$= \left(\nabla g(\gamma^{(j)}_{j+1})\right)^T (\theta^{(j)} - \theta^{(j)}) + g(\theta^{(j)}_{j+1})$$

$$= \left[\begin{array}{c}
\partial R(\gamma^{(j)}_1), \
\vdots \\
\partial R(\gamma^{(j)}_{d})
\end{array}\right] \theta^{(j)}$$

$$+ g(\theta^{(j)}_{j+1})$$

$$= [H_{j+1,1}, \ldots, H_{j+1,j}, 0, \ldots, 0] \left[\begin{array}{c}
g_1(\theta_{1}^{(j)}) \\
\vdots \\
g_d(\theta_{d}^{(j)})
\end{array}\right] + g(\theta^{(j)}_{j+1})$$

$$+ \sum_{j=0}^{d-1} \frac{1}{2L_{j+1}} g(\theta^{(j)}_{j+1})^2 + \sum_{j=0}^{d-1} \frac{1}{2L_{j+1}} \|H\|_2^2 \epsilon_{Euc}^2$$

We can bound the spectral norm $\|H\|_2$ as follows:

$$\|H\|_2 \leq \|A^{1/2} + HA^{-1/2}\|_2 \leq \|A^{1/2}\|_2 + \|HA^{-1/2}\|_2$$

$$\leq 2 \left( L_{\max} + \frac{\|H\|_2}{L_{\min}} \right)$$

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

$$\|H\|_2 \leq \sum_{j=1}^{d} \|h_j\|_2 \leq \sum_{j=1}^{d} \|\nabla g(\gamma^{(j)}_j)\|_2$$

$$\leq \sum_{j=1}^{d} \sqrt{L_{j}} \leq d \sqrt{L_{\max}}$$

Combining the previous inequality with (B8) and (B9), we find:

$$R(\theta^{(i)}) - R(\theta^{(i+1)})$$

$$\geq \frac{1}{8L_{\max} \left( 1 + d \frac{L_{\max}}{L_{\min}} \right)} \|g(\theta^{(i)})\|_2^2$$

$$- \frac{\epsilon_{Euc}^2}{2} \left( \frac{1}{L_{\min}} + 2d \left( \frac{L_{\max}}{L_{\min}} \right)^2 \right)$$

$$\geq \frac{1}{8L_{\max} \left( 1 + d \frac{L_{\max}}{L_{\min}} \right)} \|g(\theta^{(i)})\|_2^2$$

$$- \frac{\epsilon_{Euc}^2}{2} \left( \frac{1}{L_{\min}} + 2d \frac{L_{\max}}{L_{\min}} \frac{1 + d \frac{L_{\max}}{L_{\min}}}{L_{\min}} \right)$$

$$\geq \frac{1}{8L_{\max} \left( 1 + d \frac{L_{\max}}{L_{\min}} \right)} \|g(\theta^{(i)})\|_2^2 - \frac{3}{4L_{\min}} \epsilon_{Euc}^2$$

where the last step uses that $\frac{L_{\max}}{L_{\min}} \leq 1$. Using $\lambda$-strong convexity by choosing $\theta_1 = \theta^{(i)}$ in inequality (9) and mini-
mizing both sides w.r.t. \( \theta \) we obtain:

\[
R(\theta(t)) - R^* \leq \frac{1}{2\lambda} \|g(\theta(t))\|^2,
\]

which combined with the previous inequality yields the contraction inequality:

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

and after \( T \) iterations we have:

\[
R(\theta(T)) - R^* \leq (R(\theta(0)) - R^*)(1 - 2\lambda \kappa)^T + \frac{3e_{Euc}^2}{8L_{\text{min}}^2 \lambda \kappa},
\]

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 (B9) work all the same if one were to apply a permutation to the coordinates beforehand.

### B.3 Convergence of the parameter error

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

**Theorem 8** 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

\[
E\|\theta(T) - \theta^*\|_2 \leq \|\theta(0) - \theta^*\|_2 (1 - \frac{2\beta \lambda L}{d\lambda + L})^T + \sqrt{\frac{d(\lambda + L)}{\lambda L}} \|\epsilon(\delta)\|_2
\]  

(B10)

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 \( \hat{\theta}_j(\theta) \) be the estimators used and introduce the notations

\[
\hat{\theta}_j = \hat{\theta}_j(\theta(t)) \quad \text{and} \quad g_t = g_j(\theta(t)).
\]

We also condition on the event (B5) 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 \( j \), i.e. for a random variable \( \xi \) we have \( \| \xi \|_{L_2} = \sqrt{E_{\delta,j} \| \xi \|^2} \). We compute:

\[
\|\theta(t+1) - \theta^*\|_{L_2} = \|\theta(t) - \beta_j g_t e_j - \theta^*\|_{L_2} \
\leq \|\theta(t) - \beta_j g_t e_j - \theta^*\|_{L_2} + \|\beta_j (\hat{\theta}_j - g_t)\|_{L_2}.
\]  

(B11)

We first treat the first term of (B11), in the case of uniform sampling with equal step-sizes \( \beta_j = \beta \) we have:

\[
\|\theta(t) - \beta g_t e_j - \theta^*\|^2 = \|\theta(t) - \theta^*\|^2 + \beta^2 g_t^2 - 2\beta \langle g_t e_j, \theta(t) - \theta^*\rangle.
\]

By taking the expectation w.r.t. the random coordinate \( j \), we find:

\[
E_{\delta,j}\|\theta(t) - \theta^*\|^2 = E_{\delta,j}\|\theta(t) - \theta^*\|^2
\]

\[
\leq E_{\delta,j}\|\theta(t) - \theta^*\|^2 + \beta^2 E_{\delta,j}g_t^2
\]

\[
+ \beta^2 E_{\delta,j}\|g(\theta(t))\|^2(1 - \frac{1}{d})
\]

\[
\leq E_{\delta,j}\|\theta(t) - \theta^*\|^2(1 - \frac{2\beta \lambda L}{d\lambda + L}) + \beta \frac{2\beta \lambda L}{d\lambda + L} E_{\delta,j}\|g(\theta(t))\|^2(1 - \frac{1}{d})
\]

\[
= E_{\delta,j}\|\theta(t) - \theta^*\|^2(1 - \frac{2\beta \lambda L}{d\lambda + L}) + \beta \frac{2\beta \lambda L}{d\lambda + L} E_{\delta,j}\|g(\theta(t))\|^2(1 - \frac{1}{d})
\]

\[
\leq \|\theta(t) - \theta^*\|^2(1 - \frac{2\beta \lambda L}{d\lambda + L}) + \beta \frac{2\beta \lambda L}{d\lambda + L} E_{\delta,j}\|g(\theta(t))\|^2(1 - \frac{1}{d})
\]

The first inequality is obtained by applying inequality (2.1.24) from Nesterov (2004) (see also Bubeck (2015) Lemma 3.11) and the second one is due to the choice of \( \beta \). We can bound the second term as follows:

\[
\|\hat{\theta}_j - g_t\|^2_{L_2} = \|\hat{\theta}_j - gt\|^2 \leq \frac{e_{Euc}^2}{d}.
\]

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

\[
\|\theta(t+1) - \theta^*\|_{L_2} \leq \kappa \|\theta(t) - \theta^*\|_{L_2} + \frac{\beta e_{Euc}}{\sqrt{d}}.
\]

By iterating this argument on \( T \) rounds we find that:

\[
\|\theta(T) - \theta^*\|_{L_2} \leq \kappa^T \|\theta(0) - \theta^*\|_{L_2} + \frac{\beta e_{Euc}}{\sqrt{d(1 - \kappa)}}.
\]
Finally, the following inequality yields the result in the case of uniform sampling:

\[
\frac{1}{1 - \kappa} \leq \frac{1 + \sqrt{1 - \frac{2p\lambda L}{d(\alpha + L)}}}{d(\lambda + L)} \leq \frac{d(\lambda + L)}{\beta \lambda L}.
\]

\[\square\]

B.4 Proof of Lemma 1

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

\[
|\ell'(\theta^T X, Y)| \leq C_{\ell,1} + 2^{q-1} C_{\ell,2} |\theta^T X - Y|^{q}
\leq C_{\ell,1} + 2^{q-1} C_{\ell,2} |\theta^T X|^q + |Y|^q.
\]

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

\[
|\ell'(\theta^T X, Y)|^1 + \alpha
\leq 2^a \left( (C_{\ell,1}^\alpha + C_{\ell,2}^\alpha |\theta^T X - Y|^{q-1}) X_j \right)^{1 + \alpha}
\leq 2^a \left( (C_{\ell,1}^\alpha + C_{\ell,2}^\alpha |\theta^T X|^q + |Y|^q X_j) \right)^{1 + \alpha},
\]

\[
= 2^a \left( (C_{\ell,1}^\alpha + C_{\ell,2}^\alpha \sum_{k=1}^d |\theta_k(X_k^{q-1} - X_j)| + |Y|^q X_j) \right)^{1 + \alpha},
\]

\[
\leq 2^a \left( |C_{\ell,1}^\alpha X_j|^{1 + \alpha} + 2^a (C_{\ell,2}^\alpha)^{1 + \alpha} \sum_{k=1}^d |\theta_k(X_k^{q-1} - X_j)| + |Y|^q X_j \right)^{1 + \alpha}.
\]

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

\[
\mathbb{E}[|X_j^{q-1} X_j|^1 + \alpha] \leq \left( \mathbb{E}[|X_k|^{q(1 + \alpha)}]^{1/a} \mathbb{E}[|X_j|^{q(1 + \alpha)}]^{1/b} \right)^{1 + \alpha},
\]

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

B.5 Proof of Lemma 2

This proof follows a standard argument from Lugosi and Mendelson (2019a); Geoffrey et al. (2020) in which we use a Lemma from Bubeck et al. (2013) 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 \( \mathbb{E}[|\ell'(\theta^T X, Y)|^{1 + \alpha}] < \infty \) for any \( j \in [d] \). Recall that \( \hat{g}_j^k(\theta) \) stands for the block-wise empirical mean given by Eq. (15) and introduce the set of non-corrupted block indices given by \( K = \{ k \in [K] : B_k \cap O = \emptyset \} \). We will initially assume that the number of outliers satisfies \( |O| \leq (1 - \varepsilon)K/2 \) for some \( 0 < \varepsilon < 1 \). Note that since samples are i.i.d in \( B_k \) for \( k \in K \), we have \( \mathbb{E}[\hat{g}_j^k(\theta)] = g_j(\theta) \). We use the following Lemma from Bubeck et al. (2013).

Lemma 7 [Lemma 3 from Bubeck et al. (2013)] Let \( Z, Z_1, \ldots, Z_n \) be an i.i.d sequence with \( m_a = \mathbb{E}[|Z - \mathbb{E}[Z]|^{1 + \alpha}] < \infty \) for any \( \delta \in (0, 1), \) with a probability \( 1 - \delta \).

Lemma 7 entails that

\[
\frac{\hat{g}_j^k(\theta)}{\mathbb{E}[\hat{g}_j^k(\theta)]} \leq \left( \frac{3m_a}{\delta(n/K)^{a}} \right)^{1/(1 + a)} =: \eta_{j, a, \delta}(\theta)
\]

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

\[
P \left[ \left| \frac{\hat{g}_j^k(\theta)}{\mathbb{E}[\hat{g}_j^k(\theta)]} - g_j(\theta) \right| \geq \eta_{j, a, \delta}(\theta) \right]
\leq P \left[ \sum_{k \in K} \mathbb{1} \left( \left| \frac{\hat{g}_j^k(\theta)}{\mathbb{E}[\hat{g}_j^k(\theta)]} - g_j(\theta) \right| \geq \eta_{j, a, \delta}(\theta) \right) \geq K/2 \right]
\leq P \left[ \sum_{k \in K} \mathbb{1} \left( \left| \frac{\hat{g}_j^k(\theta)}{\mathbb{E}[\hat{g}_j^k(\theta)]} - g_j(\theta) \right| \geq \eta_{j, a, \delta}(\theta) \right) \right],
\]

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

\[
\sum_{k \in K} \mathbb{1} \left( \left| \frac{\hat{g}_j^k(\theta)}{\mathbb{E}[\hat{g}_j^k(\theta)]} - g_j(\theta) \right| \geq \eta_{j, a, \delta}(\theta) \right)
\]

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

\[
P \left[ \left| \frac{\hat{g}_j^k(\theta)}{\mathbb{E}[\hat{g}_j^k(\theta)]} - g_j(\theta) \right| \geq \eta_{j, a, \delta}(\theta) \right]
\]
where we used the fact that $|\Theta| \leq (1 - \varepsilon)K/2$ 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 $\delta' = 1/8$.

**B.6 Proof of Proposition 3**

**Step 1** First, we fix $\theta \in \Theta$ and try to bound $|\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\theta)|$ in terms of quantities only depending on $\tilde{\theta}$ which is the closest point to $\theta$ in an $\varepsilon$-net. Recall that $\Delta$ is the diameter of the parameter set $\Theta$ and let $\varepsilon > 0$ be a positive number. There exists an $\varepsilon$-net covering $\Theta$ with cardinality no more than $(3\Delta/2\varepsilon)^d$ i.e. a set $N_{\varepsilon}$ such that for all $\theta \in \Theta$ there exists $\tilde{\theta} \in N_{\varepsilon}$ such that $|\tilde{\theta} - \theta| \leq \varepsilon$. Consider a fixed $\theta \in \Theta$ and $j \in [d]$, we wish to bound the quantity $|\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\theta)|$.

Using the $\varepsilon$-net $N_{\varepsilon}$, there exists $\tilde{\theta}$ such that $|\tilde{\theta} - \theta| \leq \varepsilon$ which we can use as follows:

$$
|\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\theta)| \leq |\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\tilde{\theta})| + |g_{j}(\tilde{\theta}) - g_{j}(\theta)| \\
\leq |\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\tilde{\theta})| + L_j \varepsilon, \quad (B12)
$$

where we used the gradient’s coordinate Lipschitz constant to bound the second term. We now focus on the second term. Introducing the notation $\hat{g}_{j}'(\theta) = \hat{\ell}'(\tilde{\theta}^\top X_i, \gamma^\top)X_i^j$, we have

$$
\hat{g}_{j}'(\theta) = \hat{\ell}'(\tilde{\theta}^\top X_i, \gamma^\top)X_i^j + \hat{\ell}'(\tilde{\theta}^\top X_i, \gamma^\top - \hat{\ell}'(\tilde{\theta}^\top X_i, \gamma^\top))X_i^j. \quad := \Delta_i
$$

Let $(B_k)_{k \in [K]}$ be the blocks used to compute the $\text{MOM}$ estimator and associated block means $\hat{g}_{j}^{(k)}(\tilde{\theta})$ and $\hat{\mathbb{E}}_{j}^{(k)}(\tilde{\theta})$. Notice that the $\text{MOM}$ estimator is monotonous non decreasing w.r.t. to each of the entries $g_{j}'(\theta)$ when the others are fixed. Without loss of generality, assume that $\hat{g}_{j}^{(k)}(\tilde{\theta}) - g_{j}(\tilde{\theta}) \geq 0$ then we have:

$$
|\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\tilde{\theta})| \leq |\hat{\mathbb{E}}_{j}^{\text{MOM}}(\theta) - g_{j}(\tilde{\theta})|, \quad (B13)
$$

where $\hat{\mathbb{E}}_{j}^{\text{MOM}}(\tilde{\theta})$ is the $\text{MOM}$ estimator obtained using the entries $\hat{\ell}'(\tilde{\theta}^\top X_i, \gamma^\top)X_i^j + \hat{\ell}'(\tilde{\theta}^\top X_i, \gamma^\top)X_i^j$ instead of $g_{j}'(\theta)$. Note that $\hat{\mathbb{E}}_{j}^{\text{MOM}}(\tilde{\theta})$ no longer depends on $\theta$ except through the fact that $\tilde{\theta}$ is chosen in $N_{\varepsilon}$ so that $|\tilde{\theta} - \theta| \leq \varepsilon$. Indeed, using the Lipschitz smoothness of the loss function and a Cauchy-Schwarz inequality we find that:

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

**Step 2**

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

$$
|\hat{\mathbb{E}}_{j}^{(k)}(\tilde{\theta}) - g_{j}(\tilde{\theta})| \leq \left( \frac{6m_{L,a}(\tilde{\theta})}{\delta'(n/K)^{\alpha}} \right)^{1/(1+\alpha)} =: \eta_{j,a,\delta'/2}(\tilde{\theta}),
$$

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

$$
|\hat{\mathcal{L}}^{(k)} - \mathcal{L}| \leq \left( \frac{6m_{L,a}}{\delta'(n/K)^{\alpha}} \right)^{1/(1+\alpha)} =: \eta_{L,a,\delta'/2},
$$

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

$$
\mathbb{P}\left( \frac{|\hat{g}_{j}^{(k)}(\tilde{\theta}) + \varepsilon \hat{\mathcal{L}}^{(k)} - g_{j}(\tilde{\theta})|}{\eta_{j,a,\delta'/2}(\tilde{\theta}) + \varepsilon(\mathcal{L} + \eta_{L,a,\delta'/2})} \right) \leq \mathbb{P}\left( \frac{|\hat{g}_{j}^{(k)}(\tilde{\theta}) - g_{j}(\tilde{\theta})|}{\eta_{j,a,\delta'/2}(\tilde{\theta})} \right) + \mathbb{P}\left( \frac{|\hat{\mathcal{L}}^{(k)} - \mathcal{L}|}{\eta_{L,a,\delta'/2}} \right) \leq \delta'/2 + \delta'/2 = \delta'.
$$

Now defining the Bernoulli variables

$$
U_k := 1\{|\hat{g}_{j}^{(k)}(\tilde{\theta}) + \varepsilon \hat{\mathcal{L}}^{(k)} - g_{j}(\tilde{\theta})| > \eta_{j,a,\delta'/2}(\tilde{\theta}) + \varepsilon(\mathcal{L} + \eta_{L,a,\delta'/2})\}
$$

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

$$
\mathbb{P}\left( \left| \hat{\mathbb{E}}_{j}^{\text{MOM}}(\tilde{\theta}) - g_{j}(\tilde{\theta}) \right| \geq \eta_{j,a,\delta'/2}(\tilde{\theta}) + \varepsilon(\mathcal{L} + \eta_{L,a,\delta'/2}) \right) \leq \mathbb{P}\left( \sum_{k \in [K]} U_k > K/2 \right) \leq \mathbb{P}\left( \sum_{k \in K} U_k > K/2 - |\Theta| \right),
$$

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

$$
\mathbb{P}\left( \left| \hat{\mathbb{E}}_{j}^{\text{MOM}}(\tilde{\theta}) - g_{j}(\tilde{\theta}) \right| \geq \eta_{j,a,\delta'/2}(\tilde{\theta}) + \varepsilon(\mathcal{L} + \eta_{L,a,\delta'/2}) \right) \leq \mathbb{P}[S > K/2 - |\Theta|] \leq \mathbb{P}[S > K/2 - |\Theta|].
$$
\[ \mathbb{P}[S - ES > K/2 - |\mathcal{O}| - \delta' |\mathcal{K}|] \]
\[ \leq \mathbb{P}[S - ES > (\epsilon' - 2\delta')/2] \]
\[ \leq \exp\left(-K(\epsilon' - 2\delta')^2/2\right), \]
where we used the condition $|\mathcal{O}| \leq (1-\epsilon')K/2$ and $|\mathcal{K}| \leq K$ for the second inequality and the Hoeffding inequality for the last. To conclude, we choose $\epsilon' = 5/6$ and $\delta' = 1/4$ and combine (B12), (B13) and the last inequality in which we take $K = 18 \log(1/\delta)$ and use a union bound argument to obtain that with probability at least $1 - \delta$ for all $j \in [d]$

\[ \left| \hat{g}^{(k)}(\theta) - g_j(\theta) \right| \leq \left((24m_{j,1}(\theta))^{1/(1+\alpha)} + \epsilon L \right). \] (B14)

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

\[ \sup_{\theta \in \Theta} \left| \hat{g}^{(k)}(\theta) - g_j(\theta) \right| \]
\[ \leq \sup_{\theta \in \Theta} \left| \hat{g}^{(k)}(\theta(\theta)) - g_j(\theta(\theta)) \right| + \left| g_j(\theta(\theta)) - g_j(\theta) \right| \]
\[ \leq \sup_{\theta \in \Theta} \left| \hat{g}^{(k)}(\theta(\theta)) - g_j(\theta(\theta)) \right| + \varepsilon L \]
\[ = \max_{\theta \in N_\varepsilon} \left| \hat{g}^{(k)}(\theta) - g_j(\theta) \right| + \varepsilon L. \]

Here, we make a union bound argument over $\theta \in N_\varepsilon$ for the inequality (B14) and choose $\varepsilon = n^{-\alpha/(1+\alpha)}$ to obtain the final result concluding the proof of Proposition 3.

**B.7 Proof of Proposition 4**

This proof reuses arguments from the proof of Theorem 2 in Lecué et al. (2020). We wish to bound $\left| \hat{g}^{(k)}(\theta) - g_j(\theta) \right|$ with high probability and uniformly on $\theta \in \Theta$. Fix $\theta \in \Theta$ and $j \in [d]$, we have $\hat{g}^{(k)}(\theta) = \text{median}(\hat{g}^{(1)}(\theta), \ldots, \hat{g}^{(K)}(\theta))$ with $\hat{g}^{(k)}(\theta) = \frac{K}{n} \sum_{i \in B_k} \hat{g}^{(1)}(\theta)$ where the blocks $B_1, \ldots, B_K$ constitute a partition of $[n]$.

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

\[ \sup_{\theta \in \Theta} \sum_{k=1}^{K} 1\left\{ \left| \hat{g}^{(k)}(\theta) - g_j(\theta) \right| > x \right\} \]
\[ \leq \sup_{\theta \in \Theta} \sum_{k \in K} \mathbb{E}[\phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x)] + |\mathcal{O}| \]
\[ + \sup_{\theta \in \Theta} \sum_{k \in K} \left( \phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x) \right) \]
\[ - \mathbb{E}[\phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x)]. \]

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

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

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

\[ \sup_{\theta \in \Theta} \sum_{k=1}^{K} 1\left\{ |\hat{g}^{(k)}(\theta) - g_j(\theta)| > x \right\} \]
\[ \leq K \left( \frac{3M_{a,j}}{(x/2)^{1+\alpha}(n/K)^{\alpha}} + \frac{|\mathcal{O}|}{K} \right) \]
\[ + \sup_{\theta \in \Theta} \frac{1}{K} \left( \sum_{k \in K} \phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x) \right) \]
\[ - \mathbb{E}[\phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x)]. \]

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:

\[ \sup_{\theta \in \Theta} \frac{1}{K} \left( \sum_{k \in K} \phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x) \right) \]
\[ - \mathbb{E}[\phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x)]. \]

Using a simple symmetrization argument (see for instance Lemma 11.4 in Boucheron et al. (2013)) we find:

\[ \mathbb{E}\left[ \sup_{\theta \in \Theta} \frac{1}{K} \left( \sum_{k \in K} \phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x) \right) \right] \]
\[ - \mathbb{E}[\phi(2|\hat{g}^{(k)}(\theta) - g_j(\theta)|/x)] \]
\[ \leq 2\mathbb{E}\left[ \sup_{\theta \in \Theta} \frac{1}{K} \sum_{k \in K} \xi_k \phi(2|\hat{g}^{(k)}(\theta) - g(\theta)|/x) \right]. \]

where the $\xi_k$s are independent Rademacher variables. Since $\phi$ is 1-Lipschitz and satisfies $\phi(0) = 0$ we can use the contraction principle [see Theorem 11.6 in Boucheron et al. (2013)] followed by another symmetrization step to find
contraction principle adapted to our case in which functions
Denote $\phi^k(\theta) = 2\hat{g}_j^k(\theta) - g_j(\theta)$ and notice that

$$2E \sup_{\theta \in \Theta} \frac{1}{K} \sum_{k \in K} \epsilon_k \phi^2(\theta) / x$$

$$\leq 4E \sup_{\theta \in \Theta} \frac{1}{K} \sum_{k \in K} \epsilon_k \phi^2(\theta) / x$$

$$\leq \frac{8}{nx} E \left[ \sup_{\theta, \theta' \in \Theta} \sum_{i \in J} \epsilon_i \phi^2(\theta) \right] \leq \frac{8R_j(\Theta)}{nx}.$$

Taking $|\Theta| \leq (1 - \epsilon)K/2$, we found that with probability
$1 - \exp(-2y^2)$

$$\sup_{\theta \in \Theta} \sum_{k=1}^K 1 \left| \hat{g}_j^k(\theta) - g_j(\theta) \right| > x$$

$$\leq K \left( \frac{3M_{\alpha, j}}{(x/2)^{1+\alpha} (n/K)^{\alpha}} + \frac{|\Theta|}{K} + 8R_j(\Theta) \right).$$

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

$$P \left( \sup_{\theta \in \Theta} \left| \hat{g}_j^k(\theta) - g_j(\theta) \right| \left| \Theta \right| \right) \geq \max \left( \left( \frac{36M_{\alpha, j}}{(n/K)^{\alpha}} \right)^{1/(1+\alpha)} , \frac{64R_j(\Theta)}{n} \right)$$

$$\leq P \left( \sup_{\theta \in \Theta} \sum_{k=1}^K 1 \left| \hat{g}_j^k(\theta) - g_j(\theta) \right| > K/2 \right)$$

$$\leq \exp(-2(x - 1/2)^2 K/4)$$

$$\leq \exp(-K/18).$$

where the last inequality comes from the choice $\epsilon = 5/6$. 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 $I = [n]$, we first write

$$\mathcal{R}_j(\Theta) = E \left[ \sup_{\theta \in \Theta} \sum_{i=1}^n \epsilon_i \ell'(\theta^T X_i, Y_i) X_i^j \right]$$

$$= E \left[ \sum_{i=1}^n \epsilon_i \ell'(0, Y_i) X_i^j \right] + \sup_{\theta \in \Theta} \sum_{i=1}^n \epsilon_i \left( \ell'(\theta^T X_i, Y_i) - \ell'(0, Y_i) \right) X_i^j.$$

Denote $\phi_i(t) = (\ell'(t, Y_i) - \ell'(0, Y_i)) X_i^j$ and notice that

$$E \left[ \sum_{i=1}^n \epsilon_i \ell'(0, Y_i) X_i^j \right] = 0.$$
the inequality:

$$
\mathbb{E} \left[ \left( \sum_{i=1}^{n} \epsilon_{i} x_{i} \right)^{p} \right]^{1/p} \leq B_{p,\alpha} \left( \sum_{i=1}^{n} |x_{i}|^{1+\alpha} \right)^{1/(1+\alpha)}
$$

with the constant $B_{p,\alpha} := 2p \left( \frac{1+\alpha}{\alpha} \right)^{\frac{p}{1+\alpha}} \Gamma \left( \frac{\alpha p}{1+\alpha} \right)$. 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 Ledoux and Talagrand (1991) and uses similar methods. For all $\lambda > 0$ we have:

$$
\mathbb{E} \exp \left( \lambda \sum_{i} \epsilon_{i} x_{i} \right) = \prod_{i} \mathbb{E} \exp(\lambda \epsilon_{i} x_{i}) = \prod_{i} \cosh(\lambda x_{i}) 
$$

$$
\leq \prod_{i} \exp \left( \frac{\lambda |x_{i}|^{1+\alpha}}{1+\alpha} \right)
$$

$$
= \exp \left( \sum_{i} \frac{\lambda |x_{i}|^{1+\alpha}}{1+\alpha} \right),
$$

where we used the inequality $\cosh(u) \leq \exp \left( \frac{u^{1+\alpha}}{1+\alpha} \right)$ valid for all $u \in \mathbb{R}$ which can be quickly proven. Since both functions are even, fix $u > 0$ and define $f_{u}(\alpha) = \exp \left( \frac{u^{1+\alpha}}{1+\alpha} \right) − \cosh(u)$, 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 $\frac{2e^{2u}}{u^{2} + e^{2u}} \leq 2 \leq 1 + e/\mu$ 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 $(\sum_{i=1}^{n} |x_{i}|^{1+\alpha})^{1/(1+\alpha)} = 1$, we compute:

$$
\mathbb{E} \left[ \left( \sum_{i} \epsilon_{i} x_{i} \right)^{p} \right] = \int_{0}^{+\infty} \mathbb{P} \left( \sum_{i} \epsilon_{i} x_{i} > t \right) dt 
$$

$$
\leq 2 \int_{0}^{+\infty} \exp \left( \frac{\lambda u^{1+\alpha}}{1+\alpha} \right) \frac{dt}{\lambda} 
$$

$$
= 2 \int_{0}^{+\infty} \exp \left( - \frac{\alpha u^{1+\alpha}}{1+\alpha} \right) du 
$$

$$
= 2p \left( \frac{1+\alpha}{\alpha} \right)^{\frac{p}{1+\alpha}} \Gamma \left( \frac{\alpha p}{1+\alpha} \right) = B_{p,\alpha}^{\alpha},
$$

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 $\alpha$ while for $\alpha \sim 0$ it is consequence of the fact that $\Gamma(\alpha) \sim 1/\alpha$ near 0 and the limit $x^{\alpha} \to 0$ when $x \to 0^{+}$.

**B.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 $\hat{X}_{1}, \ldots, \hat{X}_{N}$, $\hat{Y}_{1}, \ldots, \hat{Y}_{N}$ denote an $\eta$-corrupted i.i.d sample with rate $\eta$ 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 $\hat{\mu}$ the $\epsilon$-trimmed mean estimator computed as $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \phi_{\alpha,\beta}(\hat{X}_{i})$ with $\phi_{\alpha,\beta}(x) = \max(\alpha, \min(x, \beta))$ and the thresholds $\alpha = \hat{Y}_{\epsilon}^{*}$ and $\beta = \hat{Y}_{\epsilon}^{*}$.

Let $1 > \delta \geq e^{-N/4}$, taking $\epsilon = 8\eta + 12 \log(4/\delta) / N$, we have

$$
|\hat{\mu} - \mu| \leq 7M^{1+\gamma} (\epsilon / 2)^{1+\gamma}
$$

with probability at least $1 - \delta$.

**Proof** This proof goes along the lines of the proof of Theorem 1 from Lugosi and Mendelson (2021) 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 $\tilde{X} = X - \mu$. Let $X_{1}, \ldots, X_{N}$, $Y_{1}, \ldots, Y_{N}$ the original uncorrupted i.i.d. sample from $X$ and let $\hat{X}_{1}, \ldots, \hat{X}_{N}$, $\hat{Y}_{1}, \ldots, \hat{Y}_{N}$ denote the corrupted sample with rate $\eta$. We define the following quantity which will intervene in the proof:

$$
\overline{X}(\epsilon, X) := \max \left\{ 1_{X \geq \alpha} \mathbb{E} \left[ |\tilde{X} - Q_{\epsilon/2}(\tilde{X})| 1_{\tilde{X} \leq Q_{\epsilon/2}(\tilde{X})} \right], \mathbb{E} \left[ |\tilde{X} - Q_{1-\epsilon/2}(\tilde{X})| 1_{\tilde{X} \geq Q_{1-\epsilon/2}(\tilde{X})} \right] \right\},
$$

(B16)

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

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

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

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

where $\min(\alpha, \min(x, \beta))$
and similarly with probability $\geq 1 - \exp(-\epsilon N/12)$ we have:

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

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

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

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 $Y_1, \ldots, Y_N$. Since $\eta \leq \epsilon/8$, if $2\eta N$ samples are corrupted we still have:

$$\left| \{ i \ : \ \tilde{Y}_i \geq \mu + Q_{1-\epsilon}(X) \} \right| \geq ((3/2)\epsilon - 2\eta) N \geq \epsilon N$$

and

$$\left| \{ i \ : \ \tilde{Y}_i \leq \mu + Q_{1-\epsilon/2}(X) \} \right| \geq (1 - (3/4)\epsilon - 2\eta) N \geq (1 - \epsilon) N$$

consequently, the two following bounds hold

$$Q_{1-\epsilon}(X) \leq \tilde{Y}^{\star}_{\epsilon} = Q_{1-\epsilon}(X) - \mu \leq Q_{1-\epsilon/2}(X)$$

and similarly

$$Q_{\epsilon/2}(X) \leq \tilde{Y}^{\star}_{\epsilon \epsilon} = \tilde{Y}^{\star}_{(1-\epsilon)N} - \mu \leq Q_{2\epsilon}(X).$$

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

**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. We write:

$$\frac{1}{N} \sum_{i=1}^{N} \phi_{\alpha, \beta}(X_i) \leq \frac{1}{N} \sum_{i=1}^{N} \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X_i)$$

$$= \mathbb{E}\left[ \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X) \right]$$

$$+ \frac{1}{N} \sum_{i=1}^{N} \left( \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X_i) \right) - \mathbb{E}\left[ \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X) \right].$$

The first term is dominated by:

$$\mathbb{E}\left[ \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X) \right] = \mathbb{E}\left[ \phi_{Q_2(X), Q_{1-\epsilon/2}(X)}(X) \right]$$

$$= \mathbb{E}\left[ Q_{2\epsilon}(X) \mathbf{1}_{X \leq Q_{2\epsilon}(X)} + X \mathbf{1}_{Q_{1-\epsilon/2}(X) \leq X < Q_{1-\epsilon/2}(X)} \right]$$

$$+ Q_{1-\epsilon/2}(X) \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)}$$

$$= \mu + \mathbb{E}\left[ (Q_{2\epsilon}(X) - X) \mathbf{1}_{X \leq Q_{2\epsilon}(X)} \right]$$

$$+ (Q_{1-\epsilon/2}(X) - X) \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)}$$

$$\leq \mu + \mathbb{E}\left[ (Q_{2\epsilon}(X) - X) \mathbf{1}_{X \leq Q_{2\epsilon}(X)} \right]$$

and lower bounded by:

$$\mathbb{E}\left[ \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X) \right]$$

$$= \mu + \mathbb{E}\left[ (Q_{2\epsilon}(X) - X) \mathbf{1}_{X \leq Q_{2\epsilon}(X)} \right]$$

$$+ (Q_{1-\epsilon/2}(X) - X) \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)}$$

$$\geq \mu + \mathbb{E}\left[ (Q_{1-\epsilon/2}(X) - X) \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)} \right]$$

$$= \mu - \mathbb{E}(\epsilon, X).$$

The sum in (B17) above has terms upper bounded by $Q_{1-\epsilon/2}(X) + \mathbb{E}(\epsilon, X)$. We need to work with the knowledge that $\mathbb{E}|X|^{1+\gamma} = M < \infty$ in order to bound their variance:

$$\mathbb{E}\left[ \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X) \right] - \mathbb{E}\left[ \phi_{\mu + Q_{2\epsilon}(X), \mu + Q_{1-\epsilon/2}(X)}(X) \right]^2$$

$$\leq \mathbb{E}\left[ (Q_{2\epsilon}(X) - X) \mathbf{1}_{X \leq Q_{2\epsilon}(X)} \right]^2$$

$$\leq \mathbb{E}\left[ (Q_{1-\epsilon/2}(X) - X) \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)} \right]^2$$

$$= \mathbb{E}\left[ Q_{2\epsilon}(X) \mathbf{1}_{X \leq Q_{2\epsilon}(X)} + X \mathbf{1}_{Q_{1-\epsilon/2}(X) \leq X < Q_{1-\epsilon/2}(X)} \right]$$

$$+ Q_{1-\epsilon/2}(X) \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)}$$

$$= \mathbb{E}\left[ Q_{2\epsilon}(X)^2 \mathbf{1}_{X \leq Q_{2\epsilon}(X)} + X^2 \mathbf{1}_{Q_{1-\epsilon/2}(X) \leq X < Q_{1-\epsilon/2}(X)} \right]$$

$$+ Q_{1-\epsilon/2}(X)^2 \mathbf{1}_{X \leq Q_{1-\epsilon/2}(X)} \right].$$

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

$$2\epsilon = \mathbb{P}(X \leq Q_{2\epsilon}(X)) \leq \mathbb{P}(|X|^{1+\gamma})$$

$$\geq |Q_{2\epsilon}(X)|^{1+\gamma} \leq \frac{M}{|Q_{2\epsilon}(X)|^{1+\gamma}},$$

(B18)

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

$$\epsilon/2 = \mathbb{P}(X \geq Q_{1-\epsilon/2}(X)) \leq \mathbb{P}(|X|^{1+\gamma})$$

$$\geq |Q_{1-\epsilon/2}(X)|^{1+\gamma} \leq \frac{M}{|Q_{1-\epsilon/2}(X)|^{1+\gamma}}, \quad \text{(B19)}$$

from (B18), we deduce that

$$\mathbb{E}\left[ Q_{2\epsilon}(X)^2 \mathbf{1}_{X \leq Q_{2\epsilon}(X)} \right] = 2\epsilon Q_{2\epsilon}(X)^2$$

$$\leq 2\epsilon \left( \frac{M}{2\epsilon} \right)^{2/(1+\gamma)} \leq 2\epsilon \left( \frac{2M}{\epsilon} \right)^{2/(1+\gamma)},$$

\( \diamond \) Springer
and from (B19) we find
\[
\mathbb{E}[Q_{1-\epsilon/2}(\overline{X})^2 1_{\overline{X} \leq Q_{1-\epsilon/2}(\overline{X})}] = Q_{1-\epsilon/2}(\overline{X})^2 / 2 \\
\leq 2\epsilon \left( \frac{2M}{\epsilon} \right)^{2/(1+\gamma)}.
\]

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

The case \(Q_{1-\epsilon/2}(\overline{X}) \leq 0\) is similarly handled. Moreover, a simple calculation yields
\[
\mathbb{E}[\overline{X}^2 1_{\overline{X} \leq Q_{1-\epsilon/2}(\overline{X})}] \leq M \max \{ |Q_{2\epsilon}(\overline{X})|, |Q_{1-\epsilon/2}(\overline{X})| \}^{1-\gamma} \leq 2\epsilon \left( \frac{2M}{\epsilon} \right)^{2/(1+\gamma)}.
\]

All in all, we have shown the inequality:
\[
\mathbb{E}[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)] \\
- \mathbb{E}[\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)]^2 \leq 6\epsilon \left( \frac{2M}{\epsilon} \right)^{2/(1+\gamma)},
\]
which we now use to apply Bernstein’s inequality on the sum in (B17) to find, conditionally on \(Y_1, \ldots, Y_n\), with probability at least \(1 - \delta/4\):
\[
\frac{1}{N} \sum_{i=1}^{N} \phi_{\mu_{\epsilon}}(X_i) \\
\leq \mu + \mathbb{E}(4\epsilon, X) + \sqrt{\frac{6\epsilon \log(4/\delta)}{N}} \left( \frac{2M}{\epsilon} \right)^{1/(1+\gamma)} \\
+ \frac{\log(4/\delta)}{3N} \left( Q_{1-\epsilon/2}(\overline{X}) + \mathbb{E}(\epsilon, X) \right) \\
\leq \mu + 2\mathbb{E}(4\epsilon, X) + \sqrt{\frac{6\epsilon \log(4/\delta)}{N}} \left( \frac{2M}{\epsilon} \right)^{1/(1+\gamma)} \\
+ \frac{\log(4/\delta)}{3N} Q_{1-\epsilon/2}(\overline{X}) \\
\leq \mu + 2\mathbb{E}(4\epsilon, X) + (3/2)M^{1/(1+\gamma)} \left( \epsilon / 2 \right)^{\gamma/(1+\gamma)},
\]
where we used (B19), the fact that \(\log(4/\delta) / N \leq \epsilon / 12\) and the assumption that \(\delta \geq \epsilon^{-N}/4\). Using the same argument on the lower tail, we obtain, on the event \(E\), that with probability at least \(1 - \delta/2\)
\[
\left| \frac{1}{N} \sum_{i=1}^{N} \phi_{\mu_{\epsilon}}(X_i) - \mu \right| \\
\leq 2\mathbb{E}(4\epsilon, X) + (3/2)M^{1/(1+\gamma)} \left( \epsilon / 2 \right)^{\gamma/(1+\gamma)}.
\]

**Step 3**

Now we show that \(\left| \frac{1}{N} \sum_{i=1}^{N} \phi_{\mu_{\epsilon}}(X_i) - \frac{1}{N} \sum_{i=1}^{N} \phi_{\mu_{\epsilon}}(\widetilde{X}_i) \right|\) 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:
\[
|\phi_{\mu_{\epsilon}}(X_i) - \phi_{\mu_{\epsilon}}(\widetilde{X}_i)| \leq |Q_{1-\epsilon/2}(\overline{X})| + |Q_{1-\epsilon/2}(\overline{X})|,
\]
and since we have \(\eta \leq \epsilon / 8\) then
\[
\left| \frac{1}{N} \sum_{i=1}^{N} \phi_{\mu_{\epsilon}}(X_i) - \frac{1}{N} \sum_{i=1}^{N} \phi_{\mu_{\epsilon}}(\widetilde{X}_i) \right| \\
\leq 2\eta \left( |Q_{1-\epsilon/2}(\overline{X})| + |Q_{1-\epsilon/2}(\overline{X})| \right) \\
\leq \frac{\epsilon}{2} \max \{ |Q_{1-\epsilon/2}(\overline{X})|, |Q_{1-\epsilon/2}(\overline{X})| \} \\
\leq M^{1/(1+\gamma)} \left( \epsilon / 2 \right)^{\gamma/(1+\gamma)},
\]
where the last step follows from (B18) and (B19). Finally, using similar arguments along with Hölder’s inequality, we show that:
\[
\mathbb{E}[\| \overline{X} - Q_{\epsilon/2}(\overline{X}) \| 1_{\overline{X} \leq Q_{\epsilon/2}(\overline{X})}] \\
\leq \mathbb{E}[\| \overline{X} \| 1_{\overline{X} \leq Q_{\epsilon/2}(\overline{X})}] + \mathbb{E}[\| Q_{\epsilon/2}(\overline{X}) \| 1_{\overline{X} \leq Q_{\epsilon/2}(\overline{X})}] \\
\leq M^{1/(1+\gamma)} \left( \epsilon / 2 \right)^{\gamma/(1+\gamma)} + |Q_{\epsilon/2}(\overline{X})| \left( \epsilon / 2 \right) \\
\leq 2M^{1/(1+\gamma)} \left( \epsilon / 2 \right)^{\gamma/(1+\gamma)},
\]
and a similar computation for \(\mathbb{E}[\| \overline{X} - Q_{1-\epsilon/2}(\overline{X}) \| 1_{\overline{X} \leq Q_{1-\epsilon/2}(\overline{X})}]\) leads to
\[
\mathbb{E}(4\epsilon, X) \leq 2M^{1/(1+\gamma)} \left( \epsilon / 2 \right)^{\gamma/(1+\gamma)}.
\]

This completes the proof of Lemma 9. □

**B.9 Proof of Proposition 5**

**Step 1.** Notice that the \(\hat{\theta}^{TM}\) 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 3. We define an \(\epsilon\)-net \(N_{\epsilon}\) on the set \(\Theta\), fix \(\theta \in \Theta\) and let \(\tilde{\theta}\) be the closest point in \(N_{\epsilon}\). We obtain, for all \(j \in \{1, \ldots, d\}\), the inequalities:
\[
|\hat{\theta}^{TM}_{j}(\theta) - g_{j}(\theta)| \leq |\hat{\theta}^{TM}_{j}(\theta) - g_{j}(\tilde{\theta})| \\
+ |g_{j}(\tilde{\theta}) - g_{j}(\theta)| \\
\leq |\hat{\theta}^{TM}_{j}(\tilde{\theta}) - g_{j}(\tilde{\theta})| + \epsilon L_{j}, \quad \text{(B20)}
\]
where \( \tilde{g}_j^{TM}(\hat{\theta}) \) is the \( \mathcal{T}M \) estimator obtained for the entries \( \ell'(\hat{\theta}^T X_i, Y_i) X_i^j + \varepsilon \gamma \|X_i\|^2 = g_j'(\hat{\theta}) + \varepsilon \gamma \|X_i\|^2 \).

**Step 2**

We use the concentration property of the \( \mathcal{T}M \) estimator to bound the previous quantity which is in terms of \( \hat{\theta} \). The terms \( \{g_j'(\hat{\theta}) + \varepsilon \gamma \|X_i\|^2\}_{i \in [n]} \) are independent and distributed according to \( Z := \ell'(\hat{\theta}^T X, Y) X^j + \gamma \|X\|^2 \).

Obviously we have \( \ell'(\hat{\theta}^T X, Y) X^j = g_j(\hat{\theta}) \). Furthermore, let \( \mathcal{L} = \mathbb{E}\|X\|^2 \), so that \( \mathbb{E}\left[\mathbb{E}_Z[g_j'(\hat{\theta}) + \varepsilon \gamma \|X\|^2]\right] = g_j(\hat{\theta}) + \varepsilon \mathcal{L} \).

We will apply Lemma 9 for \( \tilde{g}_j^{TM}(\hat{\theta}) \). Before we do so, we need to compute the centered \((1 + \alpha)\)-moment of \( Z \). Let \( m_{j,a}(\hat{\theta}) \) and \( m_{L,a} \) be the centered \((1 + \alpha)\)-moments of \( \ell'(\hat{\theta}^T X, Y) X^j \) and \( \gamma \|X\|^2 \) respectively, we have:

\[
\mathbb{E}\|Z - \mathbb{E}Z\|^{1+\alpha} \leq 2^a (m_{j,a}(\hat{\theta}) + \varepsilon^{1+\alpha} m_{L,a})
\]

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

\[
|\tilde{g}_j^{TM}(\hat{\theta}) - g_j(\hat{\theta}) - \varepsilon \mathcal{L}| 
\leq 7(m_{j,a}(\hat{\theta}) + \varepsilon^{1+\alpha} m_{L,a})^{1/(1+\alpha)} (2\varepsilon^{\alpha/(1+\alpha)})^2
\]

with \( \varepsilon_\delta = 8\varepsilon + 12\log(4/\delta)/n \). By combining with (B20) and using a union bound argument, we deduce that with the same probability, we have for all \( j \in [d] \)

\[
|\tilde{g}_j^{TM}(\hat{\theta}) - g_j(\hat{\theta})| 
\leq 7(m_{j,a}(\hat{\theta}) + \varepsilon^{1+\alpha} m_{L,a})^{1/(1+\alpha)} (4\varepsilon^{\alpha/(1+\alpha)})^{n/(1+\alpha)} + \varepsilon \mathcal{L}.
\]

**Step 3**

We use the \( \varepsilon \)-net to obtain a uniform bound. We proceed similarly as in the proof of Proposition 3. For \( \theta \in \Theta \) denote \( \hat{\theta}(\theta) \in \mathcal{N}_s \) the closest point in \( \mathcal{N}_s \) satisfying in particular \( \|\theta(\theta) - \theta\| \leq \varepsilon \), we write, following previous arguments

\[
\sup_{\theta \in \Theta} |\tilde{g}_j^{TM}(\theta) - g_j(\theta)| 
\leq \sup_{\theta \in \Theta} |\tilde{g}_j^{TM}(\theta) - g_j(\hat{\theta}(\theta))| 
+ |g_j(\hat{\theta}(\theta)) - g_j(\theta)| 
\leq \sup_{\theta \in \Theta} |\tilde{g}_j^{TM}(\theta)| 
- g_j(\hat{\theta}(\theta)) + \varepsilon L_j 
= \max_{\theta \in \mathcal{N}_s} |\tilde{g}_j^{TM}(\theta)| 
- g_j(\hat{\theta}) + \varepsilon L_j.
\]

Taking union bound over \( \hat{\theta} \in \mathcal{N}_s \) for the inequality (B21) and choosing \( \varepsilon = n^{-\alpha/(1+\alpha)} \) concludes the proof of Proposition 5.

### B.10 Proof of Corollary 1

We first write the result of Proposition 5 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:

\[
|\epsilon_j^{TM}(\theta)| \leq O\left(M_{j,a}^{1/(1+\alpha)} \left( \log(d/\delta) + d \log(n) \right)^{\alpha/(1+\alpha)} \right)
\]

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

\[
(R(\hat{\theta}^{(0)}) - R^*) \left(1 - \frac{\lambda}{\sum_{j \in [d]} L_j} \right)^T 
\leq \|\epsilon^{TM}(\delta)\|_2^2
\]

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

### B.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 Holland et al. (2019) with \( \delta/2 \) to obtain:

\[
\frac{1}{2} \left| \tilde{g}_j^{TM}(\theta) - g_j(\theta) \right| \leq C \sigma_j^2 \frac{s}{n} + s \log(4\delta^{-1})
\]

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 \( \psi \), the Gudemanian function, satisfies the previous inequality for \( C = 1/2 \). This, along with the choice of scale \( s \) according to (21) and our assumption on \( \hat{\sigma}_j^2 \) yields the announced deviation bound by a simple union bound argument.

### B.12 Proof of Proposition 6

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 \( \zeta(x, (x_i)_{i \in [n]}) \) as the unique \( x \) satisfying

\[
\sum_{i \in [n]} \psi\left( \frac{x - x_i}{s} \right) = 0.
\]
Since the function $\psi$ is increasing the previous equation has a unique solution. Moreover, for fixed scale $s$, the function $\zeta_s((x_i)_{i\in[n]})$ is monotomous non decreasing w.r.t. each $x_i$ when the others are fixed.

**Step 1**

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

$$
\left| \hat{g}_j^{\CH}(\tilde{\theta}) - g_j(\theta) \right| \leq \left| \hat{g}_j^{\CH}(\tilde{\theta}) - g_j(\theta) \right| + \varepsilon L_j,
$$

where we have the $\CH$ estimator $\hat{g}_j^{\CH}(\tilde{\theta}) = \zeta_s(\theta)((g_j^i(\theta))_{i\in[\varepsilon]})$ with scale $s(\theta)$ computed according to (21) and (22).

Assume, without loss of generality that $\hat{g}_j^{\CH}(\theta) - g_j(\theta) \geq 0$. Using the non-decreasing property of the $\CH$ estimator at a fixed scale, we find that

$$
\left| \hat{g}_j^{\CH}(\tilde{\theta}) - g_j(\theta) \right| = \left| \zeta_s(\theta)((g_j^i(\theta))_{i\in[\varepsilon]}) - g_j(\theta) \right| \leq \left| \zeta_s(\theta)((g_j^i(\Theta) + \varepsilon \gamma \|X_i^j\|^2)_{i\in[\varepsilon]}) - g_j(\theta) \right|.
$$

Indeed, one has

$$
g_j^i(\theta) = g_j^i(\theta) + (g_j^i(\Theta) - g_j^i(\theta)) \leq g_j^i(\Theta) + \varepsilon \gamma \|X_i^j\|^2 = g_j^i(\Theta) + \varepsilon \gamma \|X_i\|^2.
$$

We introduce the notation $\hat{g}_j^{\CH}(\tilde{\theta}) := \zeta_s(\theta)((g_j^i(\Theta) + \varepsilon \gamma \|X_i^j\|^2)_{i\in[\varepsilon]})$ so that:

$$
\left| \hat{g}_j^{\CH}(\tilde{\theta}) - g_j(\theta) \right| \leq \left| \hat{g}_j^{\CH}(\tilde{\theta}) - g_j(\tilde{\theta}) \right|.
$$

**Step 2**

We now use the concentration property of $\CH$ to bound the previous quantity which is in terms of $\tilde{\theta}$. We apply Lemma 1 from Holland et al. (2019) with $\delta/2$ and scale $s(\theta)$ to the samples $(g_j^i(\Theta) + \varepsilon \gamma \|X_i^j\|^2)_{i\in[\varepsilon]}$ which are independent and distributed according to the random variable $\mathcal{L}'(\Theta^j X, \gamma^j X^j + \varepsilon \gamma \|X_i\|^2)$ with expectation $g_j(\tilde{\theta}) + \varepsilon \mathbb{L}$. Using our assumptions on $\sigma_L, \sigma_j(\theta), \sigma_j(\tilde{\theta}), \tilde{\sigma}_j(\theta)$ and the definition of the scale $s(\theta)$ according to (21) we find:

$$
\frac{1}{2} \left| \hat{g}_j^{\CH}(\tilde{\theta}) - g_j(\tilde{\theta}) - \varepsilon \mathbb{L} \right| \leq \frac{1}{2} \left| \zeta(\zeta_s((g_j^i(\tilde{\theta}) + \varepsilon \gamma \|X_i^j\|^2)_{i\in[\varepsilon]})
$$

Taking union bound over $\tilde{\theta} \in N_\varepsilon$ for the inequality (B22) and using the choice $\varepsilon = 1/\sqrt{n}$ concludes the proof of Proposition 6.
B.13 Proof of Corollary 2

Under the assumptions made, the constants \((L_j)_{j \in [a]}\) are estimated using the MOM estimator and we obtain the bounds \((\hat{L}_j)_{j \in [a]}\) 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 \(\delta\) and replacing the constants \((L_j)_{j \in [a]}\) by their upper-bounds accordingly. The result then follows after a simple union bound argument.

B.14 Proof of Lemma 5

Let \(B_1, \ldots, B_K\) 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 \(K\) denote the uncorrupted block indices \(K = \{k \in [K] \mid B_k \cap \mathcal{O} = \emptyset\}\) and assume \(|\mathcal{O}| \leq (1 - \epsilon)K/2\). For \(k \in [K]\) let \(\hat{\sigma}_k^2 = \frac{1}{K} \sum_{i \in B_k} X_i^2\) 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

\[
\left|\hat{\sigma}_k^2 - \sigma^2\right| \leq \left(\frac{3 \mathbb{E}|X^2 - \sigma^2|^{1+\alpha}}{\delta N^\alpha}\right)^{\frac{1}{1+\alpha}} \\
\leq \left(\frac{3}{\delta N^\alpha}\right)^{\frac{1}{1+\alpha}} C\mathbb{E}|X^2 - \sigma^2| \leq \left(\frac{3}{\delta N^\alpha}\right)^{\frac{1}{1+\alpha}} C\sigma^2,
\]

which implies the inequality

\[
\sigma^2 \leq \left(1 - C\left(\frac{3}{\delta N^\alpha}\right)^{\frac{1}{1+\alpha}}\right)^{-1} \hat{\sigma}_k^2.
\]

Define the Bernoulli random variables \(U_k = 1\{\sigma^2 > \left(1 - C\left(\frac{3}{\delta N^\alpha}\right)^{\frac{1}{1+\alpha}}\right)^{-1} \hat{\sigma}_k^2\}\) 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:

\[
P\left(\left(1 - C\left(\frac{3}{\delta N^\alpha}\right)^{\frac{1}{1+\alpha}}\right)^{-1} \hat{\sigma}_k^2 < \sigma^2\right) \\
\leq P(S > K/2 - |\mathcal{O}|) \\
= P(S - ES > K/2 - |\mathcal{O}| - \delta|\mathcal{K}|) \\
\leq P(S - ES > K(\epsilon - 2\delta)/2) \\
\leq \exp\left(-K(\epsilon - 2\delta)^2/2\right),
\]

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

B.15 Proof of Lemma 6

Lemma 6 is a direct consequence of the following result.

\textbf{Lemma 10} Let \(X_1, \ldots, X_n\) 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 \(L^{(1+\alpha)}-L^1\) condition with constant \(C > 1\). Let \(\hat{\mu}\) be the median-of-means estimator of \(\mu\) with \(K\) blocks and \(\hat{m}_{\alpha}\) a similarly obtained estimate of \(m_{\alpha}\) from the samples \((|X_i - \hat{\mu}|^{1+\alpha})_{i \in [n]}\). Then, with probability at least \(1 - 2\exp(-K/18)\) we have

\[
\hat{m}_{\alpha} \geq (1 - \kappa) m_{\alpha},
\]

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

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

\[
\|\mu - \hat{\mu}\| > (24m_{\alpha})^{\frac{1}{1+\alpha}} \left(\frac{K}{n}\right)^{\frac{\alpha}{1+\alpha}}.
\]

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

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

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

\[
\frac{1}{N} \sum_{i \in B_j} |X_i - \hat{\mu}|^{1+\alpha} \\
= \frac{1}{N} \sum_{i \in B_j} |(X_i - \mu) + (\mu - \hat{\mu})|^{1+\alpha} \\
\geq \frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} + \frac{1}{N} (1 + \alpha) \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} \cdot |\mu - \hat{\mu}| \\
- |\mu|^{1+\alpha} \text{sign}(X_i - \mu) \cdot |\mu - \hat{\mu}| \\
\geq \frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} - (1 + \alpha) |\mu| \\
- \hat{\mu} \left[\frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha}\right] \\
\geq \frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} - (1 + \alpha) |\mu| \\
- \hat{\mu} \left[\frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha}\right]^{\frac{1}{1+\alpha}},
\]

as required.
where the last step uses Jensen’s inequality. Using Lemma 7
we have, for \( \delta > 0 \), the concentration bound
\[
P \left( \left| \frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} - m_\alpha \right| > \epsilon \right) 
\leq \frac{3 \epsilon}{\delta N^\alpha} \leq \frac{3 \times 2^\alpha \epsilon (1 + C(1+\alpha)^2) N^\alpha}{\epsilon^{1+\alpha}}.
\]
which, using that \( X \) satisfies the \( L^{(1+\alpha)} \)-condition, translates to
\[
P \left( \left| \frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} - m_\alpha \right| > \epsilon m_\alpha \right) 
\leq \frac{3 \times 2^\alpha (1 + C(1+\alpha)^2) N^\alpha}{\epsilon^{1+\alpha}}.
\]
Replacing \( \epsilon \) with \( \epsilon m_\alpha \) we find
\[
P \left( \left| \frac{1}{N} \sum_{i \in B_j} |X_i - \mu|^{1+\alpha} - m_\alpha \right| > \epsilon m_\alpha \right) 
\leq \frac{3 \times 2^\alpha (1 + C(1+\alpha)^2) N^\alpha}{\epsilon^{1+\alpha}}.
\]
Now conditioning on the event (B23) and using the previous bound with \( \epsilon = \left( \frac{3 \times 2^\alpha (1 + C(1+\alpha)^2)}{N^\alpha} \right)^{1/\alpha} \) in (B24), we obtain that
\[
P \left( \left| \frac{1}{N} \sum_{i \in B_j} |X_i - \hat{\mu}|^{1+\alpha} \leq (1 - \epsilon)m_\alpha (1 + \alpha) \right) \right.
\]
\[
\left. \left( \frac{24 m_\alpha}{N^\alpha} \right)^{1/\alpha} \right) \leq \delta
\]
\[
\implies \mathbb{P} \left( \frac{1}{N} \sum_{i \in B_j} |X_i - \hat{\mu}|^{1+\alpha} \right) \leq \left( \frac{1 - \epsilon}{N} \right) \left( \frac{1 + \epsilon}{N} \right)^{1/\alpha} m_\alpha \leq \delta.
\]
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 \( \delta \). We can use the MOM trick, assuming the number of outliers satisfies \( |O| \leq K (1 - \epsilon)/2 \) for \( \epsilon \in (0, 1) \), we have for \( S = \sum_j U_j \)
\[
P(S > K/2 - |O|) = \mathbb{P}(S - E S > K/2 - |O| - \delta |K|) \leq \mathbb{P}(S - E S > K (\epsilon - 2 \delta)/2) \leq \exp(-K (\epsilon - 2 \delta)^2/2).
\]
Taking \( \epsilon = 5/6 \) 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 \( |\mu - \hat{\mu}| \) 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
\[
m_\alpha \geq (1 - \kappa) m_\alpha
\]
holds with probability at least \( 1 - 2 \exp(-K/18) \).

**B.16 Proof of Theorem 7**

This proof is inspired from Theorem 5 in Nesterov (2012) and Theorem 1 in Shalev-Shwartz and Tewari (2011) while keeping track of the degradations caused by the errors on the gradient coordinates.

We condition on the event (B5) and denote \( \epsilon_j = \epsilon_j(\delta) \) and \( \epsilon_{Euc} = \|\epsilon(\delta)\|_2 \). We define for all \( \theta \in \Theta \)
\[
u_j(\theta) = \arg \min_{\theta \in \Theta} \hat{\mu}_j(\theta) t_j(\theta - \tau_\theta) + \frac{L_j}{2} (\theta - \tau_\theta)^2
\]
\[
\implies \mu_j(\theta) = \gamma_j(\theta - \tau_\theta) \mu_j(\theta) = \gamma_j(\theta - \tau_\theta)
\]
and denote \( \theta^{(t)} \) the optimization iterates for \( t = 0, \ldots, T \) and \( f_j \) the random coordinate sampled at step \( t \) and let \( \hat{\mu}_j(\theta) \) for brevity. We have that \( \mu_j(\theta) \) satisfies the following optimality condition
\[
\forall \theta \in \Theta_j \left( \hat{\mu}_j(t) + L_j (u_j(\theta) - \hat{\mu}_j(t)) \right)
\]
\[
\implies \mu_j(\theta) \geq 0,
\]
where \( \rho_j = \sign(u_j(\theta) - \hat{\mu}_j(t)) \). Using this condition for \( \theta = \hat{\theta}_j^{(t)} \) and the coordinate-wise Lipschitz smoothness property of \( K \) we find
\[
R(\theta^{(t+1)}) \leq R(\theta^{(t)}) + g_j(\theta^{(t)}) (u_j(\theta^{(t)}) - \theta_j^{(t)})
\]
\[
\leq R(\theta^{(t)}) + (\hat{\mu}_j + \epsilon_j \rho_j) (u_j(\theta^{(t)}) - \theta_j^{(t)})
\]
\[
\leq R(\theta^{(t)}) + \frac{L_j}{2} (u_j(\theta^{(t)}) - \theta_j^{(t)})^2
\]
\[
\leq R(\theta^{(t)}) - \frac{L_j}{2} (u_j(\theta^{(t)}) - \hat{\mu}_j(t))^2.
\]
Defining the potential $\Phi(\theta) = \sum_{j=1}^{d} L_j(\theta_j - \theta_j^*)^2$, we have:

$$
\Phi(\theta^{(t+1)}) = \Phi(\theta^{(t)}) + 2L_j(u_j(\theta^{(t)}) - \theta_j^*)^2
\leq \Phi(\theta^{(t)}) + 2L_j(u_j(\theta^{(t)}) - \theta_j^*)^2
- L_j(u_j(\theta^{(t)}) - \theta_j^*)^2
\leq \Phi(\theta^{(t)}) - 2(\delta_t + \epsilon_j, \rho_0)(u_j(\theta^{(t)}) - \theta_j^*)^2
\leq \Phi(\theta^{(t)}) - 2(\delta_t + \epsilon_j, \rho_0)(u_j(\theta^{(t)}) - \theta_j^*)^2
- 2(\delta_t + \epsilon_j, \rho_0)(u_j(\theta^{(t)}) - \theta_j^*)^2
+ 4\epsilon_j|\theta_j^* - \theta_j^*|.
$$

where the first inequality uses the optimality condition with $\theta = \theta_j^*$ and the second one uses (B25). 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

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

Recall that according to (B26), we have $R(\theta^{(t+1)}) \leq R(\theta^{(t)})$, summing over $t = 0, \ldots, T$ we find:

$$
\mathbb{E} \left[ \frac{T + 1}{d} (R(\theta^{(T)}) - R(\theta^*)) \right]
\leq \mathbb{E} \left[ \frac{1}{d} \sum_{t=0}^{T} (R(\theta^{(t)}) - R(\theta^*)) \right]
\leq \sum_{t=0}^{T} \left( \mathbb{E}[\Psi(\theta^{(t)}) - \Psi(\theta^{(t+1)})] + \frac{2\epsilon_{Euc}}{d} \mathbb{E}[\|\theta^{(t)} - \theta^*\|_2] \right)
= \mathbb{E} \left[ \Psi(\theta^{(0)}) - \Psi(\theta^{(t+1)}) \right] + \frac{2\epsilon_{Euc}}{d} \sum_{t=0}^{T} \mathbb{E}[\|\theta^{(t)} - \theta^*\|_2]
\leq \Psi(\theta^{(0)}) + \frac{2\epsilon_{Euc}}{d} \sum_{t=0}^{T} \mathbb{E}[\|\theta^{(t)} - \theta^*\|_2],
$$

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 $\|\theta^{(t+1)} - \theta^*\|_2 \leq \|\theta^{(t)} - \theta^*\|_2$. Indeed a straightforward computation yields

$$
\|\theta^{(t+1)} - \theta^*\|_2^2 = \|\theta^{(t)} - \theta^*\|_2^2 + (u_j(\theta^{(t)}) - \theta_j^*)^2
+ 2(u_j(\theta^{(t)}) - \theta_j^*)(\theta_j^* - \theta_j^*).
$$

We need to show that $\delta_t^2 \leq -2\delta_t(\theta_j^* - \theta_j^*)$ with $\delta_t = (u_j(\theta^{(t)}) - \theta_j^*)$. Notice that $\delta_t$ always has the opposite sign of $g_j(\theta^{(t)})$ (thanks to the thresholding) so by convexity of $R$ along the coordinate $j_t$ we have $\delta_t(\theta_j^* - \theta_j^*) \leq 0$ and so it is down to showing $|\delta_t| \leq 2|\theta_j^{(t)} - \theta_j^*|$ which can be seen from

$$
|\delta_t| \leq \frac{|g_j(\theta^{(t)})|}{L_{j_t}} = \frac{|g_j(\theta^{(t)}) - g_j(\theta^*)|}{L_{j_t}} \leq |\theta_j^{(t)} - \theta_j^*|,
$$

which concludes the proof of Theorem 7.

References

Acharya, A., Hashemi, A., Jain, P., Sanghavi, S., Dhillon, I.S., Topcu, U.: Robust training in high dimensions via block coordinate geometric median descent. In: International Conference on Artificial Intelligence and Statistics, pp. 11145–11168 (2022).
Alon, N., Matias, Y., Szegedy, M.: The space complexity of approximating the frequency moments. J. Comput. Syst. Sci. 58(1), 137–147 (1999).
Armijo, L.: Minimization of functions having Lipschitz continuous first partial derivatives. Pac. J. Math. 16(1), 1–3 (1966).
Audibert, J.-Y., Munos, R., Szepesvári, C.: Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoret. Comput. Sci. 410(19), 1876–1902 (2009). (Algorithmic Learning Theory)
Ballester-Ripoll, R., Paredes, E.G., Pajarola, R.: Sobol tensor trains for global sensitivity analysis. Reliab. Eng. Syst. Saf. 183, 311–322 (2019).
Barlett, P.L., Bousquet, O., Mendelson, S.: Local rademacher complexities. Ann. Stat. 33(4), 1497–1537 (2005).
Beck, A., Tsetlatsivili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. 23(4), 2037–2060 (2013).
Bhatia, K., Jain, P., Kamalaruban, P., Kar, P.: Consistent robust regression. Adv. Neural. Inf. Process. Syst. 30, 2110–2119 (2017).
Blondel, M., Seki, K., Uehara, K.: Block coordinate descent algorithms for large-scale sparse multiclass classification. Mach. Learn. 93(1), 31–52 (2013).
Boucheron, S., Lugosi, G., Massart, P., Ledoux, M.: Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford (2013).
Brownlee, C., Joly, E., Lugosi, G., et al.: Empirical risk minimization for heavy-tailed losses. Ann. Stat. 43(6), 2507–2536 (2015).
Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends® Mach. Learn. 8(3–4), 231–357 (2015).
Bubeck, S., Cesa-Bianchi, N., Lugosi, G.: Bandits with heavy tail. IEEE Trans. Inf. Theory 59(11), 7171–7177 (2013).
Candanedo, L.M., Feldheim, V.: Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy Build. 112, 28–39 (2016).
Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736 (2010)

Gao, C., et al.: Robust regression via multivariate regression depth. Bernoulli 26(2), 1139–1170 (2020)

Geer, S.A., van de Geer, S.: Empirical Processes in M-estimation, vol. 6. Cambridge University Press, Cambridge (2000)

Genkin, A., Lewis, D.D., Madigan, D.: Large-scale bayesian logistic regression for text categorization. Technometrics 49(3), 291–304 (2007)

Geoffrey, C., Guillaume, L., Matthieu, L.: Robust high dimensional learning for Lipschitz and convex losses. J. Mach. Learn. Res. 21 (2020)

Gubbs, F.E.: Procedures for detecting outlying observations in samples. Technometrics 11(1), 1–21 (1969)

Gupta, A., Kohli, S.: An MCDM approach towards handling outliers in web data: a case study using OWA operators. Artif. Intell. Rev. 46(1), 59–82 (2016)

Hampel, F.R.: A general qualitative definition of robustness. Ann. Math. Stat. 42(6), 1887–1896 (1971)

Hampel, F.R., Ronchetti, E.M., Rousseuw, P., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley-Interscience, New York (1986)

Hawkins, D.M.: Identification of Outliers, vol. 11. Springer, Berlin (1980)

Hoare, C.A.R.: Algorithm 65: find. Commun. ACM 4(7), 321–322 (1961)

Holland, M.: Robustness and scalability under heavy tails, without strong convexity. In International Conference on Artificial Intelligence and Statistics, pp. 865–873 (2021)

Holland, M., Ikeda, K.: Better generalization with less data using robust gradient descent. In International Conference on Machine Learning, pp. 2761–2770 (2019)

Holland, M.J.: Robust descent using smoothed multiplicative noise. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 703–711 (2019)

Holland, M.J., Ikeda, K.: Efficient learning with robust gradient descent. Mach. Learn. 108(8), 1523–1560 (2019)

Hopkins, S.B.: Mean estimation with sub-Gaussian rates in polynomial time. Ann. Stat. 48(2), 1193–1213 (2020)

Hsu, D., Sabato, S.: Loss minimization and parameter estimation with heavy tails. J. Mach. Learn. Res. 17(1), 543–582 (2016)

Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. 35(1), 73–101 (1964)

Huber, P.J.: The 1972 wald lecture robust statistics: a review. Ann. Math. Stat. 43(4), 1041–1067 (1972)

Huber, P.J.: Robust Statistics. Wiley, New York (1981)

Jerrum, M.R., Valiant, L.G., Vazirani, V.V.: Random generation of combinatorial structures from a uniform distribution. Theoret. Comput. Sci. 43, 169–188 (1986)

Juditsky, A., Kunluchakov, A., Tsytsev, H.: Sparse recovery by reduced variance stochastic approximation. Inf. Inference: J. IMA 12(2), 851–896 (2022)

Klivans, A., Kothari, P.K., Meka, R.: Efficient algorithms for outlier-robust regression. In: Conference On Learning Theory, pp. 1420–1430 (2018)

Klivans, A.R., Long, P.M., Servedio, R.A.: Learning halfspaces with malicious noise. J. Mach. Learn. Res. 10(12) (2009)

Knuth, D.E.: Seminumerical algorithms (the art of computer programing 2). AddisonWesley, Reading, MA, pp. 124–125 (1969)

Koklu, M., Ozkan, I.A.: Multiclass classification of dry beans using computer vision and machine learning techniques. Comput. Electron. Agric. 174, 105507 (2020)

Koltchinskii, V.: Local Rademacher complexities and oracle inequalities in risk minimization. Ann. Stat. 34(6), 2593–2656 (2006)

Kuhn, M., Johnson, K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press (2019)
Lai, K.A., Rao, A.B., Vempala, S.: Agnostic estimation of mean and covariance. In: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pp. 665–674. IEEE (2016)

Lecué, G., Lerasle, M., et al.: Robust machine learning by median-of-means: theory and practice. Ann. Stat. 48(2), 906–931 (2020)

Lecué, G., Lerasle, M., Mathieu, T.: Robust classification via MOM minimization. Mach. Learn. 109(8), 1635–1665 (2020)

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

Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes. Springer, Berlin (1991)

Lei, Z., Luh, K., Venkat, P., Zhang, F.: A fast spectral algorithm for mean estimation with sub-Gaussian rates. In: Conference on Learning Theory, pp. 2598–2612 (2020)

Li, J.: Robust sparse estimation tasks in high dimensions. arXiv preprint arXiv:1702.05860 (2017)

Li, X., Zhao, T., Arora, R., Liu, H., Hong, M.: On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. J. Mach. Learn. Res. 18(1), 6741–6764 (2017)

Liu, L., Li, T., Caramanis, C.: High dimensional robust estimation of sparse models via trimmed hard thresholding. arXiv preprint arXiv:1901.08237 (2019)

Liu, L., Shen, Y., Li, T., Caramanis, C.: High dimensional robust sparse regression. In: International Conference on Artificial Intelligence and Statistics, pp. 411–421 (2020)

Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Trans. Pattern Anal. Mach. Intell. 38(3), 447–461 (2015)

Lugosi, G., Mendelson, S.: Mean estimation and regression under heavy-tailed distributions: a survey. Found. Comput. Math. 19(5), 1145–1190 (2019)

Lugosi, G., Mendelson, S.: Sub-gaussian estimators of the mean of a random vector. Ann. Stat. 47(2), 783–794 (2019)

Lugosi, G., Mendelson, S.: Robust multivariate mean estimation: the optimality of trimmed mean. Ann. Stat. 49(1), 393–410 (2021)

Massart, P., Nédélec, É.: Risk bounds for statistical learning. Ann. Stat. 34(5), 2326–2366 (2016)

Maurer, A., Pontil, M.: Empirical Bernstein bounds and sample-variance penalization. In: COLT (2009)

Minsker, S., et al.: Geometric median and robust estimation in banach spaces. Bernoulli 21(4), 2308–2335 (2015)

Minsker, S., et al.: Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. Ann. Stat. 46(6A), 2871–2903 (2018)

Mizera, I., et al.: On depth and deep points: a calculus. Ann. Stat. 30(6), 1681–1736 (2002)

Mnih, V., Szepesvári, C., Audibert, J.-Y.: Empirical Bernstein stopping. In: Proceedings of the 25th International Conference on Machine Learning, pp. 672–679 (2008)

Nemirovskij, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience (1983)

Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer, New York (2004)

Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. 22(2), 341–362 (2012)

Owen, A.: A robust hybrid of lasso and ridge regression. Contemp. Math. 443(7), 59–72 (2007)

Paul, D., Chakraborty, S., Das, S.: Robust principal component analysis: a median of means approach. arXiv preprint arXiv:2102.03403 (2021)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830 (2011)

Pensia, A., Jog, V., Loh, P.-L.: Robust regression with covariate filtering: heavy tails and adversarial contamination. arXiv preprint arXiv:2009.12976 (2020)

Prasad, A., Balakrishnan, S., Ravikumar, P.: A robust univariate mean estimator is all you need. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, vol. 108, pp. 4034–4044 (2020)

Prasad, A., Suggala, A.S., Balakrishnan, S., Ravikumar, P.: Robust estimation via robust gradient estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(3), 601–627 (2020)

Shalev-Shwartz, S., Tewari, A.: Stochastic Methods for $\ell_1$-Regularized Loss Minimization. The Journal of Machine Learning Research 12, 1865–1892 (2011)

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

Srebro, Nathan, Srikanth, K., Tewari, Ambuj: Optimalistic methods for learning with a smooth loss. arXiv preprint arXiv:1009.3896, (2010)

Jiuyan, T., Liu, W., Mao, X., Chen, X.: Variance Reduced Median-of-Means Estimator for Byzantine-Robust Distributed Inference. J. Mach. Learn. Res. 22(84), 1–67 (2021)

Tukey, John W.: A survey of sampling from contaminated distributions. Contributions to Probability and Statistics, pp. 448–485, (1960)

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

van Erven, Tim, Sachs, Sarah, Kolen, Wouter M., Kotlowski, Wojciech: Robust Online Convex Optimization in the Presence of Outliers. In: Proceedings of Thirty Fourth Conference on Learning Theory, volume 134, pp. 4174–4194. PMLR, (2021)

Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York, NY (1999)

Vardi, Y., Zhang, C.-H.: The multivariate $L_1$-median and associated data depth. Proc. Natl. Acad. Sci. 97(4), 1423–1426 (2000)

Vergara, A., Vembu, S., Ayhan, T., Ryan, M.A., Homer, M.L., Huerta, R.: Chemical gas sensor drift compensation using classifier ensembles. Sens. Actuators, B Chem. 166, 320–329 (2012)

Wright, S.J.: Coordinate descent algorithms. Math. Program. 151(1), 3–34 (2015)

Wu, T.T., Lange, K.: Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics 2(1), 224–244 (2008)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)

Zhang, Lijun, Zhou, Zhi-Hua: $\ell_1$-regression with heavy-tailed distributions. In: Advances in Neural Information Processing Systems, pp. 1084–1094, (2018)

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

Zheng, Alice, Casari, Amanda: Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O’Reilly Media, Inc., (2018)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.