Robust Methods for High-Dimensional Linear Learning

Ibrahim Merad   IMERAD@LPSM.PARIS
LPSM, UMR 8001, Université Paris Diderot, Paris, France

Stéphane Gaïffas   STEPHANE.GAIFFAS@LPSM.PARIS
LPSM, UMR 8001, Université Paris Diderot, Paris, France and DMA, École normale supérieure

Abstract

We propose statistically robust and computationally efficient linear learning methods in the high-dimensional batch setting, where the number of features d may exceed the sample size n. We employ, in a generic learning setting, two algorithms depending on whether the considered loss function is gradient-Lipschitz or not. Then, we instantiate our framework on several applications including vanilla sparse, group-sparse and low-rank matrix recovery. This leads, for each application, to efficient and robust learning algorithms, that reach near-optimal estimation rates under heavy-tailed distributions and the presence of outliers. For vanilla s-sparsity, we are able to reach the $s \log(d)/n$ rate under heavy-tails and η-corruption, at a computational cost comparable to that of non-robust analogs. We provide an efficient implementation of our algorithms in an open-source `Python` library called `linlearn`, by means of which we carry out numerical experiments which confirm our theoretical findings together with a comparison to other recent approaches proposed in the literature.

Keywords: robust methods, heavy-tailed data, outliers, sparse recovery, mirror descent, generalization error

1. Introduction

Learning from heavy tailed or corrupted data is a long pursued challenge in statistics receiving considerable attention in literature (Huber, 1964; Hampel, 1971; Huber, 2004; Prasad et al., 2020; Diakonikolas et al., 2019; Audibert and Catoni, 2011) and gaining an additional degree of complexity in the high-dimensional setting (Lugosi and Mendelson, 2019c; Lecué and Lerasle, 2020; Dalalyan and Thompson, 2019; Balakrishnan et al., 2017; Liu et al., 2019, 2020; Fan et al., 2021). Sparsity inducing penalization techniques (Donoho et al., 2000; Bühlmann and Van De Geer, 2011) are the go-to approach for high-dimensional data and have found many applications in modern statistics (Tibshirani, 1996; Bühlmann and Van De Geer, 2011; Donoho et al., 2000; Hastie et al., 2015). A clear favourite is the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). Theoretical studies have shown that under the so-called Restricted Eigenvalue (RE) condition, the latter achieves a nearly-optimal estimation rate (Bickel et al., 2009; Bühlmann and Van De Geer, 2011; Negahban et al., 2012). Further, a rich literature extensively studies the oracle performances of LASSO in various contexts and conditions (Bunea et al., 2007; Lounici, 2008; Zhang and Huang, 2008; Zhang, 2009; Zhao and Yu, 2006; Zou, 2006; Van de Geer, 2008; Lecué and Mendelson, 2018; Bellec et al., 2018). Other penalization techniques induce different sparsity patterns or lead to different statistical guarantees, such as, to cite but a few, SLOPE (Bogdan et al., 2015; Su and Candès, 2016) which is adaptive to the unknown sparsity and leads to the optimal

estimation rate, OSCAR (Bondell and Reich, 2008) which induces feature grouping or group-ℓ_1 penalization (Yuan and Lin, 2006; Huang et al., 2010) which induces block-sparsity. Other approaches include for instance Iterative Hard Thresholding (IHT) (Blumensath and Davies, 2009, 2010; Jain et al., 2014; Shen and Li, 2017; Jain et al., 2016) whose properties are studied under the Restricted Isometry Property (RIP). Another close problem is low-rank matrix recovery, involving the nuclear norm as a low-rank inducing convex penalization (Koltchinskii et al., 2011; Candès and Recht, 2009; Candès and Plan, 2011; Rohde and Tsybakov, 2011; Negahban and Wainwright, 2011, 2012).

The high-dimensional statistical inference methods cited above are, however, not robust: theoretical guarantees are derived under light-tails (generally sub-Gaussian) and the i.i.d assumption. Unfortunately, these assumptions fail to hold in general, for instance, it is known that financial and biological data often displays heavy-tailed behaviour (Fan et al., 2021) and outliers or corruption are not uncommon when handling massive amounts of data which are tedious to thoroughly clean. Moreover, the majority of the previous references focus on the oracle performance of estimators as opposed to providing guarantees for explicit algorithms to compute them. A natural question is therefore : *can one build alternatives to such high-dimensional estimators that are robust to heavy tails and outliers, that are computationally efficient and achieve rates similar to their non-robust counterparts ?* Recent advances about robust mean estimation (Catoni, 2012; Lugosi and Mendelson, 2021; Diakonikolas et al., 2020; Depersin and Lecué, 2019; Lei et al., 2020) gave a strong impulse in the field of robust learning (Lecué et al., 2020; Prasad et al., 2020; Holland and Ikeda, 2019; Diakonikolas et al., 2019; Cherapanamjeri et al., 2020; Bakshi and Prasad, 2021), including the high-dimensional setting (Liu et al., 2019, 2020; Balakrishnan et al., 2017) which led to significant progress towards a positive answer to this question.

However, to the best of our knowledge, the solutions proposed until now are all suboptimal in one way or another. The shortcomings either lie in the obtained statistical rate: which is sometimes significantly far from optimal, or in the robustness: most works consider heavy tailed and corrupted data separately and only very limited amounts of corruption, or in computational complexity: some corruption-filtering algorithms are too heavy and do not scale to real world applications.

In this paper, we propose explicit algorithms to solve multiple sparse estimation problems with high performances in all previous aspects. In particular, our algorithm for vanilla sparse estimation enjoys a nearly optimal statistical rate (up to a logarithmic factor), is simultaneously robust to heavy tails and strong corruption (when a fraction of the data is corrupted) and has a comparable computational complexity to a non robust method.

1.1 Main contributions.

This paper combines non-Euclidean optimization algorithms and robust mean estimators of the gradient into explicit algorithms in order to achieve the following main contributions.

- We propose a framework for robust high-dimensional linear learning in the batch setting using two linearly converging stage-wise algorithms for high-dimensional optimization based on Mirror Descent and Dual Averaging. These may be applied for smooth and non-smooth objectives respectively so that most common loss functions are covered. The previous algorithms may be plugged with an appropriate gradient estimator to obtain explicit robust algorithms for solving a variety of problems.

- The central application of our framework is an algorithm for "vanilla" s-sparse estimation reaching the nearly optimal $s \log(d)/n$ statistical rate in the batch setting by combining stage-wise Mirror Descent (Section 3) with a simple trimmed mean estimator of the gradient. This algorithm is simultaneously robust to heavy-tailed distributions and η-corruption of the data.[1] This improves over previous literature which considered the two issues separately or required a very restricted value of the corruption rate η.

- In addition to vanilla sparsity, we instantiate our procedures for group sparse estimation and low-rank matrix recovery, in which different metrics on the parameter space are induced and used to measure the statistical error on the gradient (Section 5). For heavy-tailed data and η-corruption, the gradient estimator we propose for vanilla sparsity enjoys an optimal statistical rate with respect to the induced metric while the one proposed for group sparsity is nearly optimal up to a logarithmic factor. Moreover, for heavy tailed data and a limited number of outliers,[2] our proposed gradient estimator for low-rank matrix recovery is nearly optimal. Thus, our solutions to each of these problems are the most robust yet in the literature.

- Our algorithms offer a good compromise between robustness and computational efficiency with the only source of overhead coming from the robust gradient estimation component. In particular, for vanilla sparse estimation, this overhead is minimal so that the asymptotic complexity of our procedure is *equivalent* to that of non-robust counterparts. This is in contrast with previous works requiring costly sub-procedures to filter out corruption.

- We validate our results through numerical experiments using synthetic data for regression and real data sets for classification (Section 6). Our experiments confirm our mathematical results and compare our algorithms to concurrent baselines from literature.

- All algorithms introduced in this paper as well as the main baselines from literature we use for comparisons are implemented and easily accessible in a few lines of code through our `Python` library called `linlearn`, open-sourced under the BSD-3 License on `GitHub` and available here.[3]

1.2 Related Works

The general problem of robust linear learning was addressed by Gaïffas and Merad (2022) where the performance of coordinate gradient descent using various estimators was studied and experimentally evaluated. Several other works (Prasad et al., 2020; Pensia et al., 2020; Lecué et al., 2020; Holland and Ikeda, 2019) deal with this problem, however, they do not consider the high-dimensional setting.

The early work of Agarwal et al. (2012) focuses on vanilla sparse recovery in the stochastic optimization setting and uses a multistage annealed LASSO algorithm where the penalty shrinks progressively. The method reaches the nearly optimal $s \log(d)/n$ rate, however, it is not robust since the data is assumed i.i.d sub-Gaussian. The subsequent work of Sedghi et al. (2014) extends this framework to other problems such as additive sparse and low-rank matrix decomposition by changing the optimization algorithm but the sub-Gaussian assumption remains necessary.

---

1. We say that data is η-corrupted for some number $0 < \eta < 1/2$ if an η fraction of the samples is replaced by arbitrary (and potentially adversarial) outliers after data generation.
2. For low-rank matrix recovery, our estimator is based on Median-Of-Means so that the number of tolerated outliers is up to $K/2$ where K is the number of blocks used for estimation (see Section 5.3)
3. `https://github.com/linlearn/linlearn`

More recently, Juditsky et al. (2022) proposed a stochastic optimization mirror descent algorithm which computes multiple solutions on disjoint subsets of the data and aggregates them with a Median-Of-Means type procedure. The final solution achieves the rate $s \log(d)/n$ under s-vanilla sparsity with sub-Gaussian deviation and an application to low-rank matrix recovery is also developed. This aggregation method can handle some but not all heavy-tailed data. For instance, if the data follows a Pareto(α) distribution then the analysis yields a statistical error with a factor of order $d^{1/\alpha}$ which is not acceptable in a high-dimensional setting. Moreover, the presence of outliers is not considered so that the given bounds do not measure the impact of corruption. Nevertheless, the combination of the restarted mirror descent optimization procedure proposed in (Juditsky et al., 2022, Algorithm 1) with the proper robust gradient estimators yields a fast and highly robust learning algorithm in the batch setting which we present in Section 3. We also exploit the same core ideas in Section 4 to extend our framework to a wider range of objective functions.

Other works consider high-dimensional linear learning methods that are robust to corrupted data. An outlier robust method for mean and covariance estimation in the sparse η-contaminated[4] high-dimensional setting is proposed by Balakrishnan et al. (2017) along with theoretical guarantees. By extension, these also apply to several problems of interest such as sparse linear estimation or sparse GLMs. The idea is to use an SDP relaxation of sparse PCA (d'Aspremont et al., 2004) in order to adapt the filtering approach from (Diakonikolas et al., 2019), which relies on the covariance matrix to detect outliers, to the high-dimensional setting. However, the need to solve SDP problems makes the algorithm computationally costly. In addition, the true data distribution is assumed Gaussian and the considered η-contamination framework is weaker than η-corruption which allows for adversarial outliers.

The previous ideas were picked up again in the later work of Liu et al. (2020) who proposes a robust variant of IHT for sparse regression on η-corrupted data. Unfortunately, these results suffer from several shortcomings since data needs to be Gaussian with a known or sparse covariance matrix. Moreover, the gradient estimation subroutine, which is reminiscent of (Balakrishnan et al., 2017), is computationally heavy since it requires solving SDP problems as well and a number of samples scaling as s^2 instead of s is required. This seems to come from the fact that sparse gradients need to be estimated, in which case the s^2 dependence is unavoidable based on an oracle lower bound for such an estimation (Diakonikolas et al., 2017).

Recently, Dalalyan and Thompson (2019) derived oracle bounds for a robust estimator in the linear model with Gaussian design and a number of adversarially contaminated labels. Although optimal rates in terms of the corruption are achieved, this setting excludes corruption of the covariates and does not apply for heavy-tailed data distributions. Minsker et al. (2022) similarly consider the linear model and derive oracle bounds for a robustified SLOPE objective which is adaptive to the sparsity level. Remarkably, they achieve the optimal $s \log(ed/s)/n$ dependence in the heavy-tailed corrupted setting. However, corruption is restricted to the labels and the dependence of the result thereon significantly degrades if the covariates or the noise are not sub-Gaussian. In contrast, the very recent work of Sasai (2022) considers sparse estimation with heavy-tailed and η-corrupted data and derives a nearly optimal estimation bound using an algorithm which filters the data before running an ℓ_1-penalized robust Huber regression which corresponds to a similar approach to (Pensia et al., 2020) where the non sparse case was treated. Although the $s \log(d)/n$ rate is achieved with optimal robustness, this claim only applies for regression under the linear model with some restric-

---

4. η-contamination refers to the case where the data is sampled from a mixed distribution $(1 - \eta)P + \eta Q$ where P is the true data distribution and Q is an arbitrary one.

tive assumptions such as zero mean covariates. In addition, little attention is granted to the practical aspect and no experiments are carried out. A later extension (Sasai and Fujisawa, 2022) improves the statistical rate to $s \log(d/s)/n$ for sub-Gaussian covariates and, if the data covariance is known as well, better dependence on the corruption rate is obtained. Robust high-dimensional linear regression algorithms were recently surveyed by Filzmoser and Nordhausen (2021) with particular attention payed to methods based on dimension reduction, shrinkage and combinations thereof.

Finally, Liu et al. (2019) proposes an IHT algorithm using robust coordinatewise gradient estimators. These results cover the heavy-tailed and η-corrupted settings separately thanks to Median-Of-Means (Alon et al., 1999; Jerrum et al., 1986; Nemirovskij and Yudin, 1983) and Trimmed mean (Chen et al., 2013; Yin et al., 2018) estimators respectively. However, the corruption rate η is restricted to be of order at most $O(1/(\sqrt{s} \log(nd)))$ and the question of elaborating an algorithm which is simultaneously robust to both corruption and heavy tails is left open.

We summarize the settings and results of the previously mentioned works, along with ours on vanilla sparse estimation, in Table 1 which focuses on robust papers with explicit algorithms.

1.3 Agenda

The remainder of this document is structured as follows : Section 2 lays out the setting including the definition of the objective and our assumptions on the data. Sections 3 and 4 define optimization algorithms based on Mirror Descent and Dual Averaging addressing the cases of smooth and non-smooth losses respectively. Both Sections state convergence results for their respective algorithms. Section 5 considers instantiations of our general setting to vanilla sparse, group sparse and low-rank matrix estimation for a general loss. In each case, the norm $\| \cdot \|$ and dual norm $\| \cdot \|_*$ are instantiated and a robust and efficient gradient estimator is proposed so that, combined with the results of Sections 3 and 4, we obtain solutions with nearly optimal statistical rates (up to logarithmic terms) in each case. Finally, Section 6 presents numerical experiments on synthetic and real data sets which demonstrate the performance of our proposed methods and compare them with baselines from recent literature.

2. Setting, Notation and Assumptions

We consider supervised learning from a data set $(X_i, Y_i)_{i=1}^n$ from which the majority is distributed as a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ where the covariate space $\mathcal{X}$ is a high-dimensional Euclidean space and the label space $\mathcal{Y}$ is $\mathbb{R}$ or a finite set. The remaining minority of the data are called outliers and may be completely arbitrary or even adversarial. Given a loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ where $\widehat{\mathcal{Y}}$ is the prediction space, our goal is to optimize the unobserved objective

$$\mathcal{L}(\theta) = \mathbb{E}[\ell(\langle \theta, X \rangle, Y)] \tag{1}$$

over a convex set of parameters Θ, where the expectation is taken w.r.t. the joint distribution of (X, Y). Given an optimum $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} \mathcal{L}(\theta)$ (assumed unique), we are also interested in controlling the estimation error $\|\theta - \theta^\star\|$ where $\| \cdot \|$ is a norm on Θ, which will be defined according to each specific problem (see Section 5 below). Moreover, we assume that the optimal parameter is sparse according to an abstract sparsity measure $S : \Theta \to \mathbb{N}$.

Assumption 1 *The optimal solution $\theta^\star$ is s-sparse for some integer s smaller than the problem dimension i.e. $S(\theta^\star) \leq s$. Additionally, for any s-sparse vector $\theta \in \Theta$ we have the inequality $\|\theta\| \leq \sqrt{s}\|\theta\|_2$ and an upper bound $\bar{s} \geq s$ on the sparsity is known.*

5

Method	Statistical rate	Iteration Complexity	Data dist. and corruption	Loss
AMMD This paper Section 3	$O\left(\sqrt{s}\sqrt{\eta + \frac{\log(d/\delta)}{n}}\right)$ with $\mathbb{P} \geq 1 - \delta$	$O(nd)$	L_4 covariates, L_2 labels, η-corruption.	Lip. smooth, QM.
AMDA This paper Section 4	$O\left(\sqrt{s}\sqrt{\eta + \frac{\log(d/\delta)}{n}}\right)$ with $\mathbb{P} \geq 1 - \delta$	$O(nd)$	L_4 covariates, L_2 labels, η-corruption.	Lipschitz, PLM.
(Balakrishnan et al., 2017)	$O\Big(\|\theta^\star\|_2\big(\eta \log(1/\eta)+$ $s\sqrt{\log(d/\delta)/n}\big)\Big)$ with $\mathbb{P} \geq 1 - \delta$	$\Omega(nd^2+d^3)$	Gaussian, η-contamination.	LSQ, GLMs, Logit.
(Liu et al., 2020)	$O\big(\eta \vee s\sqrt{\log(d/\delta)/n}\big)$ with $\mathbb{P} \geq 1 - \delta$	$\Omega(nd^2+d^3)$	Gaussian, η-corruption.	LSQ.
(Liu et al., 2019) (MOM)	$O\left(\sqrt{s\log(d)/n}\right)$ with $\mathbb{P} \geq 1 - d^{-2}$	$O(nd)$	L_4 covariates, linear/logit model.	LSQ, Logit.
(Liu et al., 2019) (TMean)	$O\big(\eta\sqrt{s}\log(nd)$ $+\sqrt{s\log(d)/n}\big)$ with $\mathbb{P} \geq 1 - d^{-2}$	$O(nd\log(n))$	sub-Gaussian, η-corruption.	LSQ, Logit.
(Juditsky et al., 2020)	$O\big(\sqrt{s\log(d)\log(1/\delta)/n}\big)$ with $\mathbb{P} \geq 1 - \delta$	$O(d)$ (stochastic optim.)	L_2 gradient.	Lip. smooth, QM.
(Sasai, 2022)	$O\big(\sqrt{\eta}+\sqrt{s\log(d/\delta)/n}\big)$ with $\mathbb{P} \geq 1 - \delta$	Preliminary $\Omega(nd^2+d^3)$ then $O(nd)$	zero-mean & L_8 covariates, indep. noise, $n \gtrsim s^2\log(d/\delta)$.	Penalized Huber.

Table 1: Summary of the main hypotheses and results of our proposed algorithms and related works in the literature on vanilla sparse estimation. The statistical rate column gives the derived error bound on $\|\widehat{\theta} - \theta^\star\|_2$ between the estimated and true parameter and the associated confidence. In the "Data distribution and corruption" column, rows with no reference to corruption correspond to methods which do not consider it. In the "Loss" column, the following abbreviations are used : QM = quadratic minorization (Assumption 4), PLM = pseudo-linear minorization (Assumption 6), LSQ = least squares, GLM = generalized linear model, Logit = Logistic, Lip. smooth = Lipschitz smooth (gradient Lipschitz).

The precise notion of sparsity will be determined later in Section 5 through the sparsity measure m depending on the application at hand. The simplest case corresponds to the conventional notion of vector sparsity where $\mathcal{X} = \mathbb{R}^d$ for some large d, $\Theta \subset \mathbb{R}^d$ and the sparsity measure $S(\theta) = \sum_{j\in\llbracket d\rrbracket} \mathbf{1}_{\theta_j\neq 0}$ counts the number of non-zero coordinates. However, as we intend to also cover other forms of sparsity later, we do not fix this setting right away. Note that the required knowledge of $\bar{s} \geq s$ in Assumption 1 is common in the thresholding based sparse learning literature (Blumensath

and Davies, 2009, 2010; Jain et al., 2014, 2016; Juditsky et al., 2022). Adaptive methods to unknown sparsity exist (Bogdan et al., 2015; Bellec et al., 2018; Su and Candès, 2016) although designing a robust version thereof is beyond the scope of this work.

Since the objective (1) is not observed due to the distribution of the data being unknown, statistical approximation will be necessary in order to recover an approximation of $\theta^\star$. Instead of estimating the objective itself (which is of limited use for optimization), we will rather compute estimates of the gradient

$$g(\theta) := \nabla_\theta \mathcal{L}(\theta) \tag{2}$$

in order to run gradient based optimization procedures. Note that, since we consider the high-dimensional setting, a standard gradient descent approach is excluded since it would incur an error strongly depending on the problem dimension. In order to avoid this, one must use a non Euclidean optimization method as is customary for high-dimensional problems (Juditsky et al., 2022; Agarwal et al., 2012).

As commonly stated in the robust statistics literature (Catoni, 2012; Lugosi and Mendelson, 2021, 2019a), estimating an expectation using a conventional empirical mean only yields values with far from optimal deviation properties in the general case. Several estimators have been proposed which enjoy sub-Gaussian deviations from the true mean and robustness to corruption. Notable examples in the univariate case are the median-of-means (MOM) estimator (Alon et al., 1999; Jerrum et al., 1986; Nemirovskij and Yudin, 1983), Catoni's estimator (Catoni, 2012) and the trimmed mean (Lugosi and Mendelson, 2021). However, in the multivariate case (estimating the mean of a random vector), the optimal sub-Gaussian estimation rate cannot be obtained by a straightforward extension of the previous methods and a line of works (Lugosi and Mendelson, 2019d; Hopkins, 2018; Cherapanamjeri et al., 2019; Depersin and Lecué, 2019; Lugosi and Mendelson, 2021; Lei et al., 2020) has pursued elaborating efficient algorithms to achieve it. Most recently, Diakonikolas et al. (2020) managed to show that stability based estimators enjoy sub-Gaussian deviations while being robust to corruption of a fraction of the data. However, it is important to remember that all the works we just mentioned measure the estimation error using the Euclidean norm while many other choices are possible which may require the estimation algorithm to be adapted in order to achieve optimal deviations with respect to the chosen norm. This aspect was studied in (Lugosi and Mendelson, 2019b) who gave a norm-dependent formula for the optimal deviation and an algorithm to achieve it, although the latter has exponential complexity and does not consider the presence of outliers.

This is an important aspect to keep in mind in our high-dimensional setting since we will be measuring the statistical error on the gradient using the dual norm $\|\cdot\|_*$ of $\|\cdot\|$ which will never be the Euclidean one.

$$\|v\|_* = \sup_{\|x\| \leq 1} \langle v, x \rangle \tag{3}$$

Of course, apart from the way it is measured, the quality of the estimations one can obtain also crucially depends on the assumptions made on the data. We formally state ours here. We denote $|A|$ as the cardinality of a finite set A and use the notation $[\![k]\!] = \{1, \ldots, k\}$ for any integer $k \in \mathbb{N} \setminus \{0\}$.

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

the distribution P is such that :

$$\mathbb{E}\big[\|X\|_2^4\big] < +\infty, \quad \mathbb{E}\big[\|YX\|_2^2\big] < +\infty \quad and \quad \mathbb{E}\big[|Y|^2\big] < +\infty. \tag{4}$$

Moreover, the loss function ℓ admits constants $C_{\ell,1}, C_{\ell,2}, C'_{\ell,1}, C'_{\ell,2} > 0$ such that for all $z, y \in \widehat{\mathcal{Y}} \times \mathcal{Y}$:

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

where ℓ' is the derivative of ℓ in its first argument.

The above hypotheses are sufficient so that the objective function and its gradient exist for any parameter θ and the gradient admits a second moment. The distribution P is allowed to be heavy-tailed and the conditions on ℓ do not go far beyond limiting it to a quadratic behavior and are satisfied by common loss functions for regression and classification.[5] Note that depending on the loss function used and the moment requirements of gradient estimation, the previous moments assumption can be weakened as in (Gaïffas and Merad, 2022, Assumption 2) for instance. However, we stick to this version in this work for simplicity. Depending on the gradient estimator, the number of outliers $|\mathcal{O}|$ will be bounded in the subsequent statements either by a constant fraction ηn (η-corruption) of the sample for some $0 < \eta < 1/2$, or by a constant number.

In the two following sections, we will assume that we have a gradient estimator $\widehat{g}$ at our disposal such that for all $\theta \in \Theta$ we have $\widehat{g}(\theta) = g(\theta) + \epsilon(\theta)$ where $\epsilon(\theta)$ represents the random estimation error of the gradient at θ. Same as for the norm $\|\cdot\|$, this estimator will be precisely defined for each individual application in Section 5 in such a way that the error $\|\widehat{g}(\theta) - g(\theta)\|_* = \|\epsilon(\theta)\|_*$ is (nearly) optimally controlled.

In the sequel, we interchangeably use the terms *statistical rate*, *estimation rate*, *deviation rate* or simply *rate* to designate the statistical dependence of the bounds we obtain on the excess risk $\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta^\star)$ and the parameter error $\|\widehat{\theta} - \theta^\star\|_2$ for a general estimator $\widehat{\theta}$. This may lead to some confusion since the excess risk is only comparable to the square error $\|\widehat{\theta} - \theta^\star\|_2^2$ up to a constant factor. However, the reader should be able to distinguish the two situations based on context. Note that the previous terms should not be confused with *optimization rate* and *corruption rate*.

3. The Smooth Case with Mirror Descent

In this section we will assume the loss ℓ is smooth, formally :

Assumption 3 *For any $y \in \mathcal{Y}$, the loss $z \mapsto \ell(z, y)$ is convex, differentiable and γ-smooth meaning that*

$$|\ell'(z,y) - \ell'(z',y)| \le \gamma|z - z'| \tag{5}$$

for some $\gamma > 0$ and all $z, z' \in \widehat{\mathcal{Y}}$, where the derivative is taken w.r.t. the first argument.

The above assumption is stated in all generality for z, z' belonging to the prediction space $\widehat{\mathcal{Y}}$. For regression or binary classification tasks, we will have $\widehat{\mathcal{Y}} = \mathbb{R}$ and $\ell'(\cdot, y) \in \mathbb{R}$ so that the absolute values are enough to interpret the required inequality. Nonetheless, it can also be extended for K-way multiclass classification where $\widehat{\mathcal{Y}} = \mathbb{R}^K$ and $\ell'(\cdot, y) \in \mathbb{R}^K$, in which case the absolute values on both sides of the above inequality should be interpreted as Euclidean norms.

We also make the following quadratic growth assumption (Necoara et al., 2019, Definition 4).

---

5. Including the square loss, the absolute loss, Huber's loss, the logistic loss and the Hinge loss.

Assumption 4 *Let $\theta^\star \in \Theta$ be the optimum of the objective $\mathcal{L}$ and $\|\cdot\|_2$ the usual Euclidean norm. There exists a constant $\kappa > 0$ such that for all $\theta \in \Theta$:*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) \geq \kappa \|\theta - \theta^\star\|_2^2. \tag{6}$$

Assumption 4 is similar to but weaker than strong convexity because it only requires the quadratic minorization to hold around the optimum $\theta^\star$ whereas a strongly convex function is minorized by a quadratic function at every point. In the linear regression setting with samples $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, it is easy to see that the above condition holds as soon as the data follows a distribution with non-singular covariance $\Sigma = \mathbb{E}XX^\top$. A more general setting where this condition holds is also given in (Juditsky et al., 2022, Section 3.1). In comparison, the commonly used restricted eigenvalue or compatibility conditions (Bickel et al., 2009; Castillo et al., 2015) roughly require the empirical covariance $\widehat{\Sigma} = \frac{1}{n}\sum_{i\leq n} X_i X_i^\top$ to satisfy $\|\widehat{\Sigma}v\|_2 \geq \kappa \|v\|_2$ for all approximately s-sparse vectors $v \in \mathbb{R}^d$. This was shown to hold for covariates following some well known distributions (e.g. Gaussian with non singular covariance) with a sufficient sample count n (Raskutti et al., 2010). However, this is clearly more constraining than Assumption 4. Some variants of the compatibility condition are formulated in terms of the population covariance Σ (Van De Geer and Bühlmann, 2009; Bühlmann and Van De Geer, 2011) but these serve as a basis to show oracle inequalities for LASSO rather than the study of an implementable algorithm.

As a consequence of Assumptions 2 and 3, the objective gradient is L-Lipschitz continuous for some constant $L > 0$, meaning that we have :

$$\|g(\theta) - g(\theta')\|_* \leq L\|\theta - \theta'\| \quad \forall \theta, \theta' \in \Theta. \tag{7}$$

This property is necessary to establish the convergence of the Mirror Descent algorithm (Nemirovskij and Yudin, 1983) proposed in this section. Since we adopt a multistage mirror descent procedure as done in (Juditsky et al., 2022), our framework is also similar to theirs.

Definition 1 *A function $\omega : \Theta \to \mathbb{R}$ is a distance generating function if it is a real convex function over Θ which satisfies :*

1. *ω is continuously differentiable and strongly convex w.r.t. the norm $\|\cdot\|$ i.e.*

$$\langle \nabla\omega(\theta) - \nabla\omega(\theta'), \theta - \theta' \rangle \geq \|\theta - \theta'\|^2.$$

2. *We have $\omega(\theta) \geq \omega(0) = 0$ for all $\theta \in \Theta$.*

3. *There exists a constant $\nu > 0$ called the quadratic growth constant such that we have :*

$$\omega(\theta) \leq \nu\|\theta\|^2 \quad \forall \theta \in \Theta. \tag{8}$$

We shall see, for individual applications, that one needs to choose ω in such a way that it is strongly convex and the constant ν has only a light dependence on the dimension. For a reference point $\theta_0 \in \Theta$, we define $\omega_{\theta_0}(\theta) = \omega(\theta - \theta_0)$ and the associated Bregman divergence :

$$V_{\theta_0}(\theta, \theta') = \omega_{\theta_0}(\theta) - \omega_{\theta_0}(\theta') - \langle \nabla\omega_{\theta_0}(\theta'), \theta - \theta' \rangle.$$

Given a step size $\beta > 0$ and a dual vector $u \in \Theta^*$, we define the following proximal mapping :

$$\begin{aligned}
\mathrm{prox}_\beta(u, \theta; \theta_0, \Theta) &:= \operatorname*{argmin}_{\theta' \in \Theta}\{\langle \beta u, \theta'\rangle + V_{\theta_0}(\theta', \theta)\} \\
&= \operatorname*{argmin}_{\theta' \in \Theta}\{\langle \beta u - \nabla\omega_{\theta_0}(\theta), \theta'\rangle + \omega_{\theta_0}(\theta')\}.
\end{aligned}$$

The previous operator yields the next iterate of Mirror Descent for previous iterate θ, gradient u and step size β with Bregman divergence defined according to the reference point θ_0. Ideally, we would plug $g(\theta)$ as gradient u but since the true gradient is not observed, we replace it with the estimator $\widehat{g}(\cdot)$. All in all, given an initial parameter θ_0 we obtain the following iteration for Mirror Descent :

$$\theta_{t+1} = \mathrm{prox}_\beta(\widehat{g}(\theta_t), \theta_t; \theta_0, \Theta), \tag{9}$$

with a step size β to be defined later according to problem parameters. The previous proximal operator can be computed in closed form in each of the applications we consider in Section 5, see Appendix A.4 for details. We state the convergence properties of the above iteration in the following proposition.

Proposition 2 *Grant Assumptions 2 and 3 so that the objective $\mathcal{L}$ is L-Lipschitz-smooth for some $L > 0$. Let mirror descent be run with constant step size $\beta \le 1/L$ starting from $\theta_0 \in \Theta$ with $\Theta = B_{\|\cdot\|}(\theta_0, R)$ for some radius $R > 0$. Let $\theta_1, \ldots, \theta_T$ denote the resulting iterates and $\widehat{\theta}_T = \sum_{t=1}^T \theta_t/T$, then the following inequality holds :*

$$\begin{aligned}
\mathcal{L}(\widehat{\theta}_T) - \mathcal{L}(\theta^\star) &\le \frac{1}{T}\Big(\frac{1}{\beta}V_{\theta_0}(\theta^\star, \theta_0) + \sum_{t=0}^{T-1}\langle \epsilon_t, \theta^\star - \theta_{t+1}\rangle\Big) \\
&\le \frac{\nu R^2}{\beta T} + 2\bar{\epsilon}R
\end{aligned}$$

where $\epsilon_t = \widehat{g}(\theta_t) - g(\theta_t)$ and $\bar{\epsilon} = \max_{t=0\ldots T-1}\|\epsilon_t\|_*$.

Proposition 2 is proven in Appendix A.2.1 based on (Juditsky et al., 2022, Proposition 2.1) and quantifies the progress of mirror descent on the objective value while measuring the impact of the gradient errors. The original version in (Juditsky et al., 2022) considers a stochastic optimization problem in which a new sample arrives at each iteration providing an unbiased estimate of the gradient so that it is possible to obtain a bound with optimal quadratic dependence on the statistical error. Though the above result is suboptimal in this respect, we will show in the sequel that an optimal statistical rate can still be achieved using a multistage procedure.

Notice that the previous statement only provides guarantees for the average $\widehat{\theta}_T = \sum_{t=1}^T \theta_t/T$. While this is commonplace for online settings, we intuitively expect the last iterate θ_T to be the best estimate of $\theta^\star$ in our batch setting where all the data is available from the beginning.

In order to address this issue, we define a *corrected* proximal operator given an upper bound $\bar{\epsilon}$ on the statistical error :

$$\widehat{\mathrm{prox}}_\beta(u, \theta; \theta_0, \Theta) = \operatorname*{argmin}_{\theta' \in \Theta}\{\langle \beta u, \theta'\rangle + \beta\bar{\epsilon}\|\theta' - \theta\| + V_{\theta_0}(\theta', \theta)\}. \tag{10}$$

For this new operator, the following statement applies.

Proposition 3 *In the setting of Proposition 2, let mirror descent be similarly run with constant step size $\beta \leq 1/L$ starting from $\theta_0 \in \Theta$ with $\Theta = B_{\|\cdot\|}(\theta_0, R)$ for some radius $R > 0$. Let $\theta_1, \ldots, \theta_T$ denote the resulting iterates obtained through $\theta_{t+1} = \widehat{\mathrm{prox}}_\beta(\widehat{g}(\theta_t), \theta_t; \theta_0, \Theta)$ then the following inequality holds* :*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^\star) \leq \frac{1}{T}\left(\frac{1}{\beta}V_{\theta_0}(\theta^\star, \theta_0) + 2\sum_{t=0}^{T-1}\langle \epsilon_t, \theta^\star - \theta_{t+1}\rangle\right)$$

$$\leq \frac{\nu R^2}{\beta T} + 4\bar{\epsilon}R,$$

where $\bar{\epsilon} = \max_{t=0\ldots T-1}\|\epsilon_t\|_*$.

The proof of Proposition 3 is given in Appendix A.2.2 and mainly differs from that of Proposition 2 in that the introduced correction allows to show a monotonous decrease of the objective i.e. $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)$ letting us draw the conclusion on the last iterate. Nevertheless, we suspect that the correction is not really needed for this bound to hold on θ_T and consider it rather as an artifact of our proof.

Propositions 2 and 3 only state a linear dependence of the final excess risk on the statistical error $\bar{\epsilon}$ which leads to a suboptimal statistical rate of $1/\sqrt{n}$. However, the optimal rate of $1/n$ can be achieved by leveraging the sparsity condition on $\theta^\star$ (Assumption 1) and the quadratic growth condition (Assumption 4) upon running *multiple stages* of Mirror Descent (Juditsky et al., 2011, 2022). The idea is that by factoring these two assumptions in, for T big enough in Proposition 3 and given $\bar{s} \geq s$, it can be shown that the closest $\bar{s}$-sparse element $\mathrm{sparse}_{\bar{s}}(\theta_T)$ to the last iterate θ_T is such that $\Theta' = B_{\|\cdot\|}(\mathrm{sparse}_{\bar{s}}(\theta_T), R') \ni \theta^\star$ with $R' < R$. Therefore, $\mathrm{sparse}_{\bar{s}}(\theta_T)$ can serve as the starting point of a new stage of mirror descent on the smaller domain Θ'. By repeating this trick multiple times, we obtain the following multistage mirror descent algorithm.

Algorithm : Approximate Multistage Mirror Descent (AMMD)

- *Initialization*: Initial parameter $\theta^{(0)}$ and $R > 0$ such that $\theta^\star \in \Theta := B_{\|\cdot\|}(\theta_0, R)$.

  Number of stages $K > 0$. Step size $\beta \leq 1/L$. Quadratic minorization constant κ.

  High probability upperbound $\bar{\epsilon}$ on the error $\|\widehat{g}(\theta) - g(\theta)\|_*$.

  Upperbound $\bar{s}$ on the sparsity s.

- Set $R_0 = R$.

- Loop over stages $k = 1\ldots K$:

  - Set $\theta_0^{(k)} = \theta^{(k-1)}$ and $\Theta_k = B_{\|\cdot\|}(\theta_0^{(k)}, R_{k-1})$.
  - Run iteration
    $$\theta_{t+1}^{(k)} = \widehat{\mathrm{prox}}_\beta\big(\widehat{g}(\theta_t^{(k)}), \theta_t^{(k)}; \theta_0^{(k)}, \Theta_k\big),$$
    for T_k steps with $T_k = \left\lceil\frac{\nu R_{k-1}}{\beta\bar{\epsilon}}\right\rceil$.
  - Set $\theta^{(k)} = \mathrm{sparse}_{\bar{s}}(\widetilde{\theta}^{(k)})$ where $\widetilde{\theta}^{(k)} = \theta_{T_k}^{(k)}$.
  - Set $R_k = \frac{1}{2}(R_{k-1} + \frac{40\bar{s}\bar{\epsilon}}{\kappa})$.

11

- *Output*: The final stage estimate $\theta^{(K)}$.

The AMMD algorithm borrows ideas from Juditsky et al. (2011); Juditsky and Nesterov (2014); Juditsky et al. (2022) aiming to achieve linear convergence using mirror descent. The main trick lies in the fact that performing multiple stages of mirror descent allows to repeatedly restrict the parameter space into a ball of radius R_k which shrinks geometrically with each stage. In this work, we find that the radius R_k evolves following a special contraction as a result of the statistical error being factored in. Note that, although a few instructions of AMMD are stated in terms of unknown quantities, the procedure may be simplified to get around this difficulty with satisfactory results, see Section 6 for details. We show that the above procedure allows to improve the result of Proposition 3 to achieve a fast statistical rate. The following statement expresses the theoretical properties of AMMD.

Theorem 4 *Grant Assumptions 1, 2, 3 and 4. Let $L > 0$ denote the Lipschitz smoothness constant for the objective $\mathcal{L}$. Assume approximate Mirror Descent is run with step size $\beta \leq 1/L$ starting from $\theta_0 \in \Theta$ such that $\theta^\star \in B_{\|\cdot\|}(\theta_0, R)$ for some $R > 0$ and using a gradient estimator $\widehat{g}$ with error upperbound $\bar{\epsilon}$ as in Proposition 9, then after K stages we have the inequalities :*

$$\left\|\theta^{(K)} - \theta^\star\right\| \leq \sqrt{2\bar{s}}\left\|\theta^{(K)} - \theta^\star\right\|_2 \leq 2\sqrt{2\bar{s}}\left\|\widetilde{\theta}^{(K)} - \theta^\star\right\|_2 \leq 2^{-(K-1)/2}R + \frac{40\bar{s}\bar{\epsilon}}{\kappa},$$

$$\mathcal{L}(\widetilde{\theta}^{(K)}) - \mathcal{L}(\theta^\star) \leq 10\bar{\epsilon}\Big(2^{-K}R + \frac{40\bar{s}\bar{\epsilon}}{\kappa}\Big).$$

Moreover, the corresponding number of necessary iterations is bounded by :

$$T = \sum_{k=1}^{K} T_k \leq \frac{2R\nu}{\beta\bar{\epsilon}} + K\Big(1 + \frac{40\nu\bar{s}}{\kappa\beta}\Big).$$

Theorem 4 is proven in Appendix A.2.3 and may be compared to (Juditsky et al., 2022, Theorem 2.1). Both statements bound the risk in terms of the objective and parameter error by the sum of an exponentially vanishing optimization error and a statistical error term. Note that the exponential optimization rate in the number of stages K also holds in the number of iterations since successive stages contain a geometrically decreasing number of them. Theorem 4 expresses the statistical error in terms of a bound $\bar{\epsilon}$ and will thus lead to a high confidence statement when combined with a bound on $\bar{\epsilon}$ (see Section 5). In contrast, Theorem 2.1 of Juditsky et al. (2022) is a result in expectation which is later used to derive a high confidence bound on an aggregated estimate.

One can see that Theorem 4 exhibits a dependence in $\bar{\epsilon}^2$ of the excess risk upperbound so that the suboptimal statistical rate in Propositions 2 and 3 is improved into a fast rate as announced. This is accomplished by shrinking the size of the considered parameter set through the stages until it reaches the scale of the statistical error, yielding an optimal rate. This shrinkage is achieved thanks to the choice of stage-length $T_k = \Omega(R_{k-1})$ in AMMD leading to a bound in terms of R rather than R^2 in Proposition 3. Combined with Assumption 4, this implies that a square root function is applied to R_k after each stage, see the proof for further details. We can now turn to the case of a non smooth loss function ℓ.

4. The Non Smooth Case with Dual Averaging

In the previous section, we saw how sparse estimation can be performed using the Mirror Descent algorithm to optimize a smooth objective with a non Euclidean metric on the parameter space. The smoothness property is necessary for these results to hold so that many loss functions not satisfying it are left uncovered. Therefore, we propose to use another algorithm for non smooth objectives. The alternative is the Dual Averaging algorithm (Nesterov, 2009) which was already used for non smooth sparse estimation in (Agarwal et al., 2012) for instance. Since the original algorithm requires to average the iterates to obtain a parameter with provable convergence properties, we instead use a variant (Nesterov and Shikhman, 2015) for which such properties apply for individual iterates.

The smoothness condition in Assumption 3 is no longer required but we still need to replace it with a Lipschitz property :

Assumption 5 *There exists a positive constant $M > 0$ such that the objective $\mathcal{L}$ is M-Lipschitz w.r.t. the norm $\| \cdot \|$ i.e. for all $\theta, \theta' \in \Theta$ it holds that :*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta') \leq M\|\theta - \theta'\|.$$

We also replace Assumption 4 by the following weaker assumption :

Assumption 6 *There exist positive constants $\kappa, \lambda > 0$ such that the following inequality holds :*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) \geq \frac{\kappa\|\theta - \theta^\star\|_2^2}{\lambda + \|\theta - \theta^\star\|_2}.$$

We introduce this (to our knowledge) previously unknown assumption in the literature which we call the *pseudo-linear* growth assumption in order to better suit the setting of this section. Indeed, few non-smooth loss functions, if any, result in quadratically growing objectives as Assumption 4 requires. Note that the lower bound of Assumption 6 is linear away from the optimum i.e. for big $\|\theta - \theta^\star\|_2$ and behaves quadratically around it. This assumption is also weaker than a linear lower bound proportional to $\|\theta - \theta^\star\|_2$ because of its quadratic behaviour around the optimum. We will show that, for κ big enough, this minorization suffices to obtain linear convergence to a solution with fast statistical rate.

Analogously to Mirror Descent's distance generating function ω, we let $\omega : \Theta \to \mathbb{R}^+$ be the *prox-function*. We choose to denote it similarly since it plays an analogous role for Dual Averaging and has the same properties as those listed in Definition 1.

Let $(a_t)_{t\geq0}$ be a sequence of step sizes and $(\gamma_t)_{t\geq0}$ a non decreasing sequence of positive scaling coefficients. The DA procedure is defined, given an initial $\theta_0 \in \Theta$, by the following scheme :

$$s_t = \frac{1}{A_t}\sum_{i=0}^{t} a_i\widehat{g}_i \quad \text{with} \quad \widehat{g}_i = \widehat{g}(\theta_i) \text{ and } g_i = g(\theta_i) \quad \forall i = 0,\dots,T.$$

$$A_t = \sum_{i=0}^{t} a_i \quad \text{and} \quad \theta_t^+ = \operatorname*{argmin}_{\theta\in\Theta} A_t\langle s_t, \theta\rangle + \gamma_t\omega(\theta).$$

$$\theta_{t+1} = (1 - \tau_t)\theta_t + \tau_t\theta_t^+ \quad \text{where} \quad \tau_t = \frac{a_{t+1}}{A_{t+1}}.$$

13

Proposition 5 *Grant Assumption 5, let Dual Averaging be run following the above scheme, let $R > 0$ such that $\Theta \subseteq B_{\|\cdot\|}(\theta_0, R)$ and denote $\bar{\epsilon} = \max_i \|\epsilon_i\|_*$, we have the following inequality :*

$$A_t(\mathcal{L}(\theta_t) - \mathcal{L}(\theta^\star)) + \frac{\gamma_t}{2}\|\theta_t^+ - \theta^\star\|^2 \leq \gamma_t\omega(\theta^\star) + \sum_{i=0}^{t}\frac{a_i^2}{2\gamma_{i-1}}\|g_i\|_*^2 + 4A_t R\bar{\epsilon}.$$

In particular, by choosing $a_i = 1$ and $\gamma_i = \sqrt{i+1}$ for all i we get :

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta^\star) \leq \frac{1}{\sqrt{t}}\big(\omega(\theta^\star) + M^2\big) + 4R\bar{\epsilon}.$$

The proof of Proposition 5 is given in Appendix A.3.1 and is inspired from (Nesterov and Shikhman, 2015, Theorem 3.1). In this result, we manage to obtain a statement in terms of the individual iterates θ_t thanks to the running average performed in the above scheme whereas the initial study of dual averaging focused on the *average* of the iterates (Nesterov, 2009). Notice that, due to Assumption 3 being dropped, the convergence speed degrades to $1/\sqrt{t}$ as opposed to $1/t$ previously. This convergence speed is the fastest possible and cannot be improved with a different choice of a_i and γ_i. Most importantly, Proposition 5 quantifies the impact of the errors on the gradients on the quality of the optimisation result and shows that it remains controlled in this case too.

As in the previous section, the statistical rate we initially obtain is suboptimal and a multistage procedure is needed to improve it. The idea is the same as in Section 3 and consists in sparsifying the final iterate in Proposition 5 and using it as the initial point of a new optimization stage which takes place on a narrower domain. We make the resulting algorithm explicit below.

Algorithm : Approximate Multistage Dual Averaging (AMDA)

- *Initialization*: Initial parameter θ_0 and $R > 0$ such that $\theta^\star \in \Theta := B_{\|\cdot\|}(\theta_0, R)$.

  Pseudolinear minorization constants κ, λ.

  High probability upperbound $\bar{\epsilon}$ on the error $\|\widehat{g}(\theta) - g(\theta)\|_*$.

  Upperbound $\bar{s}$ on the sparsity s.

- Set $R_0 = R$ and $\tau = \frac{10\sqrt{8\bar{s}\bar{\epsilon}}}{\kappa}$ and $R^\star = \frac{80\lambda\bar{s}\bar{\epsilon}}{\kappa}$.

- Set $k = 0$ and the per stage number of iterations $T' = \left\lceil \left(\frac{\nu+M^2}{\bar{\epsilon}}\right)^2 \right\rceil$

- For $k = 1, \ldots, K$:

  - Set $\theta_0^{(k)} = \theta^{(k-1)}$ and $\Theta_k = B_{\|\cdot\|}(\theta_0^{(k)}, R_{k-1})$.
  - Run Dual averaging with prox-function $\omega_{\theta^{(k-1)}}$ and steps $a_i = R_{k-1}$ for T' iterations.
  - Set $\theta^{(k)} = \text{sparse}_{\bar{s}}(\widetilde{\theta}^{(k)})$ where $\widetilde{\theta}^{(k)} := \theta_{T'}^{(k)}$.
  - Set $R_k = \max\big(\tau R_{k-1}, \frac{1}{2}(R_{k-1} + R^\star)\big)$.

- *Output*: The final stage estimate $\theta^{(K)}$.

Similar to AMMD, the AMDA algorithm runs multiple optimisation stages through which the parameter space is repeatedly restricted allowing to obtain similar benefits regarding convergence speed and statistical performance. However, it is worth noting that these improvements are obtained under much milder conditions here since the objective may not even be smooth and is only required to satisfy the pseudo-linear growth condition of Assumption 6 whereas smoothness and quadratic minorization or strong convexity were indispensable in previous works (Juditsky et al., 2011; Juditsky and Nesterov, 2014; Juditsky et al., 2022).

As previously mentioned for AMMD, a simplified version of AMDA can be implemented which does not require knowledge of all the involved quantities, see Section 6 for details. We state the convergence guarantees for the above algorithm in the following Theorem.

Theorem 6 *Grant Assumptions 1, 2, 5, 6 and assume that $\tau = \frac{10\sqrt{8\bar{s}\bar{\epsilon}}}{\kappa} < 1$. At the end of each stage $k \geq 1$, we have :*

$$\left\|\theta^{(k)} - \theta^\star\right\| \leq \sqrt{2\bar{s}}\left\|\theta^{(k)} - \theta^\star\right\|_2 \leq 2\sqrt{2\bar{s}}\left\|\widetilde{\theta}^{(k)} - \theta^\star\right\|_2 \leq R_k, \tag{11}$$

$$\mathcal{L}(\widetilde{\theta}^{(k)}) - \mathcal{L}(\theta^\star) \leq 5\bar{\epsilon}R_{k-1}. \tag{12}$$

Moreover, the total number of necessary iterations before $R_k \leq 2R^\star = \frac{160\lambda\bar{s}\bar{\epsilon}}{\kappa}$ is at most

$$\log(R_0/R^\star)\Big(\frac{1}{\log(1/\tau)} + \frac{1}{\log(2)}\Big)\Big(\Big(\frac{\nu + M^2}{\bar{\epsilon}}\Big)^2 + 1\Big).$$

The proof of Theorem 6 is given in Appendix A.3.2 and shows that the optimization stages go through two phases: an initial *linear* phase corresponding to the linear regime of the lower bound given by Assumption 6 and a later *quadratic* phase during which the quadratic regime takes over. The success of the linear phase relies on the condition $\tau < 1$ which can be rewritten as $\bar{\epsilon} \leq O(\bar{s}\kappa)$ where the factor $\bar{s}$ is a byproduct of measuring the parameter error with the norm $\|\cdot\|$ while Assumption 6 is stated with the Euclidean one. In the linear regime, κ acts as a lower bound for the gradient norm so that the condition ensures that the error is smaller than the actual gradient allowing the optimisation to make progress. Theorem 6 states that convergence to the optimum occurs at geometrical speed through the stages of AMDA despite the absence of strong convexity. This is achieved thanks to the choice of step $a_i = R_{k-1}$ in AMDA which leads to a bound in terms of R rather than R^2 emerging from the term $\omega(\theta^\star)$ in Proposition 5, see the proof for more details.

Remark In recent work, Juditsky et al. (2019) created a common framework for the study of the Mirror Descent and Dual Averaging algorithms which they recover as special cases of a generic Unified Mirror Descent procedure. However, the distinction between the two remains necessary since they address smooth and non-smooth objectives respectively and, in each case, the attainable within-stage convergence speed differs from $1/t$ to $1/\sqrt{t}$ as seen in Propositions 3 and 5. This is reflected in Theorems 4 and 6 which display a dependence of the necessary number of iterations in $1/\bar{\epsilon}$ in the gradient error for Mirror Descent as opposed to $1/\bar{\epsilon}^2$ for Dual Averaging.

5. Applications

We now consider a few problems which may be solved using the previous optimization procedures. As said earlier, we have omitted to quantify the gradient errors $\|\epsilon\|_*$ until now. This is because the

definition of the dual norm $\| \cdot \|_*$ is problem dependent. In the next subsections, we consider a few instances and propose adapted gradient estimators for them. In each case, the existence of a second moment for the gradient random variable $G(\theta) := \ell'(\langle \theta, X \rangle, Y)X$ is required. This follows from the next Lemma proven in Appendix A.1 based on Assumption 2.

Lemma 7 *Under Assumption 2 the objective $\mathcal{L}(\theta)$ is well defined for all $\theta \in \Theta$ and we have :*

$$\mathbb{E}\big[\big\|G(\theta)\big\|_*^2\big] = \mathbb{E}\big[\big\|\ell'(X^\top\theta, Y)X\big\|_*^2\big] < +\infty.$$

In what follows, we will assume that, at each step of the optimization algorithm, the estimation of the gradient is performed with a new batch of data. For example, if the available data set contains n samples then it needs to be divided into T disjoint splits in order to make T optimization steps. This is necessary in order to guarantee that the gradient samples used for estimation at each step t are independent from θ_t, the (random) current parameter which depends on the data used before.

    This trick was previously used for example in (Prasad et al., 2020) for the same reasons. A possible alternative is to use an ϵ-net argument or Rademacher complexity in order to obtain uniform deviation bounds on gradient estimation over a compact parameter set Θ. However, this entails extra dependence on the dimension in the resulting deviation bound which we cannot afford in the high-dimensional setting. For these reasons, we prefer to use data splitting in this work and regard it more like a proof artifact rather than a true practical constraint. Note that we do not implement it later in our experimental section.

5.1 Vanilla sparse estimation

In this section, we consider the problem of optimizing an objective $\mathcal{L}(\theta) = \mathbb{E}[\ell(\langle \theta, X \rangle, Y)]$ where the covariate space $\mathcal{X}$ is simply $\mathbb{R}^d$ and the labels are either real numbers $\mathcal{Y} = \mathbb{R}$ (regression) or binary labels (binary classification). In this case, the parameter space is a subset $\Theta \subset \mathbb{R}^d$, the sparsity of a parameter $\theta \in \Theta$ is measured as its number of nonzero entries $S(\theta) = \sum_{j \in \llbracket d \rrbracket} \mathbf{1}_{\theta_j \neq 0}$, and $\| \cdot \|$ is defined as the ℓ_1 norm $\| \cdot \| = \| \cdot \|_1$ so that $\| \cdot \|_* = \| \cdot \|_\infty$. We define the distance generating function ω as :

$$\omega(\theta) = \frac{1}{2}e\log(d)d^{(p-1)(2-p)/p}\|\theta\|_p^2 \quad \text{with} \quad p = 1 + \frac{1}{\log(d)}.$$

One can check that the above definition satisfies the requirements of Definition 1. In particular, it is strongly convex w.r.t. $\| \cdot \|$ and quadratically growing with constant $\nu = \frac{1}{2}e^2\log(d)$ (see Nesterov and Nemirovski, 2013, Theorem 2.1). In others words, conditions 1 and 3 of Definition 1 are reconciled with $\nu = O(\log d)$. For the sake of achieving this compromise, the previous choice of distance generating function is common in the high-dimensional learning literature using mirror descent (Agarwal et al., 2012; Duchi et al., 2010; Shalev-Shwartz and Tewari, 2011; Nemirovskij and Yudin, 1983; Juditsky et al., 2011; Nesterov and Nemirovski, 2013) up to slight variations in p and the multiplying factor.

    We consider Assumption 2 on the data with a constant fraction of outliers $|\mathcal{O}| \leq \eta n$ for some $\eta < 1/2$ (η-corruption) so that the gradient samples $g^i(\theta) := \ell'(\theta^\top X_i, Y_i)X_i$ may be both heavy-tailed and corrupted as well. We propose to compute $\widehat{g}(\theta)$ as the coordinatewise trimmed mean of the sample gradients i.e.

$$\widehat{g}_j(\theta) = \texttt{TM}_\alpha\big(g_j^1(\theta), \ldots, g_j^n(\theta)\big), \tag{13}$$

where, assuming without loss of generality that n is even, the trimmed mean estimator with parameter α for a sample $x_1, \ldots, x_n \in \mathbb{R}$ is defined as follows

$$
\mathrm{TM}_\alpha(x_1, \ldots, x_n) = \frac{2}{n} \sum_{i=n/2+1}^{n} q_\alpha \vee x_i \wedge q_{1-\alpha},
$$

where we denoted $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$ and used the quantiles $q_\alpha := x^{([\alpha n/2])}$ and $q_{1-\alpha} = x^{([(1-\alpha)n/2])}$ with $x^{(1)} \leq \cdots \leq x^{(n/2)}$ the order statistics of $(x_i)_{i \in [\![n/2]\!]}$ and where $[\cdot]$ denotes the integer part.

The main hurdle to compute the trimmed mean estimator is to find the two previous quantiles. A naive approach for this task would be to sort all the values leading to an $O(n \log(n))$ complexity. However, this can be brought down to $O(n)$ using the median-of-medians algorithm (see for instance Cormen et al., 2009, Chapter 9) so that the whole procedure runs in linear time.

We now give the deviation bound satisfied by the estimator (13). We denote x^j as the j-th coordinate of a vector x.

Lemma 8 *Grant Assumption 2 with a fraction of outliers $|\mathcal{O}| \leq \eta n$ with $\eta < 1/8$. Fix $\theta \in \Theta$, let $\sigma_j^2 = \mathrm{Var}(\ell'(\theta^\top X, Y) X^j)$ for $j \in [\![d]\!]$ be the gradient coordinate variances and let $1 > \delta > e^{-n/2}/4$ be a failure probability and consider the coordinatewise trimmed mean estimator (13) with parameter $\alpha = 8\eta + 12\frac{\log(4/\delta)}{n}$. Denoting $\sigma_{\max}^2 = \max_j \sigma_j^2$, we have with probability at least $1 - \delta$:*

$$
\left\| \widehat{g}(\theta) - g(\theta) \right\|_\infty \leq 7\sigma_{\max} \sqrt{4\eta + 6\frac{\log(4/\delta) + \log(d)}{n}}. \tag{14}
$$

Proof This is an almost immediate application of (Gaïffas and Merad, 2022, Lemma 9) (see also Lugosi and Mendelson, 2021, Theorem 1). By an immediate application of the latter, we obtain for each $j \in [\![d]\!]$ that with probability at least $1 - \delta/d$ we have :

$$
\left| \widehat{g}_j(\theta) - g_j(\theta) \right| \leq 7\sigma_j \sqrt{4\eta + 6\frac{\log(4/\delta) + \log(d)}{n}}. \tag{15}
$$

Hence, the lemma follows by a simple union bound argument. ∎

For the sake of simplicity, this deviation bound is only stated for a square integrable gradient which yields a $\sqrt{\eta}$ dependence in the corruption rate. More generally, for a random variable admitting a finite moment of order k, one can derive a bound in terms of $\eta^{1-1/k}$ which reflects a milder dependence for greater k, see (Gaïffas and Merad, 2022, Lemma 9) for the bound in question.

In a way, the fact that the gradient error is measured with the infinity norm in this setting is a "stroke of luck" since the optimal dependence in the dimension for the statistical error becomes achievable using only a univariate estimator. This is in contrast with situations where multivariate robust estimators need to be used for which the combination of efficiency, sub-Gaussianity and robustness to η-corruption is hard to come by.

Based on Lemma 8 we obtain a gradient error of order $\bar{\epsilon} = O(\sqrt{\log(d)/n})$. Plugging this deviation rate into Theorem 4 yields the optimal $s \log(d)/n$ rate for vanilla sparse estimation. The same applies for Theorem 6 provided the condition $\tau < 1$ holds.

Corollary 9 *In the context of Theorem 4 and Lemma 8, let the AMMD algorithm be run starting from $\theta_0 \in \Theta = B_{\|\cdot\|}(\theta_0, R)$ using the coordinatewise trimmed mean estimator with sample splitting i.e. at each iteration a different batch of size $\widetilde{n} = n/T$ is used for gradient estimation with confidence $\widetilde{\delta} = \delta/T$ where T is the total number of iterations. Let K be the number of stages and $\widehat{\theta}$ the obtained estimator. Denote $\sigma_{\max}^2 = \sup_{\theta \in \Theta} \max_{j \in [\![d]\!]} \mathrm{Var}(\ell'(\theta^\top X, Y)X^j)$, with probability at least $1 - \delta$, the latter satisfies :*

$$\left\|\widehat{\theta} - \theta^\star\right\|_2 \leq \frac{2^{-K/2}R}{\sqrt{\bar{s}}} + \frac{140\sqrt{2\bar{s}}\sigma_{\max}}{\kappa}\sqrt{4\eta + 6\frac{\log(4/\widetilde{\delta}) + \log(d)}{\widetilde{n}}}.$$

Proof The result is easily obtained by combining Theorem 4 and Lemma 8 with a union bound argument over all iterations T in order to bound $\bar{\epsilon} = \max_{t=0,\dots,T-1} \|\epsilon_t\|_*$ as defined in Proposition 3. ∎

In the above upper bound, the optimisation error vanishes exponentially with the number of stages K so that the final error can be attributed in large part to the second statistical error term. The latter achieves the nearly optimal $\sqrt{s \log(d)/n}$ rate and combines robustness to heavy tails and η-corruption. Moreover, this statement holds for a *generic* loss function satisfying the assumptions of Section 3. These can be further weakened to those given in Section 4 by using Dual Averaging while preserving the same statistical rate. The statement of a result in this weaker setting is postponed to Section A.3.3 of the Appendix in order to avoid excessive repetition. To our knowledge, this is the first result with such properties for vanilla sparse estimation whereas previous results from the literature either focused on specific learning problems with a fixed loss function (Dalalyan and Thompson, 2019; Sasai, 2022) or isolated the issues of robustness by assuming the data to be either heavy tailed or Gaussian and corrupted (Balakrishnan et al., 2017; Liu et al., 2019, 2020; Juditsky et al., 2022). Furthermore, we stress that this error bound is achieved at a comparable computational cost to that of a standard non robust algorithm since, as mentioned earlier, the robust trimmed mean estimator can be computed in linear time.

A possible room for improvement is to try to remove the $\bar{s}$ factor multiplying the corruption rate η. The only works we are aware of achieving this are (Liu et al., 2020; Sasai, 2022) but both involve costly data-filtering steps. We suspect this may be an inevitable price to pay for such an improvement.

5.2 Group sparse estimation

In the group sparse case, we again consider the covariate space $\mathcal{X} = \mathbb{R}^d$ where the coordinates $[\![d]\!]$ are arranged into groups $G_1, \dots, G_{N_G}$ which form a partition of the coordinates $[\![d]\!]$ and sparsity is measured in terms of these groups i.e. $S(\theta) = \sum_{j \in [\![N_G]\!]} \mathbf{1}_{\theta_{G_j} \neq 0}$ and we assume the optimal $\theta^\star = \operatorname{argmin}_\theta \mathcal{L}(\theta)$ satisfies $S(\theta^\star) \leq s$. The norm $\|\cdot\|$ is set to be the ℓ_1/ℓ_2 norm : $\|\theta\|_{1,2} = \sum_{j \in [\![N_G]\!]} \|\theta_{G_j}\|_2$ and the dual norm is the analogous ℓ_∞/ℓ_2 norm $\|\theta\|_{\infty,2} = \max_{j \in [\![N_G]\!]} \|\theta_{G_j}\|_2$. The label set $\mathcal{Y}$ may be equal to $\mathbb{R}$ (regression) or a finite set (binary or multiclass classification).

For simplicity, we assume that the groups are of equal size m so that $d = mN_G$. This is for example the case when trying to solve a d-dimensional linear multiclass classification with K classes by estimating a parameter $\theta \in \mathbb{R}^{d \times K}$ and predicting $\operatorname{argmax}_j(\theta^\top X)_j$ for a datapoint $X \in \mathbb{R}^d$. In this case, it makes sense to consider the rows $(\theta_{i,:})_{i \in [\![d]\!]}$ as groups which are collectively determined to be zero or not depending on the importance of feature i. For simplicity, we restrict ourselves to this setting until the end of this section with no loss of generality.

Analogously to the vanilla sparse case, following Nesterov and Nemirovski (2013), the distance generating function (or prox-function) ω may be chosen in this case as:

$$\omega(\theta) = \frac{1}{2} e \log(d) d^{(p-1)(2-p)/p} \Big(\sum_{i=1}^{d} \|\theta_{i,:}\|_2^p \Big)^{2/p} \quad \text{with} \quad p = 1 + \frac{1}{\log(d)}.$$

We assume the data corresponds to Assumption 2 with η-corruption (i.e. $|\mathcal{O}| \leq \eta n$). Analogously to the vanilla case, we propose to estimate the gradient *groupwise* i.e. one group of coordinates at a time. For this task, a multivariate, sub-Gaussian and corruption-resilient estimation algorithm is needed. We suggest to use the estimator advocated in (Diakonikolas et al., 2020) for this purpose which remarkably combines these qualities. We refer to it as the DKP estimator and restate its deviation bound here for the sake of completeness.

Proposition 10 (Diakonikolas et al. 2020, Proposition 1.5) *Let T be an η-corrupted set of n samples from a distribution P in $\mathbb{R}^d$ with mean μ and covariance Σ. Let $\eta' = \Theta(\log(1/\delta)/n + \eta) \leq c$ be given, for a constant $c > 0$. Then any stability-based algorithm on input T and η', efficiently computes $\widehat{\mu}$ such that with probability at least $1 - \delta$, we have :*

$$\|\widehat{\mu} - \mu\|_2 = O\Big(\sqrt{\frac{\mathrm{tr}(\Sigma) \log \mathrm{r}(\Sigma)}{n}} + \sqrt{\|\Sigma\|_{\mathrm{op}} \eta} + \sqrt{\frac{\|\Sigma\|_{\mathrm{op}} \log(1/\delta)}{n}} \Big), \qquad (16)$$

where $\mathrm{r}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$ is the stable rank of Σ.

The above bound is almost optimal up to the $\sqrt{\log \mathrm{r}(\Sigma)}$ factor which is at most $\sqrt{\log(d)}$. Note that we are also aware that (Diakonikolas et al., 2020, Proposition 1.6) states that, by adding a Median-Of-Means preprocessing step, stability based algorithms can achieve the optimal deviation. Nevertheless, the number k of block means required needs to be such that $k \geq 100\eta n$ so that the corruption rate η is strongly restricted because necessarily $n \geq k$. Therefore, we prefer to stick with the result above.

An algorithm with the statistical performance stated in Proposition 10 is given, for instance, in (Diakonikolas et al., 2020, Appendix A.2), we omit it here for brevity. Now, we can reuse the previous section's trick by estimating the gradients *blockwise* this time to obtain the following lemma:

Lemma 11 *Grant Assumption 2 with a fraction of outliers $|\mathcal{O}| \leq \eta n$. Fix $\theta \in \Theta$ and denote $G_1(\theta), \ldots, G_n(\theta)$ the gradient samples distributed according to $G(\theta) \in \mathbb{R}^{d \times K}$ (except for the outliers). Let $\Sigma_j = \mathrm{Var}(G(\theta)_{j,:}) \in \mathbb{R}^{K \times K}$ be the gradient block variances. Consider the groupwise estimator $\widehat{g}(\theta)$ defined such that $\widehat{g}(\theta)_{j,:}$ is the DKP estimator applied to $G(\theta)_{j,:}$. Then we have with probability at least $1 - \delta$:*

$$\|\widehat{g}(\theta) - g(\theta)\|_{\infty,2} \leq O\Big(\max_j \sqrt{\frac{\mathrm{tr}(\Sigma_j) \log \mathrm{r}(\Sigma_j)}{n}} + \sqrt{\|\Sigma_j\|_{\mathrm{op}}} \Big(\sqrt{\eta} + \sqrt{\frac{\log(1/\delta) + \log(d)}{n}} \Big) \Big) \tag{17}$$

Proof Inequality (17) is straightforward to obtain using Proposition 10 and a union bound argument on $j \in [\![d]\!]$. ∎

One can easily see that, in the absence of corruption, the above deviation bound scales as $\widetilde{O}\Big(\sqrt{\frac{K}{n}} + \sqrt{\frac{\log(d)}{n}}\Big)$. Combined with the sparsity assumption given above, plugging this estimation, which applies for the gradient error $\|\epsilon_t\|_*$, into Theorems 4 or 6 yields near optimal (up to a logarithmic factor) estimation rates for the group-sparse estimation problem (Negahban et al., 2012; Lounici et al., 2009). The following corollary formalizes this statement.

Corollary 12 *In the context of Theorem 4 and Lemma 11, let the AMMD algorithm be run starting from $\theta_0 \in \Theta = B_{\|\cdot\|}(\theta_0, R)$ and using the blockwise DKP estimator with sample splitting i.e. at each iteration a different batch of size $\widetilde{n} = n/T$ is used for gradient estimation with confidence $\widetilde{\delta} = \delta/T$ where T is the total number of iterations. Let N be the number of stages and $\widehat{\theta}$ the obtained estimator. With probability at least $1 - \delta$, the latter satisfies :*

$$\|\widehat{\theta} - \theta^\star\|_2 \leq \frac{2^{-N/2} R}{\sqrt{s}} + \frac{\sqrt{s}}{\kappa} O\bigg(\sup_{\theta \in \Theta} \max_{j \in [\![d]\!]} \sqrt{\frac{\operatorname{tr}(\Sigma_{\theta,j}) \log \operatorname{r}(\Sigma_{\theta,j})}{\widetilde{n}}} +$$

$$\sqrt{\|\Sigma_{\theta,j}\|_{\operatorname{op}}} \bigg(\sqrt{\eta} + \sqrt{\frac{\log(1/\widetilde{\delta}) + \log(d)}{\widetilde{n}}} \bigg) \bigg)$$

$$\leq \frac{2^{-N/2} R}{\sqrt{s}} + \frac{\sqrt{s}}{\kappa} \widetilde{O}\bigg(\sup_{\theta \in \Theta} \max_{j \in [\![d]\!]} \sqrt{\frac{K}{\widetilde{n}}} + \bigg(\sqrt{\eta} + \sqrt{\frac{\log(1/\widetilde{\delta}) + \log(d)}{\widetilde{n}}} \bigg) \bigg),$$

where $\Sigma_{\theta,j} = \operatorname{Var}(G(\theta)_{j,:})$.

As before, the stated bound reflects a linearly converging optimisation and displays a statistical rate nearly matching the optimal rate for group-sparse estimation (Negahban et al., 2012; Lounici et al., 2009) up to logarithmic factors. In addition, robustness to heavy tails and η-corruption likely makes this result the first of its kind for group-sparse estimation since all robust works we are aware of focus on vanilla sparsity.

5.3 Low-rank matrix recovery

We also consider the variant of the problem where the covariates belong to a matrix space $\mathcal{X} = \mathbb{R}^{p \times q}$ in which case the objective $\mathcal{L}(\theta) = \mathbb{E}[\ell(\langle \theta, X \rangle, Y)]$ needs to be optimized over $\Theta \subset \mathbb{R}^{p \times q}$. In this setting, $\langle \cdot, \cdot \rangle$ refers to the Frobenius scalar product between matrices

$$\langle a, b \rangle = \operatorname{tr}(a^\top b).$$

Without loss of generality, we assume that $p \geq q$ and sparsity is meant as the number of non zero singular values i.e. for a matrix $A \in \mathbb{R}^{p \times q}$, denoting $\sigma(A) = (\varsigma_j(A))_{j \in [\![q]\!]}$ the set of its singular values we define $S(A) = \sum_{j \in [\![q]\!]} \mathbf{1}_{\varsigma_j(A) \neq 0}$. We set $\|\cdot\|$ to be the nuclear norm $\|A\| = \|\sigma(A)\|_1$ and the associated dual norm is the operator norm $\|\cdot\|_* = \|\cdot\|_{\operatorname{op}}$.

On the optimization side, an appropriate distance generating function (resp. prox-function) needs to be defined for this setting before Mirror Descent (resp. Dual Averaging) can be run. Based on previous literature (see Nesterov and Nemirovski, 2013, Theorem 2.3 and Juditsky et al. 2022), we know that the following choice satisfies the requirements of Definition 1 :

$$\omega(\theta) = 2e \log(2q) \Big(\sum_{j=1}^{q} \varsigma_j(\theta)^{1+r} \Big)^{2/(1+r)} \quad \text{with} \quad r = 1/(12 \log(2q)).$$

This yields a corresponding quadratic growth parameter $\nu = O(\log(q))$. In order to fully define our optimization algorithm for this problem, it remains to specify a robust estimator for the gradient. This turns out to be a challenging question since the estimated value is matricial and the operator norm $\| \cdot \|_* = \| \cdot \|_{\mathrm{op}}$ emerging in this case is a fairly exotic choice to measure statistical error.

In order to achieve a nearly optimal statistical rate we define a new estimator called "CM-MOM" (short for Catoni Minsker Median-Of-Means) which combines methods from (Minsker, 2018) for sub-Gaussian matrix mean estimation and ideas from (Minsker et al., 2015; Hsu and Sabato, 2016) in order to apply a Median-Of-Means approach for multivariate estimation granting robustness to outliers provided these are limited in number. We now define this estimator in detail. Let ψ be a function defined as

$$\psi(x) = \log(1 + |x| + x^2/2)$$

We consider a restricted version of Assumption 2 in which the number of outliers is limited as[6] $|\mathcal{O}| \leq K/12$ where K is an integer such that $K < n$. Provided a sample of matrices $A_1, \ldots, A_n \in \mathbb{R}^{p \times q}$ and a scale parameter $\chi > 0$, the CM-MOM estimator proceeds as follows :

- Split the sample into K disjoint blocks $B_1, \ldots, B_K$ of equal size $m = n/K$.

- Compute the dilated block means $\xi^{(j)}$ for $j = 1, \ldots, K$ as

$$\xi^{(j)} = \frac{1}{\chi m} \sum_{i \in B_j} \psi(\theta \widetilde{A}_i) \in \mathbb{R}^{(p+q) \times (p+q)},$$

  where the dilation $\widetilde{A}$ of matrix $A \in \mathbb{R}^{p \times q}$ is defined as $\widetilde{A} = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix} \in \mathbb{R}^{(p+q) \times (p+q)}$ which is symmetric and the function ψ is applied to a symmetric matrix $S \in \mathbb{R}^{d \times d}$ by applying it to its eigenvalues i.e. let $S = UDU^\top$ be its eigendecomposition with $D = \mathrm{diag}((\lambda_j)_{j \in [\![d]\!]})$ then $\psi(S) = U\psi(D)U^\top = U\mathrm{diag}((\psi(\lambda_j))_{j \in [\![d]\!]})U^\top$.

- Extract the block means $\widehat{\mu}_j \in \mathbb{R}^{p \times q}$ such that $\xi^{(j)} = \begin{pmatrix} \xi_{11}^{(j)} & \widehat{\mu}_j \\ \widehat{\mu}_j^\top & \xi_{22}^{(j)} \end{pmatrix}$.

- Compute the pairwise distances $r_{jl} = \|\widehat{\mu}_j - \widehat{\mu}_l\|_{\mathrm{op}}$ for $j, l \in [\![K]\!]$.

- Compute the vectors $r^{(j)} \in \mathbb{R}^K$ for $j \in [\![K]\!]$ where $r^{(j)}$ is the increasingly sorted version of $r_{j:}$.

- return $\widehat{\mu}_{\widehat{i}}$ where $\widehat{i} \in \mathrm{argmin}_i r_{K/2}^{(i)}$.

One may guess that the choice of the scale parameter χ plays an important role to guarantee the quality of the estimate. This aspect is inherited from Catoni's original estimator for the mean of a heavy-tailed real random variable (Catoni, 2012) from which Minsker's estimator (Minsker, 2018), which we use to estimate the block means, is inspired. The following statement gives the optimal value for χ and the associated deviation bound satisfied by CM-MOM.

---

6. In fact, one may allow up to $|\mathcal{O}| \leq K/2$ outliers at the price of worse constants in the resulting deviation bound. See the proof of Proposition 13

Proposition 13 (CM-MOM) *Let $A_1, \ldots, A_n \in \mathbb{R}^{p \times q}$ be an i.i.d sample following a random variable A with expectation $\mu = \mathbb{E}A$ such that a subset of indices $\mathcal{O} \subset [\![n]\!]$ are outliers and finite variance*

$$v(A) = \max \left(\left\| \mathbb{E}(A - \mu)(A - \mu)^\top \right\|_{\mathrm{op}}, \left\| \mathbb{E}(A - \mu)^\top (A - \mu) \right\|_{\mathrm{op}} \right) < \infty.$$

Let $\delta > 0$ be a failure probability and take $K = \lceil 18 \log(1/\delta) \rceil < n$ blocks, we assume $n = mK$. Let $\widehat{\mu}$ be the CM-MOM estimate as defined above with scale parameter

$$\chi = \sqrt{\frac{2m \log(8(p+q))}{v(A)}}.$$

Assume we have $|\mathcal{O}| \leq K/12$ then with probability at least $1 - \delta$ we have :

$$\left\| \widehat{\mu} - \mu \right\|_{\mathrm{op}} \leq 18 \sqrt{\frac{v(A) \log(8(p+q)) \log(1/\delta)}{n}}.$$

Proposition 13 is proven in Appendix A.5 and enjoys a deviation rate which scales optimally, up to logarithmic factors, as $\sqrt{p+q}$ in the dimension (Vershynin, 2018). This dependence is hidden by the factor $\sqrt{v(A)}$ which scales in that order (see for instance Tropp, 2015). Although the dependence of the optimal scale χ on the unknown value of $v(A)$ constitutes an obstacle, previous experience using Catoni-based estimators (Catoni, 2012; Holland and Ikeda, 2019; Gaïffas and Merad, 2022) has shown that the choice is lenient and good results are obtained as long as a value of the correct scale is used. Possible improvements for Proposition 13 are to derive a bound with an additive instead of multiplicative term $\log(1/\delta)$ or supporting η-corruption. However, we are not aware of a more robust solution for matrix mean estimation in the general case than the above result.

Now that we have an adapted gradient estimation procedure, we can proceed to combine its deviation bound with our optimization theorems in order to obtain guarantees on learning performance.

Corollary 14 *In the context of Theorem 4 and Proposition 13, let the AMMD algorithm be run starting from $\theta_0 \in \Theta = B_{\|\cdot\|}(\theta_0, R)$ and using the CM-MOM estimator with sample splitting i.e. at each iteration a different batch of size $\widetilde{n} = n/T$ is used for gradient estimation with confidence $\widetilde{\delta} = \delta/T$ where T is the total number of iterations. Assume that each batch contains no more than $K/12$ outliers. Let N be the number of stages and $\widehat{\theta}$ the obtained estimator. With probability at least $1 - \delta$, the latter satisfies :*

$$\left\| \widehat{\theta} - \theta^\star \right\|_2 \leq \frac{2^{-N/2} R}{\sqrt{\overline{s}}} + \sup_{\theta \in \Theta} \frac{360 \sqrt{\overline{s}}}{\kappa} \sqrt{\frac{2 v(G(\theta)) \log(8(p+q)) \log(1/\widetilde{\delta})}{\widetilde{n}}},$$

where $\| \cdot \|_2$ denotes the Frobenius norm.

Corollary 14 matches the optimal performance bounds given in classical literature for low-rank matrix recovery (Koltchinskii et al., 2011; Rohde and Tsybakov, 2011; Candès and Plan, 2011; Negahban and Wainwright, 2011) up to logarithmic factors. The previous statement is most similar to (Juditsky et al., 2022, Proposition 3.3) except that it applies for more general learning tasks and under much lighter data assumption.

6. Implementation and Numerical Experiments

In this section, we demonstrate the performance of the proposed algorithms on synthetic and real data. Before we proceed, we prefer to point out that our implementation does not *exactly* correspond to the previously given pseudo-codes. Indeed, as the reader may have noticed previously, certain instructions of AMMD and AMDA require the knowledge of quantities which are not available in practice and, even in a controlled setting, the estimation of some quantities (such as the maximum gradient error $\bar{\epsilon}$) may be overly conservative which generally impedes the proper convergence of the optimization. We list the main divergences of our implementation from the theoretically studied procedures of AMMD and AMDA given before :

1. For AMMD, we only use the conventional prox operator rather than the corrected $\widehat{\mathrm{prox}}$ operator defined in Section 3.

2. For both AMMD and AMDA, the whole data set is used at each step to compute a gradient estimate and no data-splitting is performed.

3. The radii R_k are taken constant equal to a fixed $R > 0$.

4. The stage lengths (T_k for AMMD and T' for AMDA) are fixed as constants.

5. The number of stages is determined through a maximum number of iterations but the algorithm stops after the last whole stage.

6. The within-stage step-sizes a_i in AMDA are fixed to a small constant (smaller than R) for more stability.

The constant stage-lengths are fixed using the following heuristic: run the MD/DA iteration while tracking the evolution of the empirical objective on a validation subset of the data[7] and set the stage-length as the number of steps before a plateau is reached. Reaching a plateau indicates that the current reference point $\theta^{(k)}$ has become too constraining for the optimization and more progress can be made after updating it.

The simplifications brought by points 1. and 2. are related to likely proof artifacts. Indeed as pointed out in Section 3 the use of the corrected $\widehat{\mathrm{prox}}$ operator rather than simply prox is chiefly meant to ensure objective monotonicity while the data splitting ensures the gradient deviation bounds are usable in the proofs at each iteration.

Points 3. and 4. are due to the fact that the values of the stage-lengths T_k/T' and the radii R_k used in AMMD and AMDA are based on conservative estimates from the theoretical analysis making them unfit for practical implementation. Moreover, the said estimates use constants which cannot be identified in an arbitrary setting. For example, for least squares regression, the quadratic minorization constant κ depends on the data distribution which is unknown in general.

Finally, point 5. is commonplace for batch learning where one may simply iterate until convergence and point 6. follows the wisdom that smaller step-sizes ensure more stability.

Despite these differences, the numerical experiments we present below demonstrate that our implementations perform on par with the associated theoretical results.

Clearly, optimisation using Mirror Descent should be preferred over Dual Averaging due to its faster convergence speed. This is conditioned by the smoothness of the objective $\mathcal{L}$ which holds,

---

7. To remain consistent with a robust approach, the objective is estimated using a trimmed mean here as well.

for example, when the loss function ℓ is smooth. If ℓ is not smooth but the data distribution contains no atoms, one may still use Mirror Descent since it is reasonable to expect the objective $\mathcal{L}$ to be smoothed by the expectation (1). Note however that such an objective is likely not to satisfy Assumption 4 making Theorem 4 inapplicable. Still, in this case, if the weaker Assumption 6 holds, the expected performance is as stated in Theorem 6 with an improved number of required iterations of order $1/\bar{\epsilon}$ instead of $1/\bar{\epsilon}^2$ due to faster within-stage optimisation.

6.1 Synthetic sparse linear regression

We first test our algorithms on the classic problem of linear regression. We generate n covariates $X_i \in \mathbb{R}^d$ following a non-isotropic distribution with covariance matrix Σ and labels $Y_i = X_i^\top \theta^\star + \xi_i$ for a fixed s-sparse $\theta^\star \in \mathbb{R}^d$ and simulated noise entries ξ_i. The covariance matrix Σ is diagonal with entries drawn uniformly at random in $[1, 10]$.

We use the least-squares loss $\ell(z, y) = \frac{1}{2}(z-y)^2$ in this experiment and the problem parameters are $n = 500, d = 5000, s = 40$ and a sparsity upper bound $\bar{s} = 50$ is given to the algorithms instead of the real value. The noise variables ξ_i always follow a Pareto distribution with parameter $\alpha = 2.05$. Apart from that we consider three settings :

(a) The gaussian setting : the covariates follow a gaussian distribution.

(b) The heavytailed setting : the covariates are generated from a multivariate Student distribution with $\nu = 4.1$ degrees of freedom.

(c) The corrupted setting : the covariates follow the same Student distribution and 5% of the data ((X_i, Y_i) pairs) are corrupted.

We run various algorithms:

- AMMD using the trimmed mean estimator (`AMMD`).

- AMDA using the trimmed mean estimator (`AMDA`).

- The iterative thresholding procedure defined in (Liu et al., 2019) using the MOM estimator (`LLC_MOM`).

- The iterative thresholding procedure defined in (Liu et al., 2019) using the trimmed-mean estimator (`LLC_TM`).

- Lasso with CGD solver and the trimmed mean estimator as implemented in (Gaïffas and Merad, 2022) (`Lasso_CGD_TM`).

- Lasso with CGD solver as implemented in Scikit Learn (Pedregosa et al., 2011) (`Lasso_CGD`).

Another possible baseline is the algorithm proposed in (Liu et al., 2020). Nevertheless, we do not include it here because it relies on the outlier removal algorithm inspired from (Balakrishnan et al., 2017). The latter requires to run an SDP subroutine making it excessively slow as soon as the dimension is greater than a few hundreds.

Note that the "trimmed mean" estimator used in (Liu et al., 2019) is different from ours since they simply exclude the entries below and above a pair of empirical data quantiles. On the other

hand, the estimator we define in Section 5.1 simply replaces the extreme values by the exceeded threshold before computing an average. This is also called a "Winsorized mean" and enjoys better statistical properties.

The algorithms using Lasso (Tibshirani, 1996) optimize an ℓ_1 regularized objective. The regularization is weighted by a factor $2\sigma\sqrt{\frac{2\log(d)}{n}}$ where $\sigma^2 = \mathrm{Var}(\xi)$ is the noise variance. The previous regularization weight is known to ensure optimal statistical performance, see for instance (Bickel et al., 2009). The experiment is repeated 30 times and the results are averaged. We do not display
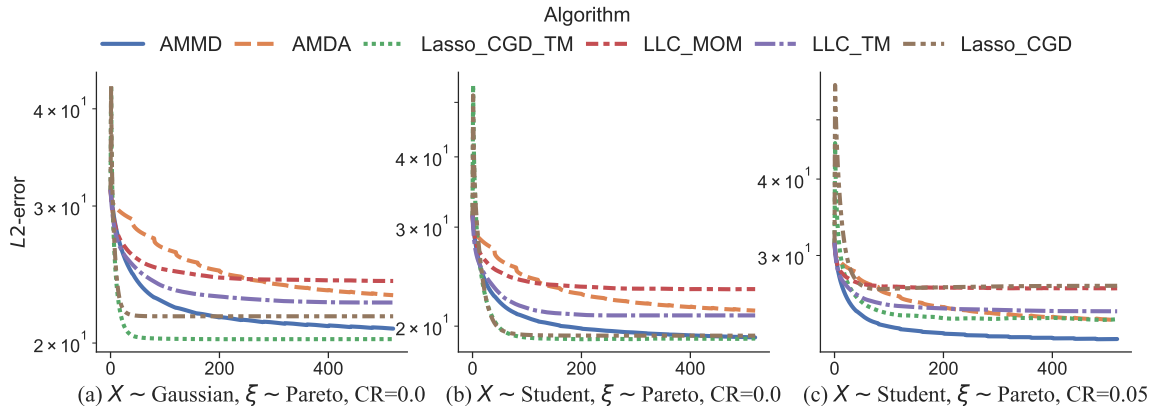

Figure 1: L2 error $\|\theta_t - \theta^\star\|_2$ (y-axis) against iterations (x-axis) for all the considered algorithms in the simulation settings.

any confidence intervals for better readability. Figure 1 displays the results. We observe that Lasso based methods quickly reach good optima in general and that the version using the robust trimmed mean estimator is sometimes superior in the presence of heavy tails and corruption in particular. `AMMD` reaches nearly equivalent optima, albeit significantly slower than Lasso methods as seen for settings (a) and (b). However, it is somehow more robust to corruption as seen on setting (c). Unfortunately, the `AMDA` algorithm struggles to closely approximate the original parameter. We mainly attribute this to slow convergence in settings (a) and (b). Nonetheless, `AMDA` is among the most robust algorithms to corruption as seen on setting (c). The remaining iterative thresholding based methods `LLC_MOM` and `LLC_TM` seem to generally stop at suboptimal optima. The Median-Of-Means variant `LLC_MOM` is barely more robust than `Lasso_CGD` in the corrupted setting (c). The `LLC_TM` variant is better but still inferior to `AMMD`. This reflects the superiority of the (Winsorized) trimmed mean used by `AMMD` and `Lasso_CGD_TM` to the conventional trimmed mean in `LLC_TM`.

6.2 Sparse classification on real data

We also carry out experiments on real high dimensional binary classification data sets. These are referred to as `gina` and `bioresponse` and were both downloaded from `openml.com`. We run `AMMD`, `AMDA`, `LLC_MOM` and `LLC_TM` with similar sparsity upperbounds and various levels of corruption and track the objective value, defined using the Logistic loss $\ell(z, y) = \log(1 + e^{-zy})$ (with $y \in \mathcal{Y} = \{\pm 1\}$), for each of them. The results are displayed on Figure 2 (average over 10 runs). In the non corrupted case, we see that all algorithms reach approximately equivalent

optima whereas they display different levels of resilience when corruption is present. In particular, `LLC_MOM` is unsurprisingly the most vulnerable since it is based on Median-Of-Means which is not robust to η-corruption. The rest of the algorithms cope better thanks to the use of trimmed mean estimators, although `LLC_TM` seems to be a little less robust which is probably due to the previously mentioned difference in its gradient estimator. Finally, Figure 2 also shows that `AMMD` and `AMDA` (respectively using Mirror Descent and Dual Averaging) tend to reach generally better final optima despite converging a bit slower than the other algorithms. They also prove to be more stable, even when high step sizes are used.
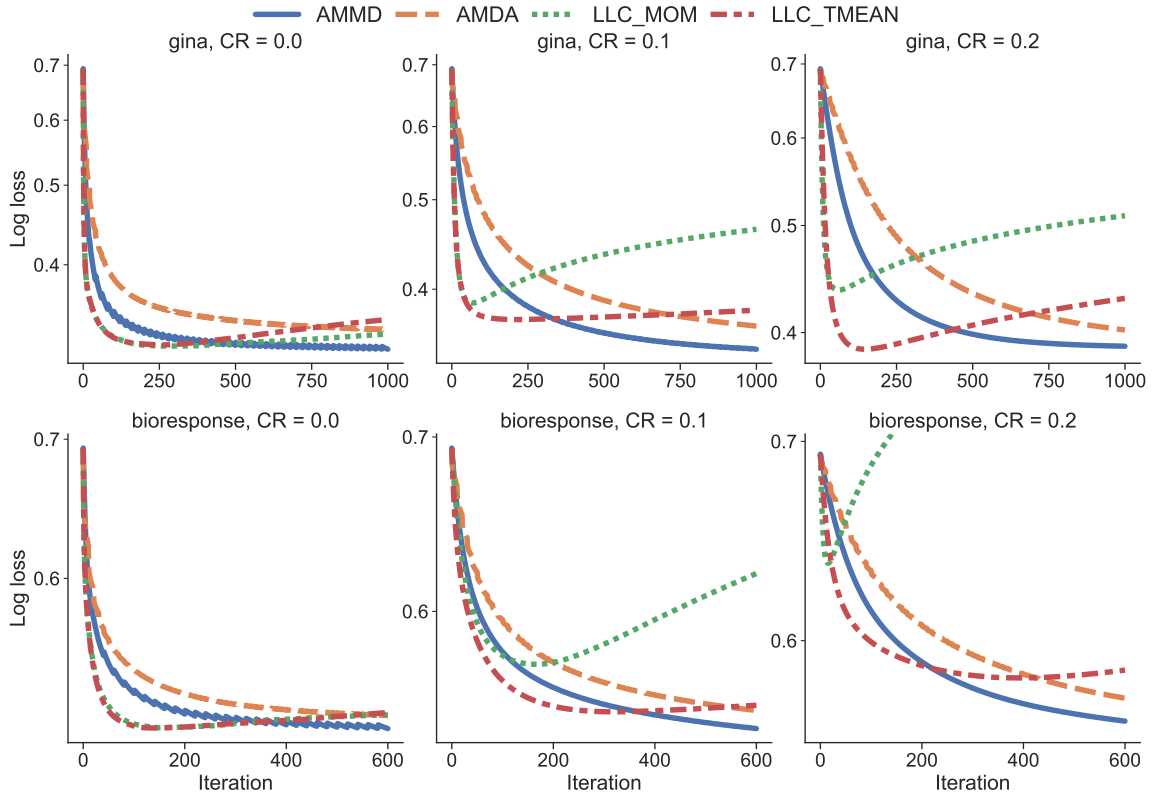

Figure 2: Log loss (y-axis) along training iterations (x-axis) on two data sets (rows) for 0% corruption (first column), 10% corruption (middle column) and 20% corruption (last column).

7. Conclusion

In this work, we address the problem of robust supervised linear learning in the high-dimensional setting. In order to cover both smooth and non-smooth loss functions, we propose two optimisation algorithms which enjoy linear convergence speeds with only a mild dependence on the dimension. We combine these algorithms with various robust mean estimators, each of them tailored for a specific variant of the sparse estimation problem. We show that the said estimators are robust to heavy-tailed and corrupted data and allow to reach the optimal statistical rates for their respective

instances of sparse estimation problems. Furthermore, their computation is efficient which favorably reflects on the computational cost of the overall procedure. We also confirm our theoretical results through numerical experiments where we evaluate our algorithms in terms of speed, robustness and performance of the final estimates. Finally, we compare our performances with the most relevant concurrent works and discuss the main differences. Perspectives for future work include considering other types of sparsity, divising an algorithm capable of reaching the optimal $s \log(d/s)/n$ rate for vanilla sparsity or considering problems beyond recovery of a single parameter such as, for example, additive sparse and low-rank matrix decomposition.

Acknowledgments

Appendix A. Proofs

A.1 Proof of Lemma 7

Let $\theta \in \Theta$, using Assumption 2 we have:

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

Taking the expectation and using Assumption 2 again shows that the objective $\mathcal{L}(\theta)$ is well defined. Next, for all $j \in [\![d]\!]$, simple algebra gives:

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

Recall that we assume $\mathbb{E}|X^j|^2 < \infty$ and $\mathbb{E}|YX^j|^2 < \infty$, moreover, using a Cauchy Schwarz inequality, we find:

$$\mathbb{E}\left|(X^k)X^j\right|^2 \leq \sqrt{\mathbb{E}|X^k|^4 \mathbb{E}|X^j|^4},$$

which is also assumed finite. This concludes the proof of Lemma 7.

A.2 Proofs for Section 3

A.2.1 PROOF OF PROPOSITION 2

We use the abbreviations $\widehat{g}_t = \widehat{g}(\theta_t)$ and $g_t = g(\theta_t)$. Let $\phi \in \Theta$ be any parameter, we first write the optimality condition of the proximal operator defining each step $\theta_{t+1} = \mathrm{prox}_\beta(\widehat{g}_t, \theta_t; \theta_0, \Theta)$. Using the convexity and smoothness properties of the objective $\mathcal{L}$, we find that :

$$
\begin{aligned}
\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\phi) &= \mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) + \mathcal{L}(\theta_t) - \mathcal{L}(\phi) \\
&\leq \langle g(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 + \langle g(\theta_t), \theta_t - \phi\rangle \\
&= \langle g_t, \theta_{t+1} - \phi\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2.
\end{aligned}
\tag{18}
$$

We have $\widehat{g}_t = g_t + \epsilon_t$ and the optimality condition says that for all $\phi \in \Theta$ we have the inequality :

$$\langle \beta\widehat{g}_t, \phi - \theta_{t+1}\rangle + \langle \nabla\omega_{\theta_0}(\theta_{t+1}) - \nabla\omega_{\theta_0}(\theta_t), \phi - \theta_{t+1}\rangle \geq 0.$$

Plugging this into (18) we get :

$$
\begin{aligned}
\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\phi) &\leq \frac{1}{\beta}\langle \nabla\omega_{\theta_0}(\theta_{t+1}) - \nabla\omega_{\theta_0}(\theta_t), \phi - \theta_{t+1}\rangle + \langle \epsilon_t, \phi - \theta_{t+1}\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= \frac{1}{\beta}(V_{\theta_0}(\phi, \theta_t) - V_{\theta_0}(\theta_{t+1}, \theta_t) - V_{\theta_0}(\phi, \theta_{t+1})) + \langle \epsilon_t, \phi - \theta_{t+1}\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&\leq \frac{1}{\beta}(V_{\theta_0}(\phi, \theta_t) - V_{\theta_0}(\phi, \theta_{t+1})) + \langle \epsilon_t, \phi - \theta_{t+1}\rangle,
\end{aligned}
$$

where the last step is due to the choice $\beta \leq 1/L$ and the strong convexity of V and the second step follows from the remarkable identity :

$$\langle \nabla_\theta V_{\theta_0}(\theta, \theta'), z - \theta \rangle = V_{\theta_0}(z, \theta') - V_{\theta_0}(\theta, \theta') - V_{\theta_0}(z, \theta) \quad \text{for all } z, \theta, \theta', \theta_0 \in \mathbb{R}^d.$$

It suffices to multiply the previous inequality by β, sum it for $t = 0 \ldots, T - 1$ and use the convexity of $\mathcal{L}$ to find that $\widehat{\theta}_T = \sum_{t=1}^{T} \theta_t / T$ satisfies :

$$\mathcal{L}(\widehat{\theta}_T) - \mathcal{L}(\phi) \leq \frac{V_{\theta_0}(\phi, \theta_0) - V_{\theta_0}(\phi, \theta_T)}{\beta T} + \frac{1}{T} \sum_{t=0}^{T-1} \langle \epsilon_t, \phi - \theta_{t+1} \rangle.$$

Then, it only remains to choose $\phi = \theta^\star$ to finish the proof.

A.2.2 PROOF OF PROPOSITION 3

We proceed similarly to the previous Proposition. As previously we have :

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\phi) \leq \langle g_t, \theta_{t+1} - \phi \rangle + \frac{L}{2} \|\theta_{t+1} - \theta_t\|^2,$$

where $\widehat{g}_t = g_t + \epsilon_t$. Let $\phi \in \Theta$, the optimality condition of $\theta_{t+1} = \widehat{\text{prox}}_\beta(\widehat{g}_t, \theta_t; \theta_0, \Theta)$ reads :

$$\langle \beta(\widehat{g}_t + \bar{\epsilon}\partial_{\|\cdot\|}(\theta_{t+1} - \theta_t)), \phi - \theta_{t+1} \rangle + \langle \nabla \omega_{\theta_0}(\theta_{t+1}) - \nabla \omega_{\theta_0}(\theta_t), \phi - \theta_{t+1} \rangle \geq 0,$$

where $\partial_{\|\cdot\|}(\theta)$ is any subgradient of $\|\cdot\|$ at θ. Plugging this into (18) we get :

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\phi) \leq \frac{1}{\beta}\langle \nabla \omega_{\theta_0}(\theta_{t+1}) - \nabla \omega_{\theta_0}(\theta_t), \phi - \theta_{t+1} \rangle + \langle \epsilon_t + \bar{\epsilon}\partial_{\|\cdot\|}(\theta_{t+1} - \theta_t), \phi - \theta_{t+1} \rangle$$

$$+ \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2$$

$$= \frac{1}{\beta}(V_{\theta_0}(\phi, \theta_t) - V_{\theta_0}(\theta_{t+1}, \theta_t) - V_{\theta_0}(\phi, \theta_{t+1}))$$

$$+ \langle \epsilon_t + \bar{\epsilon}\partial_{\|\cdot\|}(\theta_{t+1} - \theta_t), \phi - \theta_{t+1} \rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2$$

$$\leq \frac{1}{\beta}(V_{\theta_0}(\phi, \theta_t) - V_{\theta_0}(\phi, \theta_{t+1})) + \langle \epsilon_t + \bar{\epsilon}\partial_{\|\cdot\|}(\theta_{t+1} - \theta_t), \phi - \theta_{t+1} \rangle, \quad (19)$$

where the last step is due to the choice $\beta \leq 1/L$ and the strong convexity of V and the second step follows from the remarkable identity :

$$\langle \nabla_\theta V_{\theta_0}(\theta, \theta'), z - \theta \rangle = V_{\theta_0}(z, \theta') - V_{\theta_0}(\theta, \theta') - V_{\theta_0}(z, \theta) \quad \text{for all } z, \theta, \theta', \theta_0 \in \mathbb{R}^d.$$

Notice that since $\|\epsilon_t\|_* \leq \bar{\epsilon}$ for all t and using the identity $\langle \partial_{\|\cdot\|}(\theta), \theta \rangle = \|\theta\|$ which holds for any norm $\|\cdot\|$ we find :

$$\langle \epsilon_t + \bar{\epsilon}\partial_{\|\cdot\|}(\theta_{t+1} - \theta_t), \theta_t - \theta_{t+1} \rangle = \langle \epsilon_t, \theta_t - \theta_{t+1} \rangle - \bar{\epsilon}\|\theta_{t+1} - \theta_t\| \leq 0,$$

so that by taking $\phi = \theta_t$ in (19) we find that :

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t),$$

i.e. $\mathcal{L}(\theta_t)$ is monotonously decreasing with t. Using this observation, it suffices to average (19) over $t = 0 \ldots, T - 1$ to find that :

$$\mathcal{L}(\theta_T) - \mathcal{L}(\phi) \leq \frac{V_{\theta_0}(\phi, \theta_0) - V_{\theta_0}(\phi, \theta_T)}{\beta T} + \frac{1}{T} \sum_{t=0}^{T-1} \langle \epsilon_t + \bar{\epsilon} \partial_{\|\cdot\|}(\theta_{t+1} - \theta_t), \phi - \theta_{t+1} \rangle.$$

From here, the final bound is straight forward to derive by replacing $\phi = \theta^\star$ and using the fact that $\|\partial_{\|\cdot\|}(\theta)\|_* \leq 1$.

A.2.3 PROOF OF THEOREM 4

The following Lemma is needed for this proof and that of Theorem 6.

Lemma 15 (Juditsky et al. 2022, Lemma A.1) *Let $\theta^\star \in \Theta$ be s-sparse, $\theta \in \Theta$, and let $\theta_s = $ sparse$(\theta) \in \mathrm{argmin}\{\|\mu - \theta\| : \mu \in \Theta$ s-sparse$\}$. We have :*

$$\|\theta_s - \theta^\star\| \leq \sqrt{2s}\|\theta_s - \theta^\star\|_2 \leq 2\sqrt{2s}\|\theta - \theta^\star\|_2.$$

We would like to show by induction that $\|\theta^{(k)} - \theta^\star\| \leq R_k$ for $k \geq 0$. In the base case $k = 0$, we have $\|\theta^{(0)} - \theta^\star\| \leq R = R_0$. For a phase $k + 1 \geq 1$ of the approximate Mirror Descent algorithm, assuming the property holds for k, by applying Lemma 15 and Proposition 3 we find :

$$\mathcal{L}(\widetilde{\theta}^{(k+1)}) - \mathcal{L}(\theta^\star) \leq \frac{\nu R_k^2}{\beta T_{k+1}} + 4\bar{\epsilon} R_k \leq 5\bar{\epsilon} R_k, \tag{20}$$

where the last inequality uses that that $T_{k+1} = \left\lceil \frac{\nu R_k}{\beta \bar{\epsilon}} \right\rceil$. Using the quadratic growth hypothesis (Assumption 4) leads to :

$$\|\theta^{(k+1)} - \theta^\star\|^2 \leq 2\bar{s}\|\theta^{(k+1)} - \theta^\star\|_2^2 \leq 8\bar{s}\|\widetilde{\theta}^{(k+1)} - \theta^\star\|_2^2 \leq \frac{40\bar{s}\bar{\epsilon} R_k}{\kappa}. \tag{21}$$

We have just obtained the bound $\|\theta^{(k+1)} - \theta^\star\| =: \widehat{R}_{k+1} \leq h(R_k) := \sqrt{\frac{40\bar{s}\bar{\epsilon} R_k}{\kappa}}$. It is easy to check that $h(r)$ has a unique fixed point $R^\star := \frac{40\bar{s}\bar{\epsilon}}{\kappa}$ and that for $r \geq R^\star$ we have $h'(r) \leq 1/2$. Assuming that the former bound holds for $r = R_k$ (otherwise there is nothing to prove) we find :

$$\widehat{R}_{k+1} - R^\star \leq h(R_k) - h(R^\star) \leq \frac{1}{2}(R_k - R^\star)$$

$$\implies \widehat{R}_{k+1} \leq \frac{1}{2}(R_k + R^\star) = R_{k+1},$$

this finishes the induction argument. By unrolling the recursive definition of R_k, we obtain that, for all $k \geq 1$:

$$R_k \leq 2^{-k} R_0 + R^\star = 2^{-k} R_0 + \frac{40\bar{s}\bar{\epsilon}}{\kappa}.$$

The main bound of the Theorem then follows by plugging the above inequality with $k = K - 1$ into (21) and using the fact that $R_0 \geq R^\star$ and the standard inequality $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ which holds for all $a, b \geq 0$. The bound on the objective is obtained similarly.

Let us compute T, the total number of iterations necessary for this bound to hold. Given the minimum number of iterations T_k necessary for stage k we have :

$$T = \sum_{k=1}^{K} T_k \leq \sum_{k=1}^{K} \left(\frac{\nu R_{k-1}}{\beta \bar{\epsilon}} + 1 \right) \leq K + \frac{\nu}{\beta \bar{\epsilon}} \sum_{k=0}^{K-1} (2^{-k} R_0 + R^\star)$$
$$\leq \frac{2 R_0 \nu}{\beta \bar{\epsilon}} + K \left(1 + \frac{40 \nu \bar{s}}{\kappa \beta} \right).$$

This completes the proof.

A.3 Proofs for Section 4

A.3.1 PROOF OF PROPOSITION 5

For a sequence of iterates $(\theta_t)_{t=0\ldots T}$ we introduce the notations :

$$\ell_t(\theta) = \sum_{i=0}^{t} a_i (\mathcal{L}(\theta_i) + \langle \widehat{g}_i, \theta - \theta_i \rangle) \quad \text{and} \quad \psi_t^* = \min_{\theta \in \Theta} \ell_t(\theta) + \gamma_t \omega(\theta).$$

We will show the following inequality by induction :

$$A_t f(\theta_t) \leq \psi_t^* + \widehat{B}_t,$$

where we define $\widehat{B}_t = \sum_{i=0}^{t} \frac{a_i^2}{2\gamma_{i-1}} \|g_i\|_*^2 + 2 a_i R \|\epsilon_i\|_*$ with the convention $\gamma_{-1} = \gamma_0$. Assume it holds for $t \geq 0$, since $\gamma_{t+1} \geq \gamma_t$ we have :

$$\psi_{t+1}^* = \min_{\theta \in \Theta} \ell_t(\theta) + a_{t+1}(\mathcal{L}(\theta_{t+1}) + \langle \widehat{g}_{t+1}, \theta - \theta_{t+1} \rangle) + \gamma_{t+1} \omega(\theta)$$
$$\geq \min_{\theta \in \Theta} \ell_t(\theta) + a_{t+1}(\mathcal{L}(\theta_{t+1}) + \langle \widehat{g}_{t+1}, \theta - \theta_{t+1} \rangle) + \gamma_t \omega(\theta). \tag{22}$$

Note that, by definition, θ_t^+ realizes the minimum $\psi_t^* = \min_{\theta \in \Theta} \ell_t(\theta) + \gamma_t \omega(\theta) = \ell_t(\theta_t^+) + \gamma_t \omega(\theta_t^+)$, so that for all $\theta \in \Theta$ we have

$$\langle \nabla \ell_t(\theta_t^+) + \gamma_t \nabla \omega(\theta_t^+), \theta - \theta_t^+ \rangle \geq 0. \tag{23}$$

Also, using the convexity of ℓ_t and the strong convexity of $\omega(\cdot)$ we have :

$$\ell_t(\theta) + \gamma_t \omega(\theta) \geq \ell_t(\theta_t^+) + \langle \nabla \ell_t(\theta_t^+), \theta - \theta_t^+ \rangle +$$
$$\gamma_t \left(\omega(\theta_t^+) + \langle \nabla \omega(\theta_t^+), \theta - \theta_t^+ \rangle + \frac{1}{2} \|\theta - \theta_t^+\|^2 \right). \tag{24}$$

By combining Inequalities (22), (23) and (24), we find that :

$$\psi_{t+1}^* \geq \min_{\theta \in \Theta} \psi_t^* + \frac{\gamma_t}{2} \|\theta - \theta_t^+\|^2 + a_{t+1}(\mathcal{L}(\theta_{t+1}) + \langle \widehat{g}_{t+1}, \theta - \theta_{t+1} \rangle).$$

Now, using the induction hypothesis $A_t f(\theta_t) \leq \psi_t^* + \widehat{B}_t$, we compute that :

$$
\begin{aligned}
\psi_{t+1}^* &\geq \min_{\theta \in \Theta} A_t f(\theta_t) - \widehat{B}_t + \frac{\gamma_t}{2} \|\theta - \theta_t^+\|^2 + a_{t+1}(\mathcal{L}(\theta_{t+1}) + \langle \widehat{g}_{t+1}, \theta - \theta_{t+1} \rangle) \\
&\geq \min_{\theta \in \Theta} A_t(f(\theta_{t+1}) + \langle g_{t+1}, \theta_t - \theta_{t+1} \rangle) - \widehat{B}_t + \frac{\gamma_t}{2} \|\theta - \theta_t^+\|^2 + a_{t+1}(\mathcal{L}(\theta_{t+1}) + \\
&\quad \langle g_{t+1}, \theta - \theta_{t+1} \rangle) - 2Ra_{t+1}\epsilon_{t+1} \\
&\geq \min_{\theta \in \Theta} A_{t+1} f(\theta_{t+1}) - \widehat{B}_t + \frac{\gamma_t}{2} \|\theta - \theta_t^+\|^2 + a_{t+1}\langle g_{t+1}, \theta - \theta_t^+ \rangle - 2Ra_{t+1}\epsilon_{t+1} \\
&\geq \min_{\theta \in \Theta} A_{t+1} f(\theta_{t+1}) - \widehat{B}_t - \frac{a_{t+1}^2}{2\gamma_t} \|\widehat{g}_t\|_*^2 - 2Ra_{t+1}\|\epsilon_{t+1}\|_* \\
&= \min_{\theta \in \Theta} A_{t+1} f(\theta_{t+1}) - \widehat{B}_{t+1},
\end{aligned}
$$

where the penultimate inequality uses that $A_{t+1}\theta_{t+1} = A_t \theta_t + a_{t+1}\theta_t^+$. It only remains to check the base case :

$$
\begin{aligned}
\psi_0^* &= \min_{\theta \in \Theta} a_0(\mathcal{L}(\theta_0) + \langle \widehat{g}_0, \theta - \theta_0 \rangle) + \frac{\gamma_0}{2}\omega(\theta) \\
&\geq \min_{\theta \in \Theta} a_0(\mathcal{L}(\theta_0) + \langle g_0, \theta - \theta_0 \rangle) + \frac{\gamma_0}{2}\omega(\theta) + \min_{\theta \in \Theta} a_0 \langle \epsilon_0, \theta - \theta_0 \rangle \\
&\geq A_0 \mathcal{L}(\theta_0) - \frac{a_0^2}{2\gamma_{-1}} \|g_0\|_*^2 - 2a_0 R\|\epsilon_0\|_* = A_0 \mathcal{L}(\theta_0) - \widehat{B}_0,
\end{aligned}
$$

which completes the induction. Now we can compute :

$$
\begin{aligned}
A_t \langle s_t, \theta^\star \rangle + \gamma_t \omega(\theta^\star) &\geq A_t \langle s_t, \theta_t^+ \rangle + \gamma_t \omega(\theta_t^+) + \frac{\gamma_t}{2}\|\theta_t^+ - \theta^\star\|^2 \\
&\geq \psi_t^* - \sum_{i=0}^{t} a_i(\mathcal{L}(\theta_i) - \langle \widehat{g}_i, \theta_i \rangle) + \frac{\gamma_t}{2}\|\theta_t^+ - \theta^\star\|^2 \\
&\geq A_t(\mathcal{L}(\theta_t) - \mathcal{L}(\theta^\star) + \langle s_t, \theta^\star \rangle) - \widehat{B}_t + \\
&\quad \sum_{i=0}^{t} a_i(\mathcal{L}(\theta^\star) - \mathcal{L}(\theta_i) - \langle g_i, \theta^\star - \theta_i \rangle) - \sum_{i=0}^{t} a_i \langle \epsilon_i, \theta^\star - \theta_i \rangle + \frac{\gamma_t}{2}\|\theta_t^+ - \theta^\star\|^2,
\end{aligned}
$$

which, by rearranging and using the convexity of $\mathcal{L}$, leads to :

$$
\begin{aligned}
A_t(\mathcal{L}(\theta_t) - \mathcal{L}(\theta^\star)) + \frac{\gamma_t}{2}\|\theta_t^+ - \theta^\star\|^2 &\leq \gamma_t \omega(\theta^\star) + \widehat{B}_t + \sum_{i=0}^{t} a_i \langle \epsilon_i, \theta^\star - \theta_i \rangle \\
&\leq \gamma_t \omega(\theta^\star) + \sum_{i=0}^{t} \frac{a_i^2}{2\gamma_{i-1}} \|g_i\|_*^2 + 4A_t R\bar{\epsilon}.
\end{aligned}
$$

For the particular case $a_t = 1$ and $\gamma_t = \sqrt{t+1}$ we have $A_t = t+1$. Moreover, by summing the inequality $\frac{1}{\sqrt{i+1}} \leq \int_i^{i+1} \frac{du}{\sqrt{u}}$ we get that $\sum_{i=1}^{t} \frac{1}{\sqrt{i}} \leq 2\sqrt{t} - 1$. Using this estimate and the Lipschitz property of the objective quickly yields the last part of the Theorem.

A.3.2 PROOF OF THEOREM 6

We will show by induction that the inequality $\|\theta^{(k)} - \theta^\star\| \le R_k$ holds for all $k \ge 0$. The case $k = 0$ holds by assumption. Note that based on Proposition 5 with the choice $a_i = R$ and $\gamma_i = \sqrt{i+1}$ and using the quadratic growth bound (8) for ω we get :

$$\mathcal{L}(\theta_{T'}) - \mathcal{L}(\theta^\star) \le \frac{(\nu + M^2)R}{\sqrt{T'}} + 4R\bar\epsilon \le 5\bar\epsilon R, \tag{25}$$

where the last step follows from the choice of T'. At the same time, at the end of stage $k+1$, we have the alternative :

- Either $\|\widetilde\theta^{(k+1)} - \theta^\star\|_2 \le \lambda$: then using Lemma 15 we find :

$$
\begin{aligned}
\left\|\theta^{(k+1)} - \theta^\star\right\|^2 = \left\| \mathrm{sparse}_{\bar s}(\widetilde\theta^{(k+1)}) - \theta^\star \right\|^2 &\le 8\bar s \left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2^2 \\
&\le \frac{16\lambda\bar s \left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2^2}{\lambda + \left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2} \le \frac{16\lambda\bar s}{\kappa}\big(\mathcal{L}(\widetilde\theta^{(k+1)}) - \mathcal{L}(\theta^\star)\big) \\
&\le \frac{80\lambda\bar s\bar\epsilon R_k}{\kappa} = R^\star R_k,
\end{aligned} \tag{26}
$$

where the last inequality is an application of (25) since $\widetilde\theta^{(k+1)} = \theta_{T'}^{(k+1)}$.

- Or we have $\|\widetilde\theta^{(k+1)} - \theta^\star\|_2 > \lambda$: then only a linear regime holds and using Lemma 15 we get :

$$
\begin{aligned}
\left\|\theta^{(k+1)} - \theta^\star\right\| \le \sqrt{2\bar s}\left\|\theta^{(k+1)} - \theta^\star\right\|_2 &\le 2\sqrt{2\bar s}\left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2 \\
&\le \frac{4\sqrt{2\bar s}\left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2^2/\lambda}{1 + \left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2/\lambda} = \frac{4\sqrt{2\bar s}\left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2^2}{\lambda + \left\|\widetilde\theta^{(k+1)} - \theta^\star\right\|_2} \\
&\le \frac{4\sqrt{2\bar s}}{\kappa}\big(\mathcal{L}(\widetilde\theta^{(k+1)}) - \mathcal{L}(\theta^\star)\big) \le \frac{20\bar\epsilon R_k\sqrt{2\bar s}}{\kappa} = \tau R_k,
\end{aligned} \tag{27}
$$

where we used the inequality $1 \le 2x/(1+x)$ valid for all $x \ge 1$ on the quantity $\|\widetilde\theta^{(k+1)} - \theta^\star\|_2/\lambda$.

Similarly to the proof of Theorem 4, inequality (26) implies that $\left\|\theta^{(k+1)} - \theta^\star\right\| \le \frac{1}{2}(R_k + R^\star)$ so that we obtained :

$$\left\|\theta^{(k+1)} - \theta^\star\right\| \le \sqrt{2\bar s}\|\theta^{(k+1)} - \theta^\star\|_2 \le 2\sqrt{2\bar s}\|\widetilde\theta^{(k+1)} - \theta^\star\|_2 \le \max\left(\tau R_k, \frac{1}{2}(R_k + R^\star)\right) = R_{k+1},$$

which finishes the induction to show (11). Inequality (12) then follows using (25). Note that, since we assume $\tau < 1$ and $R_0 \ge R^\star$, the sequence $(R_k)_{k\ge 0}$ is decreasing. We now distinguish two phases :

- The linear phase : if $\tau R_0 > \frac{1}{2}(R_0 + R^\star)$ (which implies $\tau > 1/2$ since $R_0 \ge R^\star$) then while $\tau R_k > \frac{1}{2}(R_k + R^\star)$ we have $R_{k+1} = \tau R_k = \tau^{k+1} R_0$ and the number of stages necessary to reverse the previous inequality is :

$$\log\Big(\frac{(2\tau - 1)R_0}{R^\star}\Big)\Big/ \log(1/\tau) \le \log\Big(\frac{R_0}{R^\star}\Big)\Big/ \log(1/\tau).$$

33

- The quadratic phase : let $K_1 \geq 0$ be the first stage index such that we have $\tau R_k \leq \frac{1}{2}(R_k + R^\star)$ and so $R_{k+1} = \frac{1}{2}(R_k + R^\star)$ and by iteration $R_{l+K_1} \leq 2^{-l} R_{K_1} + R^\star$. In all cases $R_{K_1} \leq R_0$ so the number of necessary stages is :

$$\frac{\log(R_{K_1}/R^\star)}{\log(2)} \leq \frac{\log(R_0/R^\star)}{\log(2)}.$$

We have shown that the overall number of necessary stages is at most $\log(R_0/R^\star)\left(\frac{1}{\log(2)} + \frac{1}{\log(1/\tau)}\right)$. The Theorem's final claim then follows since the number of per-stage iterations is constant equal to $T' = \left\lceil \left(\frac{\nu + M^2}{\bar\epsilon}\right)^2 \right\rceil$.

A.3.3 DUAL AVERAGING FOR VANILLA SPARSE ESTIMATION

Corollary 16 *In the context of Theorem 6 and Lemma 8, let the AMDA algorithm be run starting from $\theta_0 \in \Theta = B_{\|\cdot\|}(\theta_0, R)$ using the coordinatewise trimmed mean estimator with sample splitting i.e. at each iteration a different batch of size $\widetilde{n} = n/T$ is used for gradient estimation with confidence $\widetilde{\delta} = \delta/T$ where T is the total number of iterations. Let K be the number of stages and $\widehat\theta$ the obtained estimator. Denote $\sigma_{\max}^2 = \sup_{\theta \in \Theta} \max_{j \in \llbracket d \rrbracket} \mathrm{Var}(\ell'(\theta^\top X, Y)X^j)$, with probability at least $1 - \delta$, the latter satisfies :*

$$\left\|\widehat\theta - \theta^\star\right\|_2 \leq \tau^{K \wedge K_1} 2^{(K_1 - K)\wedge 0} \frac{R}{\sqrt{2\bar s}} + \frac{280\lambda\sqrt{2\bar s}\sigma_{\max}}{\kappa}\sqrt{4\eta + 6\frac{\log(4/\widetilde\delta) + \log(d)}{\widetilde n}}.$$

with K_1 an integer such that $K_1 \leq \log\left(\frac{\kappa R}{80\lambda\bar s\bar\epsilon}\right)\Big/ \log(1/\tau)$ with $\tau = \frac{10\sqrt{8\bar s\bar\epsilon}}{\kappa} < 1$ by assumption and

$$\bar\epsilon = 7\sigma_{\max}\sqrt{4\eta + 6\frac{\log(4/\delta) + \log(d)}{n}}.$$

Proof By the proof of Theorem 6, we have $R = R_0$ and K_1 is defined as the first stage index k such that we have $\tau R_k \leq \frac{1}{2}(R_k + R^\star)$ implying that $R_{k+1} = \frac{1}{2}(R_k + R^\star)$ and hence $R_{l+K_1} \leq 2^{-l} R_{K_1} + R^\star$ for $l \geq 0$. We also had $K_1 \leq \log\left(\frac{R}{R^\star}\right)\Big/ \log(1/\tau)$ with $R^\star = \frac{80\lambda\bar s\bar\epsilon}{\kappa}$. Using Theorem 6, it follows that:

$$\sqrt{2\bar s}\left\|\theta^{(k)} - \theta^\star\right\|_2 \leq R_k \leq 2^{(K_1 - k)\wedge 0}\tau^{k \wedge K_1} R + R^\star,$$

whence the result is easily obtained by plugging the value of $R^\star$ and using Lemma 8 with a union bound argument over all iterations T in order to bound $\bar\epsilon = \max_i \|\epsilon_i\|_*$ as defined in Proposition 5. ∎

A.4 Closed form computation of the prox operator

In both Sections 3 and 4 the optimization methods are defined using a prox operator which involves solving an optimization problem of the following form:

$$\underset{\|\theta\| \leq R}{\operatorname{argmin}}\langle u, \theta\rangle + \omega(\theta)$$

for some $R > 0$ and $u \in \Theta^*$.

A.4.1 VANILLA/GROUP-SPARSE CASE

We consider the group sparse case where $\Theta \subset \mathbb{R}^{d \times K}$, the groups are the rows of $\theta \in \Theta$ and we use the norm $\| \cdot \| = \| \cdot \|_{1,2}$ and the usual scalar product $\langle u, \theta \rangle = u^\top \theta$. We use the prox function ω defined as $\omega(\theta) = C\|\theta\|_{p,2}^2 = C\left(\sum_{i=1}^d \|\theta_{i,:}\|_2^p\right)^{2/p}$ with $p = 1 + 1/\log(d)$ and $\theta_{i,:}$ the i-th row of θ. Note that, for $K = 1$ we retrieve the usual linear learning setting. The setting $K > 1$ can be used, for example, for multiclass classification.

In order to obtain a closed form solution, we start by writing the Lagrangian :

$$\mathcal{L}(\theta) = \langle u, \theta \rangle + \omega(\theta) + \lambda(\|\theta\| - R), \tag{28}$$

where we introduce the multiplier $\lambda \geq 0$. We initially assume the latter known and try to find a critical point for $\mathcal{L}$ that is $\theta \in \mathbb{R}^{d \times K}$ such that :

$$\partial\mathcal{L}(\theta) = u + \nabla\omega(\theta) + \partial\| \cdot \|_{p,2}(\theta) \ni 0,$$

where we denoted $\partial\| \cdot \|_{1,2}(\theta)$ the subdifferential of the norm $\| \cdot \|_{1,2}$ since the latter is not differentiable whenever $\theta_{i,:} = 0$ for some $i \in [\![d]\!]$.

Defining the function $h_\alpha(\theta) : \mathbb{R}^{d \times K} \to \mathbb{R}^{d \times K}$ such that $h_\alpha(\theta)_{i,j} = \frac{\theta_{i,j}}{\|\theta_{i,:}\|_2^\alpha}$ for θ such that $\theta_{i,:} \neq 0$ for all i, we can write for such θ :

$$\nabla\omega(\theta) = 2C\|\theta\|_{p,2}^{2-p} h_{2-p}(\theta) \quad \text{and} \quad \partial\| \cdot \|_{1,2}(\theta) = h_1(\theta).$$

When $\theta_{i,:} = 0$ for some i, a subgradient of $\| \cdot \|_{1,2}$ can be obtained by using this definition and plugging any subunit vector for index i. Assuming that $\theta_{i,:} \neq 0$ for all $i \in [\![d]\!]$, a critical point of the Lagrangian must satisfy for all i, j :

$$u_{i,j} + \theta_{i,j}\left(2C\left(\frac{\|\theta\|_{p,2}}{\|\theta_{i,:}\|_2}\right)^{2-p} + \frac{\lambda}{\|\theta_{i,:}\|_2}\right) = 0.$$

From here, a quick computation yields for all i that :

$$\|\theta_{i,:}\|_2 = \frac{\|u_{i,:}\|_2}{\left(2C\left(\frac{\|\theta\|_{p,2}}{\|\theta_{i,:}\|_2}\right)^{2-p} + \frac{\lambda}{\|\theta_{i,:}\|_2}\right)} \quad \text{and hence} \quad 2C\|\theta\|_{p,2}^{2-p}\|\theta_{i,:}\|_2^{p-1} = \|u_{i,:}\|_2 - \lambda.$$

Notice that this equality cannot hold when $\|u_{i,:}\|_2 < \lambda$, in this case, we deduce that $\theta_{i,:} = 0$ which satisfies the critical point condition. This leads to the relation :

$$\theta_{i,j} = \frac{-\beta_i u_{i,j}}{2C\left(\frac{\|\theta\|_{p,2}}{\|\theta_{i,:}\|_2}\right)^{2-p} + \frac{\lambda}{\|\theta_{i,:}\|_2}} \quad \text{with } \beta_i = \mathbf{1}_{\|u_{i,:}\|_2 > \lambda}.$$

From here, it is easy to figure out that $\|\theta\|_{p,2} = \left(\sum_{i=1}^d \beta_i\left(\frac{\|u_{i,:}\|_2 - \lambda}{2C}\right)^{p/(p-1)}\right)^{(p-1)/p}$. All computations are now possible knowing λ.

To find the latter's value, we plug the formula we have for θ into the constraint $\|\theta\|_{1,2} \leq R$. After a few manipulations, we find the constraint is satisfied for λ such that :

$$\frac{1}{2C}\left(\sum_{i=1}^d \beta_i\left(\|u_{i,:}\|_2 - \lambda\right)^{1/(p-1)}\right)\left(\sum_{i=1}^d \beta_i\left(\|u_{i,:}\|_2 - \lambda\right)^{p/(p-1)}\right)^{(p-2)/p} \leq R,$$

(recall that the β_is also depend on λ). It only remains to choose the smallest $\lambda \geq 0$ such that the above inequality holds.

A.4.2 LOW-RANK MATRIX CASE

In the low-rank matrix case the parameter space is $\Theta \subset \mathbb{R}^{p \times q}$ and the norm $\|\cdot\|$ represents the nuclear norm $\|\theta\| = \|\sigma(\theta)\|_1$ with $\sigma(\theta)$. The d.-g.f./proximal function is the scaled squared p-Schatten norm $\omega(\theta) = C\|\sigma(\theta)\|_p^2$ with $p = 1 + \frac{1}{12\log(q)} = 1 + r$ and the scalar product is defined as $\langle u, v \rangle = \operatorname{tr}(u^\top v)$.

We introduce the notation $h_r(\theta) = \|\sigma(\theta)\|_{1+r}$. Letting $UDV^\top$ denote an SVD of $\theta \in \mathbb{R}^{p \times q}$, the gradient of h_r at θ for $r > 0$ is given by $\nabla h_r(\theta) = U\left(\frac{D}{h_r(\theta)}\right)^r V^\top$. The nuclear norm h_0 is not differentiable but a subgradient is given by $\partial h_0(\theta) = UV^\top + W$ for any W such that $\|W\|_{\mathrm{op}} \leq 1$ and $UU^\top W = 0$ and $WVV^\top = 0$ (see Watson, 1992).

In order to define the prox operator in this setting we need to solve problem (28) again which amounts to finding a critical point $\theta = UDV^\top$ such that :

$$u + 2Ch_r(\theta)^{1-r}UD^r V^\top + \lambda UV^\top \ni 0.$$

We define θ by choosing U and V such that $u = UD_u V^\top$ is an SVD of u. Thanks to this choice, it only remains to choose $D = \operatorname{diag}(\sigma(\theta))$ properly in order to have :

$$\sigma(u) + 2C\|\sigma(\theta)\|_{1+r}^{1-r}\sigma(\theta)^r + \lambda(\mathbf{1}_{\sigma(\theta)\neq 0} + w) = 0,$$

where the power $\sigma(\theta)^r$ is computed coordinatewise, $\mathbf{1}_{\sigma(\theta)\neq 0}$ is the indicator vector of non zero coordinates of $\sigma(\theta)$ and for some vector w such that $|w_j| \leq 1$ for all j supported on the coordinates where $\sigma(\theta) = 0$.

The problem then becomes analogous to finding the proximal operator for Vanilla sparsity and after some computations we find that the solution is given by :

$$\sigma(\theta) = -\frac{\mathcal{T}_\lambda(u)^{1/r}}{2C\|\mathcal{T}_\lambda(u)\|_{(1+r)/r}^{(1-r)/r}},$$

where the soft threshold operator is defined by $\mathcal{T}_\lambda(u)_j = \operatorname{sign}(u_j)\max(0, |u_j| - \lambda)$ and λ is the smallest real number such that :

$$\frac{1}{2C}\|\mathcal{T}_\lambda(u)\|_{1/r}^{1/r} \cdot \|\mathcal{T}_\lambda(u)\|_{(r-1)/r}^{(r+1)/r} \leq R.$$

A.5 Proof of Proposition 13

Let $B_1, \ldots, B_K$ be a partition of $[\![n]\!]$ into disjoint equal sized blocks. We assume that the number of outliers is $|\mathcal{O}| \leq (1-\varepsilon)K/2$ where $0 < \varepsilon < 1$ will be fixed later. Let $\mathcal{K} = \{k \in [\![K]\!] : B_k \cap \mathcal{O} = \emptyset\}$ be the set of outlier-free blocks. Denote the block means for $j \in [\![K]\!]$ as

$$\xi^{(j)} = \frac{1}{m\chi}\sum_{i\in B_j}\psi(\chi\widetilde{A}_i),$$

where χ is temporarily chosen as $\chi = \sqrt{\frac{2m\log(2(p+q)/\delta')}{v(A)}}$ for some $0 < \delta' < 1/2$. By applying Corollary 3.1 from (Minsker, 2018), we obtain that with probability at least $1 - \delta'$

$$\left\|\widehat{\mu}^{(k)} - \mu\right\|_{\mathrm{op}} \leq \sqrt{\frac{2v(A)\log(2(p+q)/\delta'))}{m}}. \tag{29}$$

Now let $r_{jl} = \|\widehat{\mu}^{(j)} - \widehat{\mu}^{(l)}\|_{\mathrm{op}}$, denote $r^{(j)}$ the increasingly sorted version of $r_{j:}$ and let $\widehat{j} \in \mathrm{argmin}_j \, r^{(j)}_{\lceil K/2 \rceil}$ so that $\widehat{\mu} = \widehat{\mu}^{(\widehat{j})}$.

Define the events $E_j = \left\{ \|\widehat{\mu}^{(j)} - \mu\|_{\mathrm{op}} \le \sqrt{\frac{2v(A)\log(2(p+q)/\delta'))}{m}} \right\}$ and assume that we have $\sum_{k=1}^K \mathbf{1}_{E_k} > K/2$ i.e. over half of the block means satisfy Inequality (29) simultaneously. Then there exists $j' \in [\![K]\!]$ such that the block $\widehat{j}$ satisfies

$$r^{(\widehat{j})}_{K/2} = \|\widehat{\mu}^{(\widehat{j})} - \widehat{\mu}^{(j')}\|_{\mathrm{op}} \le \|\widehat{\mu}^{(\widehat{j})} - \mu\|_{\mathrm{op}} + \|\widehat{\mu}^{(j')} - \mu\|_{\mathrm{op}} \le 2\sqrt{\frac{2v(A)\log(2(p+q)/\delta'))}{m}}.$$

Moreover, among the $K/2$ block means closest to $\widehat{\mu}^{(\widehat{j})}$, at least one of them $\widehat{\mu}^{(j'')}$ satisfies $\|\widehat{\mu}^{(j'')} - \mu\|_{\mathrm{op}} \le \sqrt{\frac{2v(A)\log(2(p+q)/\delta'))}{m}}$ thus we find :

$$
\begin{aligned}
\|\widehat{\mu} - \mu\|_{\mathrm{op}} = \|\widehat{\mu}^{(\widehat{j})} - \mu\|_{\mathrm{op}} &\le \|\widehat{\mu}^{(\widehat{j})} - \widehat{\mu}^{(j'')}\|_{\mathrm{op}} + \|\widehat{\mu}^{(j'')} - \mu\|_{\mathrm{op}} \\
&\le r^{(\widehat{j})}_{K/2} + \sqrt{\frac{2v(A)\log(2(p+q)/\delta'))}{m}} \le 3\sqrt{\frac{2v(A)\log(2(p+q)/\delta'))}{m}}.
\end{aligned}
$$

Finally, let us show that $\sum_{k=1}^K \mathbf{1}_{E_k} > K/2$ happens with high probability. Observe that for $j \in \mathcal{K}$ the variables $\mathbf{1}_{\overline{E}_j}$ are stochastically dominated by Bernoulli variables with parameter δ' so that their sum is stochastically dominated by a Binomial random variable $S := \mathrm{Bin}(|\mathcal{K}|, \delta')$. We compute :

$$
\begin{aligned}
\mathbb{P}\Big(\sum_{k=1}^K \mathbf{1}_{E_k} < K/2 \Big) = \mathbb{P}\Big(\sum_{k=1}^K \mathbf{1}_{\overline{E}_k} > K/2 \Big) \\
\le \mathbb{P}\Big(|\mathcal{O}| + \sum_{k \in \mathcal{K}} \mathbf{1}_{\overline{E}_k} > K/2 \Big) \\
\le \mathbb{P}(S - \mathbb{E}S > K/2 - |\mathcal{O}| - \delta'|\mathcal{K}|) \\
\le \mathbb{P}(S - \mathbb{E}S > K(\varepsilon - 2\delta')/2) \\
\le \exp(-K(\varepsilon - 2\delta')^2/2)
\end{aligned}
$$

where we used that $|\mathcal{O}| \le (1-\varepsilon)K/2$, $|\mathcal{K}| \le K$ and Hoeffding's inequality at the end. Choosing $\varepsilon = 5/6$, $\delta' = 1/4$ and recalling the choice of K and that $m = n/K$ we finish the proof.

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. *Advances in Neural Information Processing Systems*, 25, 2012.

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

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

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

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212. PMLR, 2017.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Thomas Blumensath and Mike E Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):298–309, 2010.

Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9 (3):1103, 2015.

Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Peter Bühlmann and Sara Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57 (4):2342–2359, 2011.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. 2015.

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

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 774–782, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/chen13h.html`.

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

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

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

Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in Neural Information Processing Systems*, 17, 2004.

Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

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

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

David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1(2000):32, 2000.

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Conference on Learning Theory*, volume 10, pages 14–26. Citeseer, 2010.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of statistics*, 49(3):1239, 2021.

Peter Filzmoser and Klaus Nordhausen. Robust linear regression for high-dimensional data: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(4):e1524, 2021.

Stéphane Gaïffas and Ibrahim Merad. Robust supervised learning with coordinate gradient descent. *arXiv preprint arXiv:2201.13372*, 2022.

Frank R Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887 – 1896, 1971. doi: 10.1214/aoms/1177693054. URL `https://doi.org/10.1214/aoms/1177693054`.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on Statistics and Applied Probability*, 143:143, 2015.

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

Samuel B Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *arXiv: Statistics Theory*, 2018.

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

Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4):2282, 2010.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. ISSN 00034851. URL `http://www.jstor.org/stable/2238020`.

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

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

Prateek Jain, Nikhil Rao, and Inderjit S Dhillon. Structured sparse regression via greedy hard thresholding. *Advances in Neural Information Processing Systems*, 29, 2016.

Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(86)90174-X. URL `https://www.sciencedirect.com/science/article/pii/030439758690174X`.

Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, 30(9):121–148, 2011.

Anatoli Juditsky, Joon Kwon, and Éric Moulines. Unifying mirror descent and dual averaging. *arXiv preprint arXiv:1910.13742*, 2019.

Anatoli Juditsky, Andrei Kulunchakov, and Hlib Tsyntseus. Sparse recovery by reduced variance stochastic approximation. *Information and Inference: A Journal of the IMA*, 12(2):851–896, 11 2022. ISSN 2049-8772. doi: 10.1093/imaiai/iaac028. URL `https://doi.org/10.1093/imaiai/iaac028`.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

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

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

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

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

Liu Liu, Tianyang Li, and Constantine Caramanis. High Dimensional Robust M-Estimation: Arbitrary Corruption and Heavy Tails. *arXiv preprint arXiv:1901.08237*, 2019.

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

Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2:90–102, 2008.

Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

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

Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 175(3):957–973, 2019b.

Gábor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106, 2019c.

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

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

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

Stanislav Minsker, Mohamed Ndaoud, and Lang Wang. Robust and tuning-free sparse linear regression via square-root slope. *arXiv preprint arXiv:2210.16808*, 2022.

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

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yu Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Yurii Nesterov and Arkadi Nemirovski. On first-order algorithms for l1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

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

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

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

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Takeyuki Sasai. Robust and sparse estimation of linear regression coefficients with heavy-tailed noises and covariates. *arXiv preprint arXiv:2206.07594*, 2022.

Takeyuki Sasai and Hironori Fujisawa. Outlier robust and sparse estimation of linear regression coefficients. *arXiv preprint arXiv:2208.11592*, 2022.

Hanie Sedghi, Anima Anandkumar, and Edmond Jonckheere. Multi-step stochastic admm in high dimensions: Applications to sparse optimization and matrix decomposition. *Advances in Neural Information Processing Systems*, 27, 2014.

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

Jie Shen and Ping Li. A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691, 2017.

Weijie Su and Emmanuel Candès. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

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

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(92)90407-2. URL https://www.sciencedirect.com/science/article/pii/0024379592904072.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

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

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Tong Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.