Error bounds for sparse classifiers in high-dimensions

Antoine Dedieu
Vicarious AI

Abstract

We prove an L2 recovery bound for a family of sparse estimators defined as minimizers of some empirical loss functions – which include hinge loss and logistic loss. More precisely, we achieve an upper-bound for coefficients estimation scaling as $(k^*/n)\log(p/k^*)$: $n \times p$ is the size of the design matrix and k^* the dimension of the theoretical loss minimizer. This is done under standard assumptions, for which we derive stronger versions of a cone condition and a restricted strong convexity. Our bound holds with high probability and in expectation and applies to an L1-regularized estimator and to a recently introduced Slope estimator, which we generalize for classification problems. Slope presents the advantage of adapting to unknown sparsity. Thus, we propose a tractable proximal algorithm to compute it and assess its empirical performance. Our results match the best existing bounds for classification and regression problems. [1]

1 Introduction

Motivated by the increasing availability of very large-scale datasets, high-dimensional statistics has focused on analyzing the performance of sparse estimators. An estimator is said to be sparse if the response of an observation is given by a small number of coefficients: sparsity delivers better interpretability and often leads to computational efficiency. Statistical performance and L2 consistency for high-dimensional linear regression have been widely studied. For two polynomial-time sparse estimators, a Lasso (Tibshirani, 1996) and a Dantzig selector (Candes and Tao, 2007), Bickel et al. (2009) proved a $(k^*/n)\log(p)$ rate for the L2 estimation of the coefficients: $n \times p$ is the dimension of the input matrix and k^* the degree of sparsity of the vector used to generate the model. The optimality of this bound is essential for a theoretical understanding of the method performance. Candes and Davenport (2013) and Raskutti et al. (2011) proved a $(k^*/n)\log(p/k^*)$ lower bound for estimating the L2 norm of a sparse vector, regardless of the input matrix and estimation procedure. This optimal minimax rate is known to be achieved by a sparse but theoretically intractable BIC estimator (Bunea et al., 2007) which considers an L0 regularization. The BIC estimator adapts to unknown sparsity: the degree k^* does not have to be specified. Recently, Bellec et al. (2016) reached this optimal minimax bound for a Lasso estimator with knowledge of the sparsity k^*. They also proved that a recently introduced and polynomial-time Slope estimator (Bogdan et al., 2013) achieves this optimal rate while adapting to unknown sparsity.

Little work has been done on deriving (theoretical) upper bounds for the estimation error on high-dimensional classification problems: the literature has essentially focused on analysis of convergence (Tarigan et al., 2006; Zhang et al.). Recently, Peng et al. (2016) proved a $(k^*/n)\log(p)$ upper-bound for L2 coefficients estimation of a L1-regularized Support Vector Machines (SVM): k^* is now the sparsity of the theoretical minimizer to estimate. They recovered the rate proposed by Van de Geer (2008), which considered a weighted L1 norm for linear models. Ravikumar et al. (2010) obtained a similar bound for a L1-regularized Logistic Regression estimator in a binary Ising graph. Their frameworks and bounds are similar to the model proposed by Belloni et al. (2011) for L1-regularized Quantile Regression; this inspired us to include this problem in our framework. However, this rate of $(k^*/n)\log(p)$ is not the best known for a classification estimator: Plan and Vershynin (2013) proved a $k^*\log(p/k^*)$ error bound for estimating a single vector through sparse models – including 1-bit

[1]Antoine Dedieu's research was partially supported by the Office of Naval Research: N000141512342, by MIT and by Vicarious AI. Email: antoine@vicarious.com.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

compressed sensing and Logistic Regression – over a bounded set of vectors. Contrary to this work, our approach does not assume a generative vector and applies to a larger class of problems (SVM, Quantile Regression) and regularizations (Slope). In addition, our framework shares similarity with Section 4.4. of Negahban et al. (2009): the authors consider some subgaussian tails assumptions and restricted eigenvalue conditions to derive a restricted strong convexity condition similar to our Theorem 4. However, their results only apply to generalized linear models, and are weaker: the parameter $\tau(k)$ proposed in the tolerance function of the restricted strong convexity condition is higher than ours.

What this paper is about: In this paper, we propose a theoretical framework to analyze the properties of a general class of sparse estimators for classification problems – which includes SVM and Logistic Regression – with different regularization schemes. Our approach draws inspiration from the least squares regression case and illustrates the distinction between regression and classification studies. Our main results are first presented for a family of L1-regularized estimators. We achieve a $(k^*/n)\log(p/k^*)$ upper-bound for coefficients estimation, which holds with high probability and in expectation. In addition, we introduce a version of the Slope estimator for classification problems: we propose a proximal algorithm to compute the solution, and we prove that a tractable Slope estimator achieves a similar upper-bound while adapting to unknown sparsity. To the best of our knowledge, it is the first time any of these bounds is reached for the estimators considered.

The rest of this paper is organized as follows. Section 2 introduces and discusses common assumptions in the literature, and builds our framework of study in the case of L1-regularized estimators. Section 3 proves two essential results and derive our upper-bounds in Theorem 1 and Corollary 1. Finally, Section 4 defines and computes the Slope estimator for our class of problems and discusses its statistical performance.

2 General assumptions with an L1 regularization

We consider a set of training data $\{(\mathbf{x_i}, y_i)\}_{i=1}^n$, $(\mathbf{x_i}, y_i) \in \mathbb{R}^p \times \mathcal{Y}$ from an unknown distribution $\mathbb{P}(\mathbf{X}, \mathbf{y})$. We note our loss f and define the theoretical loss $\mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}\left(f\left(\langle \mathbf{x}, \boldsymbol{\beta}\rangle; y\right)\right)$. We consider a theoretical minimizer $\boldsymbol{\beta}^*$:

$$\boldsymbol{\beta}^* \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\mathbb{E}\left(f\left(\langle \mathbf{x}, \boldsymbol{\beta}\rangle; y\right)\right)\right\}. \qquad (1)$$

In the rest of this section, we denote by $k = \|\boldsymbol{\beta}^*\|_0$ the number of non-zeros of the theoretical minimizer and $R = \|\boldsymbol{\beta}^*\|_1$ its L1 norm. We assume $R \geq 1$. We study the L1-regularized L1-constrained problem defined as:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p:\ \|\boldsymbol{\beta}\|_1 \leq 2R} \ \frac{1}{n}\sum_{i=1}^n f\left(\langle \mathbf{x_i}, \boldsymbol{\beta}\rangle; y_i\right) + \lambda\|\boldsymbol{\beta}\|_1. \qquad (2)$$

We consider an empirical minimizer $\hat{\boldsymbol{\beta}}$, solution of Problem (2). The constraint $2R$ in Problem (2) is somewhat arbitrary: it enforces the empirical minimizer to be close enough to the theoretical minimizer, that is, $\boldsymbol{\beta}^*$: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 3R$. The L1 regularization in Lagrangian form is known to induce sparsity in the coefficients of $\hat{\boldsymbol{\beta}}$. Note that Problem (2) is fully tractable.

For a given λ, we fix a solution $\hat{\boldsymbol{\beta}}(\lambda, R)$ of Problem (2) – R is fixed throughout the paper. Our main result is an error bound – achieved for a certain λ – for the L2 norm of the difference between the empirical and theoretical minimizers $\|\hat{\boldsymbol{\beta}}(\lambda, R) - \boldsymbol{\beta}^*\|_2$. When no confusion can be made, we drop the dependence upon the parameters λ, R. Our bound is reached under standard assumptions in the literature. In particular, it is similar to those proposed by Peng et al. (2016), Ravikumar et al. (2010), Belloni et al. (2011). The rest of this section presents our framework of study.

2.1 Lipschitz loss function

Our first assumption concerns the Lipschitz-continuity of the loss f.

Assumption 1. *The loss $f(., y)$ is non-negative, convex and Lipschitz continuous with constant L, that is, $|f(t_1, y) - f(t_2, y)| \leq L|t_1 - t_2|,\ \forall t_1, t_2$. In addition, there exists $\partial f(., y)$ such that $f(t_2, y) - f(t_1, y) \geq \partial f(t_1, y)(t_2 - t_1),\ \forall t_1, t_2$.*

$\partial f(., y)$ is said to be a sub-gradient of the loss: if $f(., y)$ is differentiable, we simply consider its gradient. It trivially holds $\|\partial f(., y)\|_\infty \leq L,\ \forall y$. We list three main examples that fall into this framework.

Example 1: Support Vectors Machines We assume $\mathcal{Y} = \{-1, 1\}$ and consider the L1-regularized L1-constrained Support Vector Machines (SVM) problem. It learns a classification rule of the data of the form $\text{sign}(\langle \mathbf{x}, \boldsymbol{\beta}\rangle)$ by solving the problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p:\ \|\boldsymbol{\beta}\|_1 \leq 2R} \ \frac{1}{n}\sum_{i=1}^n \left(1 - y_i\langle \mathbf{x}_i, \boldsymbol{\beta}\rangle\right)_+ + \lambda\|\boldsymbol{\beta}\|_1. \qquad (3)$$

The hinge loss $f\left(\langle \mathbf{x}, \boldsymbol{\beta}\rangle; y\right) = \max(0, 1 - y\langle \mathbf{x}, \boldsymbol{\beta}\rangle)$ admits as a subgradient $\partial f(., y) = \mathbf{1}(1 - y. \geq 0)y.$ and satisfies Assumption 1 for $L = 1$.

Example 2: Logistic Regression Here, $\mathcal{Y} = \{-1, 1\}$ and we consider the additional assumption $\log(\mathbb{P}(y_i = 1|\mathbf{X} = \mathbf{x_i})) - \log(\mathbb{P}(y_i = -1|\mathbf{X} = \mathbf{x_i})) = \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle, \ \forall i$. The L1-regularized L1-constrained Logistic Regression estimator is a solution of the problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p: \ \|\boldsymbol{\beta}\|_1 \leq 2R} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)\right) + \lambda \|\boldsymbol{\beta}\|_1. \tag{4}$$

The logistic loss $f(\langle \mathbf{x}, \boldsymbol{\beta} \rangle; y) = \log(1 + \exp(-y\langle \mathbf{x}, \boldsymbol{\beta} \rangle))$ has a derivative with respect to its first variable $|\partial_t f(t, y)| = |1/(1 + e^{yt})| \leq 1$, hence it satisfies Assumption 1 for $L = 1$.

Example 3: Quantile Regression We now consider a class of parametric quantile estimation problems. Following Buchinsky (1998), we assume that for $\theta \in (0, 1)$ the conditional quantile of y given $\mathbf{X}$ is given by $Q_\theta(y|\mathbf{X} = \mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta}_\theta \rangle$, where the model is of the form $y = \langle \mathbf{x}, \boldsymbol{\beta}_\theta \rangle + u_\theta$, and u_θ is unkown. The L1-regularized L1-constrained θ-Quantile Regression estimator is defined as a solution of:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p: \ \|\boldsymbol{\beta}\|_1 \leq 2R} \frac{1}{n} \sum_{i=1}^{n} \rho_\theta\left(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle\right) + \lambda \|\boldsymbol{\beta}\|_1, \tag{5}$$

where $\rho_\theta(t) = (\theta - \mathbf{1}(t \leq 0))t$ is the quantile regression loss. ρ_θ satisfies Assumption 1 for $L = \max(1 - \theta, \theta)$. Note that the hinge loss is a simple translation of the quantile regression loss for $\theta = 0$.

2.2 Differentiability of the theoretical loss

The following assumption ensures the unicity of $\boldsymbol{\beta}^*$ and the twice differentiability of the theoretical loss $\mathcal{L}$. Equation (6) is equivalent to saying that the gradient of the theoretical loss is equal to the theoretical sub-gradient of the loss – defined in Assumption 1.

Assumption 2. *The theoretical minimizer is unique. In addition, the theoretical loss is twice-differentiable: we denote its gradient $\nabla \mathcal{L}(\boldsymbol{\beta})$ and its Hessian matrix $\nabla^2 \mathcal{L}(\boldsymbol{\beta})$. We also assume:*

$$\nabla \mathcal{L}(.) = \mathbb{E}\left(\partial f\left(\langle \mathbf{x}, . \rangle; y\right) \mathbf{x}\right). \tag{6}$$

Support Vectors Machines: Koo et al. (2008) studied specific conditions under which Assumption 2 holds for SVM. Let f and g denote the respective conditional densities of $\mathbf{X}$ given $y = 1$ and $y = -1$. The authors proved that if f and g are continuous with common support $\mathcal{S} \subset \mathbb{R}^p$ and have finite second moments, then the gradient $\nabla \mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}\left(\mathbf{1}\left(1 - y\langle \mathbf{x}, \boldsymbol{\beta} \rangle \geq 0\right) y\mathbf{x}\right)$ and the Hessian matrix $\nabla^2 \mathcal{L}(\boldsymbol{\beta}) = \mathbb{E}\left(\delta\left(1 - y\langle \mathbf{x}, \boldsymbol{\beta} \rangle\right) y\mathbf{x}\right)$ ($\delta(.)$ is the Dirac function) are defined and continuous.

Logistic and Quantile Regression: The regularity of $\nabla \mathcal{L}$ and $\nabla^2 \mathcal{L}$ are trivial for the logistic regression loss. Equation (6) holds as the sub-gradient is simply the gradient of the loss. For the quantile regression loss, a study similar to the hinge loss case – using Assumption D.1 from Belloni et al. (2011) – can be applied to obtain Assumption 2.

2.3 Sub-Gaussian columns

We denote $\mathbb{X}$ the design matrix, with rows $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The following assumption guarantees that some random variables of the columns $(\mathbf{X}_1, \ldots, \mathbf{X}_p)$ of $\mathbb{X}$ have their tails bounded by a sub-Gaussian random variable with variance proportional to n. We first recall the definition of a sub-Gaussian random variable (Rigollet, 2015):

Definition 1. *A random variable Z is said to be sub-Gaussian with variance $\sigma^2 > 0$ if $\mathbb{E}(Z) = 0$ and $\mathbb{P}(|Z| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \ \forall t > 0$.*

A sub-Gaussian variable will be noted $Z \sim \text{subG}(\sigma^2)$. We would like here to notice another important aspect of our contribution. Our next Theorem 3 derives a cone condition, a necesary step to prove our main bounds. Our approach draws inspiration from the regression case with Gaussian noise. However, it relies on a new study of sub-Gaussian random variables – such analysis is not needed in the regression case. Our results are derived under the following Assumption 3:

Assumption 3. *There exists $M > 0$ such that with the notations of Assumption 1:*

$$\sum_{i=1}^{n} \partial f\left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle, y_i\right) x_{ij} \sim subG(nL^2M^2), \ \forall j. \tag{7}$$

$\boldsymbol{\beta}^*$ minimizes the theoretical loss. Thus, if Assumption 2 holds, $\mathbb{E}\left[\partial f\left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle, y_i\right) x_{ij}\ \right] = 0, \forall i, j$. The next lemma gives more insight about Assumption 3. The proof is presented in Appendix A.2.

Lemma 1. *If the rows of the design matrix are independent and if all the entries $\partial f\left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle; y_i\right) x_{ij}, \ \forall i, j$ are sub-Gaussian with variance L^2M^2, then $\sum_{i=1}^{n} \partial f\left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle; y_i\right) x_{ij} \sim subG(8nL^2M^2), \forall j$.*

In particular, if $|x_{i,j}| \leq M, \ \forall i, j$, then – under Assumption 1 – Hoeffding's lemma guarantees that $\partial f\left(\langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle; y_i\right) x_{i,j} \sim \text{subG}(L^2M^2), \ \forall i, j$. Thus Assumption 3 is satisfied. Assumption 3 is also satisfied if the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently drawn from a multivariate centered Gaussian distribution. Hence, Assumption 3 is rather mild. It is considerably much weaker than Assumption (A1) by Peng et al. (2016) which imposes a finite bound on the L2 norm of each column of $\mathbb{X}$.

2.4 Restricted eigenvalue conditions

The next assumption draws inspiration from the restricted eigenvalue conditions defined for regression problems (Bickel et al., 2009; Bellec et al., 2016). In particular, for an integer k, Assumption 4.1 ensures that some random variable is upper-bounded on the set of k sparse vectors. Similarly, Assumption 4.2 ensures that the quadratic form associated to the Hessian matrix $\nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)$ is lower-bounded on a cone of $\mathbb{R}^p$.

Assumption 4. *Let $k \in \{1, \ldots, p\}$. Assumption 4.1(k) is satisfied if there exists a nonnegative constant $\mu(k)$ such that almost surely:*

$$\mu(k) \geq \sup_{\mathbf{z} \in \mathbb{R}^p:\ \|\mathbf{z}\|_0 \leq k} \frac{\sqrt{k}\|\mathbb{X}\mathbf{z}\|_1}{\sqrt{n}\|\mathbf{z}\|_1} > 0.$$

Let γ_1, γ_2 be two non-negative constants. Assumption 4.2(k, γ) holds if there exists a nonnegative constant $\kappa(k, \gamma_1, \gamma_2)$ which almost surely satisfies:

$$0 < \kappa(k, \gamma_1, \gamma_2) \leq \inf_{|S| \leq k} \inf_{\mathbf{z} \in \Lambda(S, \gamma_1, \gamma_2)} \frac{\|\mathbf{z}^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)\mathbf{z}\|_2}{\|\mathbf{z}\|_2},$$

where $\gamma = (\gamma_1, \gamma_2)$ and for every subset $S \subset \{1, \ldots, p\}$, the cone $\Lambda(S, \gamma_1, \gamma_2) \subset \mathbb{R}^p$ is defined as:

$$\Lambda(S, \gamma_1, \gamma_2) = \left\{ \mathbf{z} \in \mathbb{R}^p :\ \|\mathbf{z}_{S^c}\|_1 \leq \gamma_1 \|\mathbf{z}_S\|_1 + \gamma_2 \|\mathbf{z}_S\|_2 \right\}.$$

We refer to Assumption 4(k, γ) when both Assumptions 4.1(k) and 4.2(k, γ) are assumed to hold.

In the SVM framework, Peng et al. (2016) define Assumption (A4): it is similar to our Assumption 4.2(k, γ) but it considers a different cone of $\mathbb{R}^p$. In addition, their Assumption (A3) defines $\mu(k)$ as an upper bound of the quadratic form associated to $n^{-1/2}\mathbb{X}^T\mathbb{X}$ – restricted to the set of k sparse vectors. That is, under their definition, $\|\mathbb{X}\mathbf{z}\|_2/\sqrt{n} \leq \mu(k)\|\mathbf{z}\|_2,\ \forall \mathbf{z} : \|\mathbf{z}\|_0 \leq k$. Our Assumption 4.1($k$) is stronger: when satisfied, we can recover Assumption (A3) since that

$$\forall \mathbf{z} \in \mathbb{R}^p :\ \|\mathbf{z}\|_0 \leq k,$$
$$\|\mathbb{X}\mathbf{z}\|_2/\sqrt{n} \leq \|\mathbb{X}\mathbf{z}\|_1/\sqrt{n} \leq \mu(k)\|\mathbf{z}\|_1/\sqrt{k} \leq \mu(k)\|\mathbf{z}\|_2$$

where we have used Cauchy-Schwartz inequality on the k sparse vector $\mathbf{z}$. However, Assumption 4.1(k) uses an L1 norm, more naturally associated to the class of L1-regularized estimators studied in this work.

Similarly, in the Logistic Regression case Ravikumar et al. (2010) consider a dependency and incoherence conditions for the population Fisher information matrix (Assumptions A1 and A2). Finally, Assumption D.4 for Quantile Regression (Belloni et al., 2011) is a uniform Restricted Eigenvalue condition.

2.5 Growth condition

Since $\boldsymbol{\beta}^*$ minimizes the theoretical loss, it holds $\nabla L(\boldsymbol{\beta}^*) = 0$. In particular, under Assumption 4.2(k^*, γ), the theoretical loss evaluated on the family of cones $\Lambda(S, \gamma_1, \gamma_2)$ – where $|S| \leq k^*$ – is lower-bounded by a quadratic form around $\boldsymbol{\beta}^*$. By continuity, we define the maximal radius on which the following lower-bound holds:

$$r(k^*) = \max \left\{ r :\ \begin{array}{l} \mathcal{L}(\boldsymbol{\beta}^* + \mathbf{z}) \geq \mathcal{L}(\boldsymbol{\beta}^*) + \frac{\kappa(k^*)}{4}\|\mathbf{z}\|_2^2 \\ \forall S \subset (p) : |S| \leq k^*, \\ \forall \mathbf{z} \in \Lambda(S) : \|\mathbf{z}\|_1 \leq r \end{array} \right\}$$

where the notations $r(k^*)$, $\kappa(k^*)$ and $\Lambda(S)$ are shorthands for $r(k^*, \gamma_1, \gamma_2)$, $\kappa(k^*, \gamma_1, \gamma_2)$ and $\Lambda(S, \gamma_1, \gamma_2)$. This definition is similar to the one proposed by Belloni et al. (2011) in the proof of Lemma (3.7) . We now define a growth condition which gives a relation between the number of samples n, the dimension space p, the degree of sparsity k^*, our constants introduced in Assumption 4, and a parameter δ.

Assumption 5. *Let $\delta \in (0, 1)$ and $k \in \{1, \ldots, p\}$. We say Assumption 5.1(k) is satisfied if $p \leq k\sqrt{k}$. In addition, Assumption 5.2(k, γ, δ) is said to hold if the parameters n, p, k satisfy:*

$$\frac{\kappa(k)}{16\alpha L} r(k) \geq 3M\sqrt{\frac{k \log(2pe/k) \log(2/\delta)}{n}} + 7\mu(k)\sqrt{\frac{\log(3) + \log(p/k)/k + \log(2/\delta)}{n}}.$$

We refer to Assumption 5(k, γ, δ) when both Assumptions 5.1(k) and 5.2(k, γ, δ) hold.

Assumption 5 is similar to Equation (17) from Ravikumar et al. (2010) for Logistic Regression. Belloni et al. (2011) also require a growth condition for Theorem 2 to hold for quantile regression. Consequently, we showed that Assumptions 1-5 are common assumptions or similar to existing ones in the literature. The next section uses our framework to derive upper bounds for L2 coefficients estimation scaling with the parameters n, p, k^*.

3 Main results

This section establishes the following theorem:

Theorem 1. *Let $\delta \in (0, 1)$, $\alpha \geq 2$ and assume Assumptions 1-3, 4(k^*, γ) and 5(k^*, γ, δ) hold – where $\gamma = (\gamma_1, \gamma_2)$ and $\gamma_1 := \frac{\alpha}{\alpha - 1}$, $\gamma_2 := \frac{\sqrt{k^*}}{\alpha - 1}$.*

Then, the empirical estimator $\hat{\boldsymbol{\beta}}$, defined as a solution of Problem (2) for the regularization parameter $\lambda = 12\alpha LM\sqrt{\frac{\log(2pe/k^*)}{n}} \log(2/\delta)$, satisfies with probability *at least* $1-\delta$*:*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 \lesssim \frac{\alpha L M}{\kappa(k^*)}\sqrt{\frac{k^*\log\left(p/k^*\right)\log\left(2/\delta\right)}{n}} + \frac{\alpha L \mu(k^*)}{\kappa(k^*)}\sqrt{\frac{\log(3)+\log\left(p/k^*\right)/k^*+\log\left(2/\delta\right)/k^*}{n}}. \quad (8)$$

This upper bound scales as $\left((k^*/n)\log(p/k^*)\right)^{1/2}$. It strictly improves over existing results. Note that our estimator is not adaptative to unknown sparsity: the regularization parameter λ depends upon k^*. The proof of Theorem 1 is presented in Appendix E. It relies on two essential steps: a cone condition and a restricted strong convexity condition: these results are respectively derived in Theorems 2 and 4. The two terms of the sum in Equation (8) are related to the two parameters λ and τ respectively introduced in these theorems.

In addition, Theorem 1 holds for any $\delta \le 1$. Thus, we obtain by integration the following bound in expectation. The proof is presented in Appendix F.

Corollary 1. *If the assumptions presented in Theorem are satisfied for a small enough* δ*, then:*

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 \lesssim \frac{\alpha L}{\kappa(k^*)}\left(M\sqrt{\frac{k^*\log\left(p/k^*\right)}{n}}+\frac{\mu(k^*)}{\sqrt{n}}\right).$$

The rest of this section follows through the steps required to prove Theorem 1 and Corollary 1.

3.1 Cone condition

Similarly to the regression case (Bickel et al., 2009; Bellec et al., 2016), we first derive a cone condition which applies to the difference between the empirical and theoretical minimizers. That is, by selecting a suitable regularization parameter, we show that this difference belongs to the family of cones $\Lambda(S,\gamma_1,\gamma_2)$ of $\mathbb{R}^p$ defined in Assumption 4.

Theorem 2. *Let* $\delta \in (0,1)$ *and assume that Assumptions 1 and 3 are satisfied. Let* $\alpha \ge 2$*.*

Let $\hat{\boldsymbol{\beta}}$ *be a solution of Problem* (2) *with parameter* $\lambda = 12\alpha L M\sqrt{\frac{\log(2pe/k^*)}{n}}\log(2/\delta)$*. Then it holds with probability at least* $1-\frac{\delta}{2}$*:*

$$\mathbf{h} := \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^* \in \Lambda\left(S_0,\ \gamma_1 := \frac{\alpha}{\alpha-1},\ \gamma_2 := \frac{\sqrt{k^*}}{\alpha-1}\right),$$

where S_0 *is the subset of indices of the* k^* *highest coefficients of* $\mathbf{h}$*.*

The regularization parameter λ is selected so that it dominates the sub-gradient of the loss f evaluated at the theoretical minimizer $\boldsymbol{\beta}^*$. The proof is presented in Appendix B: it uses a new result to control the maximum of independent sub-Gaussian random variables. As a result, our cone condition is stronger than the ones proposed by Peng et al. (2016) and Ravikumar et al. (2010): their value of λ^2 is of the order of $(k^*/n)\log(p)$ whereas ours scales as $(k^*/n)\log(p/k^*)$.

3.2 A supremum result

The next Theorem 3 is an essential step to obtain our main Theorem 1. It controls the supremum of the difference between an empirical random variable and its expectation. This supremum is taken over a bounded set of sequences of k sparse vectors with supports being a partition of $\{1,\dots,p\}$. The restricted strong convexity condition derived in Theorem 4 is a consequence of Theorem 3.

To motivate this theorem, it helps considering the difference between the usual regression framework and our framework for classification problems. The linear regression case assumes the generative model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$. Therefore, with the notations of Theorem 3, $\Delta(\boldsymbol{\beta}^*,\boldsymbol{z}) = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{z}\|_2^2 - \frac{2}{n}\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{z}$. By combining a cone condition (similar to Theorem 1) with an upper-bound of the term $\epsilon^T\boldsymbol{X}\boldsymbol{z}$, we can obtain a restricted strong convexity similar to Theorem 4. However, in the classification case, $\boldsymbol{\beta}^*$ is defined as the minimizer of the theoretical risk. Two majors differences appear: (i) we cannot simplify $\Delta(\boldsymbol{\beta}^*,\boldsymbol{z})$ with basic algebra, (ii) we need to introduce the expectation $\mathbb{E}(\Delta(\boldsymbol{\beta}^*,\boldsymbol{z}))$ and to control the quantity $|\mathbb{E}(\Delta(\boldsymbol{\beta}^*,\boldsymbol{z}))-\Delta(\boldsymbol{\beta}^*,\boldsymbol{z})|$. Theorem 3 derives the cost to pay for this control.

Theorem 3. *We define* $\forall \mathbf{w},\mathbf{z}\in\mathbb{R}^p$*:*

$$\Delta(\mathbf{w},\mathbf{z}) = \frac{1}{n}\sum_{i=1}^n f\left(\langle \mathbf{x}_i, \mathbf{w}+\mathbf{z}\rangle; y_i\right) - \frac{1}{n}\sum_{i=1}^n f\left(\langle \mathbf{x}_i, \mathbf{w}\rangle; y_i\right).$$

Let $k \in \{1,\dots,p\}$ *and* $S_1,\dots S_q$ *be a partition of* $\{1,\dots,p\}$ *with* $q=\lceil p/k\rceil$ *and* $|S_j|\le k, \forall j$*.*

Let $\tau(k) = 14L\sqrt{\frac{\log(3)}{n}+\frac{\log(4p/k)}{nk}+\frac{\log(2/\delta)}{nk}}$ *and assume that Assumptions 1, 4.1(k) and 5.1(k) hold. Then, for any* $\delta\in(0,1)$*, it holds with probability at least* $1-\frac{\delta}{2}$*:*

$$\sup_{\substack{\mathbf{z}_{S_1},\dots,\mathbf{z}_{S_q}\in\mathbb{R}^p:\\ \mathrm{Supp}(\mathbf{z}_{S_j})\subset S_j\ \ \forall j\\ \|\mathbf{z}_{S_j}\|_1\le 3R\ \ \ \forall j}} \left\{ \sup_{\ell=1,\dots,q} \left\{\Omega\left(\mathbf{w}_{\ell-1},\ \mathbf{z}_{S_\ell}\right)\right\}\right\} \le 0, \text{with}$$

$$\Omega\left(\mathbf{w}_{\ell-1},\mathbf{z}_{S_\ell}\right) := \left|\Delta\left(\mathbf{w}_{\ell-1},\mathbf{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\mathbf{w}_{\ell-1},\mathbf{z}_{S_\ell}\right)\right)\right| - \tau(k)\|\mathbf{z}_{S_\ell}\|_1.$$

$\mathrm{Supp}(.)$ *refers to the support of a vector and we define* $\mathbf{w}_\ell = \boldsymbol{\beta}^* + \sum_{j=1}^{\ell}\mathbf{z}_{S_j}, \forall \ell$.

The proof is presented in Appendix C. It uses Hoeffding's inequality to obtain an upper bound of the inner supremum for a sequence of k sparse vectors. The result is extended to the outer supremum with an ϵ-net argument.

3.3 Restricted strong convexity condition

Theorem 3 applies to a sequence of k sparse vectors with disjoint supports. In particular we can fix $k = k^*$ and consider $\mathbf{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. In addition, we can exploit the minimality of $\boldsymbol{\beta}^*$ and the cone condition proved in Theorem 2. By pairing these points, we derive the next Theorem 4. It says that the loss f satisfies a restricted strong convexity (Negahban et al., 2009) with curvature $\kappa(k^*)/4$. We propose two results, respectively achieved with L1 and L2 tolerance function.

Theorem 4. *Let $\boldsymbol{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and $\delta \in (0, 1)$. Under the notations of Theorem 3, if Assumptions 1-3, 4(k^*, γ) and 5(k^*, γ, δ) are satisfied, then it holds with probability at least $1 - \delta$:*

$$\begin{aligned}\Delta\left(\boldsymbol{\beta}^*, \mathbf{h}\right) &\geq \frac{1}{4}\kappa(k^*)\left\{\|\mathbf{h}\|_2^2 \wedge r(k^*)\|\mathbf{h}\|_2\right\} - \tau(k^*)\|\mathbf{h}\|_1 \\ &\geq \frac{1}{4}\kappa(k^*)\left\{\|\mathbf{h}\|_2^2 \wedge r(k^*)\|\mathbf{h}\|_2\right\} - \frac{2\alpha}{\alpha - 1}\tau(k^*)\sqrt{k^*}\|\mathbf{h}\|_2\end{aligned} \tag{9}$$

The proof is presented in Appendix D. We convert the L1 tolerance function into an L2 norm by using the cone condition derived inTheorem 2. Let us note that the parameter $k^*\tau(k^*)^2$ used for the L2 tolerance function scales as $n^{-1}(k^* + \log(p/k^*))$, whereas Peng et al. (2016), Ravikumar et al. (2010) and Negahban et al. (2009) all propose a parameter scaling as $n^{-1}k^* \log(p)$. Hence, our restricted strong convexity condition is stronger.

3.4 Deriving Theorem 1 and Corollary 1

Our main bounds – presented in Theorem 1 and Corollary 1 – follow from the two preceding Theorems 2 and 4. The proofs are respectively presented in Appendix E and F. Our family of L1-regularized L1-constrained estimators reach a bound that strictly improve over existing results. Our rate is the best known for the classification problems considered here, and it holds both with high probability and in expectation.

4 Algorithm and upper bounds for Slope estimator

This section introduces the Slope estimator – originally presented for the linear regression case (Bogdan et al., 2013, 2015) – to our class of problems. We propose a tractable algorithm to compute it and study its statistical performance.

4.1 Introducing Slope for classification

We consider a sequence $\lambda \in \mathbb{R}^p$ such that $\lambda_1 \geq \ldots \geq \lambda_p > 0$, and we note $\mathcal{S}_p$ the set of permutations of $\{1, \ldots, p\}$. The Slope regularization is defined as:

$$|\boldsymbol{\beta}|_S = \max_{\phi \in \mathcal{S}_p} \sum_{j=1}^{p} |\lambda_j||\beta_{\phi(j)}| = \sum_{j=1}^{p} \lambda_j |\beta_{(j)}|, \tag{10}$$

where $|\beta_{(1)}| \geq \ldots \geq |\beta_{(p)}|$ is a non-increasing rearrangement of $\boldsymbol{\beta}$. Consequently for $\eta > 0$, we define the Slope estimator $\hat{\boldsymbol{\beta}}$ as the solution of the convex minimization problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f\left(\langle \mathbf{x_i}, \boldsymbol{\beta} \rangle; y_i\right) + \eta |\boldsymbol{\beta}|_S. \tag{11}$$

The approach presented herein uses a proximal gradient algorithm – with Nesterov smoothing (Nesterov, 2005) in the case of the hinge loss and quantile regression loss – to solve Problem (11), extending the original definition of Slope (Bogdan et al., 2013) to a larger class of loss functions. Recently, Dedieu and Mazumder (2019) combined similar first order methods for non-smooth convex optimization with column with constraint generation algorithms to solve Problems (2) and (11) when f is the hinge-loss and the number of samples and features are of the order of tenth of thousands.

4.2 Smoothing the hinge loss

The method described in Section 4.3 to solve Problem (11) requires $f(., y)$ to be differentiable with Lipschitz-continuous gradient. Among the loss functions considered in Section 2, only the logistic regression loss satisfies this condition.

To handle the non-smooth hinge loss, we use the smoothing scheme pioneered by Nesterov (2005). We construct a convex function g^τ with continuous Lipschitz gradient, which approximates the hinge loss for $\tau \approx 0$. Let us first note that $\max(0, x) = \frac{1}{2}(x + |x|) = \max_{|w| \leq 1} \frac{1}{2}(x + wx)$ as this maximum is achieved for $\text{sign}(x)$. Consequently the hinge loss can be expressed as a maximum over the L_∞ unit ball:

$$\frac{1}{n} \sum_{i=1}^{n} \max(z_i, 0) = \max_{\|w\|_\infty \leq 1} \frac{1}{2n} \sum_{i=1}^{n} \left[z_i + w_i z_i\right],$$

where $z_i = 1 - y_i \mathbf{x}_i^T \boldsymbol{\beta}$, $\forall i$. We apply the technique suggested by Nesterov (2005) and define for $\tau > 0$ the smoothed hinge loss:

$$g^\tau(\boldsymbol{\beta}) = \max_{\|w\|_\infty \leq 1} \frac{1}{2n} \sum_{i=1}^{n} \left[z_i + w_i z_i\right] - \frac{\tau}{2n}\|w\|_2^2. \tag{12}$$

Let $\mathbf{w}^\tau(\boldsymbol{\beta}) \in \mathbb{R}^n: \; w_i^\tau(\boldsymbol{\beta}) = \min\left(1, \frac{1}{2\tau}|z_i|\right)\text{sign}(z_i), \; \forall i$ be the optimal solution of the right-hand side of Equation (12). The gradient of g^τ is expressed as:

$$\nabla g^\tau(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^n (1 + w_i^\tau(\boldsymbol{\beta}))y_i\mathbf{x}_i \in \mathbb{R}^p, \qquad (13)$$

and its associated Lipschitz constant is derived from the next theorem.

Theorem 5. *Let $\mu_{\max}(n^{-1}\mathbb{X}^T\mathbb{X})$ be the highest eigenvalue of $n^{-1}\mathbb{X}^T\mathbb{X}$. Then ∇g^τ is Lipschitz continuous with constant $C^\tau = \mu_{\max}(n^{-1}\mathbb{X}^T\mathbb{X})/4\tau$.*

The proof is presented in Appendix G. It follows Nesterov (2005) and relies on first order necessary conditions for optimality. We mention how to adapt Theorem 5 to the quantile regression loss.

4.3 Thresholding operator for Slope

We note $g(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n f\left(\langle \mathbf{x_i}, \boldsymbol{\beta}\rangle; y_i\right)$. Problem (11) can be equivalently formulated as:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} g(\boldsymbol{\beta}) + \eta|\boldsymbol{\beta}|_S, \qquad (14)$$

We now require g to be a differentiable loss with C-Lipschitz continuous gradient. When f is the hinge or quantile regression loss we replace g with g^τ – defined in Section 4.2. For $D \geq C$, we upper-bound g around any $\boldsymbol{\alpha} \in \mathbb{R}^p$ with the quadratic form $Q_D(\boldsymbol{\alpha}, .)$ defined as the right-hand side of the equation:

$$g(\boldsymbol{\beta}) \leq g(\boldsymbol{\alpha}) + \nabla g(\boldsymbol{\alpha})^T(\boldsymbol{\beta} - \boldsymbol{\alpha}) + \frac{D}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2. \quad (15)$$

We approximate the solution of Problem (11) by considering the loss Q_D and solving the problem:

$$\begin{aligned} &\operatorname*{argmin}_{\boldsymbol{\beta}} Q_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \eta|\boldsymbol{\beta}|_S \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{\alpha} - \frac{1}{D}\nabla g(\boldsymbol{\alpha})\right)\right\|_2^2 + \frac{\eta}{D}|\boldsymbol{\beta}|_S \quad (16) \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 + \sum_{j=1}^p \tilde{\eta}_j|\beta_{(j)}|, \end{aligned}$$

where $\boldsymbol{\gamma} = \boldsymbol{\alpha} - \frac{1}{D}\nabla g(\boldsymbol{\alpha})$ and $\tilde{\eta}_j = \frac{\eta}{D}\lambda_j, \; \forall j$. To solve Problem (16), we need to derive the proximal operator for the sorted L1 norm. The next Lemma 2 does so by noting that the signs of the quantities β_j and γ_j are all identical.

Lemma 2. *Let us assume that $\tilde{\gamma}_1 \geq \ldots \geq \tilde{\gamma}_p \geq 0$. Since $\tilde{\eta}_1 \geq \ldots \geq \tilde{\eta}_p \geq 0$, the solution of Problem (16) can be derived from the solution of the problem:*

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad & \tfrac{1}{2}\|\boldsymbol{\beta} - \tilde{\boldsymbol{\gamma}}\|_2^2 + \sum_{j=1}^p \tilde{\eta}_j\beta_j \\ \text{s.t.} \quad & \beta_1 \geq \ldots \geq \beta_p \geq 0. \end{aligned} \qquad (17)$$

Bogdan et al. (2015) proposed an efficient proximal algorithm to solve Problem (17) called FastProxSL1: it is guaranteed to terminate in at most p iterations. We denote by $\mathcal{T}_{\{\tilde{\eta}_j\}}(\boldsymbol{\gamma})$ a solution for Problem (16).

4.4 First order algorithm

The following algorithm applies the accelerated gradient descent method (Beck and Teboulle, 2009) on the smoothed version of the Slope Problem (14) by using the thresholding operator $\mathcal{T}$. The iterations continue till the algorithm converges or a maximum number of iterations $T_{\max}$ is reached.

Input: $\mathbf{X}$, $\mathbf{y}$, a sequence of Slope coefficients $\{\lambda_j\}$, a regularization parameter η, a stopping criterion ϵ, a maximum number of iterations $T_{\max}$.
Output: An approximate solution $\boldsymbol{\beta}$ for the smoothed Slope Problem (14).

1. Initialize $T = 1$, $q_1 = 1$, $\boldsymbol{\beta}_1 = \boldsymbol{\delta}_0 = 0$.
2. : While $\|\boldsymbol{\beta}_T - \boldsymbol{\beta}_{T-1}\|_2 > \epsilon$ and $T < T_{\max}$ do:
   (a) Compute $\boldsymbol{\delta}_T = \mathcal{T}_{\{\eta\lambda_j/C\}}\left(\boldsymbol{\beta}_T - \frac{1}{C}\nabla g(\boldsymbol{\beta}_T)\right)$.
   (b) Define $q_{T+1} = \frac{1+\sqrt{1+4q_T^2}}{2}$ and compute $\boldsymbol{\beta}_{T+1} = \boldsymbol{\delta}_T + \frac{q_T - 1}{q_{T+1}}(\boldsymbol{\delta}_T - \boldsymbol{\delta}_{T-1})$.

4.5 Error bounds for Slope

We extend our previous case and study under our framework the theoretical properties of a Slope estimator. In particular, we consider the L1-constrained Slope-regularized estimator:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p:\ \|\boldsymbol{\beta}\|_1 \leq 2R} \frac{1}{n}\sum_{i=1}^n f\left(\langle \mathbf{x_i}, \boldsymbol{\beta}\rangle; y_i\right) + \eta|\boldsymbol{\beta}|_S. \qquad (18)$$

The study of the Slope estimator shares a lot of similarities with our work for L1-regularized estimators. We first derive the following cone condition:

Theorem 6. *Let $\delta \in (0,1)$ and $\alpha \geq 2$. We fix the Slope coefficients $\lambda_j = \sqrt{\log(2pe/j)}, \forall j$, and assume Assumptions 1 and 3 hold. Then the Slope estimator defined as a solution of Problem (18) for the regularization parameter $\eta = 14\alpha LM\sqrt{n^{-1}\log(2/\delta)}$ satisfies with probability at least $1 - \frac{\delta}{2}$:*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \Gamma\left(k^*, \omega^* = \frac{\alpha+1}{\alpha-1}\right),$$

where for every $k \in \{1, \ldots, p\}$ and $\omega > 0$, the cone $\Gamma(k, \omega)$ is defined as:

$$\Gamma(k, \omega) = \left\{\mathbf{z} \in \mathbb{R}^p : \sum_{j=k+1}^p \lambda_j|z_{(j)}| \leq \omega\sum_{j=1}^k \lambda_j|z_{(j)}|\right\}$$

with $|z_{(1)}| \geq \ldots \geq |z_{(p)}|, \; \forall \boldsymbol{z}$.

The proof is presented in Appendix H. We consequently adapt Assumption $4.2(k,\delta)$ to the new family of cones $\Gamma(k,\omega)$ introduced in Theorem 6.

Assumption 6. *Let $k \in \{1,\dots,p\}$ and $\omega > 0$. Assumption $6.2(k,\omega)$ is said to hold if there exists a non-negative constant $\tilde{\kappa}(k,\omega)$ such that:*

$$0 < \tilde{\kappa}(k,\omega) \leq \inf_{\mathbf{z}\in\Gamma(k,\omega)} \frac{\|\mathbf{z}^T\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\mathbf{z}\|_2}{\|\mathbf{z}\|_2}.$$

Similarly, we define a new growth condition – referred as Assumption $8(k,\omega,\delta)$ – which adapts Assumption 5 to Slope by replacing $\kappa(k,\gamma)$ with $\tilde{\kappa}(k,\omega)$ defined above. As a consequence, the following result holds.

Corollary 2. *Assume Assumptions 1-3, Asumptions $6(k^*,\omega^*)$ and $8(k^*,\omega^*,\delta)$ hold for a small enough δ – ω^* is defined in Theorem 6 – and that $\mu(k^*) \leq \alpha M$.*

Then the bounds presented in Theorem 1 and Corollary 1 are achieved by a Slope estimator, defined as a solution of Problem (11) for the coefficients $\lambda_j = \sqrt{\log(2pe/j)}, \forall j$ and the regularization parameter $\eta = 14\alpha L M n^{-1}\sqrt{\log(2/\delta)}$ – where $\alpha \geq 2$.

The proof is presented in Appendix I. This Slope estimator adapts to unknown sparsity while achieving the same bound than the L1-regularized estimator studied in Theorem 1 and Corollary 1.

4.6 Simulations

We finally compute a family of Slope estimators and demonstrate their empirical performance – for L2 coefficients estimations and misclassification accuracy – when compared to L1 and L2-regularized estimators.

Data Generation: We consider n independent realizations of a p dimensional multivariate normal centered distribution, with only k^* dimensions being relevant for classification. Half of the samples are from the $+1$ class and have mean $\mu_+ = (\mathbf{1}_{k^*},\ \mathbf{0}_{p-k^*})$. The other half are from the -1 class and have mean $\mu_- = -\mu_+$. We consider a covariance matrix $\Sigma_{ij} = \rho$ if $i \neq j$ and 1 otherwise. The data of both ±1 classes respectively have the distribution: $\forall i,\ x_i^{\pm} \sim \mathbb{N}(\mu_{\pm},\Sigma)$.

Competitors: Table 1 compares the performance of 3 approaches – each associated to a different regularization – for both the SVM and the Logistic Regression problems. Method **(a)** computes a family of L1-regularized estimators for a decreasing geometric sequence of regularization parameters $\eta_0 > \ldots > \eta_M$. We start from a high enough η_0 so that the solution of Problem (2) is the 0 estimator and we fix $\eta_M < 10^{-4}\eta_0$. For the hinge loss, we solve the Linear Programming L1-SVM problem with the commercial LP solver Gurobi version 6.5 with Python interface. The L1-regularized Logistic Regression is solved with scikit-learn Python package. In addition, method **(b)** returns a family of L2-regularized estimators with scikit-learn package: we start from $\eta_0 = \max_i \left\{\|\mathbf{x_i}\|_2^2\right\}$ as suggested by Chu et al. (2015). Finally, method **(c)** computes a family of Slope-regularized estimators, using the first order algorithm presented in Section 4.4 for $\tau = 0.2$. The Slope coefficients $\{\lambda_j\}$ are the ones proposed in Theorem 6; the set of parameters $\{\eta_i\}$ is identical to method **(a)**.

Metrics: Following our theoretical results, we want to find the estimator which minimizes the L2 estimation error:

$$\left\| \frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|_2} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_2} \right\|_2,$$

where $\boldsymbol{\beta}^*$ is the theoretical minimizer. $\boldsymbol{\beta}^*$ is computed on a large test set with $10,000$ samples: we solve the SVM / Logistic Regression problem with a very small regularization coefficient on the k^* columns relevant for classification. We also study the misclassification performances on this same test set. For each family returned by the methods **(a)**, **(b)** and **(c)**, we only keep the estimator with lowest misclassification error on an independent validation set of size $10,000$.

Table 1 compares the L2 estimation error (**L2-E**), and the test misclassification error (**Misc**) of these 3 estimators. Results are averaged over 10 simulations.

Table 1: Averaged L2 estimation (**L2-E**) and test misclassification error (**Misc**) for the methods **(a)**, **(b)** and **(c)** over 10 repetitions. We use varying n,p values with $k^* = n/10$ and $\rho = 0.1$. The Slope estimator shows impressive gains for estimating the theoretical minimizer $\boldsymbol{\beta}^*$, while achieving lower misclassification error.

	$n=100, p=1k$		$n=100, p=10k$		$n=1k, p=1k$		$n=1k, n=10k$	
	L2-E	Misc(%)	L2-E	Misc(%)	L2-E	Misc(%)	L2-E	Misc(%)
L1 SVM	0.57	1.67	0.52	1.54	1.12	1.17	1.01	0.15
L2 SVM	0.54	1.73	0.52	1.54	1.11	0.18	0.91	0.11
Slope SVM	**0.34**	**1.24**	**0.37**	**1.15**	**0.94**	**0.13**	**0.83**	**0.10**
L1 LR	0.48	1.40	0.46	1.37	1.04	0.18	1.04	0.16
L2 LR	0.92	3.2	1.25	**0.18**	0.82	**0.12**	0.89	0.16
Slope LR	**0.22**	**1.14**	**0.18**	1.12	**0.81**	**0.12**	**0.82**	**0.13**

References

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

Emmanuel J Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

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

Emmanuel Candes and Mark A Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Florentina Bunea, Alexandre B Tsybakov, Marten H Wegkamp, et al. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016.

Malgorzata Bogdan, Ewout van den Berg, Weijie Su, and Emmanuel Candes. Statistical estimation and testing via the sorted l1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

Bernadetta Tarigan, Sara A Van De Geer, et al. Classifiers of support vector machine type with l1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. Variable selection for support vector machines in high dimensions.

Bo Peng, Lan Wang, and Yichao Wu. An error bound for l1-norm support vector machine coefficients in ultra-high dimension. *Journal of Machine Learning Research*, 17:1–26, 2016.

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

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Alexandre Belloni, Victor Chernozhukov, et al. l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

Moshe Buchinsky. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, pages 88–126, 1998.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9(Jul):1343–1368, 2008.

Philippe Rigollet. 18. s997: High dimensional statistics. *Lecture Notes), Cambridge, MA, USA: MIT OpenCourseWare*, 2015.

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.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Antoine Dedieu and Rahul Mazumder. Solving large-scale l1-regularized svms and cousins: the surprising effectiveness of column and constraint generation. *arXiv preprint arXiv:1901.01585*, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1): 183–202, 2009.

Bo-Yu Chu, Chia-Hua Ho, Cheng-Hao Tsai, Chieh-Yen Lin, and Chih-Jen Lin. Warm start for parameter selection of linear classifiers. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 149–158. ACM, 2015.