Safe optimization algorithms for variable selection and hyperparameter tuning

Eugene Ndiaye

▶ **To cite this version:**

Eugene Ndiaye. Safe optimization algorithms for variable selection and hyperparameter tuning. Optimization and Control [math.OC]. Université Paris Saclay (COmUE), 2018. English. NNT : 2018SACLT004 . tel-01962450

HAL Id: tel-01962450
https://pastel.hal.science/tel-01962450

Submitted on 20 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

NNT : 2018SACLT004

TELECOM ParisTech
Une école de l'IMT

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Télécom ParisTech

Laboratoire d'accueil : Laboratoire Traitement et Communication de l'Information (LTCI)

Spécialité de doctorat : Mathématiques appliquées

Eugène Marc Gana Ndiaye

Safe Optimization Algorithms for Variable Selection and Hyperparameter Tuning

Date de soutenance : 04 Octobre 2018

Après avis des rapporteurs : GABRIEL PEYRE (École Normale Supérieure)
ALAIN RAKOTOMAMONJY (Université de Rouen)

Jury de soutenance :

CHLOÉ-AGATHE AZENCOTT	(Mines ParisTech) Examinatrice
PASCAL BIANCHI	(Télécom ParisTech) Président du jury
OLIVIER FERCOQ	(Télécom ParisTech) Codirecteur de thèse
GABRIEL PEYRE	(École Normale Supérieure) Rapporteur
ALAIN RAKOTOMAMONJY	(Université de Rouen) Rapporteur
JOSEPH SALMON	(Télécom ParisTech) Directeur de thèse
ICHIRO TAKEUCHI	(Nagoya Institute of Technology) Examinateur

Fondation mathématique FMJH Jacques Hadamard

Asa

Am in chains, you're in chains too
I wear uniforms, you wear uniforms too
I'm a prisoner, you're a prisoner too
... Mr Jailer

Jërëjëf !

Jërëjëf à mes deux directeurs de thèse Olivier Fercoq et Joseph Salmon. J'ai eu le luxe de pouvoir me plaindre d'être trop encadré ! Merci pour votre engagement et votre présence pendant ces années de thèse ; pour le temps passé à réfléchir ensemble au tableau, écrire et débugger du code Python et LaTeX ; pour les moments amicaux à Paris ou ailleurs dans le monde. Et sans doute, pour tous les autres moments de Science à venir au Sénégal ou ailleurs en Afrique !

A big thanks to my reviewers Gabriel Peyre and Alain Rakotomamonjy for taking the time to read my thesis in details and for making highly relevant comments. To all of my jury Chloé-Agathe Azencott, Pascal Bianchi and Ichiro Takeuchi to have accepted to be present at my PhD defense. I am very grateful.

Arigatōgozaimashita Ichiro for inviting me to an internship during my second year of thesis and giving me the opportunity to continue in postdoc in your laboratory. Looking forward to exploring new machine learning projects.

Merci à Stéphan Clemençon titulaire de la chaire qui a financé ma thèse et mes voyages très enrichissants en conférence, à Florence d'Alché-Buc pour ses recommandations et François Roueff pour ses autorisations et autres signatures pour que la thèse ait lieu. À Alexandre Gramfort qui a été très présent, tel un troisième encadrant qui ne dit pas son nom. Enfin tous les enseignants-chercheurs que j'ai pu côtoyer à Télécom ParisTech aussi bien pour des travaux pratiques que des regards croisés dans les couloirs.

Un point essentiel dans ma formation est sans doute mon passage au Cours Secondaire Sacré-Coeur. Avec Mr. Niokhor Ndong (ses astuces si précieuses sont encore utilisées dans ce manuscrit !) et Mr Diatta qui nous ont transmis les bases fondamentales en science, Mr Thiombane (sans doute le prof de français le plus matheux qui existe), Mister D.i.o.m.p.y (Yes sir, we are always trying to do better !). Chouettes rencontres à l'Université Paris-Est Marne-La-Vallée et une chance d'avoir rencontré M. Hebiri qui a eu un impact considérable dans mon cursus. Arghhh ... le département de math de l'Université Paris-Sud avec son lot de profs absolument impressionnant et passionnant. Un grand merci à Madame Hulin et Nathalie Carrière (non sans vouvoiement :p).

Pour les sourires à l'accueil de TélécomPa', la cafet' (Christian, toujours avec du bon son mais supporte l'OM) et resto (le rab, miaamm) et même la sécurité (ouais j'ai ma carte d'étudiant tkt ;-)). L'administration avec la très chère Tata Laurence Zelmar qui m'aide toujours à remplir les fiches de mission avec toute la gentillesse du monde (ça en fait même des jaloux auprès de certains collègues !), Florence Besnard et Dominique Roux pour être très compréhensifs et sans qui je n'aurai pas pu m'inscrire (contrairement à une certaine administration avec ses titres de séjour d'une durée absurdement attribuée !)

Ouais, on en croise des gens divers et cool dans les circuits d'étude. Emmanuel Babala ("gros vilain" qui est devenu bg, musclé et gentil), Kevin Yang, Zo Rakotondrafara et Levy Operman (un peu comme superman mais avec un génie : invisible en cours mais fort quand même). Toutes mes rencontres à l'aumônerie "sous le figuier" du temps de la Soeur Christiane Joly et du Père Patrick, l'accueil chaleureux et à la française de la famille Devernay. Grâce à vous, je me suis senti moins seul à Marne-La-Vallée. Puis s'ensuit des parcours sous les bois pour rentrer dans les

amphis de la fac d'Orsay, croiser des personnages emblématiques des maths, ne rien comprendre en cours ... puis en parler à Tran (copain survivor), Rafa, Justine (sans doute une excellente prof aujourd'hui), Havet (Ave Havet, grand merci pour les places d'opera et de comédies françaises), Dimitri, Zurek, Ghislain, Timothée, Jeanne, Victor, Nadège, Marion, Diane, Romain, Que-ja (dans l'abus des failles de distributeurs de canettes), B.G., Poulenard, Royer, Samuel, Toufik, Camille, Xiao, Joris, ... `Memory Error` sans doute mais un grand merci pour ces moments.

Et puis thèse, oups ! mais quand même chouchouté au Fiap par le grand chef du pestacle. C'était bien cool de chiller avec les plus grands hypster Claire Vernade en cheftaine, Albert Thomas le grand frère, Igor, Maxime Sangnier, Nico, Éric, Andrés, Chirine et Minh (merci de m'avoir fait de la place dans votre bureau anciennement E305, mon premier bureau d'ailleurs ! de m'avoir supporté toutes les fois où je me morfondais en début de thèse), Anna Korba (les échanges de derniers sons stylés pour ne surtout pas faire sa thèse, que ça ? je tairais le reste ma gueule !), Adil Salim (ps : BAI) et Moussab Djerrab pour les generationels d'Adum et de soutenances, Umut, Mastane, Magda, Tom, Romain, Hamid et Alex comme représentants de la start up $_{\text{nation}}$, Constance et Raphaël pour les `PhDrink`, Valentin (toujours des bons plans de soirée ;-)). Mais aussi Arthur, Elvis, Alexandre. Et merci à toi, cher collègue de télécomPa' que j'ai oublié de mentionner ... `Memory Error` sans doute mais un grand merci pour ces moments.

Et `E306 skrrr paw paw`, Mathus et les cailloux (Pierre et Pierre, sur ces pierres j'aperçois la belle Église de Sainte-Anne). Vous avez survécu à mon humour et mes humeurs ! La vie est belle, n'est ce pas P.L. ? On sera matinal pour un lever du soleil dans un village corse, parisien, dakarois ou même japonais !

La coloc o͡o ! Céline, Benoît, Kevish et Heythem dans une permutation stochastique, merci d'avoir accepté de vivre avec un mort d'entre les morts qui vit et meurt en même temps. De toute façon, "on meurt tous un jour, même les ballons" (et même le `ginosaaaaji` ! ! !) et puis je suis le plus fort à FIFA et puis je fais du sport, et puis je rentre tard et puis je vous aime quand même. Merci pour ces moments de maths, de rire, de ras le bol et d'amitié. God bless you !

Et comme "... on ne se fait pas d'amis d'enfance dans une maison de retraite ...", merci à Thiery et Souleymane (my brothers from another mothers) de supporter ma part de folie. JB (Tiép), Fodé, Thiam, Romuald, Lahad (l'esprit), Fatou Touré, Seydina (sage homme), Waly, Chérif, Abass, et toute la clique de la S1 ... et tous les autres encore ... Petit Jo (devenu trop costaud maintenant), Simon P. Faye, Olivier, Thierry Sossou et Momar Sylla (my brothers from another mothers), Babs Fall ... `Memory Error` sans doute mais un grand merci pour ces moments de kiffologie active.

Merci à toi Élisabeth Diouf pour m'avoir logé, accompagné et conseillé lors de mes premiers pas en France. À toute Ma famille Diouf, Merci d'être là. Au très inspirant Isaac, God bless you !

Merci à ma Ta' Diouma et Bernie pour votre amour et disponibilité. `Amouma sen fay <3`.

Merci à toi Anaïs (et ta famille), merci de m'avoir toujours soutenu en cette fin de thèse malgré mes "absences" et de bâtir de nouvelles et belles aventures avec moi. `Jërëjëf` !

Enfin un grand merci à ma famille à qui je dois absolument tout. Je suis si fier et ravi de rejoindre mes parents (Gabriel et Marie) au rang académique de Docteur (même si en tant que fœtus, je soutenais déjà une thèse en médecine !). À mes petites lovely and so inspirational sisters Catherine et Élisabeth (Vielen Dank !), et les brothers Zacharie et Gilbert toujours protecteurs. God bless you ! Cette thèse vous est entièrement dédiée.

Table des matières

Chapter 1

Motivations and Contributions

Automatic data processing has become ubiquitous in current technologies, with important applications in many scientific disciplines such as medicine, biology or meteorology but also in digital tools such as text translation, targeted advertising or spam detection. It is at the heart of problems in signal processing, information theory and statistics where one aims at understanding and summarizing the essential information in the collected data. The latter are often accompanied by prior information on their structures, which can then be used to lay predictive models and algorithms. Nevertheless, these algorithms depend heavily on mathematical optimization methods and it has become crucial to have tools that remain (computationally) effective when the size of the database increases. We investigate computational simplifications in some optimization problems arising from statistical learning with a main conductive thread: the saving of calculations based on optimality certificates and the exploitation of specific regularity structures of the problems.

In this chapter, we recall the importance of optimization in learning with a focus on convex formulation, outline the contributions of the thesis and introduce notation as well as convex analysis tools used along the manuscript.

1.1 Convex Optimization in Statistical Learning

We follow a classical formalization of statistical learning tasks as in (Hastie et al., 2009; Shalev-Shwartz and Ben-David, 2014). Let $\mathcal{X}$ (resp. $\mathcal{Y}$) be a set of input (resp. output) vectors and X (resp. Y) be a random variable valued in $\mathcal{X}$ (resp. $\mathcal{Y}$). We call *learning task* the identification of an application $h : \mathcal{X} \mapsto \mathcal{Y}$ that explains the relation between the input X and the output Y. Considering a *loss* (also called *cost* or *deviance*) function ℓ such that $\ell(y, y) = 0$, $\ell(y, y') \geqslant 0$, we want to learn a prediction function h minimizing the prediction error $\ell(h(X), Y)$ in expectation. For simplicity, we consider that h will be searched on a (pre-defined) parameterized family of functions $\mathcal{H} := \{h(\cdot, \beta) : \beta \in \mathbb{R}^p\}$ that encodes the prior knowledge we have on the data. Then, the learning task can be written as the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} R(\beta) := \mathbb{E}[\ell(h(X, \beta), Y)] \ . \tag{1.1}$$

Since the expectation is taken under the joint probability distribution $\mathbb{P}_{X,Y}$ of the variables X, Y which is assumed to be *unknown*, h cannot directly be learned that way: one should rather learn by considering a training sample that represents the observations at hand. Assuming that the observations on a given dataset $\{(x_i, y_i)\}_{i \in [n]}$ are independent and identically distributed, by the law of large numbers, the empirical law $\frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$, where the δ represents Dirac masses, approximates the true distribution $\mathbb{P}_{X,Y}$ if the number of observations n is sufficiently large. Hence the *Empirical Risk Minimization* (ERM) paradigm reads

$$\min_{\beta\in\mathbb{R}^p} R_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i,\beta),y_i) \ . \tag{1.2}$$

A popular instantiation of formulation (1.1), (1.2) is the fundamental tool in Statistics known as the *Maximum Likelihood Estimation (MLE)*. We refer to (Van der Vaart, 1998) for a comprehensive description and (Stigler, 2007) for the passionated history of this method. Interestingly, the MLE for the exponential family naturally leads to convex optimization problem (Brown, 1986).

Definition 1 (Exponential Family)**.** *Let ν be a σ-finite measure and $\lambda(\theta) = \int \exp(\theta y)\nu(dy)$ its Laplace transform with domain $N = \{\theta \in \mathbb{R}^n : \lambda(\theta) < +\infty\}$. For $P(\theta) = \log(\lambda(\theta))$, we define*

$$p_\theta(y) = \exp\left(\langle\theta, y\rangle - P(\theta)\right) \ . \tag{1.3}$$

For a convex set $\Theta \subset N$, the family of density $\{p_\theta : \theta \in \Theta\}$ is called (standard) exponential family.

The convexity of the optimization problem derived from MLE of exponential family follows directly from the convexity of the Log-Laplace transform P, we recall the proof in Chapter 4.

Theorem 1 (Brown (1986, Theorem 1.13))**.** *N is a convex set and P is a convex function on N. Furthermore, P is lower semi-continuous on $\mathbb{R}^n$ and continuous on the interior of N.*

In statistical inference, it is common to suppose that the distribution of the observations is parameterized by some $\theta_0 \in \Theta$ that is unknown. Then, the objective is to approximate and provide information on the model parameter, failing to find it exactly, from random variables distributed under this law. A classical inferential method is the MLE. For a variable y in the convex support of ν and Θ a convex subset of N, the function of the parameter $\Theta \ni \theta \mapsto p_\theta(y)$ is called *likelihood* at y. Now, assuming that the parameter θ_0 belongs to Θ and y being a random variable with distribution P_{θ_0}, the maximum likelihood estimator is defined as

$$\hat{\theta}(y) = \arg\max_{\theta\in\Theta} p_\theta(y) = \arg\min_{\theta\in\Theta} -\log(p_\theta(y)) = \arg\min_{\theta\in\Theta} P(\theta) - \langle\theta, y\rangle$$

which is a convex optimization problem with a loss function $\ell(\theta, y) := P(\theta) - \langle\theta, y\rangle$. Thus, the MLE for independent and identically distributed samples $y = (y_1, \cdots, y_n)$ can be expressed as

$$\hat{\theta}(y) \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i,\theta) \ . \tag{1.4}$$

In a learning setting, an important example is the generalization of regression leading to the family of *Generalized Linear Model* (GLM) (McCullagh and Nelder, 1989) where the statistical model contains a deterministic part given by a linear combination of the covariates $\eta = X\beta$ and the random part given by $\mu = \mathbb{E}[Y]$ where Y is assumed to belong to an exponential family, are linked by $h(\eta) = \mu$. Depending on the distribution of the observations, we recover the Least Squares and logistic estimation as canonical examples for regression and classification tasks.

Least Squares. Given an independent and identically distributed sample $y = (y_1, \cdots, y_n)$ with Gaussian law $\mathcal{N}(\mu, \sigma^2)$ *i.e.*

$$\begin{aligned} p_{(\mu,\sigma^2)}(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\langle(\theta_1,\theta_2),(y,y^2)\rangle - P(\theta_1,\theta_2)\right\} \end{aligned}$$

where $\theta = (\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ and $P(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(-\frac{\pi}{\theta_2})$. Assuming that σ^2 (hence θ_2) is *known*, the MLE (1.4) for the Gaussian model with mean $\mu = X\beta$ reads:

$$\hat{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i - x_i^\top\beta)^2 \ . \tag{1.5}$$

Logistic regression. For a binary variable $y \in \{0, 1\}$ that follows the Bernoulli distribution with mean μ *i.e.*

$$p_\mu(y) = \mu^y(1-\mu)^{1-y} = \exp\{\langle \theta, y\rangle - P(\theta)\} \ ,$$

where $\theta = \log(\frac{\mu}{1-\mu})$ and $P(\theta) = \log(\frac{1}{1+e^\theta})$. Stated in the regression setting where the deterministic part is $\theta = X\beta$, the MLE (1.4) and given an i.i.d. sample $y = (y_1, \cdots, y_n)$ of Bernoulli distribution reads:

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \log\left(1 + \exp(x_i^\top \beta)\right) - y_i x_i^\top \beta \ .$$

However, the ERM is not restricted to statistical models based on likelihood. Many learning paradigms provide a good predictor without assumptions on the underlying distribution over the data.

Hinge Loss minimization. In binary classification tasks, the *Perceptron* simply seeks a separation of the data points in two half-spaces. It can be formulated as a minimization of the following loss:

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i x_i^\top \beta) \ .$$

It is included in the *Support Vector Machine* (SVM) paradigm (when one adds a quadratic regularization) that separates data points into two halfspaces while maximizing a margin. The later guarantees better convergence properties and allows better prediction performance on unseen data.

Tradeoffs of Large Scale Learning

To understand the generalization capabilities of statistical learning methods, it is important to analyze the different sources of error that can corrupt our predictions. These errors come mainly from the assumptions that are introduced into the learning model. More precisely, let us denote $\beta_* \in \arg\min_{\beta \in \mathbb{R}^p} R(\beta)$ the minimizer of the true risk (1.1) and $\beta_n \in \arg\min_{\beta \in \mathbb{R}^p} R_n(\beta)$ the minimizer of the empirical risk (1.2). Importantly, we have to take seriously into account the fact that *"in general, optimization problems are unsolvable"* (Nesterov, 2004, Chapter 1). Hence we will assume that we only have access to an approximated solution β_n^ϵ of the ERM (1.2) *i.e.* for a targeted accuracy $\epsilon > 0$ controlling the optimization error, the vector β_n^ϵ satisfies

$$R_n(\beta_n^\epsilon) - R_n(\beta_n) \leqslant \epsilon \ .$$

Bousquet and Bottou (2008) highlighted the Approximation-Estimation-Optimization tradeoffs to characterize the complexity of the learning task and state that *«computational complexity becomes the limiting factor when one envisions large amounts of training data»*. Indeed, they provided a fundamental decomposition of the mean error in large scale learning as

$$\mathbb{E}[R(\beta_n^\epsilon)] = \mathcal{E}_{\text{approximation}} + \mathcal{E}_{\text{estimation}} + \mathcal{E}_{\text{optimization}} \ ,$$

where

- $\mathcal{E}_{\text{approximation}} := R(\beta_*)$ is the residual error made by restricting the analysis to a family of hypothesis function $\mathcal{H}$,
- $\mathcal{E}_{\text{estimation}} := \mathbb{E}[R(\beta_n) - R(\beta_*)]$ is the error resulting from the empirical approximation of the joint distribution of the data $\mathbb{P}_{X,Y}$,
- $\mathcal{E}_{\text{optimization}} := \mathbb{E}[R(\beta_n^\epsilon) - R(\beta_n)]$ is the (expected) optimization gap remaining when solving the ERM problem (1.2).

Before the "big data" era, the tradeoffs $\mathcal{E}_{\text{approximation}}$ vs. $\mathcal{E}_{\text{estimation}}$ also called *Bias-Variance* tradeoff was the most popular one, in particular among statisticians: *small scale* setting. However, when the size of the dataset increases (*large scale*), the scalability of the optimization algorithms for computing the parameters of statistical learning models becomes critical. Hence, given a predefined accuracy ϵ, the associated optimization problems must be efficiently addressed. To do so, one needs to explicitly exploit structures of the problems and to design specialized solvers.

Optimization Algorithms

In a large scale setting, the difficulties of solving the optimization problem (1.2) range from memory limitation, nonlinearity, non-smoothness to even non convex problems. In this case, the dimensionality can be so large that algorithms requiring evaluations of quantities relying on the full dataset become intractable. A popular trend in optimization for machine learning is to go back to simple methods developed with limited computational resources and popularized in the 50's, see (Bottou et al., 2016) for a recent review. Hence, algorithms that provide cheap and fast computations with "limited" information has been privileged *e.g.* incremental optimization including stochastic gradient descent (Robbins and Monro, 1951), Frank-Wolfe algorithm (Frank and Wolfe, 1956) also referred to as conditional gradient descent, (block) coordinate descent (Warga, 1963), and active set methods. We first highlight the two main optimization principles that are systematically used in this manuscript namely *Homotopy Continuation* and *Majorization-Minimization*.

Homotopy Continuation Principle

Homotopy continuation methods aim at evaluating the full curve of solutions of nonlinear equations $H(x, \lambda) = 0$ for a continuous range of parameter $\lambda \in \Lambda$ (for Λ a set to be specified later). The maps $H(x, \lambda)$ represent a continuous deformation of a nonlinear function $F(x)$ whose zeros are hard to find, see (Allgower and Georg, 2012) for a comprehensive description. It appears naturally in machine learning for improving numerical stability and preventing over-fitting. Indeed, solving the empirical risk minimization problem (1.2) is often not sufficient for finding good predictors because the problem tends to be ill conditioned in high dimensional settings. A classical approach consists in adding a regularization term that encodes additional knowledge on the problem. For instance it can be used to enforce the selection of simpler models and can be formulated as

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i, \beta), y_i) + \lambda \Omega(\beta) \ , \tag{1.6}$$

where Ω is the regularization function that penalizes complex solution and $\lambda > 0$ controls the level of inductive bias. It is usually related to the *simplicity principle* of G. Ockham in the 14th century or (Wrinch and Jeffreys, 1921). The regularization term balances between the minimization of the empirical risk and the structural simplicity of the model through the hyperparameter λ. Finding the optimal balance is crucial to achieve good prediction on unseen datasets: small λs lead to complex models that are likely to over fit on the training set while large λs lead to simplistic models with poor prediction power. A common approach to select a "good" parameter is to use *cross validation*. Essentially, this method avoids to perform training and evaluating the performance of an estimator on the same data. It was introduced in (Larson, 1931), see (Arlot and Celisse, 2010) for a comprehensive review. For simplicity, we consider here the simplified holdout version that consists in splitting the data $\{(x_i, y_i)\}_{i \in [n]}$ in two parts $(X_{\text{train}}, y_{\text{train}})$ and $(X_{\text{test}}, y_{\text{test}})$, and consider Λ a discrete set of hyperparameters. Given a validation loss function $\mathcal{L}$ that measures the prediction error on the test set, the holdout version of cross-validation corresponds to performing the two following steps:

1. solve problem (1.6) with the training data $(X_{\text{train}}, y_{\text{train}})$ for all $\lambda \in \Lambda$,

2. choose the $\lambda \in \Lambda$ that minimizes the validation error $\mathcal{L}(h(X_{\text{test}}, \hat{\beta}^{(\lambda)}), y_{\text{test}})$.

A standard grid considered in the literature is $\lambda_t = \lambda_{\max} 10^{-\delta t/(T-1)}$ with a small δ ($\delta = 10^{-2}$ or 10^{-3}), see for instance (Bühlmann and van de Geer, 2011)[2.12.1] or the `glmnet` package (Friedman et al., 2010b) and `scikit-learn` (Pedregosa et al., 2011). Choosing δ is challenging both from a statistical point of view (the performance tends to decrease as δ becomes close to zero, due to over-fitting) and from an optimization point of view since the computational burden tends to increase for small λ, the primal iterates being less and less sparse, and the problem to solve more and more ill-posed. It is customary to start from the largest regularizer $\lambda_0 = \lambda_{\max}$ and then to perform iteratively the computation of $\hat{\beta}^{(\lambda_t)}$ after the one of $\hat{\beta}^{(\lambda_{t-1})}$. This leads to computing the models (generally) in the order of increasing complexity: this allows important speed-up by benefiting of warm start initialization.

Depending on the context, several regularizers Ω were introduced to enforce regularity of the estimators. Popular examples used in our experiments are:

Ridge/Tikhonov Regularization. The regularization function $\Omega(\beta) = \|\beta\|_2^2/2$ was introduced in (Tikhonov, 1943) to improve the stability of inverse problems, and in statistics (Hoerl, 1962; Hoerl and Kennard, 1970) to reduce the mean square error of the vanilla Least Squares estimator when the design matrix is rank deficient. In machine learning, it is often viewed as a stabilizer of the learning algorithm in the sense that the prediction does not change much when the input data are slightly perturbed. As a consequence, the training error remains close to the test error and this prevents the algorithm from over-fitting (Shalev-Shwartz and Ben-David, 2014, Chapter 13.2).

While fundamental, preventing the over-fitting phenomenon is not sufficient in many applications. Often, one also needs to have a good representation of the data and to provide prediction models that are interpretable. Thus, it is crucial to be able to select the most relevant explanatory variables, which is what motivated the introduction of sparse regularization methods.

Sparse Lasso Regularization. The regularization $\Omega(\beta) = \|\beta\|_1$ was introduced in (Chen and Donoho, 1995; Tibshirani, 1996) in signal processing and statistics and follows classical methods for selecting the most important explanatory variables in multiple regression (Efroymson, 1960) for stepwise regression or (Breiman, 1995) for selection with non-negative garrote. The ℓ_1 norm penalty has the advantage of being able to select variables in a continuous way and its convex formulation allows the use of fast iterative algorithm.

Later, several extensions were proposed, notably by Zou and Hastie (2005) for the Elastic net where $\Omega(\beta) = \alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2/2$ interpolates between the Ridge and the Lasso, by Hebiri and van de Geer (2011) for the Smoothed Lasso where $\Omega(\beta) = \alpha \|\beta\|_1 + \gamma \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2$, or for more complex hierarchical group regularizations (Friedman et al., 2010a; Sprechmann et al., 2011). A survey providing a unified theory for convex structured sparsity-inducing norms was recently proposed in (Obozinski and Bach, 2016). Note that sparsity can also be incorporated into the data fitting term. This is the case of the hinge loss which can be used as a variable selection criterion as well (Guyon et al., 2002; Rakotomamonjy, 2003).

While using regularization, the generalization performance of the ERM is then strongly related to the capabilities of tuning the regularization parameter λ. This requires the computation of full solution path in the homotopy continuation framework over a range (often a discrete set) of hyperparameters Λ. Indeed, it is usually infeasible to compute the whole path in a continuous set if no closed form solution in Equation (1.6) is available. However, for problems involving piecewise quadratic loss and piecewise linear regularizations, the solution path $\{\hat{\beta}^{(\lambda)}, \lambda \in \Lambda\}$ is continuous and piecewise linear (Rosset and Zhu, 2007). This specific piecewise linearity allows to compute efficiently and exactly the entire solution path. This kind of property was rediscovered several times in the literature, for instance in (Markowitz, 1952) for portfolio selection, (Osborne, 1992)

for quantile regression problems, (Osborne et al., 2000a) for Lasso, (Efron et al., 2004) Lars, (Park and Hastie, 2007) for the GLM regularized with the ℓ_1 norm.

Majorization-Minimization Principle (MM)

MM is a generic and powerful technique for building iterative optimization algorithms. At each step, it simply minimizes a surrogate upper-bound of the objective function that is tight at the current estimate. Its description can be traced back at least to (Ortega and Rheinboldt, 1970), see also (Hunter and Lange, 2004) for a synthetic review. It has also been recently used in machine learning to derive stochastic incremental algorithm (Mairal, 2015).

Definition 2 (Surrogate)**.** *A function $\hat{P}(\cdot|\beta^{(0)})$ is said to be a surrogate of P near $\beta^{(0)}$ if the following conditions holds:*

$$\begin{cases} P(\beta) \leqslant \hat{P}(\beta|\beta^{(0)}) & \text{for all } \beta \in \mathbb{R}^p, \\ P(\beta^{(0)}) = \hat{P}(\beta^{(0)}|\beta^{(0)}) \ . \end{cases}$$

Given an iterate $\beta^{(k)}$, in order to solve $\min_{\beta\in\mathbb{R}^p} P(\beta)$, the MM algorithm update is given by $\beta^{(k+1)} = \arg\min_{\beta\in\mathbb{R}^p} \hat{P}(\beta|\beta^{(k)})$. It satisfies the appealing descent property $P(\beta^{(k+1)}) \leqslant P(\beta^{(k)})$ for all iteration k. Many optimization algorithms used to solve machine learning problems can be written under this framework. For example, the *proximal point algorithm* (Martinet, 1970; Parikh and Boyd, 2014) is an MM with the surrogate $\hat{P}(\beta|\beta^{(k)}) := P(\beta) + \frac{L}{2}\|\beta - \beta^{(k)}\|^2$ for some nonnegative constant L. For simplicity, we now suppose that the data fitting term $\frac{1}{n}\sum_{i=1}^n \ell(h(x_i,\beta), y_i)$ is represented by $f(X\beta)$ and thus the regularized ERM (1.6) reads:

$$\min_{\beta\in\mathbb{R}^p} P_\lambda(\beta) := f(X\beta) + \lambda\Omega(\beta) \ . \tag{1.7}$$

We furthermore assume that f is differentiable with L_f-Lipschitz continuous gradient. In this setting, a classical way to construct a surrogate function of P_λ is to upper bound the first order Taylor expansion of f. Indeed, given a vector $\beta^{(0)}$ and a direction η in $\mathbb{R}^p$, we have

$$f(X(\beta^{(0)} + \eta)) \leqslant f(X\beta^{(0)}) + \langle \nabla f(X\beta^{(0)}), X\eta\rangle + \frac{L_f}{2}\|X\eta\|^2 \ . \tag{1.8}$$

Using this upper bound, we can now describe two proximal algorithms that are very useful in large scale machine learning because they can handle different regularization structures in the convex optimization problem (1.7).

Proximal Gradient Algorithm. Given an iterate $\beta^{(k)}$, a direction $\eta = \beta - \beta^{(k)}$ and denoting $L := L_f\sigma_{\max}(X^\top X)$, where $\sigma_{\max}(X^\top X)$ is the largest singular value of $X^\top X$, the proximal gradient algorithm can be expressed as a MM algorithm with the surrogate based on the upper bound (1.8)

$$\hat{P}_\lambda(\beta|\beta^{(k)}) := f(X\beta^{(k)}) + \langle \nabla f(X\beta^{(k)}), X(\beta - \beta^{(k)})\rangle + \frac{L}{2}\|\beta - \beta^{(k)}\|^2 + \lambda\Omega(\beta) \ .$$

Whence the next iteration is the minimizer of $\hat{P}_\lambda(\beta|\beta^{(k)})$ *i.e.*

$$\beta^{(k+1)} = \arg\min_{\beta\in\mathbb{R}^p} \frac{\lambda}{L}\Omega(\beta) + \frac{1}{2}\left\|\beta - \left(\beta^{(k)} - \frac{1}{L}X^\top\nabla f(X\beta^{(k)})\right)\right\|^2 \ . \tag{1.9}$$

Thereby, the special case where there is no regularization *i.e.* λ or Ω equal to zero recovers the vanilla gradient descent algorithm with the iteration updates:

$$\beta^{(k+1)} = \beta^{(k)} - \frac{1}{L}X^\top\nabla f(X\beta^{(k)}) \ .$$

The proximal gradient algorithm and accelerated variants (Nesterov, 2004) have been widely used to solve linear inverse problems arising in signal/image processing; see also (Beck and Teboulle, 2009; Combettes and Pesquet, 2011) for more descriptions and analysis.

Proximal (block) Coordinate Gradient Algorithm. It is one of the flagship algorithm for dealing with ERM with a separable regularization structure (Friedman et al., 2007; Nesterov, 2012). It owes its popularity to its low iteration cost while having a rate of convergence proportional to that of the full (proximal) gradient descent. We refer to (Wright, 2015) for a recent review on coordinate descent. Assuming the regularization function Ω decomposes into separable group $\mathcal{G}$:

$$\mathbb{R}^p = \bigotimes_{g\in\mathcal{G}} \mathbb{R}^{|g|} \text{ and } \Omega(\beta) = \sum_{g\in\mathcal{G}} \Omega_g(\beta_g) \ ,$$

we can define the canonical partition associated to the group structure $\mathcal{G}$ of the unit matrix

$$\mathrm{I}_p = (\cdots, e_g, \cdots) \in \mathbb{R}^{p\times p}, \quad \text{for } g \in \mathcal{G} \text{ and where } e_g \in \mathbb{R}^{p\times|g|} \ .$$

This allow to represent $\beta = \sum_{g\in\mathcal{G}} e_g^\top \beta_g$. Then given an iterate $\beta^{(k)}$, a direction $\eta = e_g^\top(\beta_g - \beta_g^{(k)})$ and denoting $L_g := L_f \sigma_{\max}(X_g^\top X_g)$, the proximal (block) coordinate gradient algorithm can be expressed as a MM algorithm which iteratively loops over the surrogates (1.8) for each block g

$$\hat{P}_\lambda(\beta_g|\beta^{(k)}) := f(X\beta^{(k)}) + \langle \nabla f(X\beta^{(k)}), Xe_g^\top(\beta_g - \beta_g^{(k)})\rangle + \frac{L_g}{2}\|\beta_g - \beta_g^{(k)}\|^2 + \lambda\Omega_g(\beta_g) \ .$$

Denoting $X_g = Xe_g$, the next iteration proceeds by choosing a block g in $\mathcal{G}$ and minimizing the surrogate $\hat{P}_\lambda(\beta_g|\beta_g^{(k)})$ *i.e.*

$$\beta_g^{(k+1)} = \operatorname*{arg\,min}_{\beta_g\in\mathbb{R}^{|g|}} \frac{\lambda}{L_g}\Omega_g(\beta_g) + \frac{1}{2}\left\|\beta_g - \left(\beta_g^{(k)} - \frac{1}{L_g}X_g^\top \nabla f(X\beta^{(k)})\right)\right\|^2 . \tag{1.10}$$

As far as we know, of these two algorithms (1.9), (1.10), there is not one that is uniformly better than the other. However for functions with computationally cheap block coordinate derivatives such as (1.7), the Proximal (block) coordinate gradient algorithm tends to be much faster than the (full) proximal counterpart. This is especially true when smart updates of the coordinate gradient can be performed by taking benefit of favorable problem structures. In fact, one can notice that between two successive iterations, the vectors $\beta^{(k+1)}$ and $\beta^{(k)}$ differ only in their g-th block coordinates which has been selected. Hence, defining the vector $E_k := X\beta^{(k)}$, one may observe that $E_{k+1} = E_k + X_g(\beta_g^{(k+1)} - \beta_g^{(k+1)})$. Thus, denoting $\mathrm{nnz}(X_g)$ the number of non-zero entries in X_g, the direction of descent in Equation (1.10) can be updated in $O(\mathrm{nnz}(X_g))$ versus $O(\sum_{g\in\mathcal{G}} \mathrm{nnz}(X_g))$ for the (full) proximal gradient descent. While enjoying cheaper update of the iterations, Nesterov (2012) shows that coordinate descent algorithm with random selection can also have better rate of convergence than full gradient descent because its rate of convergence depends on the average of the coordinate-wise Lipschitz constant, see also (Richtárik and Takáč, 2014). Note also that calculations of coordinate descent based algorithm can be parallelized, which has the significant advantage of being able to take advantage of current computer architectures (Fercoq and Richtárik, 2015).

Beside the overall regularity of the functions involved, explicitly exploiting the structure of functions allows for designing faster optimization algorithms. One of the main contributions of this thesis is to propose additional speed-up by saving a considerable amount of the calculations made along the iterations. We will only consider convex optimization problems in the learning task defined in (1.6) where the hypothesis class $\mathcal{H}$ and the loss function ℓ are assumed to both be convex. We have seen that such a convex formulation already includes a large class of statistical learning tasks such as maximum likelihood estimation for exponential family distribution but also formulations resulting from the support vector machine paradigm.

1.2 Outline of the Contributions

The contributions of the thesis were published in machine learning conferences and journals:

Authors: E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon.
1− "Gap Safe Screening Rules for Sparse Multi-task and Multi-class Models".
Advances in Neural Information Processing Systems, 811-819, 2015.

2− "Gap Safe Screening Rules for Sparse-Group Lasso".
Advances in Neural Information Processing Systems, 388-396, 2016.

3− "Gap Safe Screening Rules for Sparsity Enforcing Penalties".
The Journal of Machine Learning Research 18 (1), 4671-4703, 2017.

Authors: E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, J. Salmon.
4− "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression".
Journal of Physics: Conference Series 904 (1), 012006, 2017.

Authors: E. Ndiaye, T. Le, O. Fercoq, J. Salmon, I. Takeuchi.
5− Safe Grid Search with Optimal Complexity. *Submitted*, 2018.

We present in details the results we obtained in the different chapters of the thesis as follows.

Chapter 2. We consider regularized ERM problems stated as the sum of a smooth term (data fitting) and a non-smooth term (penalty on the complexity of the solution, indirectly its sparsity), or vice versa. We show how to exploit a particular structure of the solutions to ignore unimportant variables in the optimization process without false exclusion and consequently leading to faster solvers. The underlying rationale is that there is no gain in performing worthless computations involved with non-influential features or observations. This strategy called *(safe) screening* follows the seminal work by El Ghaoui et al. (2012) and has rapidly led to an increasing literature in order to apply it to different instantiations of problem (1.6). We propose a unifying framework that highlights underlying structures of convex functions that are commonly exploited in previous derivations of screening rules for (separable) non-smooth regularized empirical risk minimization. Our method is based on the exploitation of first order optimality conditions and separation properties of the subdifferentials of convex functions which generalize the theoretical screening rules previously known in the literature. It applies to a large class of supervised learning tasks such as Lasso, Sparse-Group Lasso, multi-task Lasso, binary and multinomial logistic regression, support vector machine to name a few. Finally, leveraging information given by duality gap bounds, we provide theoretical results such as iteration complexity of active set identification and design new fast algorithms to discard safely more variables than previously considered safe rules, particularly for low regularization parameters. Our approach can cope with any iterative solver but are particularly well suited to (block) coordinate descent methods. We also introduce new active warm start strategies that have shown improved performance. In our numerical experiments, we report significant speed-ups compared to previously proposed safe rules on all tested datasets.

Chapter 3. We discuss approximated pathwise optimization and application in model selection. Despite the appealing property of homotopy continuation methods for providing better prediction in term of generalization performance, the selection of the optimal λ *w.r.t.* to validation error can be difficult even for problem such as Lasso where we can find an algorithm that computes exactly the entire solution path. Furthermore, the path following algorithm such as Lars (Efron et al., 2004) or predictor-corrector based methods may suffer

from numerical instabilities due to several matrix inversion and their complexity, *i.e.* the number of linear segments in the path, can be exponential in the dimension of the problem. For instance the worst case complexity for the Lasso is exactly $(3^p + 1)/2$ (Mairal and Yu, 2012) and $O(2^n)$ for the SVM (Gärtner et al., 2012). In this chapter, we revisit the techniques of approximating the solution path up to predefined tolerance in a unified framework and show that its complexity is $O(1/\sqrt[d]{\epsilon})$ for uniformly convex loss of order $d > 0$ and $O(1/\sqrt{\epsilon})$ for Generalized Self-Concordant functions. This includes examples such as Least Squares loss, but also the important example of logistic loss which, as far as we know, was not handled by previous works. Moreover, we clarify the link between the complexity of the approximated solution path and the regularity of the loss function considered in the reguralized ERM setting. Finally, we leverage our technique to provide refined bounds on the validation error and provide a practical algorithm for hyperparameter selection with stronger guarantee. More precisely, given the training and validation splitting data $(y_{\text{train}}, X_{\text{train}}, y_{\text{val}}, X_{\text{val}})$, we formulate the problem as a bi-level optimization one

$$\begin{aligned} \underset{\lambda \in [\lambda_{\min}, \lambda_{\max}]}{\arg\min} \; & E_v(\hat{\beta}^{(\lambda)}) = \mathcal{L}(y_{\text{val}}, X_{\text{val}} \hat{\beta}^{(\lambda)}) \\ & \text{s.t. } \hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \ell(X_{\text{train}}\beta, y_{\text{train}}) + \lambda \Omega(\beta) \;\; . \end{aligned}$$

Given a prescribed tolerance $\epsilon_v > 0$ of the prediction error on the validation set, we show how to design a discrete grid of parameter $\Lambda_{\text{val}}(\epsilon_v)$ included in $[\lambda_{\min}, \lambda_{\max}]$ such that:

$$\min_{\lambda_t \in \Lambda_{\text{val}}(\epsilon_v)} E_v(\beta^{(\lambda_t)}) - \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} E_v(\hat{\beta}^{(\lambda)}) \leqslant \epsilon_v.$$

Therefore, our contribution simply consists in a sequential exploration algorithm à la grid search while benefiting of global convergence guarantee for approximating the optimal hyperparameter at a validation error level ϵ_v.

Chapter 4. The maximum likelihood estimation is a classical and important statistical learning paradigm requiring a good statistical model to be specified. For instance, a linear model with Gaussian noise requires estimation of both the position and dispersion parameters (μ, σ^2). If σ is *known* and the observations $y = (y_1, \cdots, y_n)$ are independent and identically distributed the MLE leads to the classical Least Squares estimation (1.5) and the influence of σ can be discarded. However, if σ is *unknown*, estimating only μ is not sufficient to approximate the distribution P_{μ,σ^2} which leads to an incomplete model. Moreover, in high dimensional settings where the number of observations is smaller than the number of features, sparsity enforcing methods such as Lasso are very popular because they can select important variables and ease the interpretation of discriminant features. For efficiency, they rely on tuning a regularization parameter that trades data fitting versus sparsity, and should be proportional to the noise level σ (Bickel et al., 2009). Yet, the latter is often *unknown* in practice. A possible remedy is to jointly optimize over the regression parameter μ as well as over the noise level σ. A direct formulation of the MLE reads

$$\min_{\beta \in \mathbb{R}^p, \sigma > 0} \log(\sigma) + \frac{1}{2\sigma^2} \|y - X\beta\|^2$$

which fails to be jointly convex. Also, when $y = X\beta$ and σ tends to 0 *i.e.* approaching the boundary of the parameter space, the objective function tends to $-\infty$. This unboundedness (from below) makes both the statistical analysis and the global optimization problems difficult. We investigate different convex formulations that were considered in the literature *i.e.* Concomitant Lasso estimation (Huber, 1981; Owen, 2007) as well as re-parameterization methods (Städler et al., 2010). In an optimization point of view, we illustrate numerical issues of Concomitant Lasso formulation and propose a modification we

coined Smoothed Concomitant Lasso, aimed at increasing numerical stabilities. Our proposal builds upon smoothing techniques à la Nesterov (2005); Beck and Teboulle (2012) of the original problem. Leveraging screening rules and the active warm start within a homotopy continuation as developed in Chapter 2, we propose an efficient and accurate solver for joint estimation that achieves similar computational cost than the one for the Lasso. This is a significant advance over previous methods that are based on generic solvers of conic programming or iterative procedure that alternates Lasso steps and noise estimation steps.

The implementations of the algorithms proposed in this thesis are available in open source in

`https://github.com/EugeneNdiaye`.

1.3 Background on Convex Analysis

Notation. We denote by $[T]$ the set $\{1, \dots, T\}$ for any non zero integer T. Our observation vector is $y \in \mathbb{R}^n$ and the design matrix $X = [X_1, \dots, X_p]^\top \in \mathbb{R}^{n \times p}$ has n observations rowwise. we write $\|\cdot\|$ to denote any norm and $\mathcal{B}_{\|\cdot\|}$ its associated unit ball and we write $\mathcal{B}(\theta, r)$ the ball with center θ and radius r ($\mathcal{B}_2$ (resp. $\mathcal{B}_\infty$) will denote the euclidean (resp. infinite) unit ball). Given a vector $\beta \in \mathbb{R}^p$, we denote by $\mathrm{supp}(\beta)$ the support of β *i.e.* the set of indices corresponding to non-zero coefficients. We denote $(t)_+ = \max(0, t)$ and $\Pi_{\mathcal{C}}(\cdot)$ the projection operator over a closed convex set $\mathcal{C}$. The interior (resp. boundary) of a set C is denoted $\mathrm{int}C$ (resp. $\mathrm{bd}C$). The soft-thresholding operator ST_τ (at level $\tau \geqslant 0$) is defined for any $x \in \mathbb{R}^d$ by $[\mathrm{ST}_\tau(x)]_j = \mathrm{sign}(x_j)(|x_j| - \tau)_+$.

Definitions and Basic Convexity Properties. We recall some elements of convex analysis used in the derivation and analysis of the algorithms proposed in this thesis. The notions we recall here are from (Hiriart-Urruty and Lemaréchal, 2012) and (Rockafellar, 1997).

Let $P : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ be a function non identically equal to $+\infty$, its (effective) domain is the nonempty set

$$\mathrm{dom}(P) = \{z \in \mathbb{R}^q : P(z) < +\infty\} \ .$$

The epigraph of P is defined as

$$\mathrm{epi}P := \{(z, r) \in \mathbb{R}^q \times \mathbb{R} : P(z) \leqslant r\} \ . \tag{1.11}$$

Definition 3 (Convexity)**.** *A set C is said to be convex if $\alpha z + (1-\alpha)z'$ is in C whenever z and z' are in C and for α in $]0, 1[$.*

A function $P : \mathbb{R}^q \mapsto \mathbb{R} \cup \{+\infty\}$ is said to be convex if for all $z, z' \in \mathrm{dom}(P)$ and $\alpha \in (0, 1)$,

$$P(\alpha z + (1-\alpha)z') \leqslant \alpha P(z) + (1-\alpha)P(z') \ .$$

We recall that a function P is convex if and only if its epigraph $\mathrm{epi}P$ is a convex set (which can be seen as a geometric definition of convex function).

The convex functions for optimization problems that we will encounter are assumed to be proper in the sense that they are not identically equal to $+\infty$ and do not take the value $-\infty$. We also assume that they are closed *i.e.* their epigraph are closed which is equivalent to lower semi-continuity These properties are of interest because they allows to guarantee existence of minimizers (Peypouquet, 2015, Proposition 2.19).

For any convex set $C \subset \mathbb{R}^d$ the (convex) indicator function ι_C is defined by

$$\iota_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{otherwise .} \end{cases} .$$

Definition 4 (Subdifferential)**.** *The subdifferential of a convex function P at x is defined as*

$$\partial P(x) = \{v \in \mathbb{R}^q : P(z) \geqslant P(x) + \langle v, z - x\rangle, \forall z \in \mathbb{R}^q\} \ . \tag{1.12}$$

As a slight abuse of notation, we will write $\partial P(x)$ for any vector in the subdifferential of P at x.

Proposition 1. *For a convex function P, the subdifferential $\partial P(x)$ is a non-empty closed convex set for any x in* $\mathrm{dom}(P)$.

Definition 5 (Strong Convexity)**.** *A function P is μ-strongly convex for $\mu \geqslant 0$ if*

$$\forall x, y \in \mathrm{dom}(P), \quad P(x) + \langle \partial P(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \leqslant P(y). \tag{1.13}$$

Definition 6 (Smoothness)**.** *A continuously differentiable function P is ν-smooth for $\nu \geqslant 0$ if*

$$\forall x, y \in \mathrm{dom}(P), \quad P(y) \leqslant P(x) + \langle \nabla P(x), y - x\rangle + \frac{\nu}{2}\|y - x\|^2 . \tag{1.14}$$

Let us now introduce the notion of support function and polarity following (Rockafellar, 1997, Part 3). They will be used for deriving a concise theory for screening rule in Chapter 2.

Definition 7 (Support Function)**.** *Let C be a subset of $\mathbb{R}^q$. The support function of C is defined as*

$$\mathcal{S}_C : \mathbb{R}^q \longrightarrow [-\infty, +\infty] : \ x \mapsto \sup_{c \in C} \langle c, x\rangle \ . \tag{1.15}$$

Proposition 2 (Polarity)**.** *A support function is closed and sublinear. Furthermore, if C is closed, convex and contains 0, then $\mathcal{S}_C$ is a gauge* i.e. *a non-negative positively homogeneous convex function that vanishes at 0. We define its polar function as:*

$$\mathcal{S}_C^\circ(x^*) := \sup_{x \neq 0} \frac{\langle x^*, x\rangle}{\mathcal{S}_C(x)} = \sup_{\mathcal{S}_C(x)=1} \langle x^*, x\rangle = \sup_{\mathcal{S}_C(x) \leqslant 1} \langle x^*, x\rangle \ . \tag{1.16}$$

The function $\mathcal{S}_C^\circ$ is also a gauge function and we have the polar inequality:

$$\langle x^*, x\rangle \leqslant \mathcal{S}_C^\circ(x^*)\mathcal{S}_C(x) \quad \forall x^* \in \mathrm{dom}\mathcal{S}_C^\circ, x \in \mathrm{dom}\mathcal{S}_C \ . \tag{1.17}$$

Note that a gauge function is a norm when it is finite everywhere, symmetric and non zero except at the origin. Hence Equation (1.17) generalizes Cauchy-Schwartz and Hölder inequality.

Fenchel's Duality Theorem. We recall Fenchel duality in optimization that will be extensively use in this manuscript.

Definition 8 (Fenchel-Legendre Transform)**.** *For a function $P : \mathbb{R}^q \to [-\infty, +\infty]$, its conjugate P, is the function defined as*

$$P^* : \ \mathbb{R}^q \longrightarrow [-\infty, +\infty] : x^* \mapsto \sup_{x \in \mathbb{R}^q} \langle x^*, x\rangle - P(x) \ . \tag{1.18}$$

The conjugacy operation $P \mapsto P^*$ is called the Fenchel-Legendre transform.

Theorem 2 (Fenchel Duality see Rockafellar (1997, Theorem 31.3))**.** *Let f (resp. Ω) be closed proper and convex functions on $\mathbb{R}^n$ (resp. $\mathbb{R}^p$), and let X a matrix in $\mathbb{R}^{n \times p}$ and $\lambda > 0$. We define the primal and dual optimization problem as*

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \underbrace{f(X\beta) + \lambda\Omega(\beta)}_{P_\lambda(\beta)} \quad \text{(Primal)}. \tag{1.19}$$

$$\hat{\theta}^{(\lambda)} \in \arg\max_{\theta \in \mathbb{R}^q} \underbrace{-f^*(-\lambda\theta) - \lambda\Omega^*(X^\top\theta)}_{D_\lambda(\theta)} \quad \text{(Dual)}. \tag{1.20}$$

Strong duality holds i.e. $P_\lambda(\hat{\beta}^{(\lambda)}) = D_\lambda(\hat{\theta}^{(\lambda)}$ *if and only if*

$$-\lambda\hat{\theta}^{(\lambda)} \in \partial f(X\hat{\beta}^{(\lambda)}) \Longleftrightarrow X\hat{\beta}^{(\lambda)} \in \partial f^*(-\lambda\hat{\theta}^{(\lambda)}), \tag{1.21}$$

$$X^\top\hat{\theta}^{(\lambda)} \in \partial\Omega(\hat{\beta}^{(\lambda)}) \Longleftrightarrow \hat{\beta}^{(\lambda)} \in \partial\Omega^*(X^\top\hat{\theta}^{(\lambda)}). \tag{1.22}$$

Optimality conditions in Equation (1.21) and (1.21) are called Karush-Kuhn-Tucker (KKT) conditions.

Definition 9 (Duality Gap)**.** *For any primal/dual feasible pair of vector* $(\beta, \theta) \in \mathrm{dom}P_\lambda \times \mathrm{dom}D_\lambda$*, the duality gap is defined as the difference between the primal and dual objectives:*

$$\begin{aligned}\mathrm{Gap}_\lambda(\beta, \theta) &:= P_\lambda(\beta) - D_\lambda(\theta) \\ &= f(X\beta) + f^*(-\lambda\theta) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \ .\end{aligned}$$

For any such (β, θ)*, the weak duality holds* i.e. $P_\lambda(\beta) \geqslant D_\lambda(\theta)$*. This implies that*

$$P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \mathrm{Gap}_\lambda(\beta, \theta) \ .$$

Remark 1. *In this manuscript, the duality gap will be used as an optimality certificate or as a algorithmic stopping criterion for solving* (1.19)*.*

We also recall from (Hiriart-Urruty and Lemaréchal, 1993, Theorem 4.2.2, p. 83)

Proposition 3. *P is μ-strongly convex if and only if P^* is $1/\mu$-smooth.*

We introduce below the Fenchel-Young inequality and two other variants that exploit the notions of strong convexity and smoothness.

Lemma 1 (Fenchel-Young Inequalities)**.** *Let P be a convex function. For all x in* $\mathrm{dom}(P)$*, and $x^* \in \mathrm{dom}(P^*)$, we have*

$$P(x) + P^*(x^*) \geqslant \langle x^*, x\rangle \ , \tag{1.23}$$

with equalities if and only if $x^* \in \partial P(x)$. Moreover, if P is μ-strongly convex (resp. ν-smooth) Inequality (1.24) *(resp. Inequality* (1.25)*) holds:*

$$P(x) + P^*(x^*) \geqslant \langle x^*, x\rangle + \frac{\mu}{2}\left\|x - \partial P^*(x^*)\right\|^2 \ , \tag{1.24}$$

$$P(x) + P^*(x^*) \leqslant \langle x^*, x\rangle + \frac{\nu}{2}\left\|x - \nabla P^*(x^*)\right\|^2 \ . \tag{1.25}$$

Remark 2. *The Inequalities* (1.24)*,* (1.25) *are less common in the literature. They are refinements of classical Fenchel-Young inequality* (1.23) *and will be useful in establishing some technical inequalities.*

Proof. We have from the μ-strong convexity and the equality case in Fenchel-Young inequality

$$P(z) + P^*(\partial P(z)) = \langle \partial P(z), z\rangle$$

$$-P^*(\partial P(z)) + \langle \partial P(z), x\rangle + \frac{\mu}{2}\left\|x - z\right\|^2 = P(z) + \langle \nabla P(z), x - z\rangle + \frac{\mu}{2}\left\|x - z\right\|^2 \leqslant P(x).$$

We conclude by applying the inequality at $z = \partial P^*(x^*)$ and remark that $\partial P(z) = x^*$. The same proof holds for the upper bound (1.25). □

Lemma 2. *The function P is bounded from below if and only if* $\mathrm{dom}(P^*)$ *contains* 0*.*

Proof. We have $P^*(0) = -\inf_z P(z) < +\infty$. hence the result. □

We will assume the following technical condition everywhere: both f and Ω are bounded from below *i.e.* both $\mathrm{dom}f^*$ and $\mathrm{dom}\Omega^*$ contains the origin 0. Hence the support functions considered in this thesis simplify to gauge functions.

First Order Optimality Conditions

Proposition 4 (Fermat's Rule)**.** *(see (Bauschke and Combettes, 2011, Proposition 26.1) for a more general result) For any convex function $P : \mathbb{R}^q \to \mathbb{R}$, we have*

$$z^\star \in \arg\min_{z \in \mathbb{R}^q} P(z) \Longleftrightarrow 0 \in \partial P(z^\star) \ . \tag{1.26}$$

Proposition 5. *Let P be a convex and differentiable function, C be a closed and convex set.*

$$z^\star \in \arg\min_{z \in C} P(z) \text{ if and only if } \langle \nabla P(z), z - z^\star \rangle \geqslant 0, \ \forall z \in C \ . \tag{1.27}$$

Chapter 2

Safe Screening Rules

The computational burden of solving high dimensional regularized regression problem has led to a vast literature on improving algorithmic solvers in the last two decades. With the increasing popularity of ℓ_1-type regularization ranging from the Lasso (Tibshirani, 1996) to hierarchical sparse structure (Sprechmann et al., 2011), many algorithmic methods have emerged to solve the associated optimization problems (Efron et al., 2004; Koh et al., 2007; Friedman et al., 2010b; Bach et al., 2012). Our main objective in this work is to propose a technique that can speed-up any iterative solver for such learning tasks by reducing the dimensionality thanks to a safe elimination of unimportant variables.

The *safe rules* introduced by El Ghaoui et al. (2012) for supervised learning problems with sparse ℓ_1 regularization, is a set of rules allowing to eliminate features whose associated coefficients are guaranteed to be absent in the model parameter after solving the learning problem. It exploits the known sparsity of the solution by discarding features prior to starting a solver. The main building bloc are based on Fenchel duality and first order optimality conditions. Let us consider the example of the Lasso estimator $\hat{\beta}^{(\lambda)}$ to illustrate the method. Given a tuning parameter $\lambda > 0$, controlling the trade-off between data fidelity and sparsity of the solutions, it is defined as any solution of the (primal) optimization problem

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad .$$

Denoting $\Delta_X = \left\{\theta \in \mathbb{R}^n : \ |x_j^\top \theta| \leqslant 1, \forall j \in [p]\right\}$ the dual feasible set, a dual formulation of the Lasso reads

$$\hat{\theta}^{(\lambda)} = \arg\max_{\theta \in \Delta_X} \frac{1}{2} \|y\|_2^2 - \frac{\lambda^2}{2} \left\|\theta - \frac{y}{\lambda}\right\|_2^2 \quad .$$

From the Karush-Kuhn-Tucker (KKT) conditions, we have the relation

$$\forall j \in [p], \ x_j^\top \hat{\theta}^{(\lambda)} \in \begin{cases} \{\mathrm{sign}(\hat{\beta}_j^{(\lambda)})\} & \text{if} \quad \hat{\beta}_j^{(\lambda)} \neq 0, \\ [-1, 1] & \text{if} \quad \hat{\beta}_j^{(\lambda)} = 0. \end{cases}$$

This leads to the *screening rule*: $\forall j \in [p], \ |x_j^\top \hat{\theta}^{(\lambda)}| < 1 \implies \hat{\beta}_j^{(\lambda)} = 0$. It provides a correlation based screening rule that guarantee identification of irrelevant features. Such techniques are referred to in the literature as *safe rules* when they screen out coefficients guaranteed to be zero in the targeted optimal solution. Zeroing those coefficients allows to focus exclusively on the non-zero ones (likely to represent signal) and helps reducing the computational burden. Similar strategies have been used as data preprocessing before application of statistical methods (Fan and Lv, 2008). However they were not necessarily related to an optimization problem. It is worth noting that similar preprocessing steps known as *facial reduction* are used for accelerating the linear programming solvers (Markowitz, 1956; Brearley et al., 1975) and conic programming (Borwein

and Wolkowicz, 1981), we refer to (Mészáros and Suhl, 2003; Drusvyatskiy and Wolkowicz, 2017) for recent reviews. Another application can also be found in (Michelot, 1986) for projecting onto the simplex and ℓ_1 ball in (Condat, 2016). A noticeable difference between these approaches and the safe rules lies in the fact that the latter remove the variables only if they can be guaranteed to be inactive at the optimum.

The safe screening rules have been improved (Xiang et al., 2011) and extended to several statistical learning tasks for discarding non-influential observations and/or features in optimization problem including support vector machines (Ogawa et al., 2013; Shibagaki et al., 2016), logistic regression (Wang et al., 2014), constrained convex problems such as minimum enclosing ball (Raj et al., 2016) etc. To improve the screening performance, we can rely on the information provided by the computations done for a previous regularization parameter as in homotopy/continuation methods. This scenario is particularly relevant in machine learning where one computes solutions over a grid of regularization parameters, so as to select the best one, *e.g.* by cross-validation. Another interesting strategy is the *dynamic safe rules* introduced by Bonnefoy et al. (2015, 2014) who opened a promising venue by performing variable screening not only before the algorithm starts, but also along the iterations. It increases the number of variable eliminated as the algorithm progresses towards the optimal solution. However, the derivation of those rules strongly depends on each specific problem formulation and one can ask if there is an underlying common structure that can be exploited.

This chapter contains a synthesis and a unified presentation of the (safe) screening rules introduced so far in machine learning. We show that it relies on the "subdifferential separation" which is a natural property of convex functions. We put forward the *Gap Safe Rules* introduced for the Lasso in (Fercoq et al., 2015) that relies on duality gap computations and show how it extents to broad class of optimization problems with the following benefits:

— Gap Safe rules are easy to insert in existing solvers,
— they are proved to be safe and unify sequential and dynamic rules,
— they lead to improved speed-ups in practice *w.r.t.* previously known safe rules.

Our contribution in the literature consists in the "Gap Safe Screening Rules" series that was published in the following machine learning review:

Authors: E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon.

- "Gap Safe Screening Rules for Sparse Multi-task and Multi-class Models".
  Advances in Neural Information Processing Systems, 811-819, 2015.

- "Gap Safe Screening Rules for Sparse-Group Lasso".
  Advances in Neural Information Processing Systems, 388-396, 2016.

- "Gap Safe Screening Rules for Sparsity Enforcing Penalties".
  The Journal of Machine Learning Research 18 (1), 4671-4703, 2017.

Notation. The parameter to recover is a vector $\beta = (\beta_1, \dots, \beta_p)^\top$ admitting a group structure. A group of features is a subset $g \subset [p]$ and $|g|$ is its cardinality. The set of groups is denoted by $\mathcal{G}$ and we focus only on non-overlapping groups that form a partition of the set $[p]$. We denote by β_g the vector in $\mathbb{R}^{|g|}$ which is the restriction of β to the indices in g. We write $[\beta_g]_j$ the j-th coordinate of β_g and simply β_j if there is no ambiguity. We also use the notation $X_g \in \mathbb{R}^{n \times n_g}$ to refer to the sub-matrix of X assembled from the columns with indices $j \in g$ and X_j when the groups are a single feature, *i.e.* when $g = \{j\}$. Similar notation are used for the observations and the group of samples will be denoted $\mathcal{I}$.

Framework. We consider optimization problem involving a smooth function plus a separable regularization to enforce specific structure in the solution. Such a formulation often arises in statistical learning in a context of empirical risk minimization with inductive bias. Popular instantiations of this formulation are detailed in Section 2.2.4.

Under the classical Fenchel duality, we have the Primal/Dual formulation:

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{I}} f_i(x_i^\top \beta) + \lambda \sum_{g \in \mathcal{G}} \Omega_g(\beta_g) =: P_\lambda(\beta) \qquad \text{(Primal)}, \tag{2.1}$$

$$\hat{\theta}^{(\lambda)} \in \arg\max_{\theta \in \mathbb{R}^n} -\sum_{i \in \mathcal{I}} f_i^*(-\lambda\theta_i) - \lambda \sum_{g \in \mathcal{G}} \Omega_g^*(X_g^\top \theta) =: D_\lambda(\theta) \qquad \text{(Dual)}. \tag{2.2}$$

Moreover, we have the optimality conditions connecting primal and dual solutions:

$$\forall i \in \mathcal{I}, \quad -\lambda\hat{\theta}_i^{(\lambda)} \in \partial f_i(x_i^\top \hat{\beta}^{(\lambda)}) \Longleftrightarrow x_i^\top \hat{\beta}^{(\lambda)} \in \partial f_i^*(-\lambda\hat{\theta}_i^{(\lambda)}), \tag{2.3}$$

$$\forall g \in \mathcal{G}, \quad X_g^\top \hat{\theta}^{(\lambda)} \in \partial\Omega_g(\hat{\beta}_g^{(\lambda)}) \Longleftrightarrow \hat{\beta}_g^{(\lambda)} \in \partial\Omega_g^*(X_g^\top \hat{\theta}^{(\lambda)}). \tag{2.4}$$

We propose in this chapter a synthesis and a unified presentation of the screening rules for identifying active structure in convex optimization problem (2.1) which contains a broader class of learning problems under mild conditions. We require a loss with a Lipschitz continuous gradient (which allows to construct a safe region based on the duality gap) plus a regularization that is block-wise separable (which allows to decompose the problem into independent components). We present several strategy to take a large benefit from this rule in order to speed up the execution time of any iterative algorithms, specially the proximal (block) coordinate descent.

2.1 Non-Smoothness and Active Set Identification

Now we introduce an important lemma that captures a natural property of convex functions that allows us to obtain a simple and unified presentation of the screening rules introduced recently in the literature.

Lemma 3 (Separation of Subdifferentials)**.** *Let P be a convex function and $z \in \mathrm{dom}P$ such that $\mathrm{int}\partial P(z) \neq \varnothing$. Then we have $\mathrm{int}\partial P(z) \cap \partial P(z') = \varnothing$ for all $z \neq z'$.*

Proof. Let z' such that it exists g in $\mathrm{int}\partial P(z) \cap \partial P(z')$. Now g in the open set $\mathrm{int}\partial P(z)$ implies that it exists $\alpha > 0$ such that $g_\alpha := g + \alpha(z' - z) \in \partial P(z)$. Then we have

$$P(z) \geqslant P(z') + \langle g, z - z' \rangle \geqslant P(z) + \langle g_\alpha, z' - z \rangle + \langle g, z - z' \rangle = P(z) + \alpha \left\| z' - z \right\|_2^2 \ ,$$

where the first (resp. the second) inequality comes from $g \in \partial P(z')$ (resp. $g_\alpha \in \partial P(z)$). Hence $z' = z$. □

By applying Lemma 3 to the problem 2.1, we obtain the results that allows to identifies parts of the optimal solutions. In the rest of this chapter, for any group g in $\mathcal{G}$, $\beta_g^\star$ is any vector in $\mathbb{R}^{|g|}$ such that $\mathrm{int}\partial\Omega_g(\beta_g^\star)$ is non empty.

Proposition 6 (Feature-wise Screening Rule)**.**
For any group g in $\mathcal{G}$, if $X_g^\top \hat{\theta}^{(\lambda)}$ belongs to $\mathrm{int}\partial\Omega_g(\beta_g^\star)$, then $\hat{\beta}_g^{(\lambda)}$ is equal to $\beta_g^\star$.

Proof. From the optimality conditions Equation (2.4), we have for all group $g \in \mathcal{G}$, $X_g^\top \hat{\theta}^{(\lambda)} \in \partial\Omega_g(\hat{\beta}_g^{(\lambda)})$. Then from Lemma 3 we deduce that if $\hat{\beta}_g^{(\lambda)} \neq \beta_g^\star$, then $X_g^\top \hat{\theta}^{(\lambda)} \notin \mathrm{int}\partial\Omega_g(\beta_g^\star)$. Hence we conclude by contrapositive. □

This relation means that the g-th group can be discarded in the problem *i.e.* $\hat{\beta}_g^{(\lambda)}$ is identified to be equal to $\beta_g^\star$, whenever $X_g^\top \hat{\theta}^{(\lambda)}$ belongs to $\mathrm{int}\partial\Omega_g(\beta_g^\star)$. However, since $\hat{\theta}^{(\lambda)}$ is **unknown**, this rule is of limited use. Fortunately, it is often possible to construct a set $\mathcal{R}^* \subset \mathbb{R}^n$, called a *safe region*, that contains $\hat{\theta}^{(\lambda)}$. This observation leads to the following result.

Proposition 7 (Safe Feature-wise Screening Rule)**.** *Let $\mathcal{R}^*$ be a (dual) safe region* i.e. *it contains the (dual) optimal solution $\hat{\theta}^{(\lambda)}$. If $X_g^\top \mathcal{R}^*$ is included in $\mathrm{int}\partial\Omega_g(\beta_g^\star)$ then $\hat{\beta}_g^{(\lambda)}$ is equal to $\beta_g^\star$.*

Similar results are also obtained when the regularity of the loss and the regularization are flipped *i.e.* when the loss function is considered to be the non-smooth part. Exploiting the optimality condition $x_i^\top \hat{\beta}^{(\lambda)} \in \partial f_i^*(-\lambda\hat{\theta}_i^{(\lambda)})$ and using similar reasoning we have:

Proposition 8 (Sample-wise Screening Rule)**.** *Let $\theta_i^\star$ be a vector such that $\mathrm{int}\partial f_i^*(\theta_i^\star)$ is non empty. If $x_i^\top \hat{\beta}^{(\lambda)}$ is in $\mathrm{int}\partial f_i^*(\theta_i^\star)$, then $-\lambda\hat{\theta}_i^{(\lambda)}$ is equal to $\theta_i^\star$.*

Proposition 9 (Safe Sample-wise Screening Rule)**.** *Let $\theta_i^\star$ be a vector such that $\mathrm{int}\partial f_i^*(\theta_i^\star)$ is non empty and $\mathcal{R}$ be a (primal) safe region* i.e. *it contains the (primal) optimal solution $\hat{\beta}^{(\lambda)}$. If $X_i^\top \mathcal{R}$ is included in $\mathrm{int}\partial f_i^*(\theta_i^\star)$ then $-\lambda\hat{\theta}_i^{(\lambda)}$ is equal to $\theta_i^\star$.*

Often, we will restrict the discussions on the primal formulation where the loss is smooth and the regularization is non-smooth and separable since the same properties can be recovered in the dual by symmetry.

Remark 3. *The separability is required only for the non-smooth part, it helps to decompose the problem into independent groups and allows to run (proximal) coordinate descent algorithms which is known to be very efficient in large scale problems. The case where the non-smooth part us not separable, which includes important regularization functions such as total variation, overlapping Group Lasso, Sorted ℓ_1-norm, ℓ_∞-norm etc, is not covered in this work.*

Since the subdifferential $\partial\Omega_g(\beta_g^\star)$ is a closed convex set, the screening test «$X_g^\top \mathcal{R}^* \subset \mathrm{int}\partial\Omega_g(\beta_g^\star)$»can be evaluated computationally thanks to the following lemma that allows to check whether a given point c belongs to the interiors of a closed convex set C.

Lemma 4 (Hiriart-Urruty and Lemaréchal (2012, Theorem C-2.2.3))**.** *Let S be a subset of $\mathbb{R}^n$ and C be a nonempty closed convex set. Then we have $\sup_{s\in S}\langle s, d\rangle < \mathcal{S}_C(d)$ for all d in $\mathrm{bd}\mathcal{B}$ if and only if $S \subset \mathrm{int}C$.*

Proof. For $S \subset \mathrm{int}C$, it exists $\epsilon > 0$ such that $S + \epsilon d \in C$ for all d in the unit sphere $\mathrm{bd}\mathcal{B}$. Hence by Definition 7 of the support function, we have $\mathcal{S}_C(d) \geqslant \langle s + \epsilon d, d\rangle$ for all s in S. Hence $\mathcal{S}_C(d) \geqslant \sup_{s\in S}\langle s, d\rangle + \epsilon \|d\|^2$.

Reciprocally, let $S \subset \mathbb{R}^n$ such that $\sup_{s\in S}\langle s, d\rangle < \mathcal{S}_C(d)$ for any $d \in \mathrm{bd}\mathcal{B}$. Then for any $s \in S$, $\langle s, d\rangle < \mathcal{S}_C(d)$. Now Defining $\epsilon := \inf\{\mathcal{S}_C(d) - \sup_{s\in S}\langle s, d\rangle : d \in \mathrm{bd}\mathcal{B}\} > 0$, we have $\langle s, d\rangle + \epsilon \leqslant \mathcal{S}_C(d)$ for all $d \in \mathrm{bd}\mathcal{B}$, thus taking any u such that $\|u\| < \epsilon$, we have

$$\langle s + u, d\rangle = \langle s, d\rangle + \langle u, d\rangle \leqslant \langle s, d\rangle + \epsilon \leqslant \mathcal{S}_C(d)$$

which implies $s + u \in C$ hence $s \in \mathrm{int}C$. □

By applying the Lemma 4 to the (closed convex) sets $C = \partial\Omega_g(\beta_g^\star)$ and $S = X_g^\top \mathcal{R}^*$, the screening rule can be performed by checking if $\max_{\theta\in\mathcal{R}^*}\langle X_g^\top\theta, d\rangle < \mathcal{S}_{\partial\Omega_g(\beta_g^\star)}(d)$ for all d in the unit sphere $\mathrm{bd}\mathcal{B}$ which implies that $\hat{\beta}_g^{(\lambda)}$ is equal to $\beta_g^\star$. By denoting

$$\Omega_g^\circ(X_g^\top\theta, \beta_g^\star) := \mathcal{S}^\circ_{\partial\Omega_g(\beta_g^\star)}(X_g^\top\theta) = \sup_{d\neq 0} \frac{\langle X_g^\top\theta, d\rangle}{\mathcal{S}_{\partial\Omega_g(\beta_g^\star)}(d)} \ , \tag{2.5}$$

the computational safe screening rule can be subsumed as follow

Theorem 3. *Let $\mathcal{R}^*$ be a (dual) closed and convex set that contains the (dual) optimal solution $\hat{\theta}^{(\lambda)}$. For all group g in $\mathcal{G}$ and for any vector $\beta_g^\star$ such that $\mathrm{int}\partial\Omega_g(\beta_g^\star)$ is non empty, the screening test reads:*

$$\textbf{\textit{Screening Rules:}} \quad \textit{if } \Omega_g^\circ(X_g^\top \hat{\theta}^{(\lambda)}, \beta_g^\star) < 1 \textit{ then } \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ . \tag{2.6}$$

$$\textbf{\textit{Safe Screening Rules:}} \quad \textit{if } \max_{\theta \in \mathcal{R}^*} \Omega_g^\circ(X_g^\top \theta, \beta_g^\star) < 1 \textit{ then } \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ . \tag{2.7}$$

In the following, we will drop the dependency in $\beta_g^\star$ and simply note $\Omega_g^\circ(X_g^\top \theta)$ if their is no ambiguity.

The *safe screening* rule consists in removing the g-th group from the optimization process whenever the previous test is satisfied, since then $\hat{\beta}_g^{(\lambda)}$ is guaranteed to be equal to $\beta_g^\star$. Should $\mathcal{R}^*$ be small enough to screen many groups, one can observe considerable speed-ups in practice as long as the testing can be performed efficiently. Thus a natural goal is to find safe regions as narrow as possible: smaller safe regions can only increase the number of screened out variables. To have useful screening procedures one needs:

— the region $\mathcal{R}^*$ should be as small as possible and contains $\hat{\theta}^{(\lambda)}$,
— the computations needed to check if $X_g^\top \mathcal{R}^* \subset \mathrm{int}\partial\Omega_g(\beta_g^\star)$ should be cheap.

The later means that safe regions should be "simple" geometric objects, since otherwise, evaluating the test could lead to a computational burden limiting the benefits of screening.

Note that any time a safe rule is performed thanks to a safe region $\mathcal{R}^*$, one can associate a *safe active set* consisting of the features that cannot be removed yet by the safe screening test.

Safe Active Set and Converging Region

Definition 10 (Feature-wise (Safe) Active Sets)**.** *Let $\beta_g^\star$ be a vector such that $\mathrm{int}\partial\Omega_g(\beta_g^\star)$ is non empty. The set of (group) active features at $\beta_g^\star$ is defined as:*

$$\mathcal{A}^{(\lambda)} := \left\{ g \in \mathcal{G} : \ X_g^\top \hat{\theta}^{(\lambda)} \notin \mathrm{int}\partial\Omega_g(\beta_g^\star) \text{ i.e. } \Omega_g^\circ(X_g^\top \hat{\theta}^{(\lambda)}) \geqslant 1 \right\} \ . \tag{2.8}$$

Moreover, if $\mathcal{R}^*$ is a safe region, its corresponding set of (group) safe active features at $\beta_g^\star$ is defined as

$$\mathcal{A}_{\mathcal{R}^*} := \left\{ g \in \mathcal{G} : \ X_g^\top \mathcal{R}^* \not\subset \mathrm{int}\partial\Omega_g(\beta_g^\star) \text{ i.e. } \max_{\theta \in \mathcal{R}^*} \Omega_g^\circ(X_g^\top \theta) \geqslant 1 \right\} \ . \tag{2.9}$$

Their complements i.e. *the set of non active group, will be denoted $\mathcal{Z}^{(\lambda)}$ and $\mathcal{Z}_{\mathcal{R}^*}$.*

Let us now describe the notion of converging safe regions introduced in (Fercoq et al., 2015, Definition 1) that help to reach exact active set identification in a finite number of steps.

Definition 11 (Converging Safe Region)**.** *Let $(\mathcal{R}_k)_{k\in\mathbb{N}}$ be a sequence of closed convex sets containing the optimal solution $\hat{\theta}^{(\lambda)}$. It is a converging sequence of safe regions if the diameters of the sets converge to zero. The associated safe screening rules are referred to as converging.*

The following proposition asserts that after a finite number of steps, the active set is exactly identified. Such a property is sometimes referred to as finite identification of the support (Liang et al., 2017) and is summarized in the following proposition. Yet, note that the (primal) optimal support can be strictly smaller than the active set, see the case of the Lasso (Tibshirani, 2013) (where the active set is called equicorrelation set). For clarity, links between optimal support and (safe) active sets are illustrated in Figure 2.1.

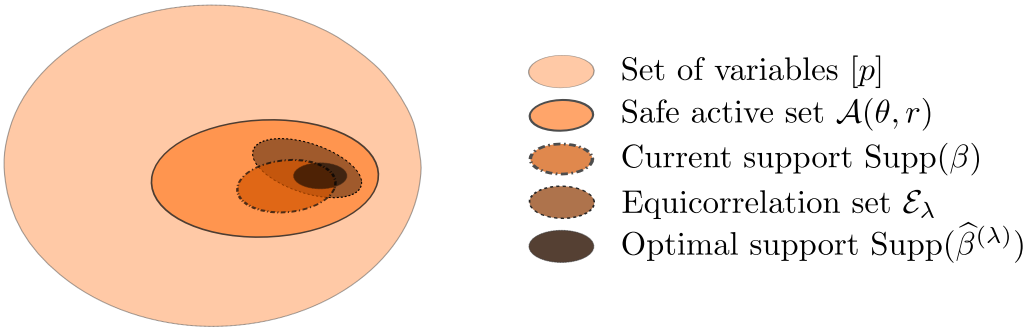


Figure 2.1 – Illustration of the inclusions between several remarkable sets: $\mathrm{supp}(\beta) \subseteq \mathcal{A}(\theta, r) \subseteq [p]$ and $\mathrm{supp}(\hat{\beta}^{(\lambda)}) \subseteq \mathcal{E}_\lambda \subseteq \mathcal{A}(\theta, r) \subseteq [p]$, where β, θ is a primal/dual pair.

Proposition 10 (Active Set Identification)**.** *Let $(\mathcal{R}_k^*)_{k\in\mathbb{N}}$ be a sequence of closed convex set containing $\hat{\theta}^{(\lambda)}$ for each k in $\mathbb{N}$. If $\mathcal{R}_k^* \longrightarrow \{\hat{\theta}^{(\lambda)}\}$, then $\mathcal{A}_{\mathcal{R}_k^*} \longrightarrow \mathcal{A}^{(\lambda)}$.*

Proof. We proceed in two times:

First we show that $\max_{\theta\in\mathcal{R}_k^*} \Omega_g^\circ(X_g^\top\theta) \longrightarrow_k \Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)})$. Indeed, for any $k \in \mathbb{N}$ and θ in $\mathcal{R}_k^*$ we have from the sublinearity and positive homogeneity of Ω_g°,

$$\Omega_g^\circ(X_g^\top\theta) \leqslant \Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)}) + \Omega_g^\circ\left(X_g^\top \frac{\theta - \hat{\theta}^{(\lambda)}}{\|\theta - \hat{\theta}^{(\lambda)}\|}\right) \|\theta - \hat{\theta}^{(\lambda)}\| \ .$$

Since $\hat{\theta}^{(\lambda)}$ in $\mathcal{R}_k^*$, then

$$\Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)}) \leqslant \max_{\theta\in\mathcal{R}_k^*} \Omega_g^\circ(X_g^\top\theta) \leqslant \Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)}) + \sup_{\|u\|=1} \Omega_g^\circ(X_g^\top u)\mathrm{diam}(\mathcal{R}_k^*) \ . \tag{2.10}$$

The conclusion follows the fact $\mathcal{R}_k^*$ is a converging sequence *i.e.* $\lim_k \mathrm{diam}(\mathcal{R}_k^*) = 0$.

Second, we proceed by double inclusion. First remark that $\mathcal{A}^{(\lambda)} = \mathcal{A}_{\mathcal{R}_\infty^*}$ where $\mathcal{R}_\infty^* := \{\hat{\theta}^{(\lambda)}\}$. So for all $k \in \mathbb{N}$, we have $\mathcal{A}^{(\lambda)} \subseteq \mathcal{A}_{\mathcal{R}_k^*}$ since $(\mathcal{A}_{\mathcal{R}_k^*})_k$ are nested sequence of sets. Reciprocally, suppose that there exists a non active group $g \in \mathcal{G}$ *i.e.* $\Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)}) < 1$ that remains in the active set $\mathcal{A}_{\mathcal{R}_k^*}$ for all iterations *i.e.* $\forall k \in \mathbb{N}$, $\max_{\theta_k\in\mathcal{R}_k^*} \Omega_g^\circ(X_g^\top\theta_k) \geqslant 1$. Since $\lim_{k\to\infty} \max_{\theta_k\in\mathcal{R}_k^*} \Omega_g^\circ(X_g^\top\theta_k) = \Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)})$, we obtain $\Omega_g^\circ(X_g^\top\hat{\theta}^{(\lambda)}) \geqslant 1$ by passing to the limit. Hence, by contradiction, there exits an integer $k_0 \in \mathbb{N}$ such that $[p]\backslash\mathcal{A}^{(\lambda)} \subseteq \mathcal{A}^c_{\mathcal{R}_k^*}$ for all $k \geqslant k_0$.

□

One can note that the rate of identification of the active set is strongly related to the rate at which the sequence of diameters $\mathrm{diam}(\mathcal{R}_k^*)$ goes to zero. We show in the next section how to construct such a converging sequence that uses the information gained during the optimization process with explicit rates.

We end this section by clarifying the differences and links between the *screening rule* and the identification of non-smooth structure of a convex function.

Definition 12 (Kink of a Function)**.** *A point z at which $\partial P(z)$ has more than one element* - i.e. *at which P is not differentiable - is called a* kink *(or corner-point) of P.*

For a convex function P defined on the real line, if the subdifferential at z is not a singleton *i.e.* P is non differentiable at z, then its interior is a non trivial interval and so is non empty. Hence

in the previous propositions one can replace "$\beta_j^\star$ such that $\mathrm{int}\partial\Omega_j(\beta_j^\star)$ is non empty" by "Ω_j is non differentiable at $\beta_j^\star$". Thus $X_j^\top\hat{\theta}^{(\lambda)} \in \mathrm{int}\partial\Omega_j(\beta_j^\star)$ implies that $\hat{\beta}_j^{(\lambda)}$ is a kink of Ω_j. Whence the kink of Ω_j coincide with the variables screened. Nevertheless, this is not always the case when the dimension of the group is larger that one. Indeed, let us take the example of $\Omega(\beta) = \|\beta\|_1 + \|\beta\|_2$ where we have only one group $[p] = \mathcal{G}$. The only knowledge that β is a kink of Ω is not sufficient to decide if $\beta = 0$ since any vector β such that it exits a coordinate $\beta_j = 0$ is also a kink of Ω.

2.2 Gap Safe Screening Rules: from Theory to Practice

Various shapes have been considered in practice for the safe region $\mathcal{R}$. Here we consider for simplicity "*sphere regions*" following the terminology introduced by El Ghaoui et al. (2012) choosing a ball $\mathcal{R}^* = \mathcal{B}(c, r)$ as a safe region.

Since the function Ω_g° is sublinear and positively homogeneous, we have:

$$\begin{aligned}\max_{\theta\in\mathcal{B}(c,r)} \Omega_g^\circ(X_g^\top\theta) &\leqslant \Omega_g^\circ(X_g^\top c) + \max_{\theta\in\mathcal{B}(c,r)} \Omega_g^\circ\left(X_g^\top\frac{(\theta-c)}{\|\theta-c\|}\right)\|\theta-c\| \\ &\leqslant \Omega_g^\circ(X_g^\top c) + r\max_{\|u\|=1}\Omega_g^\circ(X_g^\top u) \ .\end{aligned}$$

Denoting the subordinate operator associated to Ω_g°, $\Omega_g^\circ(X_g) := \max_{\|u\|=1}\Omega_g^\circ(X_g^\top u)$, we obtain the

$$\text{Safe Sphere Test:} \qquad \Omega_g^\circ(X_g^\top c) + r\Omega_g^\circ(X_g) < 1 \Longrightarrow \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ . \tag{2.11}$$

The associated *safe active set* $\mathcal{A}_{\mathcal{B}(c,r)}$, consisting of the features that cannot be removed yet by the test in Equation (2.11), is then given by

$$\mathcal{A}(c,r) := \mathcal{A}_{\mathcal{B}(c,r)} = \{g \in \mathcal{G} : \Omega_g^\circ(X_g^\top c) + r\Omega_g^\circ(X_g) \geqslant 1\} \ . \tag{2.12}$$

Note that it contains the true support of $\hat{\beta}^{(\lambda)}$.

2.2.1 Smoothness and Dual Safe Region

Finding a Radius

Theorem 4 (Gap Safe Sphere)**.** *Assuming that f_i has $1/\gamma$-Lipschitz gradient, we have*

$$\forall\beta\in\mathbb{R}^p, \forall\theta\in\Delta_X, \quad \left\|\hat{\theta}^{(\lambda)} - \theta\right\| \leqslant \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\beta,\theta)}{\gamma\lambda^2}} \ . \tag{2.13}$$

Whence the set $\mathcal{R}^* = \mathcal{B}(\theta, \sqrt{2\,\mathrm{Gap}_\lambda(\beta,\theta)/\gamma\lambda^2}$ is a safe region for any $\beta \in \mathbb{R}^p$ and $\theta \in \Delta_X$.

Proof. Remember that $\forall i \in [n]$, f_i is differentiable with a $1/\gamma$-Lipschitz gradient. As a consequence, $\forall i \in [n]$, f_i^* is γ-strongly convex (see Proposition 3) and so the dual function D_λ is $\gamma\lambda^2$-strongly concave:

$$\forall(\theta_1,\theta_2)\in\mathbb{R}^n\times\mathbb{R}^n, \quad D_\lambda(\theta_2) \leqslant D_\lambda(\theta_1) + \langle\nabla D_\lambda(\theta_1), \theta_2-\theta_1\rangle - \frac{\gamma\lambda^2}{2}\|\theta_1-\theta_2\|^2 \ .$$

Specifying the previous inequality for $\theta_1 = \hat{\theta}^{(\lambda)}, \theta_2 = \theta \in \Delta_X$, one has

$$D_\lambda(\theta) \leqslant D_\lambda(\hat{\theta}^{(\lambda)}) + \langle\nabla D_\lambda(\hat{\theta}^{(\lambda)}), \theta-\hat{\theta}^{(\lambda)}\rangle - \frac{\gamma\lambda^2}{2}\|\hat{\theta}^{(\lambda)}-\theta\|^2 \ .$$

By definition, $\hat{\theta}^{(\lambda)}$ maximizes D_λ on Δ_X, so, $\langle \nabla D_\lambda(\hat{\theta}^{(\lambda)}), \theta - \hat{\theta}^{(\lambda)} \rangle \leqslant 0$. This implies

$$D_\lambda(\theta) \leqslant D_\lambda(\hat{\theta}^{(\lambda)}) - \frac{\gamma\lambda^2}{2}\|\hat{\theta}^{(\lambda)} - \theta\|^2.$$

By weak duality, we have $\forall \beta \in \mathbb{R}^p$, $D_\lambda(\hat{\theta}^{(\lambda)}) \leqslant P_\lambda(\beta)$, hence $\forall \beta \in \mathbb{R}^p$ and $\forall \theta \in \Delta_X$, $D_\lambda(\theta) \leqslant P_\lambda(\beta) - \frac{\gamma\lambda^2}{2}\|\hat{\theta}^{(\lambda)} - \theta\|^2$ and the conclusion follows. □

Remark 4. *To build a Gap Safe region as in Equation* (2.13)*, we only need strong convexity in the dual which is equivalent to smoothness of the loss function whereas the screening property* (3)*, requires group separability of the non smooth regularizer. Hence our framework of Gap Safe screening rule automatically applies for a large class of problems.*

Construction of (dual) Feasible Vector

To build a center for the safe sphere, we map a primal vector onto the dual space thanks to the gradient mapping $\nabla f(\cdot)$. However, the obtained dual vector are not necessarily feasible for the dual problem. A generic procedure consists in rescaling it so that it belongs to the dual set because the projection on the feasible set can be hard. More precisely, we want to build $\theta \in \mathbb{R}^n$ such that

$$\forall i \in \mathcal{I},\ -\lambda\theta_i \in \mathrm{dom} f_i^* \text{ and } \forall g \in \mathcal{G},\ X_g^\top \theta \in \mathrm{dom}\Omega_g^* \ . \tag{2.14}$$

Given a vector z in $\mathbb{R}^n$, the rescaled point is denoted by $\Theta(z)$ and is defined by

$$\Theta(z) := \begin{cases} z, & \text{if } s_z \leqslant 1 \ , \\ \frac{z}{s_z}, & \text{otherwise,} \end{cases} \qquad \text{where } s_z := \mathcal{S}^\circ_{\mathrm{dom}\Omega^*}(X^\top z) \ . \tag{2.15}$$

A candidate often considered for computing a dual point is the (generalized) residual term $z = -\nabla f(X\beta)/\lambda$. This choice is motivated by the primal-dual optimality (2.3): $\hat{\theta}^{(\lambda)} = -\nabla f(X\hat{\beta}^{(\lambda)})/\lambda$.

Proposition 11. *If f and Ω are bounded from below, then the dual vector $\theta := \Theta(-\nabla f(X\beta)/\lambda)$ in $\mathbb{R}^n$ satisfies the feasibility condition* (2.14)*.*

Proof. From Lemma 2 if Ω is bounded from below then $\mathrm{dom}\Omega^*$ contains 0. Since it is also closed and convex, we have $\mathcal{S}^\circ_{\mathrm{dom}\Omega^*}$ is positively homogeneous. Hence the vector

$$\theta := \Theta\left(\frac{-\nabla f(X\beta)}{\lambda}\right) = \frac{-\nabla f(X\beta)}{\max(\lambda, \mathcal{S}^\circ_{\mathrm{dom}\Omega^*}(X^\top \nabla f(X\beta)))} \ , \tag{2.16}$$

satisfies $\mathcal{S}^\circ_{\mathrm{dom}\Omega^*}(X^\top\theta) \leqslant 1$ which is equivalent to $X^\top\theta$ in $\mathrm{dom}\Omega^*$. Moreover, by denoting $\alpha = \lambda/s \in [0,1]$, we have $-\lambda\theta = \alpha\nabla f(X\beta) = \alpha\nabla f(X\beta) + (1-\alpha)0$. Since $\mathrm{dom} f^*$ is convex, it remains to show that it contains the vectors $\nabla f(X\beta)$ and 0, thus it will necessarily contains $-\lambda\theta$ by convex combination.

From Lemma 2, $0 \in \mathrm{dom} f^*$ since f is bounded from below. Moreover, the equality case in the Fenchel-Young inequality shows that $f(X\beta) + f^*(\nabla f(X\beta)) = \langle \nabla f(X\beta), X\beta \rangle < +\infty$. Hence $f^*(\nabla f(X\beta))$ is also finite. □

Remark 5 (Faster Evaluation of the Duality Gap)**.** *By definition of the active sets Definition 10, we have $\forall g \in \mathcal{A}_{\mathcal{R}*}$, $\Omega_g^\circ(X_g^\top\theta) \geqslant 1$ while $\forall g \in \mathcal{Z}_{\mathcal{R}*}, \Omega_g^\circ(X_g^\top\theta) < 1$. Furthermore, for $\beta_g^\star$ such that $\mathrm{int}\partial\Omega_g(\beta_g^\star)$ is nonempty, we assume that $\mathcal{S}^\circ_{\mathrm{dom}\Omega_g^*} = \mathcal{S}^\circ_{\partial\Omega_g(\beta_g^\star)} =: \Omega_g^\circ$ (which is true for most of the examples in this chapter). Let us denote $\alpha = \max(\lambda, \mathcal{S}^\circ_{\mathrm{dom}\Omega^*}(X^\top\nabla f(X\beta)))$, then we have from the separability of Ω*

$$\begin{aligned} \mathcal{S}^\circ_{\mathrm{dom}\Omega^*}(X^\top\nabla f(X\beta)) &= \alpha \max_{g\in\mathcal{G}} \Omega_g^\circ(X_g^\top\theta) = \alpha \max\left(\max_{g\in\mathcal{A}_{\mathcal{R}*}} \Omega_g^\circ(X_g^\top\theta), \max_{g\in\mathcal{Z}_{\mathcal{R}*}} \Omega_g^\circ(X_g^\top\theta)\right) \\ &= \alpha \max_{g\in\mathcal{A}_{\mathcal{R}*}} \Omega_g^\circ(X_g^\top\theta) = \max_{g\in\mathcal{A}_{\mathcal{R}*}} \Omega_g^\circ(X_g^\top\nabla f(X\beta)) \ . \end{aligned}$$

Hence in practice the evaluation of the dual gap is therefore $\mathcal{O}(nq)$ where q is the size of $\mathcal{A}_{\mathcal{R}*}$. In other words, using a safe screening rule also speeds up the evaluation of the stopping criterion.

2.2.2 Complexity of Active Set Identification

Dynamic safe screening rules have practical benefits since they increase the number of screened out variables as the algorithm proceeds. In this section, it is shown that Gap Safe rules allow to have sharper and sharper dual regions along the iterations, accelerating support identification. Before this, the following proposition states that if one relies on a primal converging algorithm, then the dual sequence we propose is also converging. Note that the convergence is maintained to the same primal solution when the primal solution is non-unique.

Lemma 5 (Convergence of the Dual Points)**.** *Let β_k be a current estimate of a primal solution $\hat{\beta}^{(\lambda)}$ and $\theta_k = \Theta(-\nabla f(X\beta_k)/\lambda)$ be the current estimate of $\hat{\theta}^{(\lambda)}$. Then, $\lim_{k\to+\infty} \beta_k = \hat{\beta}^{(\lambda)}$ implies $\lim_{k\to+\infty} \theta_k = \hat{\theta}^{(\lambda)}$.*

Proof. Let $\alpha_k = \max(\lambda, \mathcal{S}^{\circ}_{\mathrm{dom}\Omega*}(X^\top \nabla f(X\beta_k)))$, we have:

$$\begin{aligned}\left\|\theta_k - \hat{\theta}^{(\lambda)}\right\|_2 &= \left\|\frac{\nabla f(X\hat{\beta}^{(\lambda)})}{\lambda} - \frac{\nabla f(X\beta_k)}{\alpha_k}\right\|_2 \\ &\leqslant \left|\frac{1}{\lambda} - \frac{1}{\alpha_k}\right| \|\nabla f(X\beta_k)\|_2 + \frac{1}{\lambda}\left\|\nabla f(X\hat{\beta}^{(\lambda)}) - \nabla f(X\beta_k)\right\|_2.\end{aligned}$$

If $\beta_k \to \hat{\beta}^{(\lambda)}$, then $\alpha_k \to \max(\lambda, \mathcal{S}^{\circ}_{\mathrm{dom}\Omega*}(X^\top \nabla f(X\hat{\beta}^{(\lambda)}))) = \max(\lambda, \lambda\mathcal{S}^{\circ}_{\mathrm{dom}\Omega*}(X^\top\hat{\theta}^{(\lambda)})) = \lambda$ since $\nabla f(X\hat{\beta}^{(\lambda)}) = -\lambda\hat{\theta}^{(\lambda)}$ thanks to the optimality condition (2.3) and since $\hat{\theta}^{(\lambda)}$ is feasible, then $\mathcal{S}^{\circ}_{\mathrm{dom}\Omega*}(X^\top\hat{\theta}^{(\lambda)}) \leqslant 1$. Hence, both terms in the previous inequality converge to zero. □

When $\theta_k = \Theta(-\nabla f(X\beta_k)/\lambda)$, Lemma 5 the sequence of radius $r_k = (2\,\mathrm{Gap}_\lambda(\beta_k, \theta_k)/(\gamma\lambda^2))^{1/2}$ converges to 0 with k by strong duality, hence the sequence $\mathcal{B}(\theta_k, r_k)$ converges to $\{\hat{\theta}^{(\lambda)}\}$. Hence we deduce the following proposition.

Proposition 12. *The Gap Safe Sphere is a converging safe region.*

Following the results and proofs in (Dünner et al., 2016), we present their primal/dual bound on the optimality certificates. This result is important for deriving the complexity of active set identification that is algorithmic independent. The next lemma is just a slight modification that take into account the dual rescaling.

Lemma 6. *Let f be ν_f-smooth and Ω be μ_Ω-strongly convex. We assume that the vectors θ, u are chosen as $\theta = -\nabla f(X\beta)/\alpha$ with α such that for all group g in $\mathcal{G}$, $X_g^\top\theta \in \mathrm{dom}\Omega_g^*$ and $u \in \partial\Omega^*(X^\top\theta)$. Then for any s in $[0,1]$, we have:*

$$P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \geqslant s(\mathrm{Gap}_\lambda(\beta,\theta) + \Delta_\alpha) + s^2\left[\frac{(1-s)\mu_\Omega}{s}\|\beta - u\|^2 - \frac{\nu_f}{2}\|X(u-\beta)\|^2\right]$$

where $\Delta_\alpha = \left(\frac{\alpha}{\lambda} - 1\right)[f(X\beta) + f^*(-\lambda\theta)] + \langle Xu, -\lambda\theta - \nabla f(X\beta)\rangle - \frac{\alpha\nu_f}{2\lambda}\|X\beta - \nabla f^*(-\lambda\theta)\|^2$.

Proof. By optimality of $\hat{\beta}^{(\lambda)}$, we have

$$\begin{aligned}P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) &\geqslant P_\lambda(\beta) - \min_{s\in[0,1]} P_\lambda(\beta + s(u-\beta)) \text{ for } u \in \partial\Omega^*(X^\top\theta) \\ &\geqslant \max_{s\in[0,1]} f(X\beta) + \lambda\Omega(\beta) - f(X(\beta + s(u-\beta))) - \lambda\Omega(\beta + s(u-\beta)) \\ &\geqslant f(X\beta) - f(X(\beta + s(u-\beta))) + \lambda(\Omega(\beta) - \Omega(\beta + s(u-\beta))) \ \forall s \in [0,1].\end{aligned}$$

By applying the smoothness (resp. strong convexity) inequality of f (resp. Ω), we have

$$P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \geqslant s\Gamma + s^2 \left[\frac{(1-s)\mu_\Omega}{s} \|\beta - u\|^2 - \frac{\nu_f}{2} \|X(u-\beta)\|^2 \right],$$

$$\text{where } \Gamma := \lambda[\Omega(\beta) - \Omega(u)] - \langle \nabla f(X\beta), X(u-\beta) \rangle. \tag{2.17}$$

Since $u \in \partial\Omega^*(X^\top\theta)$ is equivalent to $\Omega(u) + \Omega^*(X^\top\theta) = \langle u, X^\top\theta \rangle$ and the dual vector is given by $\theta = -\nabla f(X\beta)/\alpha$, then

$$\begin{aligned}
\Gamma &= \lambda[\Omega(\beta) + \Omega^*(X^\top\theta)] + \langle Xu, -\lambda\theta - \nabla f(X\beta) \rangle + \frac{\alpha}{\lambda} \langle -\lambda\theta, X\beta \rangle \\
&\geqslant \lambda[\Omega(\beta) + \Omega^*(X^\top\theta)] + \langle Xu, -\lambda\theta - \nabla f(X\beta) \rangle \\
&\qquad + \frac{\alpha}{\lambda} (f(X\beta) + f^*(-\lambda\theta) - \frac{\nu_f}{2} \|X\beta - \nabla f^*(-\lambda\theta)\|^2) \ ,
\end{aligned}$$

where the last inequality comes from the (reversed) Fenchel-Young inequality (1.25). Finally we obtain the result by simply rearranging the inequalities. □

Restricting to cases where $\mu_\Omega \neq 0$ *i.e.* Ω strongly convex, we have $\mathrm{dom}\Omega^* = \mathbb{R}^n$ and we can choose $\alpha = \lambda$ whence $\Delta_\alpha = 0$. Now choosing $s = \frac{\mu_\Omega}{\sigma_X \nu_f + \mu_\Omega}$ where σ_X is the spectral norm of the design matrix X (see also Dünner et al. (2016)), then the last term vanishes. Thus

$$\frac{\mu_\Omega}{\sigma_X \nu_f + \mu_\Omega} \mathrm{Gap}_\lambda(\beta_{k+1}, \theta_{k+1}) \leqslant P_\lambda(\beta_{k+1}) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \mathrm{Gap}_\lambda(\beta_{k+1}, \theta_{k+1}) \ . \tag{2.18}$$

Assuming that we have a linearly convergent algorithm, we have for some $\kappa > 0$:

$$\frac{\mu_\Omega}{\sigma_X \nu_f + \mu_\Omega} \mathrm{Gap}_\lambda(\beta_{k+1}, \theta_{k+1}) \leqslant (1-\kappa)^k (P_\lambda(\beta_0) - P_\lambda(\hat{\beta}^{(\lambda)})) \ . \tag{2.19}$$

Denoting $C = \frac{\sigma_X \nu_f + \mu_\Omega}{\mu_\Omega} (P_\lambda(\beta_0) - P_\lambda(\hat{\beta}^{(\lambda)}))$, we obtain the following bound on the radius of the Gap Safe Sphere:

$$\sqrt{\frac{2\,\mathrm{Gap}_\lambda(\beta_{k+1}, \theta_{k+1})}{\gamma\lambda^2}} \leqslant \sqrt{\frac{2(1-\kappa)^k C}{\gamma\lambda^2}} \tag{2.20}$$

Combining this with the inequality (2.10), we deduce that the identification of the active set occurs when $\Omega_g^\circ(X_g^\top \hat{\theta}^{(\lambda)}) + \Omega_g^\circ(X_g)\mathrm{diam}(\mathcal{R}_k^*) < 1$ for all group g in $\mathcal{Z}^{(\lambda)}$ *i.e.*

$$\mathrm{diam}(\mathcal{R}_k^*) < \frac{1 - \Omega_g^\circ(X_g^\top \hat{\theta}^{(\lambda)})}{\Omega_g^\circ(X_g)}, \quad \forall g \in \mathcal{Z}^{(\lambda)} \ . \tag{2.21}$$

Exploiting the bound on the radius of the Gap Safe Sphere, we can finally say that the identification occurs when

$$\sqrt{\frac{2(1-\kappa)^k C}{\gamma\lambda^2}} < \min_{g \in \mathcal{Z}^{(\lambda)}} \frac{1 - \Omega_g^\circ(X_g^\top \hat{\theta}^{(\lambda)})}{\Omega_g^\circ(X_g)} =: \delta_{\mathcal{Z}^{(\lambda)}}$$

Proposition 13 (Complexity of the Active Set Identification)**.** *For any linearly converging primal algorithm, the active set will be identified after at most k_0 iterations where*

$$k_0 := \begin{cases} \frac{1}{\kappa} \log \left(\frac{2C_{f,\Omega,X}}{\delta^2_{\mathcal{Z}^{(\lambda)}}} \times \frac{P_\lambda(\beta_0) - P_\lambda(\hat{\beta}^{(\lambda)})}{\gamma\lambda^2} \right) & \text{if } \gamma\lambda^2\delta^2_{\mathcal{Z}^{(\lambda)}} \leqslant 1, \\ 0 & \text{otherwise} \ , \end{cases} \tag{2.22}$$

for some nonnegative constant κ and the constant $C_{f,\Omega,X} := (\sigma_X \nu_f + \mu_\Omega)/\mu_\Omega$ depends only on the conditioning of the optimization problem.

Algorithm 1 Pathwise Algorithm with Active Warm Start

Input : $X,\ y,\ \epsilon,\ K,\ f^{\text{ce}},\ (\lambda_t)_{t\in[T-1]}$

for $t \in [T-1]$ **do**

$\beta = \beta^{(\lambda_{t-1})}$ and `// Get previous ϵ-solution`

Get an initial (safe or not) support estimator $\mathcal{S} = \mathcal{S}(\beta^{(\lambda_{t-1})})$

$\beta_{\mathcal{S}} = \text{Solver}\,(X_{\mathcal{S}},\ y,\ \beta_{\mathcal{S}},\ \epsilon,\ K,\ f^{\text{ce}},\ \lambda_t)$ `// Active warm start`

$\beta^{(\lambda_t)} = \text{Solver}\,(X,\ y,\ \beta,\ \epsilon,\ K,\ f^{\text{ce}},\ \lambda_t)$ `// Solve over all variables`

Output: $\left(\beta^{(\lambda_t)}\right)_{t\in[T-1]}$

This generalizes the existing result in activity identification in (Liang et al., 2017; Nutini et al., 2017) for separable regularization in the sense that our result is true for any primal converging algorithm not only Forward-Backward. We can also quantify the importance of the initialization for fast convergence/fast identification of the active set. Indeed, we have a logarithmic dependence on $\delta_{\mathcal{Z}^{(\lambda)}}$ and on the initial Gap Safe radius since

$$\frac{2(P_\lambda(\beta_0) - P_\lambda(\hat{\beta}^{(\lambda)}))}{\gamma\lambda^2} \leqslant \frac{2\,\text{Gap}_\lambda(\beta_0, \theta_0)}{\gamma\lambda^2} \quad .$$

Note that our reasoning easily adapts to other convergence regimes, it suffices to modify Equation (2.19) with the appropriate rate.

2.2.3 Homotopy Acceleration Strategies

When designing a supervised learning algorithm with sparsity enforcing penalties, the tuning of the parameter λ in Problem (2.1) is crucial and is usually done by cross-validation which requires evaluation over a grid of parameter values. A standard grid considered in the literature is $\lambda_t = \lambda_{\max} 10^{-\delta t/(T-1)}$ with a small δ, say $\delta = 10^{-2}$ or 10^{-3}, see for instance (Bühlmann and van de Geer, 2011)[2.12.1], the `glmnet` package (Friedman et al., 2010b) or the `scikit-learn` package (Pedregosa, 2016). The parameter δ has an important influence on the computational burden: computing time tends to increase for small λ, the primal iterates being less and less sparse, and the problem to solve more and more ill-posed. It is customary to start from the largest regularizer $\lambda_0 = \lambda_{\max}$ and then to perform iteratively the computation of $\hat{\beta}^{(\lambda_t)}$ after the one of $\hat{\beta}^{(\lambda_{t-1})}$. This leads to computing the models in the order of increasing complexity: this allows important speed-up by benefiting of warm start strategies. Here we propose a simple pathwise algorithm divided in two step:

— **Active Warm Start**: improve solver initialization by solving the problem restricted to an initial estimation of the support based on sequential information along the regularization path.

— **Dynamic Gap Safe Screening**: use the information gained during the iterations of the algorithm to obtain a smaller safe region therefore a greater elimination of inactive variables.

See below the details on the various strategies investigated. We summarize our strategy for solving the problem given by Equation (2.1) in Algorithm 1 and 2. The notation Solver $(\ldots)$ refers to any numerical solver that produces an approximation of the solution of Problem 2.1 and SolverUpdate $(\ldots)$ is the updating scheme of the current vector along the iterations [1]. We consider solvers that can use a (primal) warm start point.

We now describe the simplest safe rule strategy, which we refer to as the static strategy.

1. For our experiments we have focused on (block) coordinate descent solvers

Algorithm 2 Iterative Solver with Gap Safe Rules: Solver $(X,\ y,\ \beta,\ \epsilon,\ K,\ f^{\text{ce}},\ \lambda)$

Input : $X,\ y,\ \beta,\ \epsilon,\ K,\ f^{\text{ce}},\ \lambda$ // Warm start is authorized here through β

for $k \in [K]$ **do**

  if $k \mod f^{\text{ce}} = 1$ **then**

    Compute a dual variable $\theta = \frac{-\nabla f(X\beta)}{\max(\lambda, \mathcal{S}^{\circ}_{\text{dom}\Omega^*}(X^\top \nabla f(X\beta)))}$

    Stop if $\text{Gap}_\lambda(\beta, \theta) \leqslant \epsilon$

    $r = \sqrt{\frac{2\,\text{Gap}_\lambda(\beta,\theta)}{\gamma\lambda^2}}$ // Get Gap Safe radius as in Equation (2.13)

    $\mathcal{A} = \left\{g \in \mathcal{G} : \Omega^{\circ}_g(X_g^\top \theta) + r\Omega^{\circ}_g(X_g) \geqslant 1\right\}$ // Get Safe active set as in Equation (2.12)

  $\beta_{\mathcal{A}} = \text{SolverUpdate}\,(X_{\mathcal{A}},\ y,\ \beta_{\mathcal{A}},\ \lambda)$ // Solve on current Safe active set

Output: $\beta^{(\lambda)}$

Static Safe Rules. The first static safe rule, introduced by El Ghaoui et al. (2012) for ℓ_1 regularization, discards variables before any computation. Here, the (safe) sphere is fixed once and for all, hence the name static. The static rule reads:

$$\begin{aligned}
\text{Static sphere rule:} \quad & \text{If} \quad \Omega^{\circ}_g(X_g^\top \theta_{\max}) + r_{\max}\Omega^{\circ}_g(X_g) < 1, \quad \text{then} \quad \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ , \\
\text{Center:} \quad & \theta_{\max} := -\nabla f(X\beta_{\max})/\lambda_{\max} \ , \\
\text{Radius:} \quad & r_{\max} := \sqrt{\frac{2}{\gamma\lambda^2}\,\text{Gap}_\lambda(\beta_{\max}, \theta_{\max})} \ .
\end{aligned}$$

There is a threshold λ_{critic} such that for any λ smaller than λ_{critic} the test from the Static sphere rule can never be satisfied. This phenomenon appears clearly in the numerical experiments presented in Section 4.4. In simple cases a closed form for λ_{critic} can even be provided. For instance, in the case of the Group Lasso, (El Ghaoui et al., 2012) proposed to use $r_{\max} = \left|\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right| \|y\|_2$, and simple calculation gives:

$$\lambda_{\text{critic}} := \lambda_{\max} \times \min_{g \in \mathcal{G}} \frac{\|y\|_2\, \Omega^{\circ}_g(X_g)}{\lambda_{\max} + \|y\|_2\, \Omega^{\circ}_g(X_g) - \Omega^{\circ}_g(X_g \nabla f(X\beta_{\max}))} \ .$$

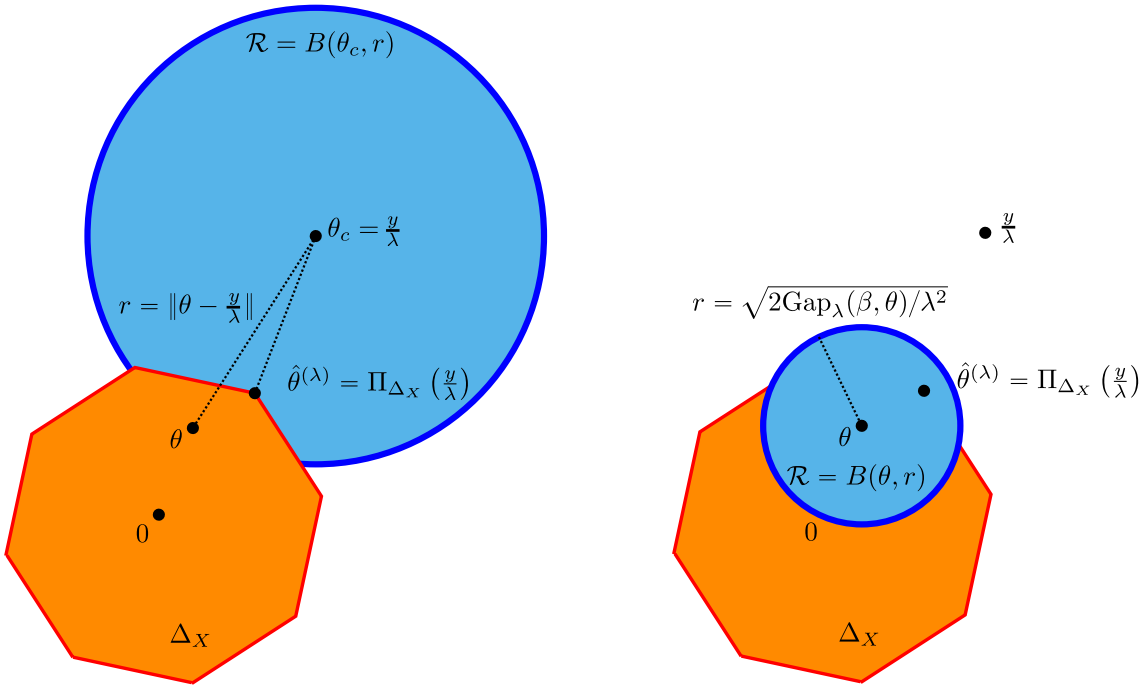
Sequential Safe Rules. Provided that the λ's are close enough along the regularization parameters, knowing an estimate of $\hat{\beta}^{(\lambda_{t-1})}$ gives a clever initialization to compute $\hat{\beta}^{(\lambda_t)}$. To initialize the solver for a new λ_t, a natural choice is to set the primal variable equal to $\beta^{(\lambda_{t-1})}$, an approximation of $\hat{\beta}^{(\lambda_{t-1})}$ output by the solver (at a prescribed precision). This popular strategy is referred to as *warm start* in the literature (Friedman et al., 2007). This leads to the sequential strategy to screen for a new λ_t:

$$\begin{aligned}
\text{Sequential sphere rule:} \quad & \text{If} \quad \Omega^{\circ}_g(X_g^\top \theta^{(\lambda_{t-1})}) + r_t\Omega^{\circ}_g(X_g) < 1, \quad \text{then} \quad \hat{\beta}_g^{(\lambda_t)} = \beta_g^\star \ , \\
\text{Center:} \quad & \theta^{(\lambda_{t-1})} := \Theta(-\nabla f(X\beta^{(\lambda_{t-1})})/\lambda_{t-1}) \ , \\
\text{Radius:} \quad & r_t := \sqrt{\frac{2}{\gamma\lambda_t^2}\,\text{Gap}_{\lambda_t}(\beta^{(\lambda_{t-1})}, \theta^{(\lambda_{t-1})})} \ .
\end{aligned}$$

Sequential screening is motivated by the idea that the duality gap growth continuously *w.r.t.* to the regularization parameter. The variation in the Gap Safe radius can be quantified by using the warm start bounds in Lemma 12. For simplicity, we describe it only when the dual loss f^* is smooth and we refer to Chapter 2 for more details and extensions.

Proposition 14 (Sequential Bounds)**.** *Assuming that f^* is ν-smooth and given a primal/dual feasible vector (β, θ), we have:*

$$r^2_{\lambda_t}(\beta, \theta) \leqslant \frac{\lambda_t}{\lambda_{t-1}} r^2_{\lambda_{t-1}}(\beta, \theta) + \left(1 - \frac{\lambda_t}{\lambda_{t-1}}\right) \frac{2}{\lambda_t^2}\Delta_t + \left(1 - \frac{\lambda_t}{\lambda_{t-1}}\right)^2 \frac{2\nu}{\lambda_t^2} \|\lambda_{t-1}\theta\|^2 \ ,$$



(a) Bonnefoy *et al.* (2014) safe region

(b) Gap Safe region

Figure 2.2 – Illustration of safe region differences between Bonnefoy *et al.* (2014) and Gap Safe strategies for the Lasso case; note that $\gamma = 1$ in this case. Here β is a primal point, θ is a dual feasible point (the feasible region Δ_X is in orange, while the respective safe balls $\mathcal{R}$ are in blue).

where $\Delta_t := f(X\beta) - f^*(\nabla f^*(-\lambda_t\theta))$ *can be made as small as desired.*

Dynamic Safe Rules. Another road to speed up solvers using screening rules was proposed by Bonnefoy et al. (2014, 2015) under the name "dynamic safe rules". For a fixed λ, it consists in performing screening along with the iterations of the optimization algorithm used to solve Problem (2.1). Let us consider a sequence (β_k) that converges to a primal solution $\hat{\beta}^{(\lambda)}$. For creating a dual feasible point, we apply the rescaling introduced in Equation (2.15) to $z = -\nabla f(X\beta_k)/\lambda$ and the dynamic strategy can be summarized by

$$\begin{aligned} \text{Dynamic sphere rule:} \quad & \text{If} \quad \Omega_g^\circ(X_g^\top \theta_k) + r_k \Omega_g^\circ(X_g) < 1, \quad \text{then} \quad \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ , \\ \text{Center:} \quad & \theta_k := \Theta(-\nabla f(X\beta_k)/\lambda) \ , \\ \text{Radius:} \quad & r_k := \sqrt{\frac{2}{\gamma\lambda^2}\,\mathrm{Gap}_\lambda(\beta_k, \theta_k)} \ . \end{aligned}$$

In practice the computation of the duality gap can be expensive due to the matrix vector operations needed to compute $X^\top \nabla f(X\beta_k)$. For instance in the Lasso case, a dual gap computation requires almost as much computation as a full pass of coordinate descent over the data. Hence, it is recommended to evaluate the dynamic (safe) rule only every few passes over the data set. In all our experiments, we have set this screening frequency parameter to $f^{ce} = 10$.

Dual Extrapolation. In the same way we can accelerate the convergence of an iterative algorithm by means of extrapolation of its iterates (Scieur et al., 2016), we can improve the estimation of $\hat{\theta}^{(\lambda)}$ by extrapolating the residual before applying the scaling in Proposition 11. This was proposed by Massias et al. (2018b) for the Lasso and have shown interesting speed up. Given a

number of iterations K (default being $K = 5$) and $\rho_k = y - X\beta_k$, let

$$\rho_k^{\text{accel}} := \begin{cases} \rho_k & \text{if } k \leqslant K, \\ \sum_{k=1}^{K} c_k \rho_{k+1-K} & \text{if } k > K, \end{cases}$$

where $c = (c_1, \cdots, c_K)$ in $\mathbb{R}^K$ is defined as $z/z^\top 1_K$ and z is a solution of $U_k^\top U_k z = 1_K$ with $U_k = [\rho_{k+1-K} - \rho_{k-K}, \cdots, \rho_k - \rho_{k-1}]$ in $\mathbb{R}^{n \times K}$. Then the new dual vector is given by

$$\theta_k^{\text{accel}} := \frac{\rho_k^{\text{accel}}}{\max(\lambda, \|X^\top \rho_k^{\text{accel}}\|_\infty)} \ . \tag{2.23}$$

See (Massias et al., 2018b) for more details and numerical experiments.

Active Warm Start. An another variant to further reduce running time in the *active warm start*, recently introduced by Ndiaye et al. (2017a) for speeding-up concomitant Lasso computations. Instead of simply leveraging the previous primal solution, the *active warm start* strategy also makes use of the previous safe active set $\mathcal{A}(\theta^{(\lambda_{t-1})}, r_{t-1})$, with $r_{t-1} = r_{\lambda_{t-1}}(\beta^{(\lambda_{t-1})}, \theta^{(\lambda_{t-1})})$. The idea is to take as a new primal warm start point, the (approximate) minimizer of P_{λ_t} under the additional constraint that its support is included in the safe active set $\mathcal{A}(\theta^{(\lambda_{t-1})}, r_{t-1})$ *i.e.*

$$\beta^{(\lambda_{t-1}, \lambda_t)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, f(X\beta) + \lambda_t \Omega(\beta) \text{ s.t. } \operatorname{supp}(\beta) \subseteq \mathcal{A}(\theta^{(\lambda_{t-1})}, r_{t-1}) \ . \tag{2.24}$$

In (2.24), we still choose $\beta^{(\lambda_{t-1})}$ as a standard warm start initialization with the same number of inner loops and/or accuracy as in (2.1) (to avoid the multiplication of parameters to be set by the user). Note that un-safe estimators of the active set can be used as for active warm start. In practice, we can use the (un-safe) strong active set provided by the Strong rules introduced by Tibshirani et al. (2012). This *Strong Warm Start* strategy is detailed below.

2.2.4 Application to Popular Estimators

Least Squares Lasso. For the Lasso estimator (Tibshirani, 1996), the data-fitting term is the standard least square, *i.e.* $f(X\beta) = \|y - X\beta\|_2^2/2 = \sum_{i=1}^n (y_i - x_i^\top \beta)^2/2$ (meaning that $f_i(z) = (y_i - z)^2/2$). The regularization term enforces sparsity at the feature level and is defined by $\Omega(\beta) = \|\beta\|_1$.

Group Lasso. For the Group Lasso estimator (Yuan and Lin, 2006), the data-fitting term is the same $f(X\beta) = \|y - X\beta\|_2^2/2$ but the penalty considered enforces group sparsity. Hence, we consider the norm $\Omega(\beta) = \Omega_w(\beta)$, often referred to as an ℓ_1/ℓ_2 norm, defined by $\Omega_w(\beta) := \sum_{g \in \mathcal{G}} w_g \|\beta_g\|_2$ where $w = (w_g)_{g \in \mathcal{G}}$ are some weights satisfying $w_g > 0$ for all group g in $\mathcal{G}$.

Elastic Net. For the Elastic Net estimator (Zou and Hastie, 2005) the data fitting term is the same than for the Lasso *i.e.* $f(X\beta) = \|y - X\beta\|_2^2/2$ and the regularization is $\Omega(\beta) = \eta \|\beta\|_1 + (1 - \eta) \|\beta\|_2^2/2$ interpolates between the ℓ_1 penalty and the ridge penalty. The regularizer is feature-wise separable with $\Omega_j(\beta_j) = \eta|\beta_j| + (1 - \eta)\beta_j^2$ for any j in $[p]$ and

$$\partial \Omega_j(\beta_j) = \begin{cases} [-\eta, \eta] & \text{if } \beta_j = 0, \\ \eta \frac{\beta_j}{|\beta_j|} + (1 - \eta)\beta_j & \text{if } \beta_j \neq 0 \end{cases}$$

$$\Omega_j^*(\xi_j) = \frac{1}{2\eta} \left[(|\xi_j| - (1 - \eta))_+\right]^2 \ .$$

ℓ_1 **Regularized Logistic Regression.** Here, we consider the formulation given in (Bühlmann and van de Geer, 2011, Chapter 3) for the two-class logistic regression. In such a context, one observes for each $i \in [n]$ a class label $l_i \in \{1, 2\}$. This information can be recast as $y_i = \mathbb{1}_{\{l_i=1\}}$ (where $\mathbb{1}$ is the indicator function), and it is then customary to minimize (2.1) where

$$f(X\beta) = \sum_{i=1}^{n} \left(-y_i x_i^\top \beta + \log\left(1 + \exp\left(x_i^\top \beta\right)\right)\right), \tag{2.25}$$

with $f_i(z) = -y_i z + \log(1 + \exp(z))$, and the penalty is simply the ℓ_1 norm: $\Omega(\beta) = \|\beta\|_1$. Let us introduce Nh, the (binary) negative entropy function defined by:

$$\mathrm{Nh}(x) = \begin{cases} x\log(x) + (1-x)\log(1-x), & \text{if } x \in [0,1] \ , \\ +\infty, & \text{otherwise} \ . \end{cases} \tag{2.26}$$

We use the convention $0\log(0) = 0$, and one can check that $f_i^*(z_i) = \mathrm{Nh}(z_i + y_i)$ and $\gamma = 4$. Note that we have privileged the formulation with the label $y \in \{0, 1\}^n$ instead of $y \in \{+1, -1\}^n$ in order to be consistent with the multinomial cases below. One can simply switch from one formulation to the other thanks to the mapping $\widetilde{y} = 2y - 1$.

ℓ_1/ℓ_2 **Multi-task Regression.** The multi-task Lasso is a regression problem where the parameters form a matrix $\mathrm{B} \in \mathbb{R}^{p\times q}$. Denoting n the number of observations for each task $k \in [q]$, it is defined as

$$\min_{\mathrm{B}\in\mathbb{R}^{p\times q}} \frac{1}{2}\|Y - X\mathrm{B}\|_F^2 + \lambda \sum_{j=1}^{p} \|\mathrm{B}_{j,:}\|_2 \,, \tag{2.27}$$

where $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$. Here we assume that the explanatory variables X are shared among the tasks however the Gap Safe rules would readily apply to the non-shared design formulation as in (Lee et al., 2010) or in (Liu et al., 2009) since the loss is still smooth, *cf.* Remark 4.

Introducing the vec operator that vectorizes a matrix by stacking its columns to form a column vector, and the Kronecker product $\otimes$ of two matrices, the multi-task Lasso can be rewritten as a special case of Group Lasso. In fact, we have n class of observations $c_i = (i + (k-1)n)_{k\in[q]}$ of size q for each $i \in [n]$ (the overall number of observations is $n' = nq$) and p groups $g_j = (j + (k-1)p)_{k\in[q]}$ such that $|g_j| = q$ for $j \in [p]$. The design matrix $\tilde{X} = I_q \otimes X \in \mathbb{R}^{n'\times p'} = \mathbb{R}^{nq\times pq}$ is a q-block diagonal matrix defined as $\tilde{X} = \mathrm{diag}(X, \ldots, X)$, $y = \mathrm{vec}(Y)$ and $\beta = \mathrm{vec}(\mathrm{B})$, we have:

$$\min_{\beta\in\mathbb{R}^{p'}} \frac{1}{2}\sum_{i=1}^{n} \left\|y_{c_i} - \tilde{x}_i^\top \beta\right\|_2^2 + \lambda \sum_{j=1}^{p} \left\|\beta_{g_j}\right\|_2 , \tag{2.28}$$

i.e. $f_i(z) = \|y_{c_i} - z\|_2^2/2$. is that it can be concisely written using the matrix forms of y and β, without the need to actually construct the large matrix X'. This is particularly appealing for the implementation. In signal processing, this model is also referred to as the Multiple Measurement Vector (MMV) problem. It allows to jointly select the same features for multiple regression tasks, see (Argyriou et al., 2006, 2008; Obozinski et al., 2010). This estimator has been used in various applications such as prediction of the location of a protein within a cell (Xu et al., 2011) or in neuroscience (Gramfort et al., 2012), for instance to diagnose Alzheimer's disease (Zhang et al., 2012) and biological data (Bellon et al., 2016; Playe et al., 2018).

ℓ_1/ℓ_2 **Multinomial Logistic Regression.** We adapt the formulation given in (Bühlmann and van de Geer, 2011, Chapter 3) for the multinomial regression. In such a context, one observes for each $i \in [n]$ a class label $l_i \in [q]$. This information can be recast into a matrix $Y \in \mathbb{R}^{n\times q}$ filled by 0's and 1's: $Y_{i,k} = \mathbb{1}_{\{l_i=k\}}$ (where $\mathbb{1}$ is the indicator function). In the same spirit as for the

multi-task Lasso, a matrix $\mathrm{B} \in \mathbb{R}^{p \times q}$ is formed by q vectors encoding the hyperplanes for the linear classification. Thus the multinomial ℓ_1/ℓ_2 regularized regression reads:

$$\min_{\mathrm{B} \in \mathbb{R}^{p \times q}} \sum_{i=1}^{n} \left(\sum_{k=1}^{q} -Y_{i,k} x_i^\top \mathrm{B}_{:,k} + \log \left(\sum_{k=1}^{q} \exp\left(x_i^\top \mathrm{B}_{:,k} \right) \right) \right) + \lambda \sum_{j=1}^{p} \|\mathrm{B}_{j,:}\|_2 \,. \tag{2.29}$$

Using a similar reformulation as in the multi-task regression, we define $c_i = (i + (k-1)n)_{k \in [q]}$ for each $i \in [n]$ and $g_j = (j + (k-1)p)_{k \in [q]}$ for each $j \in [p]$. The ℓ_1/ℓ_2 multinomial logistic regression can be cast into our framework as:

$$\min_{\beta \in \mathbb{R}^{p'}} \sum_{i=1}^{n} f_i\left(\tilde{x}_i^\top \beta\right) + \lambda \sum_{j=1}^{p} \left\|\beta_{g_j}\right\|_2 , \tag{2.30}$$

with $f_i : \mathbb{R}^q \to \mathbb{R}$ such that $f_i(z) = -y_{c_i}^\top z + \log\left(\sum_{k=1}^{q} \exp\left(z_k\right)\right)$. Note that generalizing (2.1) to functions $f_i : \mathbb{R}^q \to \mathbb{R}$ does not bear difficulties, see (Ndiaye et al., 2015). Let us introduce NH, the negative entropy function defined by

$$\mathrm{NH}(x) = \begin{cases} \sum_{i=1}^{q} x_i \log(x_i), & \text{if } x \in \Sigma_q = \{x \in \mathbb{R}_+^q : \sum_{i=1}^{q} x_i = 1\}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.31}$$

We use the convention $0 \log(0) = 0$, and one can check that $f_i^*(z) = \mathrm{NH}(z + Y_i)$ and $\gamma = 1$.

Remark 6. *The intercept has been neglected in our models for simplicity. The Gap Safe framework can also handle such a feature to the cost of more technical details (by adapting the results from (Koh et al., 2007) for instance). However, in practice, the intercept can be handled in the present formulation by adding a constant column to the design matrix X. The intercept is then regularized. However, if the constant is set high enough, regularization is small and experiments show that it has little to no impact for high-dimensional problems. This is the strategy used in the* Liblinear *package by Fan et al. (2008).*

Another alternative could be to handle the constant term as is performed by El Ghaoui et al. (2012). For the Lasso, the bias can also be treated implicitly as follows. Define $e = (1, \ldots, 1)^\top \in \mathbb{R}^n$

$$\min_{\nu \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2} \|y - (X\beta + \nu e)\|_2^2 + \lambda \|\beta\|_1 .$$

By setting to zero the derivative w.r.t. *to ν of the objective function we get $\nu = \bar{y} - \bar{X}\beta$ where $\bar{y}$ is the mean of y and $\bar{X}$ is the column-wise mean of X. Hence, we get rid of the bias term by solving*

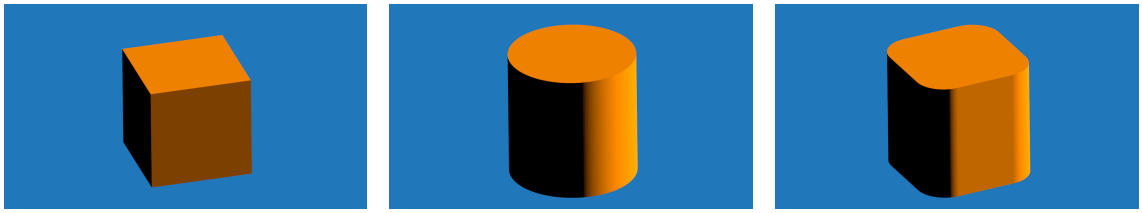
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \left\|(y - \bar{y}e) - (X - \bar{X}e^\top)\beta\right\|_2^2 + \lambda \|\beta\|_1 .$$

In cases where the bias term has to be explicit, one can just iteratively perform an (un-regularized) gradient descent on the bias component.

Smoothed SVM. For the smoothed SVM (Shalev-Shwartz and Zhang, 2014), the loss is $f(X\beta) = \sum_{i \in [n]} f_i(x_i^\top \beta)$ where the f_i is the smoothed hinge loss defined as

$$f_i(z_i) := \begin{cases} 0 & \text{if } y_i z_i > 1, \\ 1 - y_i z_i - \frac{\gamma}{2} & \text{if } y_i z_i < 1 - \gamma, \\ \frac{1}{2\gamma}(1 - y_i z_i)^2 & \text{otherwise} \end{cases}$$

$$f_i^*(\xi_i) := \begin{cases} \frac{\gamma}{2}\xi_i^2 + y_i \xi_i & \text{if } y_i \xi_i \in [-1, 0], \\ +\infty & \text{otherwise} \end{cases} .$$



(a) Lasso dual ball $\mathcal{B}_{\Omega^D}$ for $\Omega^D(\theta) = \|\theta\|_\infty$.

(b) Group Lasso dual ball $\mathcal{B}_{\Omega^D}$ for $\Omega^D(\theta) = \max(\sqrt{\theta_1^2 + \theta_2^2}, |\theta_3|)$.

(c) Sparse-Group Lasso dual ball $\mathcal{B}_{\Omega^D} = \{\theta : \forall g \in \mathcal{G}, \|\mathrm{ST}_\tau(\theta_g)\|_2 \leqslant (1-\tau)w_g\}$.

Figure 2.3 – Lasso, Group Lasso and Sparse-Group Lasso dual unit balls: $\mathcal{B}_{\Omega^D} = \{\theta : \Omega^D(\theta) \leqslant 1\}$. For the illustration, the group structure is chosen such that $\mathcal{G} = \{\{1,2\},\{3\}\}$, *i.e.* $g_1 = \{1,2\}, g_2 = \{3\}$, $n = p = 3$, $w_{g_1} = w_{g_2} = 1$ and $\tau = 1/2$.

which can be combined with ridge regularization. Note that in this case the dual loss is non smooth and the regularization is smooth. When combined with Elastic net regularization, it leads to a doubly sparse model with feature-wise and sample-wise sparsity. Simultaneous screening rule for this case was introduced in (Shibagaki et al., 2016). The subdifferential of the dual loss splits in two part:

case $y_i = 1$:

$$\partial f_i^*(-\xi_i) = \begin{cases} [1, \infty[& \text{if } \xi_i = 0 \\ ]\infty, 1-\gamma] & \text{if } \xi_i = 1 \\ -\gamma x + 1 & \text{if } \xi_i \in]0, 1[\end{cases}$$

case $y_i = -1$:

$$\partial f_i^*(-\xi_i) = \begin{cases} [\infty, -1[& \text{if } \xi_i = 0 \\ ]-1+\gamma, \infty] & \text{if } \xi_i = 1 \\ -\gamma x - 1 & \text{if } \xi_i \in]0, 1[\end{cases}$$

Similar results also holds for the vanilla SVM as well as smoothed ϵ-insensitive loss function. Screening rule for this function is used in (Sangnier et al., 2017).

Sparse-Group Lasso. In the Sparse-Group Lasso case, we also have for $\beta \in \mathbb{R}^p$, $f(X\beta) = \|y - X\beta\|_2^2/2$ and the regularization $\Omega(\beta) = \Omega_{\tau,w}(\beta)$ is defined by

$$\Omega(\beta) := \tau\|\beta\|_1 + (1-\tau)\sum_{g\in\mathcal{G}} w_g \|\beta_g\|_2 ,$$

for $\tau \in [0,1], w = (w_g)_{g\in\mathcal{G}}$ with $w_g \geqslant 0$ for all $g \in \mathcal{G}$. Note that we recover the Lasso if $\tau = 1$, and the Group Lasso if $\tau = 0$; the case where $w_g = 0$ for some $g \in \mathcal{G}$ together with $\tau = 0$ is excluded (Ω is not a norm in such a case). This estimator was introduced by Simon et al. (2013) to enforce sparsity both at the feature and at the group level, and was used in different applications such as brain imaging in (Gramfort et al., 2013) or in genomics in (Peng et al., 2010). Other hierarchical norms have also been proposed in (Sprechmann et al., 2010) or (Jenatton et al., 2011) and could be handled in our framework modulo additional technical details.

For the Sparse-Group Lasso, the geometry of the dual feasible set Δ_X is more complex (*cf.* Figure 2.3 for a comparison w.r.t. Lasso and Group Lasso). As a consequence, additional geometrical insights are needed to derive efficient safe rules, especially to compute the dual norm required by Equation (2.15) and the computation of the safe screening rules given in (3). We now introduce the ϵ-norm (denoted $\|\cdot\|_\epsilon$) as it has a connection with the Sparse-Group Lasso norm Ω. The ϵ-norm was first proposed by Burdakov (1988) for other purposes, see also (Burdakov and

Merkulov, 2001). For any $\epsilon \in [0,1]$ and any $x \in \mathbb{R}^d$, $\|x\|_\epsilon$ is defined as the unique nonnegative solution ν of the following equation (for $\epsilon = 0$, we define $\|x\|_{\epsilon=0} := \|x\|_\infty$):

$$\sum_{i=1}^{d} (|x_i| - (1-\epsilon)\nu)_+^2 = (\epsilon\nu)^2. \tag{2.32}$$

Using soft-thresholding, this is equivalent to solve in ν the equation $\|\mathrm{ST}_{(1-\epsilon)\nu}(x)\|_2 = \epsilon\nu$. Moreover, its dual norm is given by; see (Burdakov and Merkulov, 2001, Equation (42)):

$$\|\xi\|_\epsilon^D = \epsilon\|\xi\|_2^D + (1-\epsilon)\|\xi\|_\infty^D = \epsilon\|\xi\|_2 + (1-\epsilon)\|\xi\|_1 \ . \tag{2.33}$$

This allows to express the Sparse-Group Lasso norm $\Omega_{\tau,w}$ using the dual ϵ-norm. We now derive an explicit formulation for the dual norm of the Sparse-Group Lasso, originally proposed in (Ndiaye et al., 2016, Prop. 4). The proofs are recalled in the appendix Section 2.6.

Proposition 15 (Properties of Sparse-Group Lasso)**.** *For all groups g in $\mathcal{G}$, let us introduce*

$$\epsilon_g := \frac{(1-\tau)w_g}{\tau + (1-\tau)w_g} \ .$$

Then, the Sparse-Group Lasso norm satisfies the following properties for any β and ξ in $\mathbb{R}^p$,

$$\begin{aligned}
\Omega(\beta) &= \sum_{g\in\mathcal{G}} (\tau + (1-\tau)w_g)\|\beta_g\|_{\epsilon_g}^D \\
\Omega^\circ(\xi) = \Omega^D(\xi) &= \max_{g\in\mathcal{G}} \frac{\|\xi_g\|_{\epsilon_g}}{\tau + (1-\tau)w_g}. \\
\Omega^*(\xi) = \iota_{\mathcal{B}_{\Omega^D}}(\xi) &= \sum_{g\in\mathcal{G}} \iota_{\mathcal{B}}\left(\frac{\mathrm{ST}_\tau(\xi_g)}{(1-\tau)w_g}\right) \\
\partial\Omega(\beta) &= \left\{\xi \in \mathbb{R}^p : \forall g \in \mathcal{G},\ \xi_g \in \tau\partial\|\cdot\|_1(\beta_g) + (1-\tau)w_g\partial\|\cdot\|_2(\beta_g)\right\}.
\end{aligned}$$

Remark 7. *The dual formulation* (2.2) *for the Sparse-Group Lasso is a constrained optimization where the dual feasible set can be characterized as*

$$\begin{aligned}
\Delta_X &= \left\{\theta \in \mathbb{R}^n : \forall g \in \mathcal{G}, \|X_g^\top\theta\|_{\epsilon_g} \leqslant \tau + (1-\tau)w_g\right\} \\
&= \left\{\theta \in \mathbb{R}^n : \forall g \in \mathcal{G}, \|\mathrm{ST}_\tau\left(X_g^\top\theta\right)\|_2 \leqslant (1-\tau)w_g\right\} \ .
\end{aligned}$$

The first (resp. second) expression corresponds to (Fenchel) conjugation (resp. polar) duality.

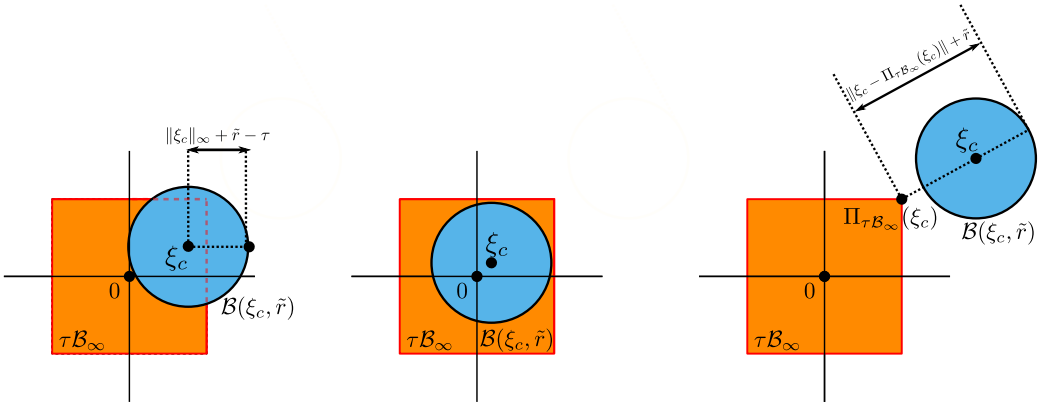
The Sparse-Group Lasso benefits from two levels of screening: the safe rules can detect both group-wise zeros and coordinate-wise zeros in the remaining groups: for any group g in $\mathcal{G}$ and any safe sphere $\mathcal{B}(\theta, r)$, Equation (3) and the sub-differential of the Sparse-Group Lasso norm in Proposition 15 give

$$\begin{aligned}
&\textbf{\textit{Group level safe screening rule:}} && \max_{\theta\in\mathcal{B}(c,r)} \frac{\|X_g^\top\theta\|_{\epsilon_g}}{\tau + (1-\tau)w_g} < 1 \Rightarrow \hat{\beta}_g^{(\lambda)} = 0 \ . \\
&\textbf{\textit{Feature level safe screening rule:}} && \forall j \in g,\ \max_{\theta\in\mathcal{B}(c,r)} |X_j^\top\theta| < \tau \Rightarrow \hat{\beta}_j^{(\lambda)} = 0 \ .
\end{aligned}$$

Proof. For $\beta_g^\star = 0$, we have $\partial\Omega_g(\beta_g^\star) = \mathcal{B}_{\Omega_g^D}$ and $\Omega_g^\circ = \Omega_g^D$. Then, using the expression of the dual norm of the Sparse-Group Lasso in Proposition 15, the group-wise screening test (2.6) consists in testing whether $\|X_g^\top\hat{\theta}^{(\lambda)}\|_{\epsilon_g} < \tau + (1-\tau)w_g$.

For $\beta_g^\star \neq 0$, we have $\partial\Omega_g(\beta_g^\star) = \tau\partial\left\|\cdot\right\|_1(\beta_g^\star) + (1-\tau)w_g\left\{\frac{\beta_g^\star}{\|\beta_g^\star\|_2}\right\}$. Hence for $j \in g$ such that $\beta_j^\star = 0$, we have $\partial\Omega_j(\beta_j^\star) = \tau\mathcal{B}_\infty$ and the feature-wise screening test consists in $|X_j^\top\hat{\theta}^{(\lambda)}| < \tau$.

□



(a) $\mathcal{B}(\xi_c, \tilde{r}) \cap \tau\mathcal{B}_\infty \neq \varnothing$; $\xi_c \in \tau \mathrm{int}\mathcal{B}_\infty$

(b) $\mathcal{B}(\xi_c, \tilde{r}) \subset \tau\mathcal{B}_\infty$

(c) $\mathcal{B}(\xi_c, \tilde{r}) \cap \tau\mathcal{B}_\infty = \varnothing$; $\xi_c \notin \tau \mathrm{int}\mathcal{B}_\infty$

Noting that $\|\mathrm{ST}_\tau(x)\|_2 = (1-\tau)w_g \iff \|x\|_{\epsilon_g} = \tau + (1-\tau)w_g$, the group level safe screening rule can be rewritten as

$$\max_{\theta \in \mathcal{B}(\theta, r)} \left\|\mathrm{ST}_\tau(X_g^\top \theta)\right\|_2 < (1-\tau)w_g \Longrightarrow \hat{\beta}_g^{(\lambda)} = 0 \ .$$

The advantage of this formulation is that one can easily derive a tight upper-bound of the non-convex optimization problem in the left hand side of the preceding test. Indeed, we have $\mathrm{ST}_\tau(x) = x - \Pi_{\tau\mathcal{B}_\infty}(x)$ which brings us finally into a geometric problem easier to solve. We recall from (Ndiaye et al., 2016, Prop. 1) that for any center $\theta \in \Delta_X$, any group $g \in \mathcal{G}$ and any $j \in g$, we have the following upper-bound

Proposition 16.

$$\max_{\theta \in \mathcal{B}(\theta, r)} |X_j^\top \theta| \leqslant |X_j^\top \theta| + r \left\|X_j\right\|_2 ,$$

$$\max_{\theta \in \mathcal{B}(\theta, r)} \left\|\mathrm{ST}_\tau(X_g^\top \theta)\right\|_2 \leqslant T_g := \begin{cases} \left\|\mathrm{ST}_\tau(X_g^\top \theta)\right\|_2 + r \left\|X_g\right\|_2 , & \text{if } \left\|X_g^\top \theta\right\|_\infty > \tau, \\ (\left\|X_g^\top \theta\right\|_\infty + r \left\|X_g\right\|_2 - \tau)_+, & \text{otherwise.} \end{cases}$$

Proof. $|X_j^\top \theta| \leqslant |[X_g^\top(\theta - \theta_c)]_j| + |X_j^\top \theta_c| \leqslant r\|X_j\| + |X_j^\top \theta_c|$ as soon as $\theta \in \mathcal{B}(\theta_c, r)$.

Since $\theta \in \mathcal{B}(\theta_c, r)$ implies that $X_g^\top \theta \in \mathcal{B}(X_g^\top \theta_c, r\|X_g\|)$, we have $\max_{\theta \in \mathcal{B}(\theta_c, r)} \|\mathrm{ST}_\tau(X_g^\top \theta)\| \leqslant \max_{\xi \in \mathcal{B}(\xi_c, \tilde{r})} \|\mathrm{ST}_\tau(\xi)\|$ where $\xi_c = X_g^\top \theta_c$ and $\tilde{r} = r\|X_j\|$. From now, we just have to show how to compute $\max_{\xi \in \mathcal{B}(\xi_c, \tilde{r})} \|\mathrm{ST}_\tau(\xi)\|$.

In the case where $\xi_c \in \mathrm{int}(\tau\mathcal{B}_\infty)$, if $\|\xi_c\|_\infty + \tilde{r} \leqslant \tau$ (*i.e.* $\mathcal{B}(\xi_c, \tilde{r}) \subset \tau\mathcal{B}_\infty$), we have $\Pi_{\tau\mathcal{B}_\infty}(\xi) = \xi$ and thus, $\max_{\xi \in \mathcal{B}(\xi_c, \tilde{r})} \|\mathrm{ST}_\tau(\xi)\| = \max_{\xi \in \mathcal{B}(\xi_c, \tilde{r})} \|\xi - \Pi_{\tau\mathcal{B}_\infty}(\xi)\| = 0$.

Otherwise if $\xi_c \in \mathrm{int}(\tau\mathcal{B}_\infty)$ and $\|\xi_c\|_\infty + \tilde{r} > \tau$, for any vector $\xi \in \partial\mathcal{B}(\xi_c, \tilde{r}) \cap (\tau\mathcal{B}_\infty)^c$ and any vector $\tilde{\xi} \in \partial\tau\mathcal{B}_\infty \cap [\xi, \xi_c]$, $\|\xi - \Pi_{\tau\mathcal{B}_\infty}(\xi)\| \leqslant \|\xi - \tilde{\xi}\| = \tilde{r} - \|\tilde{\xi} - \xi_c\|$. Hence

$$\begin{aligned} \max_{\xi \in \mathcal{B}(\xi_c, \tilde{r})} \|\xi - \Pi_{\tau\mathcal{B}_\infty}(\xi)\| &\leqslant \max_{\substack{\xi \in \partial\mathcal{B}(\xi_c, \tilde{r}) \cap (\tau\mathcal{B}_\infty)^c \\ \tilde{\xi} \in \partial\tau\mathcal{B}_\infty \cap [\xi, \xi_c]}} \tilde{r} - \left\|\tilde{\xi} - \xi_c\right\| \\ &\leqslant \tilde{r} - \min_{\xi \in \partial\tau\mathcal{B}_\infty} \|\xi - \xi_c\| = \tilde{r} - \tau + \|\xi_c\|_\infty . \end{aligned}$$

This upper bound is attained. Indeed, $\max_{\theta \in \mathcal{B}(\xi_c, \tilde{r})} \|\xi - \Pi_{\tau\mathcal{B}_\infty}(\xi)\| = \tilde{r} - \|\Pi_{\tau\mathcal{B}_\infty}(\hat{\xi}) - \xi_c\| = \tilde{r} - \tau + \|\xi_c\|_\infty$ where $\hat{\xi}$ is a vector in $\partial\mathcal{B}(\xi_c, \tilde{r})$ such that $\Pi_{\tau\mathcal{B}_\infty}(\hat{\xi}) = \xi_c + e_{j^\star}(\tau - \|\xi_c\|_\infty)$ and

$j^\star \in \arg\max_{j\in[p]} |(\xi_c)_j|$.

If $\xi_c \notin \operatorname{int}\tau\mathcal{B}_\infty$, since the projection operator on a convex set is a contraction, we have

$$\begin{aligned}\forall \xi \in \partial\mathcal{B}(\xi_c, \tilde{r}),\ \|\xi - \Pi_{\tau\mathcal{B}_\infty}(\xi)\| &\leqslant \|\xi - \Pi_{\tau\mathcal{B}_\infty}(\xi_c)\| \\ &\leqslant \|\xi_c - \Pi_{\tau\mathcal{B}_\infty}(\xi_c)\| + \|\xi - \xi_c\| \\ &= \|\xi_c - \Pi_{\tau\mathcal{B}_\infty}(\xi_c)\| + \tilde{r}.\end{aligned}$$

Moreover, it is straightforward to see that the vector $\tilde{\xi} := \tilde{\gamma}\xi_c + (1-\tilde{\gamma})\Pi_{\tau\mathcal{B}_\infty}(\xi_c)$ where $\tilde{\gamma} = 1 + \frac{\tilde{r}}{\|\xi_c\| + \|\Pi_{\tau\mathcal{B}_\infty}(\xi_c)\|}$ belongs to $\partial\mathcal{B}(\xi_c, \tilde{r})$; it verifies $\Pi_{\tau\mathcal{B}_\infty}(\xi_c) = \Pi_{\tau\mathcal{B}_\infty}(\tilde{\xi})$ and it attains this bound. □

From the bounds in Proposition 16, we derive the two level of safe screening rule:

Proposition 17 (Safe Screening rule for the Sparse-Group Lasso)**.**

Group level screening: $\forall g \in \mathcal{G}, \quad$ *if* $T_g < (1-\tau)w_g$, *then* $\hat{\beta}_g^{(\lambda)} = 0$.

Feature level screening: $\forall g \in \mathcal{G}, \forall j \in g$, *if* $|X_j^\top \theta| + r\,\|X_j\|_2 < \tau$, *then* $\hat{\beta}_j^{(\lambda)} = 0$.

In the same spirit than Proposition 10, for any safe region $\mathcal{R}$, *i.e.* a set containing $\hat{\theta}^{(\lambda)}$, we define two levels of active sets, one for the group level and one for the feature level:

$$\begin{aligned}\mathcal{A}_{\text{gp}}(\mathcal{R}) &:= \{g \in \mathcal{G},\ \max_{\theta\in\mathcal{R}} \left\|\mathrm{ST}_\tau(X_g^\top\theta)\right\|_2 \geqslant (1-\tau)w_g\}\ , \\ \mathcal{A}_{\text{ft}}(\mathcal{R}) &:= \bigcup_{g\in\mathcal{A}_{\text{gp}}(\mathcal{R})} \{j \in g :\ \max_{\theta\in\mathcal{R}} |X_j^\top\theta| \geqslant \tau\}\ .\end{aligned}$$

If one considers sequence of converging regions, then the next proposition (see (Ndiaye et al., 2016, Prop. 3)) states that we can identify in finite time the optimal active sets defined as follows:

$$\mathcal{E}_{\text{gp}} := \left\{g \in \mathcal{G} :\ \|\mathrm{ST}_\tau(X_g^\top\hat{\theta}^{(\lambda)})\|_2 = (1-\tau)w_g\right\},\ \mathcal{E}_{\text{ft}} := \bigcup_{g\in\mathcal{E}_{\text{gp}}} \left\{j \in g :\ |X_j^\top\hat{\theta}^{(\lambda)}| \geqslant \tau\right\}.$$

Proposition 18. *Let $(\mathcal{R}_k)_{k\in\mathbb{N}}$ be a sequence of safe regions whose diameters converge to 0. Then,* $\lim_{k\to\infty} \mathcal{A}_{gp}(\mathcal{R}_k) = \mathcal{E}_{gp}$ *and* $\lim_{k\to\infty} \mathcal{A}_{ft}(\mathcal{R}_k) = \mathcal{E}_{ft}$.

2.3 Computation of Support Function

The support functions inevitably intervene in the formulation of the screening rules Theorem 3 and in the rescaling procedure Proposition 11 to obtain a dual feasible vector. Their evaluations need to be performed multiple times during the algorithm and they must be computed efficiently. Fortunately, closed form expressions are available in many cases.

Given an interval $[a, b]$ on the real line, we have $\mathcal{S}_{[a,b]}(c) = \max(\langle a, c\rangle, \langle b, c\rangle)$. Although simple, this covers a large number of examples. Indeed, when we consider feature-wise convex separable function $P(z) = \sum_{j\in[p]} P_j(z_j)$, its subdifferential is a cartesian product of intervals $\partial P(z) = \Pi_{j\in[p]}\partial P_j(z_j)$ since for all j, $\partial P_j(\cdot)$ is a convex set on the real line. Hence it covers for instance the ℓ_1 regularization and the Elastic Net.

As we saw in the Section 1.3, a classical example of support function is the norm. It constitutes an important example since it is widely used as sparsity inducing penalties in machine learning.

For the Lasso,

$$\Omega(\beta) = \|\beta\|_1 \quad \text{and} \quad \Omega^\circ(\xi) = \Omega^D(\xi) = \max_{j\in[p]} |\xi_j|\ .$$

Algorithm 3 Computation of $\Lambda(x, \alpha, R)$.

Input:
$x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d,\ \alpha \in [0, 1],\ R \geqslant 0$
Output: $\Lambda(x, \alpha, R)$

if $\alpha = 0$ and $R = 0$ **then**
  $\Lambda(x, \alpha, R) = \infty$
else if $\alpha = 0$ and $R \neq 0$ **then**
  $\Lambda(x, \alpha, R) = \|x\|/R$
else if $R = 0$ **then**
  $\Lambda(x, \alpha, R) = \|x\|_\infty/\alpha$
else
  Get $I := \left\{ i \in [d] : |x_i| > \frac{\alpha\|x\|_\infty}{\alpha+R} \right\}$
  $n_I := \text{Card}(I)$
  Sort $x_{(1)} \geqslant x_{(2)} \geqslant \cdots \geqslant x_{(n_I)}$
  $S_0 = 0,\ S_0^{(2)} = 0,\ a_0 = 0$
  for $k \in [1, n_I - 1]$ **do**
    $S_k = S_{k-1} + x_{(k)}; \quad S_k^{(2)} = S_{k-1}^{(2)} + x_{(k)}^2$
    $a_{k+1} = \frac{S_k^{(2)}}{x_{(k+1)}^2} - 2\frac{S_k}{x_{(k+1)}} + k + 1$
    if $\frac{R^2}{\alpha^2} \in [a_k, a_{k+1}[$ **then**
      $j_0 = k + 1$
      break
  if $\alpha^2 j_0 - R^2 = 0$ **then**
    $\Lambda(x, \alpha, R) = \frac{S_{j_0}^2}{2\alpha S_{j_0}}$
  else
    $\Lambda(x, \alpha, R) = \frac{\alpha S_{j_0} - \sqrt{\alpha^2 S_{j_0}^2 - S_{j_0}^{(2)}(\alpha^2 j_0 - R^2)}}{\alpha^2 j_0 - R^2}$

For the Group Lasso,

$$\Omega_w(\beta) := \sum_{g \in \mathcal{G}} w_g \left\|\beta_g\right\|_2 \quad \text{and} \quad \Omega_w^\circ(\xi) = \Omega_w^D(\xi) = \max_{g \in \mathcal{G}} \frac{\|\xi_g\|_2}{w_g} \ .$$

For the Sparse-Group Lasso,

$$\Omega(\beta) := \tau\|\beta\|_1 + (1-\tau)\sum_{g \in \mathcal{G}} w_g \left\|\beta_g\right\|_2 \quad \text{and} \quad \Omega^\circ(\xi) = \Omega^D(\xi) = \max_{g \in \mathcal{G}} \frac{\|\xi_g\|_{\epsilon_g}}{\tau + (1-\tau)w_g} \ .$$

The following proposition shows how to compute exactly the dual norm of the Sparse-Group Lasso and the ϵ-norm. This is turned into an efficient procedure in Algorithm 3 (see the Section 2.6 for details and proofs).

Proposition 19. *For $\alpha \in [0, 1], R \geqslant 0$ and $x \in \mathbb{R}^d$, the equation $\sum_{i=1}^d \mathrm{ST}_{\nu\alpha}(x_i)^2 = (\nu R)^2$ has a unique solution $\nu := \Lambda(x, \alpha, R) \in \mathbb{R}_+$, that can be computed in $O(d \log d)$ operations in the worst case. With $n_I = \mathrm{Card}\{i \in [d] : |x_i| > \alpha\|x\|_\infty/(\alpha + R)\}$, the complexity of Algorithm 3 is $n_I + n_I \log(n_I)$, which is comparable to the ambient dimension d.*

We can explicit the critical parameter $\lambda_{\max}$ for the Sparse-Group Lasso that is

$$\lambda_{\max} = \max_{g \in \mathcal{G}} \frac{\Lambda(X_g^\top y, 1 - \epsilon_g, \epsilon_g)}{\tau + (1-\tau)w_g} = \Omega^D(X^\top y), \tag{2.34}$$

and get a dual feasible point (2.15), since

$$\Omega^D(X^\top \rho) = \max_{g \in \mathcal{G}} \frac{\Lambda(X_g^\top \rho, 1 - \epsilon_g, \epsilon_g)}{\tau + (1-\tau)w_g} \ .$$

2.4 Others Safe Regions and Alternative Acceleration Strategies

Previously we have restricted the discussion on the Gap Safe Sphere, here we show there is several others ways to build safe region.

The Seminal Safe Regions. The first Safe Screening rules introduced by El Ghaoui et al. (2012) can be generalized to Problem (2.1) as follows. Take $\hat{\theta}^{(\lambda_0)}$ the optimal solution of the dual problem (2.2) with a regularization parameter λ_0. Since $\hat{\theta}^{(\lambda)}$ is optimal for problem (2.2) one obtains $\hat{\theta}^{(\lambda)} \in \{\theta : D_\lambda(\theta) \geqslant D_\lambda(\hat{\theta}^{(\lambda_0)})\}$. This set was proposed as a safe region by El Ghaoui et al. (2012). In the regression case (where $f_i(z) = (y_i - z)^2/2$), it is straightforward to see that it corresponds to the safe sphere $\mathcal{R}_1^* := \mathcal{B}(y/\lambda, \|y/\lambda - \hat{\theta}^{(\lambda_0)}\|_2)$. Note that, $\hat{\theta}^{(\lambda_0)}$ can be replaced by any dual feasible vector in the definition of $\mathcal{R}_1^*$. In particular, one can use a rescaling gradient mapping in Equation (2.15). Using another first order optimality condition namely Proposition 5 on the dual problem, we have $\hat{\theta}^{(\lambda_0)}$ is optimal at λ_0 if and only if $\langle \nabla D_{\lambda_0}(\hat{\theta}^{(\lambda_0)}), (\theta - \hat{\theta}^{(\lambda_0)}) \rangle \leqslant 0$ for all θ in Δ_X. Since $\hat{\theta}^{(\lambda)}$ is feasible, we deduce that $\hat{\theta}^{(\lambda)} \in \{\theta \in \Delta_X : \langle \nabla D_{\lambda_0}(\hat{\theta}^{(\lambda_0)}), (\theta - \hat{\theta}^{(\lambda_0)}) \rangle \leqslant 0\} =: \mathcal{R}_2^*$. Finally, intersecting the two safe region above allows to obtain a smaller set $\mathcal{R}^* := \mathcal{R}_1^* \cap \mathcal{R}_2^*$

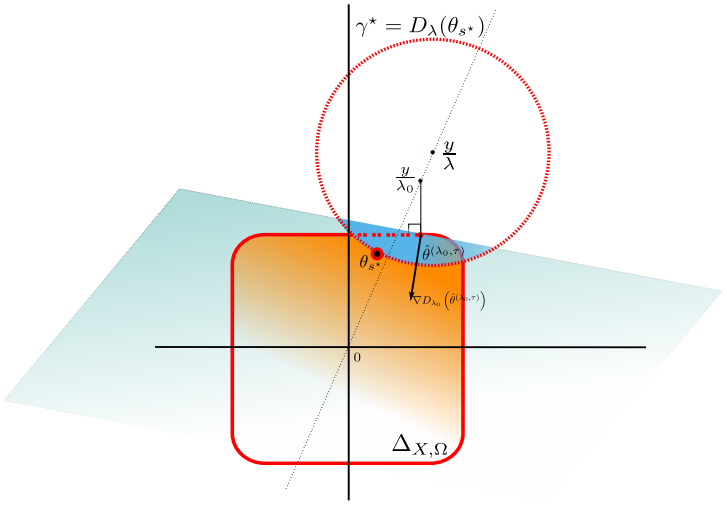


Figure 2.4 – Illustration of safe region in blue proposed in (El Ghaoui et al., 2012). The dual feasible set is represented in orange.

$$\mathcal{R}^* = \left\{\theta \in \Delta_X : D_\lambda(\theta) \geqslant D_\lambda(\hat{\theta}^{(\lambda_0)}), \langle \nabla D_{\lambda_0}(\hat{\theta}^{(\lambda_0)}), (\theta - \hat{\theta}^{(\lambda_0)}) \rangle \leqslant 0\right\} \ . \tag{2.35}$$

While interesting, this last safe region depends on the exact solution $\hat{\theta}^{(\lambda_0)}$ which is unknown in practice and a direct replacement with an approximated solution can leads to unsafe rules.

Projection Based Rules for the Quadratic Loss. A refined sphere rule can be obtained in the regression case by exploiting geometric information in the dual space. Let $g_\star \in \arg\max_{g \in \mathcal{G}} \Omega_g^D(X_g^\top(y - X\beta))$ (note that if $\lambda \leqslant \Omega^D(X^\top(y - X\beta))$, then $\Omega_{g_\star}^D(X_{g_\star}^\top(y - X\beta)) = \alpha$), and let us define

$$\mathcal{V}_\star := \{\theta \in \mathbb{R}^n : \Omega_{g_\star}^D(X_{g_\star}^\top \theta) \leqslant 1\} \text{ and } \mathcal{H}_\star := \{\theta \in \mathbb{R}^n : \Omega_{g_\star}^D(X_{g_\star}^\top \theta) = 1\} \ .$$

Note that for any $g \in \mathcal{G}$, we have $\Omega_g^D(X_g^\top \hat{\theta}^{(\lambda)}) \leqslant 1$, hence $\hat{\theta}^{(\lambda)} \in \mathcal{V}_\star$. Defining $\theta = (y - X\beta)/\alpha$ in Δ_X, we assume that the dual norm is differentiable at $X_{g_\star}^\top \theta$. Let $\eta := X_{g_\star} \nabla \Omega_{g_\star}^D(X_{g_\star}^\top \theta)$ be the vector normal to $\mathcal{V}_\star$ at θ, $\theta' \in \Delta_X$ is any dual feasible vector and define

$$\theta_c := \Pi_{\mathcal{H}_\star}\left(\frac{y}{\lambda}\right) = \frac{y}{\lambda} - \frac{\langle \frac{y}{\lambda}, \eta \rangle - 1}{\|\eta\|_2^2}\eta \text{ and } r_{\theta'} := \sqrt{\left\|\frac{y}{\lambda} - \theta'\right\|_2^2 - \left\|\frac{y}{\lambda} - \theta_c\right\|_2^2} \ .$$

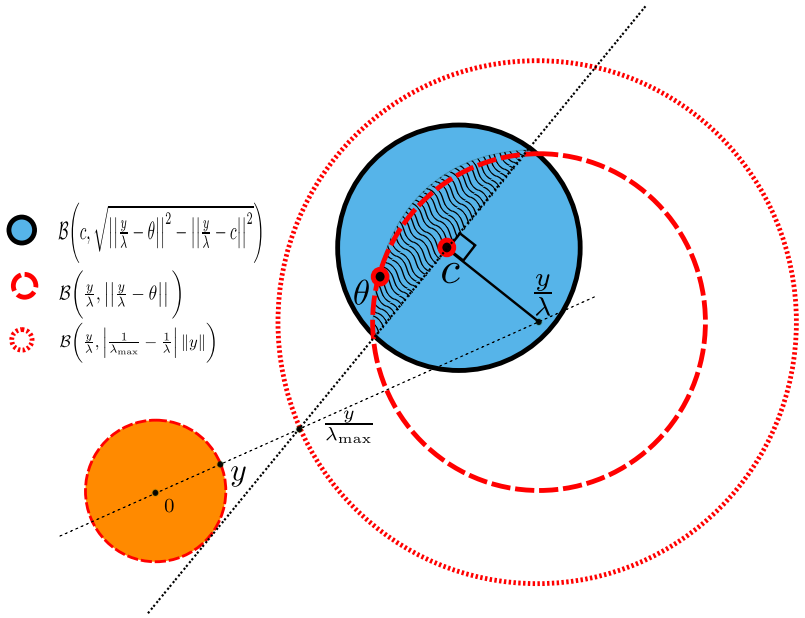


Figure 2.5 – Illustration of the projection based safe region in blue proposed in (Bonnefoy et al., 2014, 2015). The center c of the ball is the projection of y/λ onto the halfspace containing the dual feasible set.

Proof. We set $\mathcal{H}_\star^-$ the negative half-space induced by the hyperplane $\mathcal{H}_\star$. Since $\hat{\theta}^{(\lambda,\Omega)} \in \mathcal{V}_\star \subset \mathcal{H}_\star^-$ and $\mathcal{B}\left(\frac{y}{\lambda}, \|\frac{y}{\lambda}-\theta'\|\right)$ is a safe region, then $\hat{\theta}^{(\lambda,\Omega)} \in \mathcal{H}_\star^- \cap \mathcal{B}\left(\frac{y}{\lambda}, \|\frac{y}{\lambda}-\theta'\|\right)$. Moreover, for any $\theta \in \mathcal{H}_\star^- \cap \mathcal{B}\left(\frac{y}{\lambda}, \|\frac{y}{\lambda}-\theta'\|\right)$, we have:

$$\begin{aligned}\left\|\frac{y}{\lambda}-\theta'\right\|^2 \geqslant \left\|\frac{y}{\lambda}-\theta\right\|^2 &= \left\|\left(\frac{y}{\lambda}-\theta_c\right)+(\theta_c-\theta)\right\|^2 \\ &= \left\|\frac{y}{\lambda}-\theta_c\right\|^2 + \|\theta_c-\theta\|^2 + 2\left\langle\frac{y}{\lambda}-\theta_c, \theta_c-\theta\right\rangle \ .\end{aligned}$$

Since $\theta_c = \Pi_{\mathcal{H}_\star^-}(\frac{y}{\lambda})$ and $\mathcal{H}_\star^-$ is convex, then $\langle\theta_c - \frac{y}{\lambda}, \theta_c - \theta\rangle \leqslant 0$. Thus

$$\left\|\frac{y}{\lambda}-\theta'\right\|^2 \geqslant \left\|\frac{y}{\lambda}-\theta_c\right\|^2 + \|\theta_c-\theta\|^2, \text{ hence } \|\theta-\theta_c\| \leqslant \sqrt{\left\|\frac{y}{\lambda}-\theta'\right\|^2 - \left\|\frac{y}{\lambda}-\theta_c\right\|^2} =: r_{\theta'}.$$

Which show that $\mathcal{H}_\star^- \cap \mathcal{B}\left(\frac{y}{\lambda}, \|\frac{y}{\lambda}-\theta'\|\right) \subset \mathcal{B}(\theta_c, r_{\theta'})$. Hence the result. □

The special case where $\beta = 0$ and $\theta = y/\lambda_{\max}$ corresponds to the original ST3 introduced in Xiang et al. (2011) for the Lasso. A further improvement can be obtained by choosing dynamically $\theta = \theta_k$ along the iterations of an algorithm, this strategy corresponding to DST3 introduced in Bonnefoy et al. (2014, 2015) for the Lasso and Group Lasso and in Ndiaye et al. (2016) for the Sparse-Group Lasso. Now we can choose sequentially $\beta = \hat{\beta}^{(\lambda_{t-1})}$ or dynamically $\beta = \beta_k$ which lead to a center θ_c that is closer to the dual optimal solution.

Dual Polytope Projection. In the regression case, Wang et al. (2015) explore other geometric properties of the dual solution. Their method is based on the non-expansiveness of projection operators. Indeed, for $\hat{\theta}^{(\lambda)}$ (resp. $\hat{\theta}^{(\lambda_0)}$)) being optimal dual solution of (2.2) with parameter λ (resp. λ_0), one has: $\|\hat{\theta}^{(\lambda)} - \hat{\theta}^{(\lambda_0)}\|_2 = \|\Pi_\Delta(y/\lambda) - \Pi_\Delta(y/\lambda_0)\|_2 \leqslant \|y/\lambda - y/\lambda_0\|_2$ and hence $\hat{\theta}^{(\lambda)} \in \mathcal{B}(\hat{\theta}^{(\lambda_0)}, \|y/\lambda - y/\lambda_0\|_2)$. The authors also proved an enhanced version of this safe region by using the firm non-expansiveness of the projection operator. Assume that $\lambda_{t-1} < \lambda_t$, then the

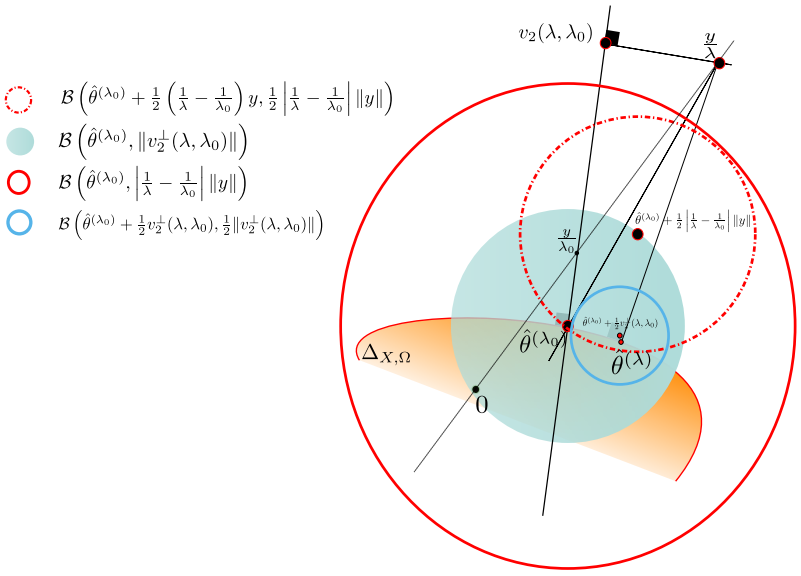


Figure 2.6 – Illustration of the region in blue obtained by dual polytope projection proposed in (Wang et al., 2015). Note that this set is unsafe due to approximation error of $\hat{\theta}^{(\lambda_0)}$. However it can be used as strong active set.

dual optimal solution of the group-Lasso with parameter λ_t, satisfies (see illustration in Figure 2.5)

$$\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\left(\hat{\theta}^{(\lambda_{t-1})} + \frac{1}{2} v^{\perp}(\lambda_{t-1}, \lambda_t), \frac{1}{2}\left\|v^{\perp}(\lambda_{t-1}, \lambda_t)\right\|_2\right),$$

$$\text{where } v^{\perp}(\lambda_{t-1}, \lambda_t) = \frac{y}{\lambda_t} - \hat{\theta}^{(\lambda_{t-1})} - \alpha[\hat{\theta}^{(\lambda_{t-1})}](\frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})})$$

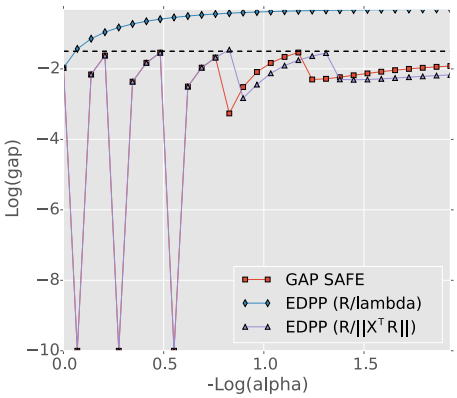
$$\alpha[\hat{\theta}^{(\lambda_{t-1})}] := \underset{\alpha \in \mathbb{R}_+}{\arg\min} \left\| \frac{y}{\lambda_t} - \hat{\theta}^{(\lambda_{t-1})} - \alpha(\frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})}) \right\|_2$$

$$= \frac{\langle \frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})}, \frac{y}{\lambda_t} - \hat{\theta}^{(\lambda_{t-1})} \rangle}{\| \frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})} \|_2^2} .$$

Note that the rule proposed by Wang et al. (2015), as pointed out by Bonnefoy et al. (2015), relies on the exact knowledge of a dual optimal solution for a previously solved Lasso problem. This is impossible to obtain in practice and even if it is possible to find accurate solutions, the search for high accuracy may hinder the benefits of the screening when it was not actually needed. Using inaccurate solutions may lead to discarding variables that should have been active and so the screened optimization algorithm will not converge to a solution of the original problem. We illustrate this issue on Figure 2.7. Knowing an approximation β to the optimal primal point, returned by the optimization algorithm at the previous regularization parameter λ_{t-1}, we need to choose an approximation θ to the optimal dual point to run EDPP.

— If we choose to approximate the dual optimal point by $\theta = \frac{1}{\lambda_{t-1}}(y - X\beta)$ (blue curve with diamonds), then the result is catastrophic. Indeed, at λ_1, $\beta = 0$ is a valid ϵ-solution for $\epsilon = 10^{-1.5}$ and the screening rule tries to perform a division by 0 when computing $\alpha[\theta]$.

— If we choose to approximate the dual optimal point by $\frac{1}{\max(\lambda_{t-1}, \|X^\top(y - X\beta)\|_\infty)}(y - X\beta)$, we have a better behavior (purple curve with triangles) but we may still have an algorithm which does not converge to an ϵ-solution. Here, for the 13^{th} Lasso problem a variable is erroneously removed and the problem can only be solved to accuracy $0.03515 > 10^{-1.5} \approx 0.03162$. This may look like a small issue but when the stopping criterion is based on the duality gap, this causes the algorithm to continue until the maximum number of iterations is reached.



$$X = \begin{bmatrix} 1/\sqrt{2} & \sqrt{2}/\sqrt{3} \\ 0 & -1/\sqrt{6} \\ -1/\sqrt{2} & -1/\sqrt{6} \end{bmatrix}, \qquad y = \begin{bmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -\sqrt{2}/\sqrt{3} \end{bmatrix}.$$

Figure 2.7 – EDPP is not safe. We run GAP SAFE and two interpretations of EDPP (described in the main text) to solve the Lasso path on the dataset defined by X and y above with target accuracy $10^{-1.5}$. For each Lasso problem, we plot the final duality gap returned by the optimization solver.

We propose in the appendix a solution for taking into account the approximation error.

Remark 8. *The preceding spheres are mainly based on the fact that $\hat{\theta}^{(\lambda)} = \Pi_\Delta(y/\lambda)$ which is limited to the regression case. Thus, those methods are not appropriate for more general data fitting term which greatly reduces the scope of such rules.*

Remark 9. *The radius of the regions above do not converge to zero even in the dynamic case (DST3), and the (fixed) center of the preceding sphere can be far from $\hat{\theta}^{(\lambda)}$ when λ gets small. Thus, those regions are not converging and are irrelevant for dynamic screening.*

Safe Screening with Variational Inequalities. As for the initial safe region proposed in (El Ghaoui et al., 2012), Liu et al. (2014) exploits the optimality condition in Proposition 5 successively at parameter λ_0 and λ reads:

$$\langle \nabla D_{\lambda_0}(\hat{\theta}^{(\lambda_0)}), (\theta - \hat{\theta}^{(\lambda_0)}) \rangle \leqslant 0, \quad \forall \theta \in \Delta_X \ , \tag{2.36}$$

$$\langle \nabla D_{\lambda}(\hat{\theta}^{(\lambda)}), (\theta - \hat{\theta}^{(\lambda)}) \rangle \leqslant 0, \quad \forall \theta \in \Delta_X \ . \tag{2.37}$$

Then setting $\theta = \hat{\theta}^{(\lambda)}$ in Equation (2.36) and $\theta = \hat{\theta}^{(\lambda_0)}$ in Equation (2.37) we have that $\hat{\theta}^{(\lambda)}$ belongs to $\mathcal{R}^*_{\text{sasvi}}$ where

$$\mathcal{R}^*_{\text{sasvi}} := \left\{ \theta \in \Delta_X : \langle \nabla D_\lambda(\theta), (\hat{\theta}^{(\lambda_0)} - \theta) \rangle \leqslant 0, \ \langle \nabla D_{\lambda_0}(\hat{\theta}^{(\lambda_0)}), (\theta - \hat{\theta}^{(\lambda_0)}) \rangle \leqslant 0 \right\} \ . \tag{2.38}$$

Note that the optimality condition in Proposition 5 states that $\hat{\theta}^{(\lambda)}$ is optimal at parameter λ if and only if Equation (2.37) holds if and only if $D_\lambda(\hat{\theta}^{(\lambda)}) \geqslant D_\lambda(\theta)$ for all θ in Δ_X, we find that the region $\mathcal{R}^*_{\text{sasvi}}$ coincides with the one in Equation (2.35) already proposed in (El Ghaoui et al., 2012).

Approximate Dictionaries. Matrix multiplications often dominate calculation costs in iterative algorithms. Hence by replacing the design matrix X by an (more structured) approximation $\tilde{X}$ which is easier to manipulate, Dantas and Gribonval (2017, 2018) have proposed an extension of the safe screening techniques. We describe it in our framework. Assuming that $X = \tilde{X} + E$, where $E = [\varepsilon_1, \cdots, \varepsilon_p]$ is a known approximation error, one have from the sublinearity of both Ω°_g and $\mathcal{S}^\circ_{\text{dom}\Omega^*}$:

$$\Omega^\circ_g((X_g + \varepsilon_g)^\top \theta) + r\Omega^\circ_g(X_g) \leqslant \Omega^\circ_g(\tilde{X}_g^\top \theta) + \Omega^\circ_g(\varepsilon_g^\top \theta) + r\Omega^\circ_g(X_g) =: T_g \ ,$$

$$\mathcal{S}^\circ_{\text{dom}\Omega^*}(X^\top \nabla f(\tilde{X}\beta)) \leqslant \mathcal{S}^\circ_{\text{dom}\Omega^*}(\tilde{X}^\top \nabla f(\tilde{X}\beta)) + \mathcal{S}^\circ_{\text{dom}\Omega^*}(E^\top \nabla f(\tilde{X}\beta)) =: \tilde{s} \ .$$

Hence, a safe elimination of the g-th variable can be done using only the approximated matrix $\tilde{X}$ if $T_g < 1$ and a dual feasible vector can be computed as $\tilde{\theta} = -\nabla f(\tilde{X}\beta)/\max(\lambda, \tilde{s}) \in \Delta_X$.

Note that when both the number of observations and features are large, performing a screening rule can be expensive. In this case, the extension to approximate dictionaries seems more scalable since one can set $\tilde{X}$ as a subsampling of X as it is usual in stochastic optimization. However, if the approximation error is not known exactly, it could be challenging to keep the safety and we are not aware of any such rules.

Gradient Based Region. In Section 2.2.1, we presented the Gap sphere region by exploiting the regularity of the loss function. An important point was that its radius goes to zero when the algorithm converges and it is a function of the duality gap. This later can be used as an optimality certificates. Similarly, we can build another safe region based on the gradient. We consider the setting where we want to solve

$$\min_{\beta \in \mathbb{R}^p} P_\lambda(\beta) = f(\beta) + \lambda\Omega(\beta) \ ,$$

and suppose that f is ν-smooth and Ω is non smooth. Given $L \geqslant \nu/2$ and an iterate $\beta^{(0)}$ in $\mathbb{R}^p$, we define the proximal step and the composite gradient direction as

$$\begin{aligned} \beta_L &:= \arg\min_{\beta \in \mathbb{R}^p} \lambda\Omega(\beta) + \frac{L}{2}\|\beta - (\beta^{(0)} - \frac{1}{L}\nabla f(\beta^{(0)}))\|^2 \ , \\ g_L(\beta) &:= L(\beta^{(0)} - \beta_L) \ . \end{aligned}$$

Proposition 20 (Gradient Safe Sphere Nesterov (2007, Lemma 2))**.** *Let P_λ be μ-strongly convex and $L \geqslant \nu/2$, then we have*

$$\|\hat{\beta}^{(\lambda)} - \beta_L\| \leqslant \frac{1}{\mu}\left(1 + \frac{\nu}{L}\right)\|g_L(\beta)\|_* \ . \tag{2.39}$$

Similar safe region have been used recently in (Yoshida et al., 2018) for deriving screening rules for some metric learning problems.

Strong Rules. The Strong rules were introduced in (Tibshirani et al., 2012) as a *heuristic* extension of the safe rules. It consists in relaxing the *safe* properties to discard features more aggressively, and can be formalized as follows. Assume that the gradient of the data fitting term ∇F is group-wise non-expansive w.r.t. the dual norm $\Omega_g^\circ(\cdot)$ along the regularization path *i.e.* that for any $g \in \mathcal{G}$, any $\lambda > 0, \lambda' > 0$, $\Omega_g^\circ\big(\nabla_g F(\hat{\beta}^{(\lambda)}) - \nabla_g F(\hat{\beta}^{(\lambda')})\big) \leqslant |\lambda - \lambda'|$. When choosing two regularization parameters such that $\lambda < \lambda'$ one has:

$$\begin{aligned} \lambda\Omega_g^\circ\left(X_g^\top\hat{\theta}^{(\lambda)}\right) = \Omega_g^\circ\left(\nabla_g F(\hat{\beta}^{(\lambda)})\right) &\leqslant \Omega_g^\circ\left(\nabla_g F(\hat{\beta}^{(\lambda')})\right) + \Omega_g^\circ\left(\nabla_g F(\hat{\beta}^{(\lambda)}) - \nabla_g F(\hat{\beta}^{(\lambda')})\right) \\ &\leqslant \Omega_g^\circ\left(\nabla_g F(\hat{\beta}^{(\lambda')})\right) + |\lambda - \lambda'| \\ &= \lambda'\Omega_g^\circ\left(X_g^\top\hat{\theta}^{(\lambda')}\right) + \lambda' - \lambda \ . \end{aligned}$$

Combining this with the screening rule (3), one obtains:

$$\Omega_g^\circ\left(X_g^\top\hat{\theta}^{(\lambda')}\right) < \frac{2\lambda - \lambda'}{\lambda'} \Longrightarrow \Omega_g^\circ\left(X_g^\top\hat{\theta}^{(\lambda)}\right) < 1 \Longrightarrow \hat{\beta}_g^{(\lambda)} = 0. \tag{2.40}$$

The set of variables not eliminated is called the *strong active set* and is defined as:

$$\mathcal{STG}(\hat{\theta}^{(\lambda')}, \lambda, \lambda') := \left\{g \in \mathcal{G} : \Omega_g^\circ\left(X_g^\top\hat{\theta}^{(\lambda')}\right) \geqslant \frac{2\lambda - \lambda'}{\lambda'}\right\}. \tag{2.41}$$

Note that Strong rules are un-safe because the non-expansiveness condition on the (gradient of the) data fitting term is usually not satisfied without stronger assumptions on the design matrix X; see discussion in (Tibshirani et al., 2012, Section 3). It requires the exact knowledge of $\hat{\theta}^{(\lambda')}$ which is not available in practice. Using such rules, the authors advised to check the KKT condition [2] a posteriori, to avoid removing wrongly some features. To overcome this limitation, we propose to use the strong active set $\mathcal{STG}(\hat{\theta}^{(\lambda_{t-1})}, \lambda_t, \lambda_{t-1})$ defined by Equation (2.41) for an active warm start strategy. We compare below this strategy with the one using $\mathcal{A}(\theta_{t-1}, r_{t-1})$ in Equation (2.24) as initial active set. A similar strategy is also used in the "big lasso" package by Zeng and Breheny (2017) as a hybrid screening strategy that *"alleviates the computational burden of KKT post-convergence checking for the strong rules by not checking features that can be safely eliminated"*. However, our warm start strategy (active or strong) does not require post-processing steps.

Correlation Based Rule. Previous works in statistics have proposed various model-based screening methods to select important variables. Those methods discard variables with small correlation between the features and response variables. For instance Sure Independence Screening (SIS) by Fan and Lv (2008) reads: for a chosen critical threshold γ (such that the number of selected variables is smaller than a prescribed proportion of the features),

$$\text{If } \Omega_g^\circ(X_g^\top y) < \gamma \text{ then remove } X_g \text{ from the problem.}$$

It is a marginal oriented variable selection method and it is worth noting that SIS can be recast as a static sphere test in linear regression scenarios:

$$\text{If } \Omega_g^\circ(X_g^\top y) < \gamma = \lambda \left(1 - r\Omega_g^\circ(X_g)\right) \text{ then } \hat{\beta}_g^{(\lambda)} = 0 \,(\text{remove } X_g).$$

Other refinements can also be found in the literature such as iterative screening (ISIS) (Fan and Lv, 2008), that bears some similarities with dynamic sphere safe tests.

Working Set. To avoid having too conservative rules of elimination, similar to strong rules, working set algorithms are strategies that relax the safe rules. Considering constrained convex optimization problem, Johnson and Guestrin (2015) introduced Blitz, a meta algorithm based on the duality gap for sequentially prioritizing relevant constraints. We describe their rules in the case of the lasso and its dual where we seek to maximizes $\|y\|_2^2/2 - \|\lambda\theta - y\|_2^2/2$ for θ such that $\|X_j^\top \theta\|_\infty \leqslant 1$ for all j in $[p]$. Given a primal/dual feasible vector (β, θ), Blitz construct a working set by selecting only the features that satisfies:

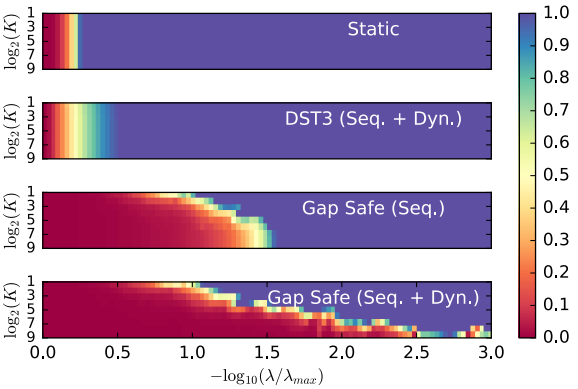
$$|X_j^\top \theta| + \|X_j\|_2 \sqrt{\frac{2}{\lambda^2}(1-\delta)^3 \operatorname{Gap}_\lambda(\beta, \theta)} \leqslant 1 \quad \text{for some } \delta \in [0, 1) \ . \tag{2.43}$$

Note that for $\delta = 0$, the working set rule (2.43) coincides with the gap safe sphere test. These rules therefore take a symmetrical strategy to ours by introducing a factor $(1-\delta)^3$ in order to aggressively eliminate more variables and optimizing onto a nested sequence of small constraint sets. Similar methods was later adopted in (Massias et al., 2017) and generalized in (Johnson and Guestrin, 2016).

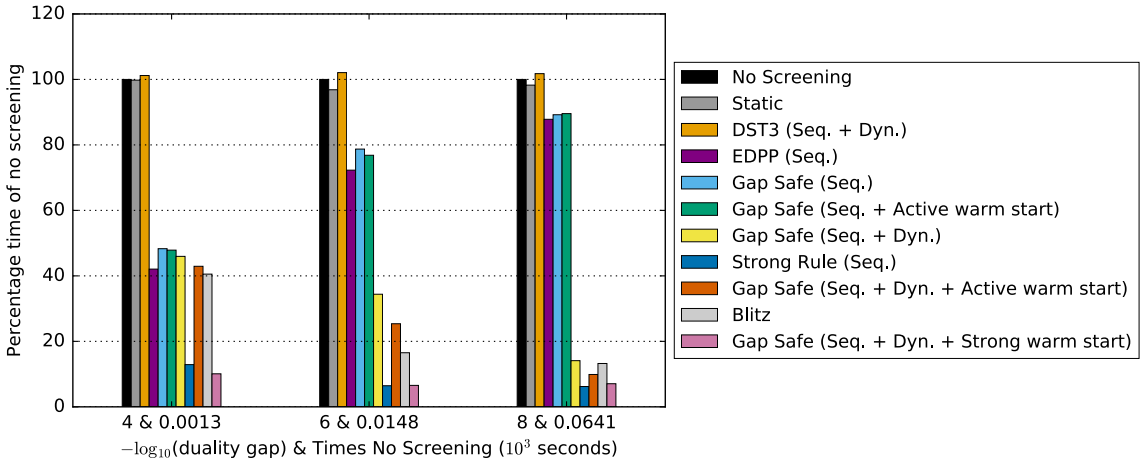
2.
The post-processing for the Lasso adds back variables violating the approximated KKT conditions

$$KKT_\epsilon : \begin{cases} |X_j^\top \theta| \leqslant 1 + \epsilon, & \text{if } \beta_j = 0, \\ |X_j^\top \theta - \operatorname{sign}(\beta_j)| \leqslant \epsilon, & \text{if } \beta_j \neq 0. \end{cases} \tag{2.42}$$

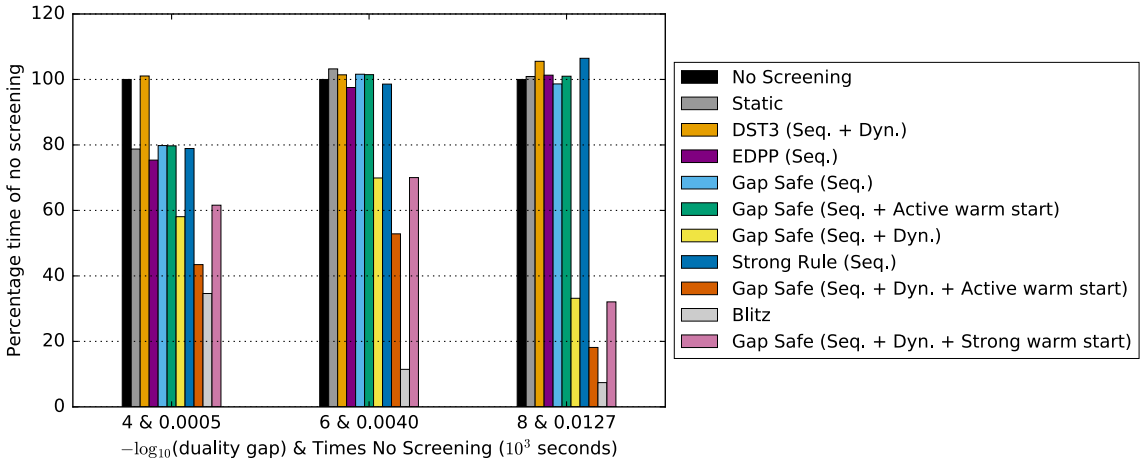
One can show that $\operatorname{Gap}_\lambda(\beta, \theta) \leqslant (1 - \lambda/\alpha)^2 \|y - X\beta\|^2/2 + \lambda\epsilon \|\beta\|_1$ where $\alpha = \max(\lambda, \left\|X^\top(y - X\beta)\right\|_\infty)$. Hence choosing $\epsilon = \epsilon'/P_\lambda(\beta) - (1 - \lambda/\alpha)^2$ imply an ϵ'-duality gap.



(a) Fraction of the variables that are active. Each line corresponds to a fixed number of iterations for which the algorithm is run.



(b) Dense grid with 100 values of λ.



(c) Coarse grid with 10 values of λ.

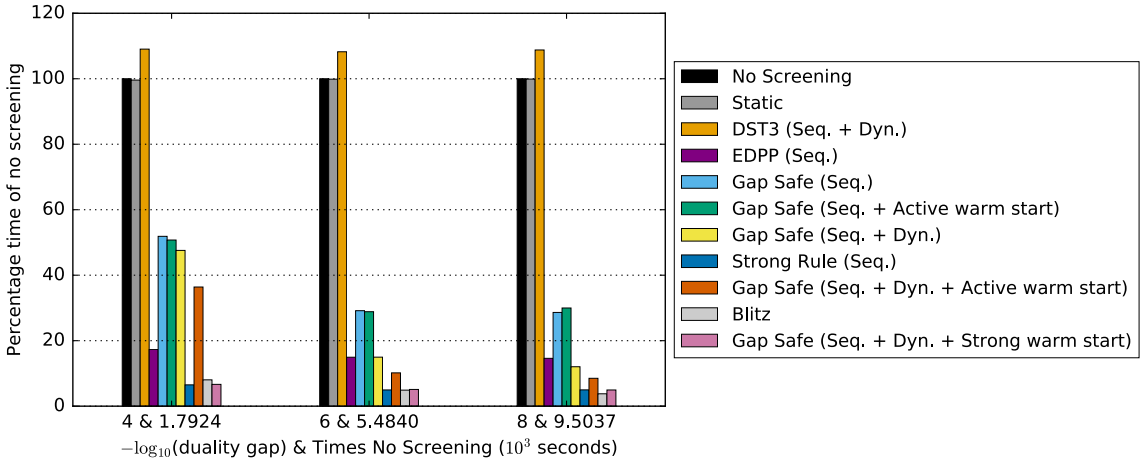
Figure 2.8 – Lasso on the Leukemia (dense data with $n = 72$ observations and $p = 7129$ features). Computation times needed to solve the Lasso regression path to desired accuracy for a grid of λ from $\lambda_{\max}$ to $\lambda_{\max}/10^3$.

2.5 Numerical Experiments

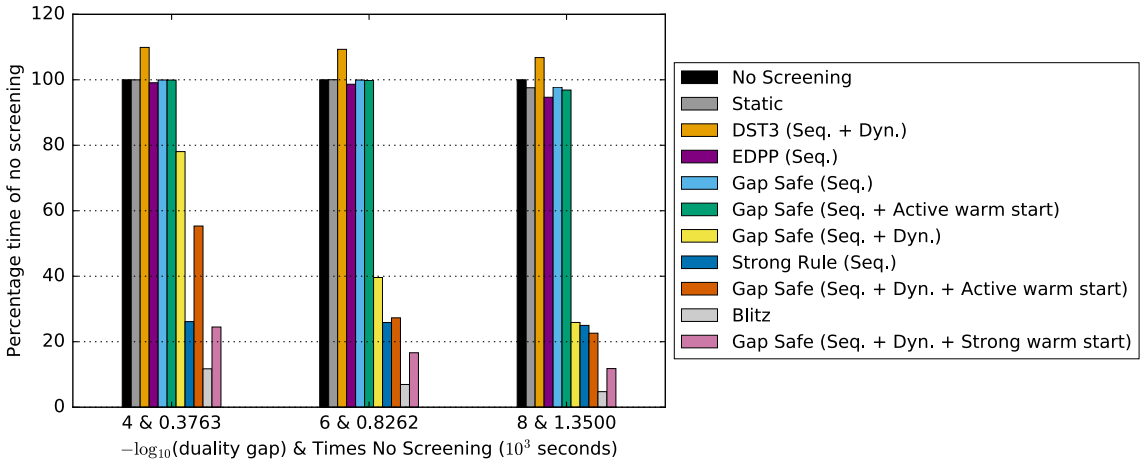
In this section we present results obtained with the Gap Safe rules on various data sets. Implementation has been done in Python and Cython (Behnel et al., 2011) for low level critical parts. A

coordinate descent algorithm is used with a scaled dual gap stopping criterion *i.e.* we normalize the targeted accuracy ϵ (in the stopping criterion) in order to have a running time that is independent from the data scaling, *i.e.* $\epsilon \leftarrow \epsilon \|y\|_2^2$ for the regression cases and $\epsilon \leftarrow \epsilon \min(n_1, n_2)/n$ where n_i is the number of observations in the class i, for the logistic cases.

Note that in the Lasso case, to compare our method with the un-safe *strong rules* by Tibshirani et al. (2012) and with the sequential screening rule such as the *eddp+* by Wang et al. (2015), we have added an approximated KKT post-processing step. We do this following Footnote 2, since they require the previous (exact) dual optimal solution which is not available in practice. The same limit holds true for the *TLFre* approach of Wang and Ye (2014) addressing the Sparse-Group Lasso formulation, as well as for the method explored by Lee and Xing (2014) to handle overlapping groups and *slores* by Wang et al. (2014) for the binary logistic regression. We have compared our method to various known safe screening rules (El Ghaoui et al., 2012; Xiang et al., 2011; Bonnefoy et al., 2014). For the Sparse-Group Lasso, such rules did not exist, so we have proposed natural extensions (Ndiaye et al., 2016) thanks to exact computation of the dual norm in Proposition 15. For the Lasso estimator, we have also compared our implementation with the *Blitz* algorithm (Johnson and Guestrin, 2015) which combines Gap Safe screening rules, *Prox-Newton* coordinate descent and an active set strategy.



(a) Dense grid with 100 values of λ.



(b) Sparse grid with 10 values of λ.

Figure 2.9 – Lasso on financial data E2006-log1p (sparse data with $n = 16087$ observations and $p = 1668737$ features). Computation times needed to solve the Lasso regression path to desired accuracy for a grid of λ from $\lambda_{\max}$ to $\lambda_{\max}/20$.

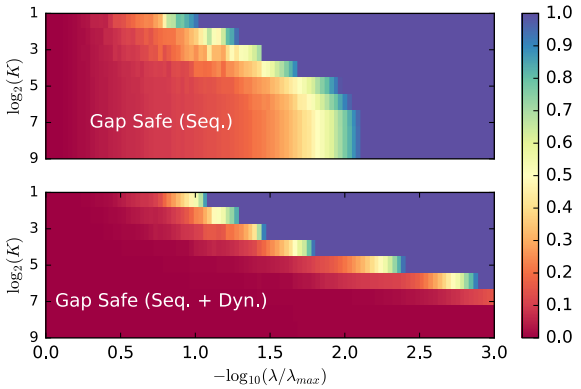
ℓ_1 Lasso Regression. We have evaluated the computing time for the Gap Safe rules with and without active warm start, and compared with the static rule El Ghaoui et al. (2012) and the refined dynamic rule DST3 by Xiang et al. (2011), as well as Bonnefoy et al. (2015). We used the classic dense data set Leukemia, and the large sparse financial data set E2006-log1p available from LIBSVM We have normalized the column of X and standardized y to have zero mean and unit variance. The experiments on Figure 2.8(a) focuses on the Leukemia data set. The screening performance for a fixed number of iterations, from 2 to 2^9, is investigated for each λ. It demonstrates that increasing the number of iterations benefits to the dynamic screening rule. Also, the closer the estimate is from the global minimum, the better the screening. This is inline with the results in running time in the benchmark on Figure 2.8(b). Note that the dynamic Gap Safe rule is the only rule that significantly improves the running time of the Lasso. Results presented in the financial data set in Figure 2.9 are inline with the results on Leukemia. We observe that the *Blitz* algorithm (Johnson and Guestrin, 2015), also achieves a significant speed-up with gains in the same order of magnitude than our dynamic Gap Safe implementation combined with active or strong warm start. One advantage of our approach though, is the simplicity to insert it in any iterative algorithm as shown in Algorithm 1 and 2.

To demonstrate the limitations of the strong rules, we report in Figure 2.8(c) results with a coarse grid with only 10 values of λ from $\lambda_{\max}$ to $\lambda_{\max}/10^3$ such that $2\lambda_t < \lambda_{t-1}$. The strong rules become then useless since the screening test (2.40) selects all variables, *i.e.* $\mathcal{STG}(\hat{\theta}^{(\lambda_{t-1})}, \lambda_t, \lambda_{t-1}) = \mathcal{G}$. Overall, the greater the gap between grid points, the lower the benefits of (active) warm start. In the experiment in Figure 2.9(b), we have stopped the grid at $\lambda_{\max}/20$ leading to a sparse solution with 1562 active variables. We obtain an important speed-up for both coarse and dense grids demonstrating the consistent efficiency of the active warm start strategy specially in a sparse regime. Finally, with an extremely coarse grid, we therefore recommend the active warm start with the previous safe active set (which performance is only affected through the initialization point) rather than the strong active set (*cf.* Figure 2.8(c)).

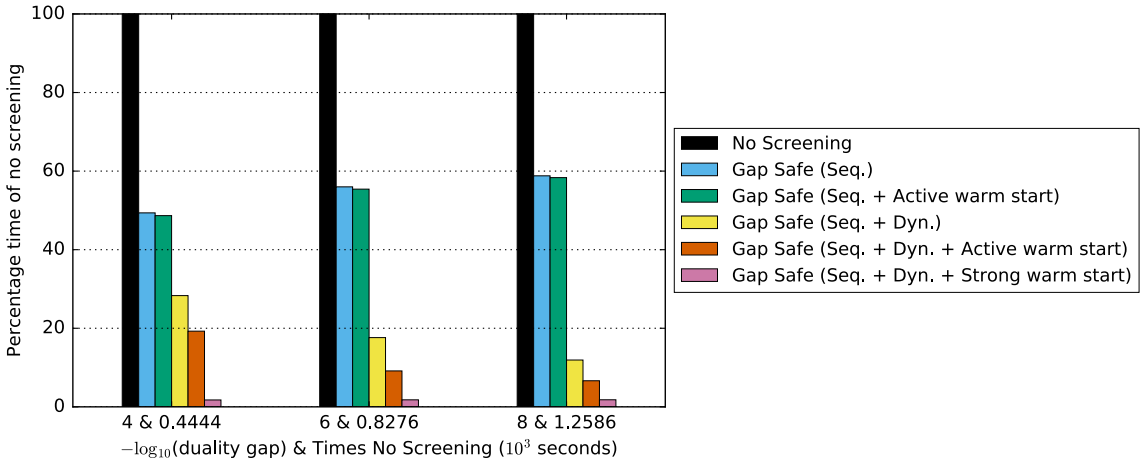
ℓ_1 Binary Logistic Regression. Results on the Leukemia data set for standard logistic regression are reported in Figure 2.10. We compare the dynamic strategy of Gap Safe to the sequential strategy. Results demonstrate the clear benefit of the dynamic rule in terms of high number of screened out variables. This is reflected in the graph of running times, which shows that dynamic Gap Safe rule with strong warm start can yield up to a $30\times$ speed-up compared to sequential rule and even more compared to an absence of screening (up to $50\times$ speed-up).

ℓ_1/ℓ_2 Multi-task Regression. To demonstrate the benefit of the Gap Safe screening rules for a multi-task Lasso problem we have considered neuroimaging data. Electroencephalography (EEG) and magnetoencephalography (MEG) are brain imaging modalities that allow to identify active brain regions. The problem to solve is a multi-task regression problem with squared loss where every task corresponds to a time instant. Using a multi-task Lasso one can constrain the recovered sources to be identical during a short time interval (Gramfort et al., 2012). This corresponds to a temporal stationary assumption. In this experiment we used a joint MEG/EEG data with 301 MEG and 59 EEG sensors leading to $n = 360$. The number of possible sources is $p = 22,494$ and the number of time instants is $q = 20$. With a $1\,$kHz sampling rate it is equivalent to say that the sources stay the same for $20\,$ms.

Results are presented in Figure 2.11. The Gap Safe rule is compared with the dynamic safe rule from Bonnefoy et al. (2015). Figure 2.11(a) shows the fraction of active variables. It demonstrates that the Gap Safe rule screens out much more variables than the competitors. Thanks to its converging nature, the more iterations are performed the more variables are screened out. On Figure 2.11(b), the computation time confirms the effective speed-up. We significantly improves the computation time for duality gap tolerances from 10^{-2} to 10^{-8}, especially when accurate



(a) Fraction of the variables that are active. Each line corresponds to a fixed number of iterations for which the algorithm is run.



(b) Computation times needed to solve the logistic regression path to desired accuracy with 100 values of λ from $\lambda_{\max}$ to $\lambda_{\max}/10^3$.

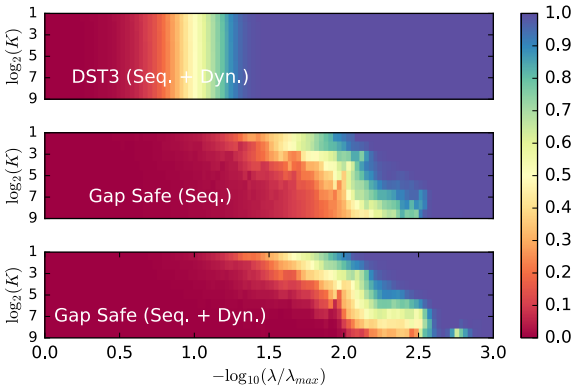
Figure 2.10 – ℓ_1 regularized binary logistic regression on the Leukemia (dense data with $n = 72$ observations and $p = 7129$ features). Sequential and full dynamic screening Gap Safe rules are compared.

estimates are required, *e.g.* for feature selection.

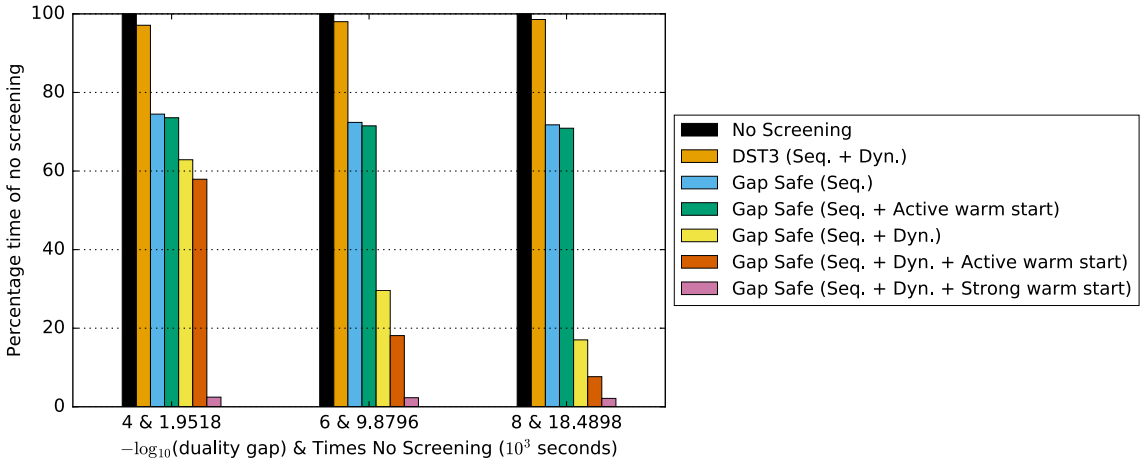
Sparse-Group Lasso Regression.

— Synthetic dataset: We use a common framework (Tibshirani et al., 2012; Wang and Ye, 2014) based on the model $y = X\beta + 0.01\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_n)$, $X \in \mathbb{R}^{n\times p}$ follows a multivariate normal distribution such that $\forall(i,j) \in [p]^2, \mathrm{corr}(X_i, X_j) = \rho^{|i-j|}$. We fix $n = 100$ and break randomly $p = 10000$ in 1000 groups of size 10 and select γ_1 groups to be active and the others are set to zero. In each selected groups, γ_2 coordinates are drawn with $[\beta_g]_j = \mathrm{sign}(\xi) \times U$ for U is uniform in $[0.5, 10]$), ξ uniform in $[-1, 1]$.

— `Real dataset: NCEP/NCAR Reanalysis 1` Kalnay et al. (1996): which contains monthly means of climate data measurements spread across the globe in a grid of $2.5^\circ \times 2.5^\circ$ resolutions (longitude and latitude 144×73) from $1948/1/1$ to $2015/10/31$. Each grid point constitutes a group of 7 predictive variables (*Air Temperature, Precipitable water, Relative humidity, Pressure, Sea Level Pressure, Horizontal Wind Speed* and *Vertical Wind Speed*) whose concatenation across time constitutes our design matrix $X \in \mathbb{R}^{814\times 73577}$. Such data have therefore a natural group structure, with seven features per group. As target variable $y \in \mathbb{R}^{814}$, we use the values of *Air Temperature* in



(a) Fraction of active variables as a function of λ and the number of iterations K. The Gap Safe strategy has a much longer range of λ with (red) small active sets.



(b) Computation time to reach convergence using different screening strategies. We have run the algorithm with 100 values of λ from $\lambda_{\max}$ to $\lambda_{\max}/10^3$.

Figure 2.11 – Experiments on MEG/EEG brain imaging data set (dense data with $n = 360$ observations, $p = 22494$ features and $q = 20$ time instants).

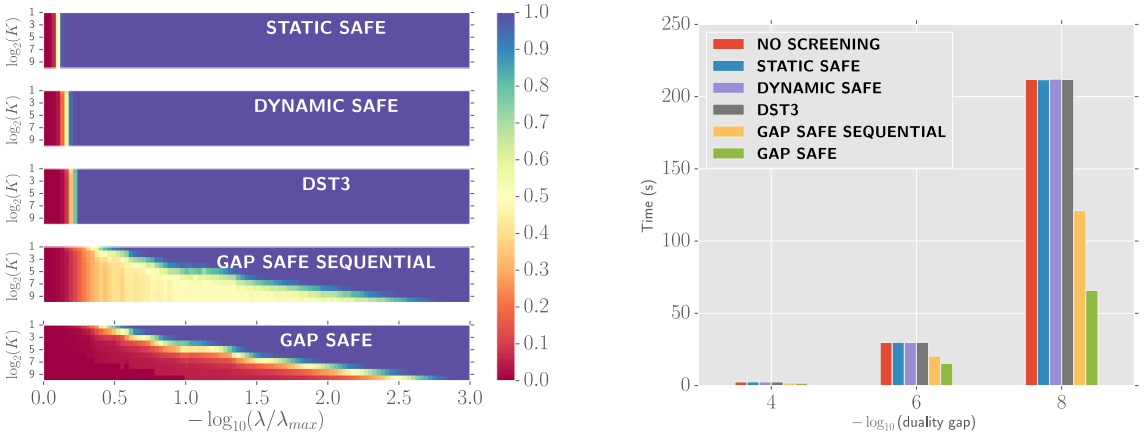


Figure 2.12 – Experiments on a synthetic dataset ($\rho = 0.5, \gamma_1 = 10, \gamma_2 = 4, \tau = 0.2$).
(a) Proportion of active variables, *i.e.* variables not safely eliminated, as a function of parameters (λ_t) and the number of iterations K. More red, means more variables eliminated and better screening. (b) Time to reach convergence w.r.t the accuracy on the duality gap, using various screening strategies.

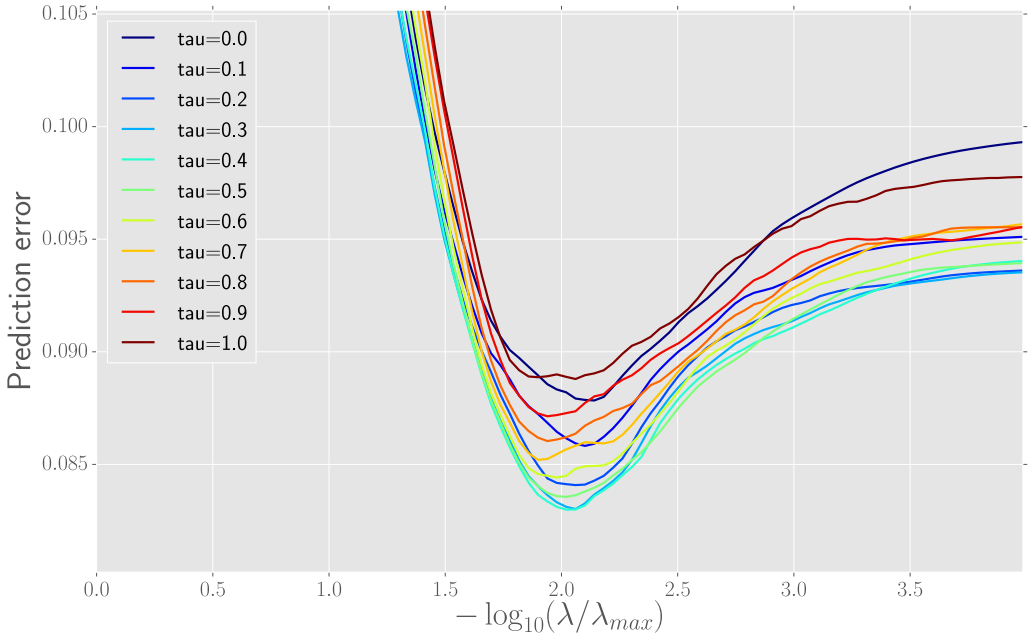


Figure 2.13 – Experiments on NCEP/NCAR Reanalysis 1 ($n = 814, p = 73577$): prediction error for the Sparse-Group Lasso path with 100 values of λ and 11 values of τ (best : $\tau^\star = 0.4$).

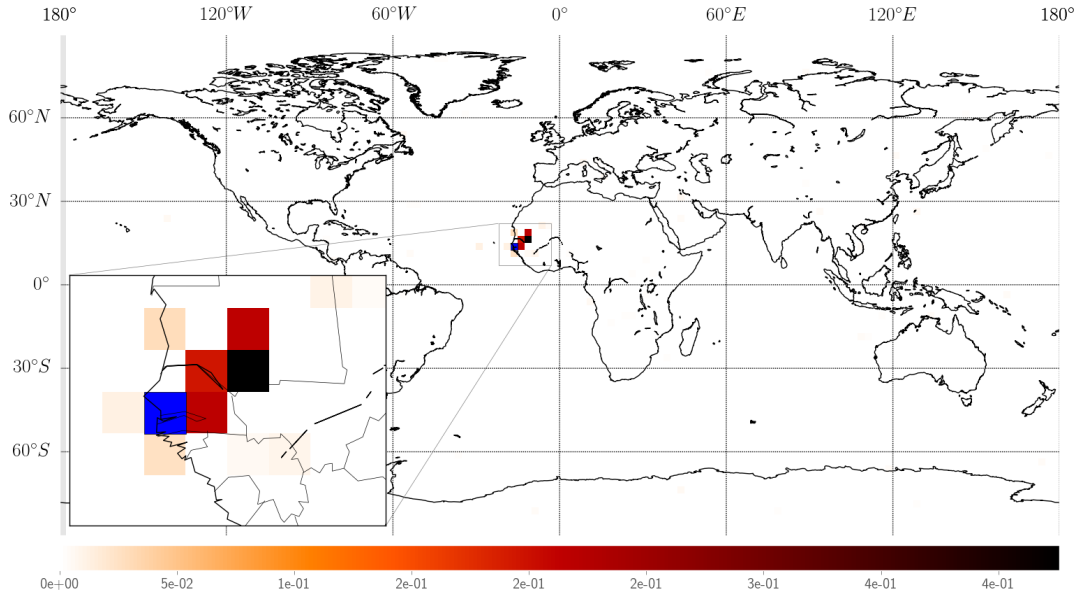
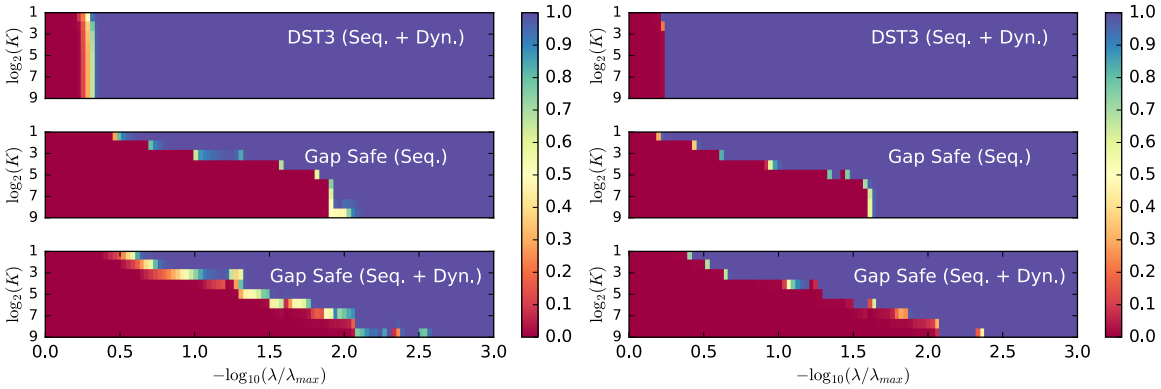


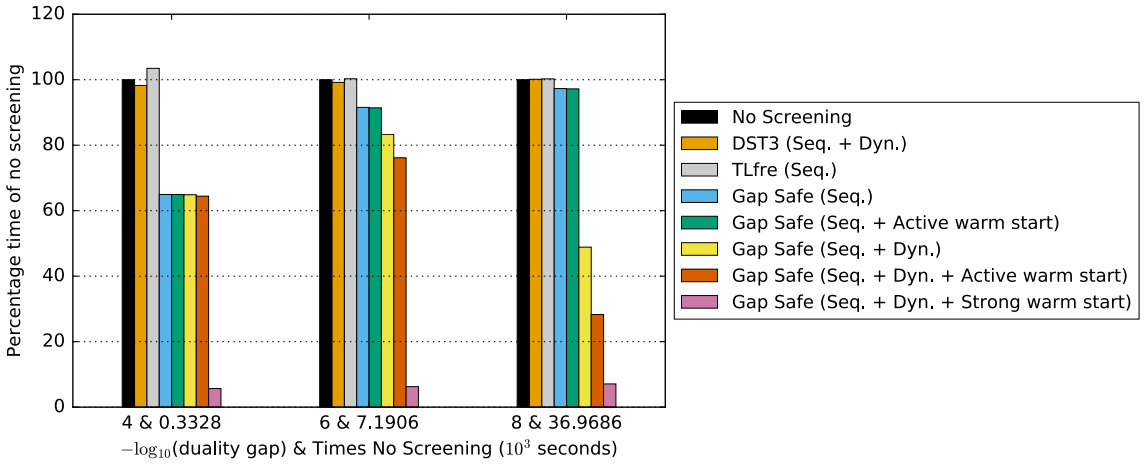
Figure 2.14 – Experiments on NCEP/NCAR Reanalysis 1 ($n = 814, p = 73577$): Active groups to predict Air Temperature in a neighborhood of Dakar (in blue). Cross validation was run over 100 values for λ's and 11 for τ's. At each location, the highest absolute value among the seven coefficients is displayed.

a neighborhood of Dakar. For preprocessing, we remove the seasonality (we center the data month by month) and the trend (we remove the linear trend obtained by least squares) present in the data set. We then standardize the data so that each feature has a variance of one. This preprocessing is usually done in climate analysis to prevent some bias in the regression estimates. Similar data have been used in the past by Chatterjee et al. (2012), demonstrating that the Sparse-Group Lasso estimator is well suited for prediction in such climatology applications. Indeed, thanks to the sparsity structure, the estimates delineate via their support some predictive regions at the group level, as well as predictive feature



(a) Proportion of active coordinate-wise variables as a function of parameters (λ_t) and the number of iterations K.

(b) Proportion of active group variables as a function of parameters (λ_t) and the number of iterations K.



(c) Time to reach convergence as a function of increasing prescribed accuracy, using various screening strategies and a logarithmic grid from $\lambda_{\max}$ to $\lambda_{\max}/10^{2.5}$.

Figure 2.15 – Sparse-Group Lasso experiments on climate data NCEP/NCAR Reanalysis 1 (dense data with $n = 814$ observations and $p = 73577$ features) with $\tau = 0.4$ chosen by cross-validation.

via coordinate-wise screening.

We choose the parameter τ in the set $\{0, 0.1, \ldots, 0.9, 1\}$ by splitting in half the observations, and run a training-test validation procedure. For each value of τ, we require a duality gap of 10^{-8} on the training set and pick the best one in term of prediction accuracy on the test set. Since the prediction error degrades increasingly for $\lambda \leqslant \lambda_{\max}/10^{-2.5}$, we fix $\delta = 2.5$. We have fixed the weight! $w_g = 1$ since all groups have the same size. The computational time benchmark is presented in Figure 2.15(c). Here also, we observe a significant gain by using a dynamic Gap Safe screening rule, which is further improved by the active warm start.

Conclusion

We have proposed a unified presentation of the Gap Safe screening rules for accelerating algorithms solving supervised learning problems under sparsity constraints. The proposed approach applies to many popular estimators that boil down to convex optimization problems where the data fitting term has a Lipschitz gradient and the regularization term is a separable sparsity enforcing function. We have shown that our methodology is more flexible than previously known safe rules as it conveniently unifies both regression and classification settings. The efficiency of the Gap Safe rules along with the new *active/strong warm start* strategies was demonstrated on multiple experiments using real high dimensional data set, suggesting that Gap Safe screening rules are always helpful to speed-up solvers targeting sparse regularization.

2.6 Appendix

Proposition 21 (Dual Norm for Separable Norm)**.** *Le Ω be separable norm $\Omega(\beta) = \sum_{g\in\mathcal{G}} \Omega_g(\beta_g)$, its dual norm can be expressed as $\Omega^D(\xi) = \max_{g\in\mathcal{G}} \Omega_g^D(\xi_g)$ where Ω_g^D is the dual norm of Ω_g.*

Proof. The definition of the dual norm reads $\Omega^D(\xi) = \max_{\beta:\Omega(\beta)\leqslant 1} \beta^\top \xi$. Then

$$
\begin{aligned}
\Omega^D(\xi) &= \sup_{\beta:\Omega(\beta)\leqslant 1} \langle \beta, \xi\rangle = \sup_{\beta} \inf_{\mu>0} \langle \beta, \sum_{g\in\mathcal{G}} \xi_g\rangle - \mu\left(\sum_{g\in\mathcal{G}} \Omega_g(\beta_g) - 1\right) \\
&= \inf_{\mu>0}\left\{\sum_{g\in\mathcal{G}} \sup_{\beta_g} \left[\langle \beta_g, \xi_g\rangle - \mu\Omega_g(\beta_g)\right] + \mu\right\} \\
&= \inf_{\mu>0}\left\{\sum_{g\in\mathcal{G}} \mu\Omega_g^*\left(\frac{\xi_g}{\mu}\right) + \mu\right\} = \inf_{\mu>0}\left\{\sum_{g\in\mathcal{G}} \iota_{\mathcal{B}_{\Omega_g^D}}\left(\frac{\xi_g}{\mu}\right) + \mu\right\} \\
&= \inf_{\mu>0}\left\{\max_{g\in\mathcal{G}} \iota_{\mathcal{B}_{\Omega_g^D}}\left(\frac{\xi_g}{\mu}\right) + \mu\right\} = \max_{g\in\mathcal{G}} \inf_{\mu>0}\left\{\Omega_g^*\left(\frac{\xi_g}{\mu}\right) + \mu\right\} \\
&= \max_{g\in\mathcal{G}} \inf_{\mu>0} \sup_{\beta_g} \langle \beta_g, \frac{\xi_g}{\mu}\rangle - \Omega_g(\beta_g) + \mu \\
&= \max_{g\in\mathcal{G}} \inf_{\mu>0} \sup_{u_g} \langle u_g, \xi_g\rangle - \mu(\Omega_g(u_g) - 1) \quad (\text{ with } \mu u_g = \beta_g) \\
&= \max_{g\in\mathcal{G}} \sup_{u_g:\Omega_g(u_g)\leqslant 1} \langle u_g, \xi_g\rangle = \max_{g\in\mathcal{G}} \Omega_g^D(\xi_g).
\end{aligned}
$$

□

Proposition 22 (Properties of Sparse-Group Lasso)**.** *For all groups g in $\mathcal{G}$, let us introduce*

$$
\epsilon_g := \frac{(1-\tau)w_g}{\tau + (1-\tau)w_g} \ .
$$

Then, the Sparse-Group Lasso norm satisfies the following properties for any β and ξ in $\mathbb{R}^p$,

$$
\begin{aligned}
\Omega(\beta) &= \sum_{g\in\mathcal{G}} (\tau + (1-\tau)w_g)\|\beta_g\|_{\epsilon_g}^D \\
\Omega^\circ(\xi) &= \Omega^D(\xi) = \max_{g\in\mathcal{G}} \frac{\|\xi_g\|_{\epsilon_g}}{\tau + (1-\tau)w_g}. \\
\Omega^*(\xi) &= \iota_{\mathcal{B}_{\Omega^D}}(\xi) = \sum_{g\in\mathcal{G}} \iota_{\mathcal{B}}\left(\frac{\mathrm{ST}_\tau(\xi_g)}{(1-\tau)w_g}\right) \\
\partial\Omega(\beta) &= \left\{\xi \in \mathbb{R}^p : \forall g \in \mathcal{G},\ \xi_g \in \tau\partial\|\cdot\|_1(\beta_g) + (1-\tau)w_g\partial\|\cdot\|_2(\beta_g)\right\}.
\end{aligned}
$$

Proof. For all β in $\mathbb{R}^p$, we have

$$
\begin{aligned}
\Omega(\beta) &= \tau\|\beta\|_1 + (1-\tau)\sum_{g\in\mathcal{G}} w_g\|\beta_g\| = \sum_{g\in\mathcal{G}} \left(\tau\|\beta_g\|_1 + (1-\tau)w_g\|\beta_g\|\right) \\
&= \sum_{g\in\mathcal{G}} (\tau + (1-\tau)w_g)\left[\frac{\tau}{\tau + (1-\tau)w_g}\|\beta_g\|_1 + \frac{(1-\tau)w_g}{\tau + (1-\tau)w_g}\|\beta_g\|\right] \\
&= \sum_{g\in\mathcal{G}} (\tau + (1-\tau)w_g)\|\beta_g\|_{\epsilon_g}^D.
\end{aligned}
$$

The Proposition 21 and the separability of Ω yields

$$\begin{aligned}\Omega^D(\xi) &= \max_{g\in\mathcal{G}} \Omega_g^D(\xi_g) = \max_{g\in\mathcal{G}} \sup_{u_g:\Omega_g(u_g)\leqslant 1} \langle u_g, \xi_g\rangle \\ &= \max_{g\in\mathcal{G}} \sup_{u_g} \langle u_g, \xi_g\rangle \quad \text{s.t. } (\tau+(1-\tau)w_g)\,\|u_g\|_{\epsilon_g}^D \leqslant 1 \\ &= \max_{g\in\mathcal{G}} \sup_{u_g:\Omega_g(u_g)\leqslant 1} \langle u_g, \xi_g\rangle = \max_{g\in\mathcal{G}} \sup_{u'_g:\|u'_g\|_{\epsilon_g}^D\leqslant 1} \frac{{u'_g}^\top \xi_g}{\tau+(1-\tau)w_g} = \max_{g\in\mathcal{G}} \frac{\|\xi_g\|_{\epsilon_g}}{\tau+(1-\tau)w_g}.\end{aligned}$$

We recall the proof of Wang and Ye (2014) for the expression of the Fenchel transform of Ω. First let us write $\Omega(\beta) = \Omega_1(\beta) + \Omega_2(\beta)$, where $\Omega_1(\beta) = \tau\|\beta\|_1$ and $\Omega_2(\beta) = (1-\tau)\sum_{g\in\mathcal{G}} w_g\|\beta_g\|_2$. Since Ω_1 and Ω_2 are continuous everywhere, we have (see (Hiriart-Urruty, 2006, Theorem 1)):

$$\Omega^*(\xi) = (\Omega_1+\Omega_2)^*(\xi) = \min_{a+b=\xi}[\Omega_1^*(a)+\Omega_2^*(b)] = \min_a[\Omega_1^*(a)+\Omega_2^*(\xi-a)]\ ,$$

which is also the inf-convolution (see (Bauschke and Combettes, 2011, Chapter 12)) of these two norms. Using the Fenchel conjugate of the ℓ_1 norm ($\Omega_1^* = \iota_{\tau\mathcal{B}_\infty}$) and of the ℓ_2 norm ($\Omega_2^* = \iota_{\mathcal{B}}$), we have

$$\Omega^*(\xi) = \sum_{g\in\mathcal{G}} \min_{a_g} \iota_{\tau\mathcal{B}_\infty}(a_g) + \iota_{\mathcal{B}}\left(\frac{\xi_g - a_g}{(1-\tau)w_g}\right) = \sum_{g\in\mathcal{G}} \iota_{\mathcal{B}}\left(\frac{\xi_g - \Pi_{\tau\mathcal{B}_\infty}(\xi_g)}{(1-\tau)w_g}\right).$$

Hence the indicator of the unit dual ball is $\iota_{\mathcal{B}_{\Omega^D}}(\xi) = \sum_{g\in\mathcal{G}} \iota_{(1-\tau)w_g\mathcal{B}}\left(\xi_g - \Pi_{\tau\mathcal{B}_\infty}(\xi_g)\right)$ and using $\mathrm{ST}_\tau(\xi_g) = \xi_g - \Pi_{\tau\mathcal{B}_\infty}$, we have:

$$\mathcal{B}_{\Omega^D} = \left\{\xi\in\mathbb{R}^p : \Omega^D(\xi)\leqslant 1\right\} = \left\{\xi\in\mathbb{R}^p : \forall g\in\mathcal{G},\ \|\mathrm{ST}_\tau(\xi_g)\| \leqslant (1-\tau)w_g\right\}.$$

□

Two Level of Active Set Convergence for Sparse-Group Lasso

Proposition 23. *Let $(\mathcal{R}_k)_{k\in\mathbb{N}}$ be a sequence of safe regions whose diameters converge to 0. Then, $\lim_{k\to\infty} \mathcal{A}_{gp}(\mathcal{R}_k) = \mathcal{E}_{gp}$ and $\lim_{k\to\infty} \mathcal{A}_{ft}(\mathcal{R}_k) = \mathcal{E}_{ft}$.*

Proof. We proceed by double inclusion. First let us prove that $\exists k_0$ s.t. $\forall k\geqslant k_0, \mathcal{A}_{\mathrm{gp}}(\mathcal{R}_k)\subset\mathcal{E}_{\mathrm{gp}}$. Indeed, since the diameter of $\mathcal{R}_k$ converges to zero, for any $\epsilon > 0$ there exist $k_0\in\mathbb{N}, \forall k\geqslant k_0, \forall\theta\in\mathcal{R}_k, \|\theta-\hat\theta^{(\lambda,\Omega)}\|\leqslant\epsilon$. The triangle inequality implies that

$$\forall g\notin\mathcal{E}_{\mathrm{gp}},\quad \|\mathrm{ST}_\tau(X_g^\top\theta)\| \leqslant \|\mathrm{ST}_\tau(X_g^\top\theta) - \mathrm{ST}_\tau(X_g^\top\hat\theta^{(\lambda,\Omega)})\| + \|\mathrm{ST}_\tau(X_g^\top\hat\theta^{(\lambda,\Omega)})\|.$$

Since the soft-thresholding operator is 1-Lipschitz, we have:

$$\left\|\mathrm{ST}_\tau(X_g^\top\theta)\right\| \leqslant \left\|X_g(\theta-\hat\theta^{(\lambda,\Omega)})\right\| + \left\|\mathrm{ST}_\tau(X_g^\top\hat\theta^{(\lambda,\Omega)})\right\| \leqslant \epsilon\,\|X_g\| + \left\|\mathrm{ST}_\tau(X_g^\top\hat\theta^{(\lambda,\Omega)})\right\|,$$

as soon as $k\geqslant k_0$. Moreover, $\forall g\notin\mathcal{E}_{\mathrm{gp}}$,

$$\left\|\mathrm{ST}_\tau(X_g^\top\theta)\right\| \leqslant \max_{g\notin\mathcal{E}_{\mathrm{gp}}}\left\|\mathrm{ST}_\tau(X_g^\top\theta)\right\| \leqslant \epsilon\max_{g\notin\mathcal{E}_{\mathrm{gp}}}\|X_g\| + \max_{g\notin\mathcal{E}_{\mathrm{gp}}}\left\|\mathrm{ST}_\tau(X_g^\top\hat\theta^{(\lambda,\Omega)})\right\|.$$

It suffices to choose ϵ such that

$$\epsilon\max_{g\notin\mathcal{E}_{\mathrm{gp}}}\|X_g\| + \max_{g\notin\mathcal{E}_{\mathrm{gp}}}\left\|\mathrm{ST}_\tau(X_g^\top\hat\theta^{(\lambda,\Omega)})\right\| < (1-\tau)w_g,$$

that is to say $\epsilon < \frac{(1-\tau)w_g - \max_{g\notin\mathcal{E}_{\text{gp}}}\|\mathrm{ST}_\tau(X_g^\top\hat{\theta}^{(\lambda,\Omega)})\|}{\max_{g\notin\mathcal{E}_{\text{gp}}}\|X_g\|}$, to remove the group g. Then for any $k \geqslant k_0$, $\mathcal{E}_{\text{gp}}^c = \{g \in \mathcal{G} : \|\mathcal{S}_\tau(X_g^\top\hat{\theta}^{(\lambda)})\| < (1-\tau)w_g\} \subset \mathcal{A}_{\text{gp}}(\mathcal{R}_k)^c$, the set of variables removed by our screening rule. This proves the first inclusion.

Now we show that $\forall k \in \mathbb{N}, \mathcal{A}_{\text{gp}}(\mathcal{R}_k) \supset \mathcal{E}_{\text{gp}}$. Indeed, for all $g^\star \in \mathcal{E}_{\text{gp}}$, $\|\mathrm{ST}_\tau(X_{g^\star}^T\hat{\theta}^{(\lambda,\Omega)})\| = (1-\tau)w_{g^\star}$. Since for all k in $\mathbb{N}$, $\hat{\theta}^{(\lambda,\Omega)} \in \mathcal{R}_k$ then $\max_{\theta\in\mathcal{R}_k}\|\mathrm{ST}_\tau(X_g^\top\theta)\| \geqslant \|\mathrm{ST}_\tau(X_{g^\star}^T\hat{\theta}^{(\lambda,\Omega)})\| = (1-\tau)w_{g^\star}$ hence the second inclusion holds. We have shown that $\forall k \geqslant k_0$, $\mathcal{A}_{\text{gp}}(\mathcal{R}_k) = \mathcal{E}_{\text{gp}}$ and so $\mathcal{A}_{\text{ft}}(\mathcal{R}_k) \subset \bigcup_{g\in\mathcal{E}_{\text{gp}}}\{j \in g : \max_{\theta\in\mathcal{R}_k}|X_j^\top\theta| \geqslant \tau\}$. Moreover, the same reasoning yields $\forall g \in \mathcal{G}$, $\{j \in g : \max_{\theta\in\mathcal{R}_k}|X_j^\top\theta| \geqslant \tau\} \subset \{j \in g : |X_j^\top\hat{\theta}^{(\lambda,\Omega)}| \geqslant \tau\}$. Hence $\forall k \geqslant k_0, \mathcal{A}_{\text{ft}}(\mathcal{R}_k) \subset \mathcal{A}_{\text{ft}}$. The reciprocal inclusion is straightforward. □

Exact Computation of the Dual Norm of Sparse-Group Lasso

Proposition 24. *For $\alpha \in [0,1], R \geqslant 0$ and $x \in \mathbb{R}^d$, the equation $\sum_{i=1}^d \mathrm{ST}_{\nu\alpha}(x_i)^2 = (\nu R)^2$ has a unique solution $\nu := \Lambda(x,\alpha,R) \in \mathbb{R}_+$, that can be computed in $O(d\log d)$ operations in the worst case. With $n_I = \mathrm{Card}\{i \in [d] : |x_i| > \alpha\|x\|_\infty/(\alpha+R)\}$, the complexity of Algorithm 3 is $n_I + n_I\log(n_I)$, which is comparable to the ambient dimension d.*

Proof. Dividing by ν^2, which is positive as soon as $x \neq 0$, we get that $\sum_{j=1}^d \mathrm{ST}_{\nu\alpha}(x_j)^2 = (\nu R)^2$ is equivalent to $\sum_{j=1}^d \mathrm{ST}_\alpha(x_j/\nu)^2 = R^2$. Note that $\sum_{j=1}^d \mathrm{ST}_\alpha(x_j/\nu)^2 = \sum_{j=1}^d \mathrm{ST}_\alpha(|x_j|/\nu)^2$ so without loss of generality we assume $x \in \mathbb{R}_+^d$.

The case $\alpha = 0$ and $R = 0$ corresponds to the situation where all x_j are equal to zero or we impose ν equals to infinity. So we avoid this trivial case.

If $\alpha = 0$ and $R \neq 0$, $\nu = \|x\|/R$. Indeed,

$$\sum_{j=1}^d \mathrm{ST}_0(x_j/\nu)^2 = R^2 \iff \sum_{j=1}^d (x_j/\nu)^2 = R^2 \iff \frac{\|x\|_2^2}{\nu^2} = R^2 \text{ hence the result.}$$

If $\alpha \neq 0$ and $R = 0$, we have :

$$\begin{aligned}\sum_{j=1}^d \mathrm{ST}_\alpha\left(\frac{x_j}{\nu}\right)^2 = 0 &\iff \forall j \in [d], \left(\frac{x_j}{\nu} - \alpha\right)_+ = 0 \\ &\iff \forall j \in [d], \frac{x_j}{\nu} \leqslant \alpha \\ &\iff \nu \geqslant \frac{\max_{j\in[d]} x_j}{\alpha}.\end{aligned}$$

So we choose the smallest ν *i.e.* $\nu = \|x\|_\infty/\alpha$. In all the above cases, the computation is done in $O(d)$.

Otherwise $\alpha \neq 0$ and $R \neq 0$. The function $\nu \mapsto \sum_{j=1}^d \mathrm{ST}_\alpha(x_j/\nu)^2$ is a non-increasing continuous function with limit $+\infty$ (resp. 0) when $\nu \to 0$ (resp. $\nu \to +\infty$). Hence, there is a unique solution to $\sum_{j=1}^d \mathrm{ST}_\alpha(x_j/\nu)^2 = R^2$.

We denote by $x_{(1)}, \ldots, x_{(d)}$ the coordinates of x ordered in decreasing order (with the convention $x_{(0)} = +\infty$ and $x_{(d+1)} = 0$). Note that $\sum_{j=1}^d \mathrm{ST}_\alpha(x_j/\nu)^2 = \sum_{j=1}^d \mathrm{ST}_\alpha(x_{(j)}/\nu)^2$. Then, there exists an index $j_0 \in [p]$ such that

$$R^2 \in \left[\sum_{j=0}^d \mathrm{ST}_\alpha\left(\alpha\frac{x_{(j)}}{x_{(j_0)}}\right)^2, \sum_{j=0}^d \mathrm{ST}_\alpha\left(\alpha\frac{x_{(j)}}{x_{(j_0+1)}}\right)^2\right). \tag{2.44}$$

For such a j_0, one can check that $\nu \in (x_{(j_0+1)}/\alpha, x_{(j_0)}/\alpha]$. The definition of the soft-thresholding operator yields

$$\mathrm{ST}_\alpha(x_j/\nu)^2 = \begin{cases} (x_j/\nu - \alpha)^2 & \text{if } x_j \geqslant \nu\alpha, \\ 0 & \text{if } x_j < \nu\alpha. \end{cases} \tag{2.45}$$

It can be simplified thanks to $x_j \geqslant x_{(j_0)} \Rightarrow x_j \geqslant \nu\alpha$ and $x_j \leqslant x_{(j_0+1)} \Rightarrow x_j < \nu\alpha$. Hence, $R^2 = \sum_{j=1}^{j_0}(x_{(j)}/\nu - \alpha)^2 = \sum_{j=1}^{j_0}(x_{(j)}/\nu)^2 + \alpha^2 \sum_{j=1}^{j_0} 1 - 2\alpha \sum_{j=1}^{j_0} x_{(j)}/\nu$ so solving $\sum_{j=1}^{p} \mathrm{ST}_\alpha(x_{(j)}/\nu)^2 = R^2$ is equivalent to solve on $\mathbb{R}_+$

$$(\alpha^2 j_0 - R^2)\nu^2 - \left(2\alpha \sum_{j=1}^{j_0} x_{(j)}\right)\nu + \sum_{j=1}^{j_0} x_{(j)}^2 = 0. \tag{2.46}$$

If $(\alpha^2 j_0 - R^2) = 0$, then $\nu = \sum_{j=1}^{j_0} x_{(j)}^2/(2\alpha \sum_{j=1}^{j_0} x_{(j)})$. Otherwise ν is the unique solution lying in $(x_{(j_0+1)}/\alpha, x_{(j_0)}/\alpha]$ of the quadratic equation stated in Eq. (2.46).

In the worst case, to compute $\Lambda(x, \alpha, R)$, one needs to sort a vector of size d, what can be done in $O(d\log(d))$ operations, and finding j_0 thanks to (2.44) requires $O(d^2)$ if we apply a naive algorithm.

In the following, we show that one can easily reduce the complexity to $O(d\log(d))$ in the worst case.

For all j in $[d]$, $\mathrm{ST}_\alpha\left(\alpha\frac{x_j}{x_{j_0}}\right) = 0$ as soon as $x_j \leqslant x_{j_0}$. This implies that (2.44) is equivalent to

$$R^2 \in \left[\sum_{j=0}^{j_0-1} \mathrm{ST}_\alpha\left(\alpha\frac{x_{(j)}}{x_{(j_0)}}\right)^2, \sum_{j=0}^{j_0} \mathrm{ST}_\alpha\left(\alpha\frac{x_{(j)}}{x_{(j_0+1)}}\right)^2\right). \tag{2.47}$$

Denoting $S_{j_0} := \sum_{j=1}^{j_0} x_{(j)}$ and $S_{j_0}^{(2)} := \sum_{j=1}^{j_0} x_{(j)}^2$, a direct calculation show that (2.47) can be rewritten as

$$R^2 \in \alpha^2\left[\frac{S_{j_0-1}^{(2)}}{x_{(j_0)}^2} - 2\frac{S_{j_0-1}}{x_{(j_0)}} + j_0,\ \frac{S_{j_0}^{(2)}}{x_{(j_0+1)}^2} - 2\frac{S_{j_0}}{x_{(j_0+1)}} + j_0 + 1\right). \tag{2.48}$$

Finally, solving $\sum_{j=1}^{p} \mathrm{ST}_\alpha(x_{(j)}/\nu)^2 = R^2$ is equivalent to finding the solution of the equation $(\alpha^2 j_0 - R^2)\nu^2 - (2\alpha S_{j_0})\nu + S_{j_0}^{(2)} = 0$ lying in $(x_{(j_0+1)}/\alpha, x_{(j_0)}/\alpha]$. Hence,

$$\Lambda(x, \alpha, R) = \frac{\alpha S_{j_0} - \sqrt{\alpha^2 S_{j_0}^2 - S_{j_0}^{(2)}(\alpha^2 j_0 - R^2)}}{\alpha^2 j_0 - R^2} =: \nu_1 \tag{2.49}$$

or

$$\Lambda(x, \alpha, R) = \frac{\alpha S_{j_0} + \sqrt{\alpha^2 S_{j_0}^2 - S_{j_0}^{(2)}(\alpha^2 j_0 - R^2)}}{\alpha^2 j_0 - R^2} =: \nu_2. \tag{2.50}$$

— If $\alpha^2 j_0 - R^2 < 0$, then $\nu_2 < 0$ and so it cannot be a solution since $\Lambda(x, \alpha, R)$ must be positive.

— Otherwise, we have

$$\nu_2 \geqslant \frac{\alpha S_{j_0}}{\alpha^2 j_0 - R^2} = \frac{1}{\alpha(j_0 - \frac{R^2}{\alpha^2})} \sum_{j=1}^{j_0} x_{(j)} > \frac{1}{\alpha j_0} \sum_{j=1}^{j_0} x_{(j)} \geqslant \frac{x_{(j_0)}}{\alpha},$$

where the second inequality results from the fact that $j_0 > j_0 - R^2/\alpha^2$. And again ν_2 cannot be a solution since $\Lambda(x, \alpha, R)$ belongs to $(x_{(j_0+1)}/\alpha, x_{(j_0)}/\alpha]$.

Hence, in all cases, the solution is given by ν_1.

Moreover, we can significantly reduce the cost of the sorting. Indeed, for all ν, we have

$$\|\mathrm{ST}_{\alpha\nu}(x)\| \geqslant \|\mathrm{ST}_{\alpha\nu}(x)\|_\infty = \max_{j\in[d]}(|x_j| - \nu\alpha)_+.$$

Hence, $\|\mathrm{ST}_{\alpha\nu}(x)\| - \nu R \geqslant \|x\|_\infty - \nu\alpha - \nu R > 0$ if and only if $\nu < \|x\|_\infty/(\alpha + R)$. Combining this with Equation (2.45), we take into account only the coordinates which have an absolute value greater than $\alpha\|x\|_\infty/(\alpha + R)$. Indeed, by contrapositive, if ν is a solution then $\nu \geqslant \|x\|_\infty/\alpha + R$ hence $x_j < \alpha\|x\|_\infty/\alpha + R \Rightarrow x_j < \nu\alpha \overset{(2.45)}{\Rightarrow} \mathrm{ST}_\alpha(x_j/\nu) = 0$.

Finally, computing $\Lambda(x, \alpha, R)$ can be done by applying Algorithm 3. Note that this algorithm is similar to (Burdakov and Merkulov, 2001, Algorithm 4). □

Properties of the ϵ-norm

We describe, for completeness, some properties of the ϵ-norm. The following material is inspired by Burdakov and Merkulov (2001).

Lemma 7. *For all $\xi \in \mathbb{R}^d$, the ϵ-decomposition reads: $\xi = \xi^\epsilon + \xi^{1-\epsilon}$, where $\xi^\epsilon := \mathrm{ST}_{(1-\epsilon)\|\xi\|_\epsilon}(\xi)$ and $\xi^{1-\epsilon} := \xi - \xi^\epsilon$. Moreover, $\|\xi^\epsilon\| = \epsilon\|\xi\|_\epsilon$ and $\left\|\xi^{1-\epsilon}\right\|_\infty = (1-\epsilon)\|\xi\|_\epsilon$. Hence, the following decomposition holds for the ϵ-norm: $\|\xi\|_\epsilon = \|\xi^\epsilon\| + \left\|\xi^{1-\epsilon}\right\|_\infty$.*

Proof. $\|\xi^\epsilon\| = \|\mathrm{ST}_{(1-\epsilon)\|\xi\|_\epsilon}(\xi)\| = \epsilon\|\xi\|_\epsilon$ by definition of the ϵ-norm $\|\xi\|_\epsilon$. Moreover,

$$\xi^{1-\epsilon} = \sum_{i=1}^d [\xi_i - \mathrm{sign}(\xi_i)(|\xi_i| - (1-\epsilon)\|\xi\|_\epsilon)_+] = \sum_{i=1}^d \mathrm{sign}(\xi_i)\left[|\xi_i| - (|\xi_i| - (1-\epsilon)\|\xi\|_\epsilon)_+\right].$$

Thus, using the symbol $a \vee b$ to represent $\max(a, b)$, one has

$$\begin{aligned}
\left\|\xi^{1-\epsilon}\right\|_\infty &= \max_{i\in[d]} |\mathrm{sign}(\xi_i)\left[|\xi_i| - (|\xi_i| - (1-\epsilon)\|\xi\|_\epsilon)_+\right]| \\
&= \max_{\substack{i\in[d] \\ |\xi_i|\leqslant(1-\epsilon)\|\xi\|_\epsilon}} ||\xi_i| - (|\xi_i| - (1-\epsilon)\|\xi\|_\epsilon)_+| \vee \max_{\substack{i\in[d] \\ |\xi_i|>(1-\epsilon)\|\xi\|_\epsilon}} ||\xi_i| - (|\xi_i| - (1-\epsilon)\|\xi\|_\epsilon)_+| \\
&= \max_{\substack{i\in[d] \\ |\xi_i|\leqslant(1-\epsilon)\|\xi\|_\epsilon}} |\xi_i| \vee (1-\epsilon)\|\xi\|_\epsilon = (1-\epsilon)\|\xi\|_\epsilon .
\end{aligned}$$

□

Lemma 8. *Define the sets*

$$\begin{aligned}
U(\|\xi\|_\epsilon) &:= \{u \in \mathbb{R}^d : \|u\| \leqslant \epsilon\|\xi\|_\epsilon\} , \\
V(\|\xi\|_\epsilon) &:= \{v \in \mathbb{R}^d : \|v\|_\infty \leqslant (1-\epsilon)\|\xi\|_\epsilon\} ,
\end{aligned}$$

we have

$$\xi^{(1-\epsilon)} = \underset{\substack{u\in U(\|\xi\|_\epsilon) \\ \xi=u+v}}{\arg\min} \|v\|_\infty \text{ and } \xi^\epsilon = \underset{\substack{v\in V(\|\xi\|_\epsilon) \\ \xi=u+v}}{\arg\min} \|u\| .$$

Proof.

• **Existence and uniqueness of the solutions**

It is clear that

$$\underset{\substack{u\in U(\|\xi\|_\epsilon) \\ \xi=u+v}}{\arg\min} \|v\|_\infty = \underset{\xi - U(\|\xi\|_\epsilon)}{\arg\min} \|v\|_\infty,$$

and

$$\underset{\substack{v\in V(\|\xi\|_\epsilon)\\ \xi=u+v}}{\arg\min}\|u\| = \underset{\xi-V(\|\xi\|_\epsilon)}{\arg\min}\|u\|.$$

Thus, these two problems have unique solution because we minimize strict convex functions onto convex sets.

• **Uniqueness of the ϵ-decomposition**

From Lemma 7 we have $\xi = \xi^\epsilon + \xi^{1-\epsilon}$ where $\|\xi^\epsilon\| = \epsilon\|\xi\|_\epsilon$ and $\|\xi^{1-\epsilon}\|_\infty = (1-\epsilon)\|\xi\|_\epsilon$. Hence $\xi^\epsilon \in U(\|\xi\|_\epsilon)$ and $\xi^{1-\epsilon} \in V(\|\xi\|_\epsilon)$. Now it suffices to show that this ϵ-decomposition is unique.

Suppose $\xi \neq 0$ (the uniqueness is trivial otherwise) and $v \in V(\|\xi\|_\epsilon)$. We show that for any $u \in \mathbb{R}^d$ such that $\xi = u + v$, $v \neq \xi^{1-\epsilon}$ implies $u \notin U(\|\xi\|_\epsilon)$.

$$\|u\|^2 = \|\xi - v\|^2 = \left\|\xi^\epsilon + (\xi^{1-\epsilon} - v)\right\|^2 = \|\xi^\epsilon\|^2 + 2\langle\xi^\epsilon, \xi^{1-\epsilon} - v\rangle + \left\|\xi^{1-\epsilon} - v\right\|^2,$$

hence $\|u\|^2 > \epsilon^2\|\xi\|_\epsilon^2 + 2\langle\xi^\epsilon,\, \xi^{1-\epsilon} - v\rangle$ because $\|\xi^\epsilon\| = \epsilon\|\xi\|_\epsilon$ and $\|\xi^{1-\epsilon} - v\| > 0$ ($v \neq \xi^{1-\epsilon}$). Moreover,

$$\begin{aligned}
\langle\xi^\epsilon,\, \xi^{1-\epsilon} - v\rangle &= \sum_{i=1}^{d} \left[\operatorname{sign}(\xi_i)(|\xi_i| - (1-\epsilon)\,\|\xi\|_\epsilon)_+\right]\left[\operatorname{sign}(\xi_i)(|\xi_i| - (|\xi_i| - (1-\epsilon)\,\|\xi\|_\epsilon)_+) - v_i\right] \\
&= \sum_{i=1}^{d} \left[(|\xi_i| - (1-\epsilon)\,\|\xi\|_\epsilon)_+\right]\left[(|\xi_i| - (|\xi_i| - (1-\epsilon)\,\|\xi\|_\epsilon)_+) - v_i \operatorname{sign}(\xi_i)\right] \\
&\geqslant \sum_{\substack{i=1\\ |\xi_i|>(1-\epsilon)\|\xi\|_\epsilon}} \left[|\xi_i| - (1-\epsilon)\,\|\xi\|_\epsilon\right]\left[(1-\epsilon)\,\|\xi\|_\epsilon - v_i \operatorname{sign}(\xi_i)\right] \geqslant 0.
\end{aligned}$$

The last inequality hold because $v \in V(\|\xi\|_\epsilon)$ *i.e.* $\forall i \in [d]$, $v_i \leqslant (1-\epsilon)\|\xi\|_\epsilon$. Finally, $\|u\|^2 > \epsilon^2\|\xi\|_\epsilon^2$ hence the result. □

Lemma 9. $\left\{\xi \in \mathbb{R}^d : \|\xi\|_\epsilon \leqslant \nu\right\} = \left\{u + v : u, v \in \mathbb{R}^d, \|u\| \leqslant \epsilon\nu, \|v\|_\infty \leqslant (1-\epsilon)\nu\right\}.$

Proof. Thanks to Lemma 7, we have $\xi = \xi^\epsilon + \xi^{1-\epsilon}$, $\|\xi^\epsilon\| = \epsilon\|\xi\|_\epsilon$ and $\|\xi^{1-\epsilon}\|_\infty = (1-\epsilon)\|\xi\|_\epsilon$. Hence, $\|\xi\|_\epsilon \leqslant \nu$ implies $\|\xi^\epsilon\| \leqslant \epsilon\nu$ and $\|\xi^{1-\epsilon}\|_\infty \leqslant (1-\epsilon)\nu$.

Suppose $\xi = u + v$ such that $\|u\| \leqslant \epsilon\nu$ and $\|v\|_\infty \leqslant (1-\epsilon)\nu$. From the ϵ-decomposition, we have $\|\xi\|_\epsilon = \|\xi^\epsilon\| + \|\xi^{1-\epsilon}\|_\infty$. Moreover, $\|\xi^\epsilon\| \leqslant \|u\|$ and $\|\xi^{1-\epsilon}\|_\infty \leqslant \|v\|_\infty$ thanks to Lemma 8. Hence $\|\xi^\epsilon\| \leqslant \|u\| + \|v\|_\infty \leqslant \epsilon\nu + (1-\epsilon)\nu = \nu$. □

Lemma 10 (Dual norm of the ϵ-norm)**.** *Let $\xi \in \mathbb{R}^d$, then $\|\xi\|_\epsilon^D = \epsilon\|\xi\| + (1-\epsilon)\|\xi\|_1$.*

Proof.

$$\begin{aligned}
\|\xi\|_\epsilon^D &:= \max_{\|x\|_\epsilon\leqslant 1} \xi^\top x = \max_{\substack{\|u\|\leqslant\epsilon\\ \|v\|_\infty\leqslant 1-\epsilon}} \xi^\top(u+v) \text{ thanks to Lemma 9} \\
&= \epsilon \max_{\|u\|\leqslant 1} \xi^\top u + (1-\epsilon) \max_{\|v\|_\infty\leqslant 1} \xi^\top v = \epsilon\,\|\xi\|^D + (1-\epsilon)\,\|\xi\|_\infty^D.
\end{aligned}$$ □

Lemma 11. *Let $\xi \in \mathbb{R}^d\backslash\{0\}$. Then $\nabla\|\cdot\|_\epsilon(\xi) = \frac{\xi^\epsilon}{\|\xi^\epsilon\|_\epsilon^D}$.*

Proof. Let us define $h : \mathbb{R} \times \mathbb{R}^d \mapsto \mathbb{R}$ by $h(\nu, \xi) = \|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\| - \epsilon\nu$. Then we have

$$\begin{aligned}
\frac{\partial h}{\partial \nu}(\nu, \xi) &= \frac{\mathrm{ST}_{(1-\epsilon)\nu}(\xi)^\top}{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|} \frac{\partial \mathrm{ST}_{(1-\epsilon)\nu}(\xi)}{\partial \nu} - \epsilon = -\frac{\mathrm{ST}_{(1-\epsilon)\nu}(\xi)^\top}{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|}(1-\epsilon)\,\mathrm{sign}(\xi) - \epsilon \\
&= -\frac{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|_1}{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|}(1-\epsilon) - \epsilon = -\frac{(1-\epsilon)\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|_1 + \epsilon\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|}{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|} \\
&= -\frac{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|_\epsilon^D}{\left\|\mathrm{ST}_{(1-\epsilon)\nu}(\xi)\right\|} \quad \text{(thanks to Lemma 10).}
\end{aligned}$$

By definition of the ϵ-norm, $h(\|\xi\|_\epsilon, \xi) = 0$. Since $\frac{\partial h}{\partial \nu}(\|\xi\|_\epsilon, \xi) = -\frac{\|\xi^\epsilon\|_\epsilon^D}{\epsilon\|\xi\|_\epsilon} \neq 0$, we obtain by applying the Implicit Function Theorem

$$\nabla \|\cdot\|_\epsilon(\xi) \times \frac{\partial h}{\partial \nu}(\|\xi\|_\epsilon, \xi) + \frac{\partial h}{\partial \xi}(\|\xi\|_\epsilon, \xi) = 0 \text{ hence } \nabla \|\cdot\|_\epsilon(\xi) = -\frac{\frac{\partial h}{\partial \xi}(\|\xi\|_\epsilon, \xi)}{\frac{\partial h}{\partial \nu}(\|\xi\|_\epsilon, \xi)}.$$

Moreover, $\frac{\partial h}{\partial \xi}(\|\xi\|_\epsilon, \xi) = \frac{\mathrm{ST}_{(1-\epsilon)\|\xi\|_\epsilon}(\xi)}{\|\mathrm{ST}_{(1-\epsilon)\|\xi\|_\epsilon}(\xi)\|} = \frac{\xi^\epsilon}{\|\xi^\epsilon\|} = \frac{\xi^\epsilon}{\epsilon\|\xi\|_\epsilon}$ hence the result: $\nabla\|\cdot\|_\epsilon(\xi) = \frac{\xi^\epsilon}{\|\xi^\epsilon\|_\epsilon^D}$. □

EDPP is not safe

In the two last sections, we present a study on the EDDP method (Wang et al., 2015), a screening rule that relies on the dual optimal point obtained for the previous λ in the path. Note that the same conclusion would hold true for generalization of the sequential approach given in (Wang et al., 2014), as well as for any other screening rule that needs exact dual solution at one step. In the remainder we consider $\lambda_0 = \lambda_{\max}$ and a non-increasing sequence of $T-1$ tuning parameters $(\lambda_t)_{t\in[T-1]}$ in $(0, \lambda_{\max})$. In practice, we choose the common grid (Bühlmann and van de Geer, 2011)[2.12.1]): $\lambda_t = \lambda_0 10^{-\delta t/(T-1)}$. Wang et al. (2015) proposed a sequential screening rule based on properties of the projection onto a convex set. Their rule is based on the exact knowledge of the true optimal solution for the previous parameter. Such a rule can be used to compute $\hat{\theta}^{(\lambda_1)}$ since $\hat{\theta}^{(\lambda_0)} = y/\lambda_0$ $(= y/\lambda_{\max})$ is known. However for $t > 1$, $\hat{\theta}^{(\lambda_t)}$ is only known approximately and the rules introduced in (Wang et al., 2015) are not safe anymore: some active groups may be wrongly disregarded if one does not use the exact value of $\hat{\theta}^{(\lambda_t)}$.

We first recall the property they proved we propose to modify their rule in order to make it safe in all cases.

Proposition 25 ((Wang et al., 2015, Theorem 19))**.** *Assume that $\lambda_{t-1} < \lambda_{\max}$, then the dual optimal solution of the group-Lasso with parameter λ_t, satisfies*

$$\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\big(\hat{\theta}^{(\lambda_{t-1})} + \frac{1}{2}v^\perp(\lambda_{t-1}, \lambda_t), \frac{1}{2}\left\|v^\perp(\lambda_{t-1}, \lambda_t)\right\|_2\big) \tag{2.51}$$

where

$$v^\perp(\lambda_{t-1}, \lambda_t) = \frac{y}{\lambda_t} - \hat{\theta}^{(\lambda_{t-1})} - \alpha[\hat{\theta}^{(\lambda_{t-1})}](\frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})})$$

and

$$\begin{aligned}
\alpha[\hat{\theta}^{(\lambda_{t-1})}] &:= \operatorname*{arg\,min}_{\alpha\in\mathbb{R}_+} \left\| \frac{y}{\lambda_t} - \hat{\theta}^{(\lambda_{t-1})} - \alpha(\frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})}) \right\|_2 \\
&= \frac{\langle \frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})}, \frac{y}{\lambda_t} - \hat{\theta}^{(\lambda_{t-1})} \rangle}{\|\frac{y}{\lambda_{t-1}} - \hat{\theta}^{(\lambda_{t-1})}\|_2^2}.
\end{aligned} \tag{2.52}$$

Making EDDP screening rule safe

The simpler screening rule

In the present paper, we give computable guarantees on the distance between the current dual feasible point and the solution of the problem. We show here how we can combine our result with Wang *et al.* 's in order to make their screening rule work even with approximate solutions to the previous Lasso problem.

For simplicity, we first consider the initial version of Wang *et al.* 's sphere test:

$$\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\big(\hat{\theta}^{(\lambda_{t-1})}, \big\|v^{\perp}(\lambda_{t-1}, \lambda_t)\big\|_2\big), \tag{2.53}$$

proved in (Wang et al., 2015, Theorem 7). As we do not know $\hat{\theta}^{(\lambda_{t-1})}$, we cannot readily use this ball. However, we can modify it to make it a safe screening rules as follows:

Proposition 26. *Assume that $\lambda_{t-1} < \lambda_{\max}$, $\theta \in \Delta_X$ is a dual feasible point and $r_{\lambda_{t-1}} > 0$ is a radius satisfying $\hat{\theta}^{(\lambda_{t-1})} \in \mathcal{B}(\theta, r_{\lambda_{t-1}})$, then*

$$\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\Big(\theta, r_{\lambda_{t-1}}(1 + |1 - \alpha[\theta]|) + \|v_t(\theta) - \alpha[\theta]v_{t-1}(\theta)\|_2\Big), \tag{2.54}$$

where

$$v_t(\theta) := \frac{y}{\lambda_t} - \theta \tag{2.55}$$

$$\alpha[\theta] := \operatorname*{arg\,min}_{\alpha \in \mathbb{R}_+} \|v_t(\theta) - \alpha v_{t-1}(\theta)\|_2 = \left(\frac{\langle v_{t-1}(\theta), v_t(\theta)\rangle}{\|v_{t-1}(\theta)\|_2^2}\right)_+ . \tag{2.56}$$

Proof. Start first by noting that (2.53) implies

$$\hat{\theta}^{(\lambda_t)} \in \bigcup_{\theta' \in \mathcal{B}(\theta, r_{\lambda_{t-1}})} \mathcal{B}\Big(\theta', \min_{\alpha \in \mathbb{R}_+} \big\|v_t(\theta') - \alpha v_{t-1}(\theta')\big\|_2\Big).$$

Let us denote

$$H = \max_{\theta' \in \mathcal{B}(\theta, r_{\lambda_{t-1}})} \min_{\alpha \in \mathbb{R}_+} \big\|v_t(\theta') - \alpha v_{t-1}(\theta')\big\|_2 ,$$

then $\hat{\theta}^{(\lambda_t)} \in \mathcal{B}(\theta, r_{\lambda_{t-1}} + H)$. We now need to upper bound H. A simple choice is to take α to be $\alpha[\theta]$ defined in Eq. (2.56) The motivation for such a choice is because it is optimal when $r_{\lambda_{t-1}} = 0$. This provides the following bound on H:

$$\begin{aligned} H &\leqslant \max_{\theta' \in \mathcal{B}(\theta, r_{\lambda_{t-1}})} \big\|v_t(\theta') - \alpha[\theta]v_{t-1}(\theta')\big\|_2 , \\ &= \left\| v_t(\theta) - \alpha[\theta]v_{t-1}(\theta) + r_{\lambda_{t-1}}(\alpha[\theta] - 1)\frac{v_t(\theta) - \alpha[\theta]v_{t-1}(\theta)}{\|v_t(\theta) - \alpha[\theta]v_{t-1}(\theta)\|} \right\|_2 , \\ &\leqslant r_{\lambda_{t-1}}|\alpha[\theta] - 1| + \|v_t(\theta) - \alpha[\theta].v_{t-1}(\theta)\| . \end{aligned} \tag{2.57}$$

Hence, after some simplifications:

$$\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\Big(\theta, r_{\lambda_{t-1}}(1 + |1 - \alpha[\theta]|) + \|v_t(\theta) - \alpha[\theta]v_{t-1}(\theta)\|_2\Big). \qquad \square$$

Remark 10. *In the case that $\|y/\lambda_{t-1}\| \leqslant \|y/\lambda_{t-1} - \theta\| \leqslant 1$ then with the definition of $\alpha[\theta]$ and the Cauchy-Schwartz inequality one has that $1 + |\alpha[\theta] - 1| \leqslant \frac{\lambda_{t-1}}{\lambda_t}$. This means that the multiplicative ratio in front of $r_{\lambda_{t-1}}$ is λ_{t-1}/λ_t. In (Fercoq et al., 2015, Proposition 3), the bound obtained would only lead to the smaller ratio: $\sqrt{\lambda_{t-1}/\lambda_t}$.*

Remark 11. *From the proof of Theorem 7 in Wang et al. (2015), it holds that for $\lambda < \lambda_{\max}$ then*

$$\left\|\hat{\theta}^{(\lambda)}\right\| \leqslant \frac{\|y\|}{\lambda} \Leftrightarrow \hat{\theta}^{(\lambda)} \in B\left(0, \frac{\|y\|}{\lambda}\right). \tag{2.58}$$

The complete screening rule (EDDP+)

Let us now consider the EDDP+ screening rule Wang et al. (2015) relying on the property (2.51): $\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\big(\hat{\theta}^{(\lambda_{t-1})} + \frac{1}{2}v^\perp(\lambda_{t-1},\lambda_t), \frac{1}{2}\left\|v^\perp(\lambda_{t-1},\lambda_t)\right\|_2\big)$. Using the same technique as for Proposition 26, we can strengthen our previous proposition with the following result.

Proposition 27. *Assume that $\lambda_{t-1} < \lambda_{\max}$, $\theta \in \Delta_X$ is a dual feasible point and $r_{\lambda_{t-1}} > 0$ is a radius satisfying $\hat{\theta}^{(\lambda_{t-1})} \in \mathcal{B}(\theta, r_{\lambda_{t-1}})$. Define $\alpha[\theta]$ as in* (2.56)*,*

$$\begin{aligned} r_{\lambda_t} =& \frac{|1-\alpha[\theta]|+1+\alpha[\theta]}{2} r_{\lambda_{t-1}} + \frac{1}{2}\left\|v_t(\theta) - \alpha[\theta]v_{t-1}(\theta)\right\|_2 \\ &+ \frac{\left\|\frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}}\right\|_2 r_{\lambda_{t-1}}}{2\|v_{t-1}(\theta)\|_2^2}\Big(3\left\|v_{t-1}(\theta)\right\|_2 + 2r_{\lambda_{t-1}}\Big) \end{aligned}$$

and

$$v^\perp(\theta,\lambda_{t-1},\lambda_t) = v_t(\theta) - \alpha[\theta]v_{t-1}(\theta). \tag{2.59}$$

Then $\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\Big(\theta + \frac{1}{2}v^\perp(\theta,\lambda_{t-1},\lambda_t), r_{\lambda_t}\Big)$.

Proof. As before, we do not know exactly $\hat{\theta}^{(\lambda_{t-1})}$ but we know that denoting

$$v^\perp(\theta',\lambda_{t-1},\lambda_t) = v_t(\theta') - \alpha[\theta']v_{t-1}(\theta') \tag{2.60}$$

with

$$\alpha[\theta'] = \left(\frac{\langle v_{t-1}(\theta'), v_t(\theta')\rangle}{\|v_{t-1}(\theta')\|_2^2}\right)_+, \tag{2.61}$$

we have

$$\hat{\theta}^{(\lambda_t)} \in \bigcup_{\theta' \in \mathcal{B}(\theta, r_{\lambda_{t-1}})} \mathcal{B}\Big(\theta' + \frac{1}{2}v^\perp(\theta',\lambda_{t-1},\lambda_t), \frac{1}{2}\left\|v^\perp(\theta',\lambda_{t-1},\lambda_t)\right\|_2\Big).$$

Our goal is to find a ball centered at $\theta + \frac{1}{2}v^\perp(\theta,\lambda_{t-1},\lambda_t)$ that contains all these balls, thus containing $\hat{\theta}^{(\lambda_t)}$. First, reminding (2.57)

$$\begin{aligned} \left\|v^\perp(\theta',\lambda_{t-1},\lambda_t)\right\|_2 &= \min_{\alpha\in\mathbb{R}_+}\left\|v_t(\theta') - \alpha v_{t-1}(\theta')\right\|_2 \\ &\leqslant \max_{\theta'\in\mathcal{B}(\theta,r_{\lambda_{t-1}})}\min_{\alpha\in\mathbb{R}_+}\left\|v_t(\theta') - \alpha v_{t-1}(\theta')\right\|_2 \\ &\leqslant r_{\lambda_{t-1}}|1-\alpha[\theta]| + \left\|v_t(\theta) - \alpha[\theta]v_{t-1}(\theta)\right\|_2. \end{aligned}$$

We continue as

$$\begin{aligned} \theta' + \frac{1}{2}v^\perp(\theta',&\lambda_{t-1},\lambda_t) - \theta - \frac{1}{2}v^\perp(\theta,\lambda_{t-1},\lambda_t) \\ &= (\theta'-\theta) + \frac{1}{2}\Big(v_t(\theta') - \alpha[\theta']v_{t-1}(\theta') - \frac{y}{\lambda_t} + \theta + \alpha[\theta]v_{t-1}(\theta)\Big) \\ &= \frac{1}{2}\Big(\theta' - \theta - (\alpha[\theta'] - \alpha[\theta])v_{t-1}(\theta') + \alpha[\theta](\theta'-\theta)\Big). \end{aligned}$$

Taking the norm on both sides of the previous display,

$$\begin{aligned} \left\|\theta' + \frac{1}{2}v^\perp(\theta',\lambda_{t-1},\lambda_t) - \theta - \frac{1}{2}v^\perp(\theta,\lambda_{t-1},\lambda_t)\right\|_2 \leqslant& \frac{1+\alpha[\theta]}{2}\left\|\theta'-\theta\right\|_2 \\ &+ \frac{|\alpha[\theta'] - \alpha[\theta]|}{2}\left\|v_{t-1}(\theta')\right\|_2. \end{aligned}$$

Now, reminding that $x \mapsto (x)_+$ is a 1-Lipschitz function, and denoting $\Gamma = |\alpha[\theta'] - \alpha[\theta]|$, we have:

$$\begin{aligned}
\Gamma &\leqslant \left| \frac{\langle v_{t-1}(\theta'), v_t(\theta') \rangle}{\|v_{t-1}(\theta')\|_2^2} - \frac{\langle v_{t-1}(\theta), v_t(\theta) \rangle}{\|v_{t-1}(\theta)\|_2^2} \right| \\
&= \left| \frac{\langle v_{t-1}(\theta'), \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \rangle}{\|v_{t-1}(\theta')\|_2^2} + 1 - \frac{\langle v_{t-1}(\theta), \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \rangle}{\|v_{t-1}(\theta)\|_2^2} - 1 \right| \\
&= \left| \frac{\langle \|v_{t-1}(\theta)\|_2^2 v_{t-1}(\theta') - \|v_{t-1}(\theta')\|_2^2 v_{t-1}(\theta), \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \rangle}{\|v_{t-1}(\theta')\|_2^2 \|v_{t-1}(\theta)\|_2^2} \right| \\
&\leqslant \frac{\left\| \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \right\|_2}{\|v_{t-1}(\theta')\|_2^2 \|v_{t-1}(\theta)\|_2^2} \Big(\|v_{t-1}(\theta')\|_2 \left| \|v_{t-1}(\theta)\|_2^2 - \|v_{t-1}(\theta')\|_2^2 \right| + \left\| \theta - \theta' \right\|_2 \|v_{t-1}(\theta')\|_2^2 \Big) \\
&\leqslant \frac{\left\| \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \right\|_2}{\|v_{t-1}(\theta')\|_2 \|v_{t-1}(\theta)\|_2^2} \Big(2 \| \frac{y}{\lambda_{t-1}} - \frac{\theta' + \theta}{2} \|_2 \|\theta - \theta'\|_2 + \left\| \theta - \theta' \right\|_2 \|v_{t-1}(\theta')\|_2 \Big) \\
&\leqslant \frac{\left\| \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \right\|_2 \|\theta - \theta'\|_2}{\|v_{t-1}(\theta')\|_2 \|v_{t-1}(\theta)\|_2^2} \Big(2\|v_{t-1}(\theta)\|_2 + \left\| \theta - \theta' \right\|_2 + \|v_{t-1}(\theta)\|_2 + \left\| \theta - \theta' \right\|_2 \Big) \quad . \quad (2.62)
\end{aligned}$$

where the second inequality comes from the triangle inequality and the Cauchy-Schwartz Inequality, and the third is obtained by factorizing the difference of squares. Plugging this in the former, we get:

$$\begin{aligned}
&\left\| \theta' + \frac{1}{2} v^\perp(\theta', \lambda_{t-1}, \lambda_t) - \theta - \frac{1}{2} v^\perp(\theta, \lambda_{t-1}, \lambda_t) \right\|_2 \\
&\qquad \leqslant \frac{1 + \alpha[\theta]}{2} \left\| \theta' - \theta \right\|_2 + \frac{\left\| \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \right\|_2 \|\theta - \theta'\|_2}{2\|v_{t-1}(\theta)\|_2^2} \Big(3 \left\| v_{t-1}(\theta) \right\|_2 + 2 \left\| \theta - \theta' \right\|_2 \Big).
\end{aligned}$$

One could check that there exists $\theta' \in \mathcal{B}(\theta, r_{\lambda_{t-1}})$ satisfying

$$\hat{\theta}^{(\lambda_t)} \in \mathcal{B}\big(\theta' + \frac{1}{2} v^\perp(\theta', \lambda_{t-1}, \lambda_t), \frac{1}{2} \left\| v^\perp(\theta', \lambda_{t-1}, \lambda_t) \right\|_2 \big)$$

and so combining the last inequality with (2.62)

$$\begin{aligned}
\left\| \hat{\theta}^{(\lambda_t)} - \theta - \frac{1}{2} v^\perp(\theta, \lambda_{t-1}, \lambda_t) \right\|_2 &\leqslant \left\| \hat{\theta}^{(\lambda_t)} - \theta' - \frac{1}{2} v^\perp(\theta', \lambda_{t-1}, \lambda_t) \right\|_2 \\
&\quad + \left\| \theta' + \frac{1}{2} v^\perp(\theta', \lambda_{t-1}, \lambda_t) - \theta - \frac{1}{2} v^\perp(\theta, \lambda_{t-1}, \lambda_t) \right\|_2 \\
&\leqslant \frac{|1 - \alpha[\theta]| + 1 + \alpha[\theta]}{2} r_{\lambda_{t-1}} + \frac{1}{2} \left\| v_t(\theta) - \alpha[\theta] v_{t-1}(\theta) \right\|_2 \\
&\quad + \frac{\left\| \frac{y}{\lambda_t} - \frac{y}{\lambda_{t-1}} \right\|_2 r_{\lambda_{t-1}}}{2\|v_{t-1}(\theta)\|_2^2} \Big(3 \left\| v_{t-1}(\theta) \right\|_2 + 2 r_{\lambda_{t-1}} \Big) \qquad \square
\end{aligned}$$

Chapter 3

Pathwise Optimization and Hyperparameter Selection

Various machine learning problems are formulated as a minimization of an empirical loss function f, regularized by a term Ω whose calibration and complexity is controlled by a non negative hyperparameter λ. The (optimal) choice of regularization parameter λ is crucial since it directly influences the generalization performance of the estimator, *i.e.* its score on unseen data set. One of the most popular method in such a context is cross-validation (CV), see (Arlot and Celisse, 2010) for a detailed review. For simplicity, we investigate here the holdout version that consists in splitting the data in two parts: on the first part (*training set*) the method is trained for a pre-defined collection of candidates $\Lambda_T := \{\lambda_0, \dots, \lambda_{T-1}\}$, and on the second part (*validation set*), the best parameter is selected.

For a piecewise quadratic loss f and a piecewise linear regularization Ω (*e.g.* for the Lasso estimator), Osborne et al. (2000b); Rosset and Zhu (2007) show that the set of solutions follows a piecewise linear curve *w.r.t.* to the parameter λ. There are several algorithms that can generate the full path — by maintaining optimality conditions when the regularization parameter changes — this is what LARS is performing for Lasso (Efron et al., 2004), but similar approaches exist for SVM (Hastie et al., 2004) or for generalized linear models (Park and Hastie, 2007). Unfortunately, these methods have some drawbacks that can be critical in many situations:

— they have a worst case complexity, *i.e.* the number of linear segments, that is exponential in the dimension p of the problem (Gärtner et al., 2012) leading to unpractical algorithms. Even in favorable case, a complexity that is linear in p can be expensive to compute when p is large.

— they suffer from numerical instabilities due to multiple and expensive inversion of ill-conditioned matrix. As a result, these algorithms may fail before exploring the entire path, a crucial issue whenever the regularization decreases.

— they lack flexibility when it comes to incorporating different statistical learning tasks because they usually rely on specific algebra to handle the structure of the regularization and loss functions. As far as we know, they apply to a limited number of cases and we are not aware of a general framework that bypasses these problems.

— they do not benefited of early stopping. As shown in (Bousquet and Bottou, 2008), it is not necessary to optimize below the statistical error to enjoy good generalization property. By nature, exact regularization path algorithms must maintain the optimality conditions when the hyperparameter changes, which is demanding in computational time.

To overcome these issues, an approximation of the solution path up to accuracy $\epsilon > 0$ was proposed. An optimal complexity was proven to be $O(1/\epsilon)$ by Giesen et al. (2010) in a fairly general setting. A noticeable contribution was proposed by Mairal and Yu (2012), that come up with an algorithm whose complexity is $O(1/\sqrt{\epsilon})$ for the Lasso case. The later result was then

	Lasso	Logistic regr.
$f_i(z)$	$(y_i - z)^2/2$	$\log(1 + \mathrm{e}^z) - y_i z$
$f_i^*(u)$	$((u - y_i)^2 - y_i^2)/2$	$\mathrm{Nh}(u + y_i)$
$\mathcal{V}_{f^*,x}(u)$	$\|u\|_2^2/2$	$w_4(\|u\|_x^2/\|u\|_2)\|u\|_x^2$

Table 3.1 – $w_4(\tau) := \frac{(1-\tau)\log(1-\tau)+\tau}{\tau^2}$ and $\mathrm{Nh}(x) := x\log(x) + (1-x)\log(1-x)$ for $x \in [0, 1]$.

extended by (Giesen et al., 2012) to objective function that has a quadratic lower bound while providing a lower and upper bound of order $O(1/\sqrt{\epsilon})$. Unfortunately, these assumptions fail to hold for a large class of problems, including logistic regression or Huber loss for instances.

Following such ideas, (Shibagaki et al., 2015) have proposed, for classification problems, to approximate the regularization path on the hold-out cross-validation error. Indeed, the later is a more natural criterion to monitor when one aims at selecting a hyperparameter guaranteed to achieve the best validation error. The main idea is to construct an upper and lower bound on the validation error as simple functions of the regularization parameter. Hence by sequentially varying the parameters, one can estimate a range of parameter for which the validation error is smaller than an accuracy $\epsilon_v > 0$ (where v stands for validation).

In this chapter, we revisit the approximation of the solution and validation path in a unified framework, under general regularity assumptions that are commonly satisfied in machine learning. We encompass both classification and regression problems and provide a complexity analysis along with optimality guarantees. We discuss the relationship between the regularity of the loss function and the complexity of the approximation path. We prove that its complexity is $O(1/\sqrt[d]{\epsilon})$ for uniformly convex loss of order $d > 0$ *i.e.* uniformly convex with modulus $\mu \|\cdot\|^d / d$ (for $\mu > 0$) (see Bauschke and Combettes (2011, Definition 10.5)) and $O(1/\sqrt{\epsilon})$ for the logistic loss thanks to a refined measure of its curvature throughout its Generalized Self-Concordant properties (Sun and Tran-Dinh, 2017). Finally, we provide an algorithm with global convergence property for selecting a hyperparameter with a validation error ϵ_v-close to the best possible hyperparameter on a given range.

The contents of this chapter are based on a collaboration that led to the technical report

Authors: E. Ndiaye, T. Le, O. Fercoq, J. Salmon, I. Takeuchi.

- Safe Grid Search with Optimal Complexity. *To be submitted*, 2018.

Notation. Given a proper, closed and convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, If f is twice continuously differentiable with positive definite Hessian $\nabla^2 f(x)$ at any $x \in \mathbb{R}^n$, we denote

$$\|z\|_x := \|z\|_{\nabla^2 f(x)} = \sqrt{\langle \nabla^2 f(x) z, z \rangle} \ .$$

Problem Setup. Let us consider the class of convex optimization problems of the form

$$\hat{\beta}^{(\lambda)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \underbrace{f(X\beta) + \lambda \Omega(\beta)}_{P_\lambda(\beta)} \quad \text{(Primal)}. \tag{3.1}$$

It includes many supervised learning problems such as generalized linear models including Least Squares and logistic regressions for instances. For simplicity, we focus only on these two canonical examples where the loss functions are written as an empirical risk $f(X\beta) = \sum_{i \in [n]} f_i(x_i^\top \beta)$ recalled in Table 3.1. The penalty term is often used to incorporate prior knowledge by enforcing a certain regularity on the solutions. For instance, choosing a Ridge penalty (Hoerl, 1962) $\Omega(\cdot) =$

$\|\cdot\|_2^2/2$ improves the stability of the resolution of inverse problems while $\Omega(\cdot) = \|\cdot\|_1$ imposes sparsity at the feature level, a motivation that led to the Lasso estimator (Tibshirani, 1996). We refer to (Bach et al., 2012) for more evolved penalties enforcing structured sparsity.

In practice, obtaining the exact solution $\hat{\beta}^{(\lambda)}$ of Problem (3.1) is unpractical and one needs to rely on an approximation at a prescribed precision.

Definition 13 (ϵ-solution)**.** *For any tolerance $\epsilon \geqslant 0$, a primal vector β is called ϵ-solution at parameter λ if its error for the objective value is smaller than ϵ,* i.e.

$$\mathcal{E}_\lambda(\beta) := P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \epsilon \ . \tag{3.2}$$

We recall the notion of approximate path, following the terminology from (Giesen et al., 2012):

Definition 14 (ϵ-path)**.** *A set $\mathcal{P}_\epsilon \subset \mathbb{R}^p$ is called an ϵ-path for the parameter range $[\lambda_{\min}, \lambda_{\max}]$ if it contains an ϵ-approximation $\beta^{(\lambda,\epsilon)}$ for any parameter λ,* i.e.

$$\forall \lambda \in [\lambda_{\min}, \lambda_{\max}], \exists\, \beta^{(\lambda)} \in \mathcal{P}_\epsilon \text{ s.t. } \mathcal{E}_\lambda(\beta^{(\lambda,\epsilon)}) \leqslant \epsilon. \tag{3.3}$$

We call path complexity for Problem (3.1) *the cardinality of the smallest ϵ-path: $T_\epsilon := \min |\mathcal{P}_\epsilon|$.*

Evaluating the approximation error (3.2) is often infeasible since it requires the unknown (exact) solution $\hat{\beta}^{(\lambda)}$ (see the discussion on optimization complexity in Nesterov (2004, Chapter 1)). Fortunately, when f is a convex function, one can often rely on the notion of *duality gap* to measure the quality of an estimate. First, we recall the classical Fenchel duality (Rockafellar, 1997, Chapter 31) recalled in Theorem 2

$$\hat{\theta}^{(\lambda)} \in \arg\max_{\theta \in \mathbb{R}^n} \underbrace{-f^*(-\lambda\theta) - \lambda\Omega^*(X^\top\theta)}_{D_\lambda(\theta)} \quad \text{(Dual)}. \tag{3.4}$$

Definition 15 (Duality Gap)**.** *For any primal/dual feasible pair of vector $(\beta, \theta) \in \mathbb{R}^p \times \mathbb{R}^n$, the duality gap is defined as the difference between the primal and dual objectives:*

$$\mathrm{Gap}_\lambda(\beta, \theta) := f(X\beta) + f^*(-\lambda\theta) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \ .$$

By weak duality, for any primal/dual feasible vector (β, θ), we have $P_\lambda(\beta) \geqslant D_\lambda(\theta)$. Hence, $\mathcal{E}_\lambda(\beta) \leqslant \mathrm{Gap}_\lambda(\beta, \theta)$ explaining the interest of the duality gap as an optimality certificate or as a stopping criterion for solving (3.1).

3.1 Duality Gap based Approximation Path

In this section, we introduce our framework and show how to efficiently compute an ϵ-path in Definition 14 for Problem (3.1).

Definition 16. *Given a real valued function f defined on $\mathbb{R}^n$ and x in $\mathrm{dom}(f)$, let $\mathcal{U}_{f,x}(\cdot)$ and $\mathcal{V}_{f,x}(\cdot)$ be non negative functions that vanish at 0. We say that f is $\mathcal{U}_{f,x}$-convex (resp. $\mathcal{V}_{f,x}$-smooth) at x when Inequality* (3.5) *(resp.* (3.6)*) is satisfied for any z in $\mathrm{dom}(f)$*

$$\mathcal{U}_{f,x}(z - x) \leqslant f(z) - f(x) - \langle \nabla f(x), z - x \rangle \ , \tag{3.5}$$

$$\mathcal{V}_{f,x}(z - x) \geqslant f(z) - f(x) - \langle \nabla f(x), z - x \rangle \ . \tag{3.6}$$

This framework is inspired from the widely used notion of μ-strong-convexity (resp. ν-smoothness), *cf.* (Nesterov, 2004), where $\mathcal{U}_{f,x}(z - x) = \mu \|z - x\|_2^2/2$ (resp. $\nu \|z - x\|_2^2/2$). Moreover, it is flexible enough to encompass:

Uniform Convexity (resp. *Uniform Smoothness*) (Azé and Penot, 1995):

$$\mathcal{U}_{f,x}(z-x) = \mathcal{U}(\|z-x\|) \ , \tag{3.7}$$
$$\mathcal{V}_{f,x}(z-x) = \mathcal{V}(\|z-x\|) \ , \tag{3.8}$$

where $\mathcal{U}(\cdot)$ and $\mathcal{V}(\cdot)$ are increasing mappings from $[0,+\infty)$ to $[0,+\infty]$ that vanish at 0 and are independent of any x. Examples of such functions are $\mathcal{U}(t) = \frac{\mu}{d}t^d$ and $\mathcal{V}(t) = \frac{\nu}{d}t^d$ where d, μ and ν are positive constants. The case $d = 2$ corresponds to the classical definition of strong convexity and smoothness; in general they are called *Uniformly Convex of order* d, see (Juditski and Nesterov, 2014) or (Bauschke and Combettes, 2011, Ch. 10.2 and 18.5) for further details. Also note that the norm $\|\cdot\|$ can be replaced by any positively homogeneous function that vanishes at zero. Such functions $\mathcal{U}_f, \mathcal{V}_f$ are known as *gauge like* functions (Rockafellar, 1997, Chapter 15).

Generalized Self-Concordant functions.

Definition 17 (Sun and Tran-Dinh (2017))**.** *A three time differentiable convex function f is called (M_f,ν)-generalized self-concordant of order $\nu \geqslant 2$ and $M_f \geqslant 0$ if for any $x \in \mathrm{Dom}(f)$ and any $u, v \in \mathbb{R}^p$,*

$$\left|\langle \nabla^3 f(x)[v]u, u\rangle\right| \leqslant M_f \|u\|_x^2 \|v\|_x^{\nu-2} \|v\|_2^{3-\nu}. \tag{3.9}$$

Proposition 28 (Prop. 10 Sun and Tran-Dinh (2017))**.** *If (M_f,ν)-generalized self concordant, then*

$$w_\nu(-d_\nu(x,z))\|z-x\|_x^2 \leqslant f(z) - f(x) - \langle \nabla f(x), z-x\rangle \leqslant w_\nu(d_\nu(x,z))\|z-x\|_x^2$$

where the right-hand side inequality holds if $d_\nu(x,z) < 1$ for the case $\nu > 2$ and where

$$d_\nu(x,z) := \begin{cases} M_f\|z-x\|_2 & \text{if } \nu = 2, \\ \left(\frac{\nu}{2}-1\right) M_f \|z-x\|_2^{3-\nu}\|z-x\|_x^{\nu-2} & \text{if } \nu > 2, \end{cases}$$

and

$$w_\nu(\tau) := \begin{cases} \frac{e^\tau - \tau - 1}{\tau^2} & \text{if } \nu = 2, \\ \frac{-\tau - \log(1-\tau)}{\tau^2} & \text{if } \nu = 3, \\ \frac{(1-\tau)\log(1-\tau)+\tau}{\tau^2} & \text{if } \nu = 4, \\ \left(\frac{\nu-2}{4-\nu}\right)\frac{1}{\tau}\left[\frac{\nu-2}{2(3-\nu)\tau}\left((1-\tau)^{\frac{2(3-\nu)}{2-\nu}}-1\right)-1\right] & \text{otherwise.} \end{cases}$$

In this case, the previous proposition show that

$$\mathcal{U}_{f,x}(z-x) = w_\nu(-d_\nu(x,z))\|z-x\|_x^2 \ ,$$
$$\mathcal{V}_{f,x}(z-x) = w_\nu(d_\nu(x,z))\|z-x\|_x^2 \ ,$$

where the second holds if $d_\nu(x,y) < 1$ for the case $\nu > 2$. This class of functions includes many important examples widely used in machine learning such as logistic and quadratic loss. In the proposed algorithm, the dual loss intervenes strongly and the dual of the logistic loss is Generalized Self-Concordant with $M_{f*} = 1, \nu = 4$ while for quadratic loss, $M_{f*} = 0$ and we can take any $\nu > 0$. We refer to (Sun and Tran-Dinh, 2017) for more details.

To simplify the notation, we will drop the subscript x in $\mathcal{U}_{f,x}$ and simply write $\mathcal{U}_f$; the meaning shall be clear depending on the context.

3.1.1 Bounding the Gap of the Homotopic Initialization

In the context of homotopy continuation recalled in Chapter 1, we solve problem (3.1) with different values of λ prefixed in a grid of hyperparameter. In this section, we provide fine bounds on the duality gap that control the optimization error when moving from one parameter to another.

Suppose we have at our disposal a primal/dual pair of vector $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ computed as outputs of an optimization algorithm at regularization parameter $\lambda_t > 0$ and t an integer, we denote

$$\mathrm{Gap}_t := \mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}), \quad \Delta_t := f(X\beta^{(\lambda_t)}) - f(\nabla f^*(z_t)) \tag{3.10}$$

for $z_t := -\lambda_t \theta^{(\lambda_t)}$ and for any function $\phi : \mathbb{R}^n \to [0, +\infty]$ that vanishes at 0,

$$Q_{t,\phi}(\rho) := \mathrm{Gap}_t + \rho \cdot (\Delta_t - \mathrm{Gap}_t) + \phi(-\rho \cdot z_t) \ . \tag{3.11}$$

In the following lemma, we propose to bound the duality gap by simply transferring the inequalities obtained from the regularity of the loss function f. It aims at controlling the growth of the duality gap *w.r.t.* the parameter λ.

Lemma 12 (Bounding the Warm Start Error)**.** *We assume that $-\lambda\theta^{(\lambda_t)} \in \mathrm{dom}(f^*)$ and $X^\top \theta^{(\lambda_t)} \in \mathrm{dom}(\Omega^*)$. If f^* is $\mathcal{V}_{f*}$-smooth (resp. $\mathcal{U}_{f*}$-convex), then, for $\rho = 1 - \lambda/\lambda_t$, the right (resp. left) hand side of Inequality* (3.12) *holds true*

$$Q_{t,\mathcal{U}_{f*}}(\rho) \leqslant \mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant Q_{t,\mathcal{V}_{f*}}(\rho) \ . \tag{3.12}$$

Proof. For simplicity, we denote

$$\mathrm{Gap}_\lambda^{\lambda_t} := \mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \quad \text{and} \quad \Gamma_t := \Omega(\beta^{(\lambda_t)}) + \Omega^*(X^\top \theta^{(\lambda_t)}) \ .$$

By Definition 15 of the duality gap at parameter λ_t, we have

$$\frac{1}{\lambda_t}[\mathrm{Gap}_t - f(X\beta^{(\lambda_t)}) - f^*(-\lambda_t\theta^{(\lambda_t)})] = \Gamma_t \ . \tag{3.13}$$

Hence using Equality (3.13) in the definition of $\mathrm{Gap}_\lambda^{\lambda_t}$, we have:

$$\begin{aligned}
\mathrm{Gap}_\lambda^{\lambda_t} &= f(X\beta^{(\lambda_t)}) + f^*(-\lambda\theta^{(\lambda_t)}) + \lambda\Gamma_t \\
&\overset{(3.13)}{=} \frac{\lambda}{\lambda_t}\mathrm{Gap}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)[f(X\beta^{(\lambda_t)}) + f^*(-\lambda_t\theta^{(\lambda_t)})] + f^*(-\lambda\theta^{(\lambda_t)}) - f^*(-\lambda_t\theta^{(\lambda_t)}).
\end{aligned}$$

Let us write the proof for the upper bound (the proof for the lower bound is similar). We apply the smoothness inequality (3.44) to the function $f^*(\cdot)$ with $z = -\lambda\theta^{(\lambda_t)}$ and $x = z_t := -\lambda_t\theta^{(\lambda_t)}$ to obtain

$$\mathrm{Gap}_\lambda^{\lambda_t} \leqslant \frac{\lambda}{\lambda_t}\mathrm{Gap}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)\Delta_t + \mathcal{V}_{f*,z_t}(-\lambda\theta^{(\lambda_t)} + \lambda_t\theta^{(\lambda_t)}) \ ,$$

where we have used the equality case in the Fenchel-Young inequality (3.42) to get:

$$\Delta_t = f(X\beta^{(\lambda_t)}) + f^*(-\lambda_t\theta^{(\lambda_t)}) + \langle \nabla f^*(z_t), -z_t \rangle = f(X\beta^{(\lambda_t)}) - f(\nabla f^*(z_t)) \ .$$

□

The function ϕ — chosen as $\mathcal{V}_{f*}$ (resp. $\mathcal{U}_{f*}$) for the upper bound (resp. lower bound) — essentially captures the regularity needed to approximate the duality gap at parameter λ when using previous primal/dual vector $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$. Note also that in the case where the function satisfies both inequalities, the tightness of the bound can be related to the conditioning number $\mathcal{K}_{f*} = \mathcal{U}_{f*}/\mathcal{V}_{f*}$ of the dual loss f^*. Hence we have an equality for the least-squares example ($\mathcal{U}_{f*} \equiv \mathcal{V}_{f*} \equiv \|\cdot\|^2/2$).

Algorithm 4 ϵ_{p}-Path on Training Set: `Training_path`

Input: $f, \Omega, \epsilon_{\mathrm{p}}, [\lambda_{\min}, \lambda_{\max}]$
Initialize $t = 0$, $\lambda_0 = \lambda_{\max}$, $\Lambda = \{\lambda_{\max}\}$.
repeat
  Solve Problem (3.1) for $\lambda = \lambda_t$ up to accuracy $\epsilon_{\mathbf{o}} < \epsilon_{\mathrm{p}}$
  Compute $\rho_t^\ell = \max\{\rho \text{ s.t. } Q_{t,\mathcal{V}_{f^*}}(\rho) \leqslant \epsilon_{\mathrm{p}}\}$
  Set $\lambda_{t+1} = \lambda_t \times (1 - \rho_t^\ell)$
  $\Lambda \leftarrow \Lambda \cup \{\lambda_{t+1}\}$ and $t \leftarrow t + 1$
until $\lambda_{t+1} \leqslant \lambda_{\min}$
Return: $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda\}$

Interpretation: In the Lasso example the inequalities (3.12) are tight and can be rewritten as

$$\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) = \frac{\lambda}{\lambda_t}\,\mathrm{Gap}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)\Delta_t + \frac{\|z_t\|_2^2}{2}\left(1 - \frac{\lambda}{\lambda_t}\right)^2 \ . \tag{3.14}$$

Then the result in Lemma 12 can be seen as a decomposition of the initialization error into *optimization* and *approximation* error. In fact, the two first terms involve Gap_t and Δ_t that correspond to the optimization error at parameter λ_t and the last term accounts for the price to pay when approximating λ_t by λ.

Adaptive Grid for a Fixed Precision. From Lemma 12, we have $\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant \epsilon$ as soon as $Q_{t,\mathcal{V}_{f^*}}(\rho) \leqslant \epsilon$ where $\rho = 1 - \lambda/\lambda_t$ varies with λ. Then, we proceed by choosing (at each grid point λ_t), ρ_t as the largest ρ such that the upper bound in (3.12) remains below the error ϵ. Hence, we obtain the following proposition that allows to track the regularization path for an arbitrary precision $\epsilon > 0$ on the training set by mean of the duality gap; see Algorithm 4.

Proposition 29 (Grid to Achieve Prescribed Precision)**.** *Assume we have solved Problem* (3.1) *for parameter λ_t up to precision $\mathrm{Gap}_t < \epsilon$, then $\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant \epsilon$ for all*

$$\lambda \in \lambda_t \times \left[1 - \rho_t^\ell(\epsilon),\, 1 + \rho_t^r(\epsilon)\right] \ ,$$

where $\rho_t^\ell(\epsilon)$ (resp. $\rho_t^r(\epsilon)$) is the largest non-negative ρ s.t. $Q_{t,\mathcal{V}_{f^*}}(\rho) \leqslant \epsilon$ (resp. $Q_{t,\mathcal{V}_{f^*}}(-\rho) \leqslant \epsilon$).

We will often drop the dependencies in ϵ for simplicity.

Adaptive Precision for a Fixed Grid. Conversely, given a grid of T points $\Lambda_T := \{\lambda_0, \ldots, \lambda_{T-1}\}$ (we assume a decreasing order: $\lambda_{t+1} < \lambda_t$), we define the error of the approximation path for a given range $[\lambda_{\min}, \lambda_{\max}]$ by using a piece-wise constant approximation of the duality gap $\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ over the grid:

$$\epsilon_{\Lambda_T} := \sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \min_{\lambda_t \in \Lambda_T} \mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \ . \tag{3.15}$$

This error is difficult to evaluate in practice so we rely on the tight upper bound based on inequalities in Lemma 12 that are easier to compute and for which closed-form are often available.

Proposition 30 (Precision for a Given Grid)**.** *For any grid of points Λ_T, the approximation error of the objective path is bounded as follows: $\epsilon_{\Lambda_T} \leqslant \max_{t \in [T]} Q_{t,\mathcal{V}_{f^*}}(1 - \lambda_t^\star/\lambda_t)$ where for all $t \in [T-1]$, $\lambda_t^\star$ is the largest $\lambda \in [\lambda_{t+1}, \lambda_t]$ such that $Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t) \geqslant Q_{t+1,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_{t+1})$.*

Proof. From the upper bound $\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant Q_{t,\mathcal{V}_{f^*}}(1-\lambda/\lambda_t)$ for all λ and λ_t, and since $[\lambda_{\min}, \lambda_{\max}] = \cup_{t\in[0:T-1]}[\lambda_{t+1}, \lambda_t]$ we have

$$\begin{aligned}\epsilon_{\Lambda_t} &\leqslant \max_{t\in[0:T-1]} \sup_{\lambda\in[\lambda_{t+1},\lambda_t]} \min_{\lambda_t\in\Lambda_T} Q_{t,\mathcal{V}_{f^*}}(1-\lambda/\lambda_t) \\ &\leqslant \max_{t\in[0:T-1]} \sup_{\lambda\in[\lambda_{t+1},\lambda_t]} \min_{t'\in\{t+1,t\}} Q_{t',\mathcal{V}_{f^*}}(1-\lambda/\lambda_{t'}) \ .\end{aligned}$$

where the last inequality holds since $\{\lambda_{t+1}, \lambda_t\}$ is a subset of Λ_T. Let us define

$$\forall\lambda\in[\lambda_{t+1},\lambda_t], \quad \psi_t(\lambda) := \min\{Q_{t+1,\mathcal{V}_{f^*}}(1-\lambda/\lambda_{t+1}), Q_{t,\mathcal{V}_{f^*}}(1-\lambda/\lambda_t)\} \ .$$

For $Q_{t+1,\mathcal{V}_{f^*}}(1-\lambda/\lambda_{t+1})$ (resp. $Q_{t,\mathcal{V}_{f^*}}(1-\lambda/\lambda_t)$) that is monotonically increasing *w.r.t.* λ (resp. decreasing) the $\sup_{\lambda\in[\lambda_{t+1},\lambda_t]}\psi_t(\lambda)$ is reached at the largest λ such that

$$Q_{t,\mathcal{V}_{f^*}}(1-\lambda/\lambda_t) \geqslant Q_{t+1,\mathcal{V}_{f^*}}(1-\lambda/\lambda_{t+1}) \ .$$

□

Finding the Step Sizes ρ. Following Proposition 29, finding the solution of equations of the form $Q_{t,\mathcal{V}_{f^*}}(\rho) = \epsilon$ is of high interest to obtain an ϵ-path. This can be done efficiently at high machine precision by various numerical solvers since this problem is one dimensional. Explicit solution are often available, for instance when $f^*(\cdot)$ is $\frac{\nu}{2}\|\cdot\|^2$-smooth, the step size is given by the solution of the quadratic inequality $Q_{t,\mathcal{V}_{f^*}}(\rho) \leqslant \epsilon$.

Proposition 31 (Quadratic Approximation Step)**.** *Given a primal/dual vector $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$, the left and right quadratic step sizes defined in Proposition 29 have the closed-form expressions:*

$$\rho_t^\ell(\epsilon) = \frac{\sqrt{2\nu R_t^2\delta_t + \tilde{\delta}_t^2} - \tilde{\delta}_t}{\nu R_t^2} \text{ and } \rho_t^r(\epsilon) = \frac{\sqrt{2\nu R_t^2\delta_t + \tilde{\delta}_t^2} + \tilde{\delta}_t}{\nu R_t^2}$$

where $\delta_t = \epsilon - \mathrm{Gap}_t$, $\tilde{\delta}_t = \Delta_t - \mathrm{Gap}_t$, $R_t = \|z_t\|$ and $\nu = \nu_{f^*}$ is the smoothness constant of the dual loss $f^*(\cdot)$.

As in (Mairal and Yu, 2012; Giesen et al., 2012), we recover as a special case the quadratic step size already known for the Lasso where the loss function is the quadratic loss $\|y-\cdot\|^2/2$. In this case, $\nu = 1$ and denoting $N_\infty := \|X^\top(y - X\beta^{(\lambda_t)})\|_\infty$, a direct calculation with the dual vector (3.16) reads:

$$\begin{aligned} z_t &:= -\lambda_t\theta^{(\lambda_t)} = \frac{\lambda_t}{\max(\lambda_t, N_\infty)}(y - X\beta^{(\lambda_t)}) \ , \\ \Delta_t &:= f(X\beta^{(\lambda_t)}) - f(\nabla f^*(z_t)) = \frac{1}{2}\left\|y - X\beta^{(\lambda_t)}\right\|_2^2 \times \left(1 - \left(\frac{\lambda_t}{\max(\lambda_t, N_\infty)}\right)^2\right) \ . \end{aligned}$$

Construction of a Feasible Vector. Given a primal vector $\beta^{(\lambda_t)}$, one can obtain a dual feasible vector by using a projected gradient mapping on the domain of the dual problem. Nevertheless, this operation can be expensive or impractical in which case we simply propose a rescaling of the gradient of the loss function *i.e.* $\theta^{(\lambda_t)} = -\alpha\nabla f(X\beta^{(\lambda_t)})$, with α s.t. $X^\top\theta^{(\lambda_t)} \in \mathrm{dom}(\Omega^*)$. Often, there is a simple expression for finding such a scaling factor α. Following the generic construction in Chapter 2, we choose

$$\theta^{(\lambda_t)} := \frac{-\nabla f(X\beta^{(\lambda_t)})}{\max(\lambda_t, \mathcal{S}^\circ_{\mathrm{dom}(\Omega^*)}(X^\top\nabla f(X\beta^{(\lambda_t)}))} \ . \tag{3.16}$$

This choice of dual point guarantees that the duality gap $\mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ will converge to 0 when $\beta^{(\lambda_t)}$ converges to a solution $\hat{\beta}^{(\lambda_t)}$ of (3.1). For any converging algorithm, one can make Δ_t arbitrarily small; for instance when $\Omega(\cdot) = \|\cdot\|$ is a norm, $\Delta_t = 0$ as soon as $\lambda_t \geqslant \|X^\top \nabla f(X\beta^{(\lambda_t)})\|_*$. The scaling is not needed if there is no constraint in the dual; for instance in the Elastic Net (Zou and Hastie, 2005) where $\Omega(\beta) = \eta \|\beta\|_1 + (1-\eta)\|\beta\|_2^2/2$ for $\eta \geqslant 0$, we can choose $-\lambda_t \theta_i^{(\lambda_t)} = \nabla f_i(x_i^\top \beta^{(\lambda_t)})$ for any primal candidate $\beta^{(\lambda_t)}$, and thus $\Delta_t = 0$. Note that, at optimality, $-\lambda_t \hat{\theta}_i^{(\lambda_t)} = \nabla f_i(x_i^\top \hat{\beta}^{(\lambda_t)})$ for all $i \in [n]$, so $\nabla f^*(-\lambda_t \hat{\theta}^{(\lambda_t)}) = X\hat{\beta}^{(\lambda_t)}$ and $\Delta_t = 0$.

3.1.2 Sampling Strategies

Adaptive Unilateral

For sparse regression methods, it is customary to start from the largest regularizer $\lambda_0 = \lambda_{\max}$ and then to iteratively compute $\hat{\beta}^{(\lambda_{t+1})}$ after having computed $\hat{\beta}^{(\lambda_t)}$. This is popular as it generally leads to computing the models in the order of increasing complexity: this allows important speed-ups by benefiting of *warm start* strategies (Friedman et al., 2007) provided that the parameters λ's are close enough from one another. This leads to the first strategy that we call *Unilateral* that consists in computing a new λ using only ρ_t^ℓ in Proposition 29. This strategy is illustrated in Section 3.1.2 for approximating the solution path of the Lasso. It has the advantage of combining simplicity and generality in the sense that it adapts simultaneously to both uniformly convex and generalized self-concordant function.

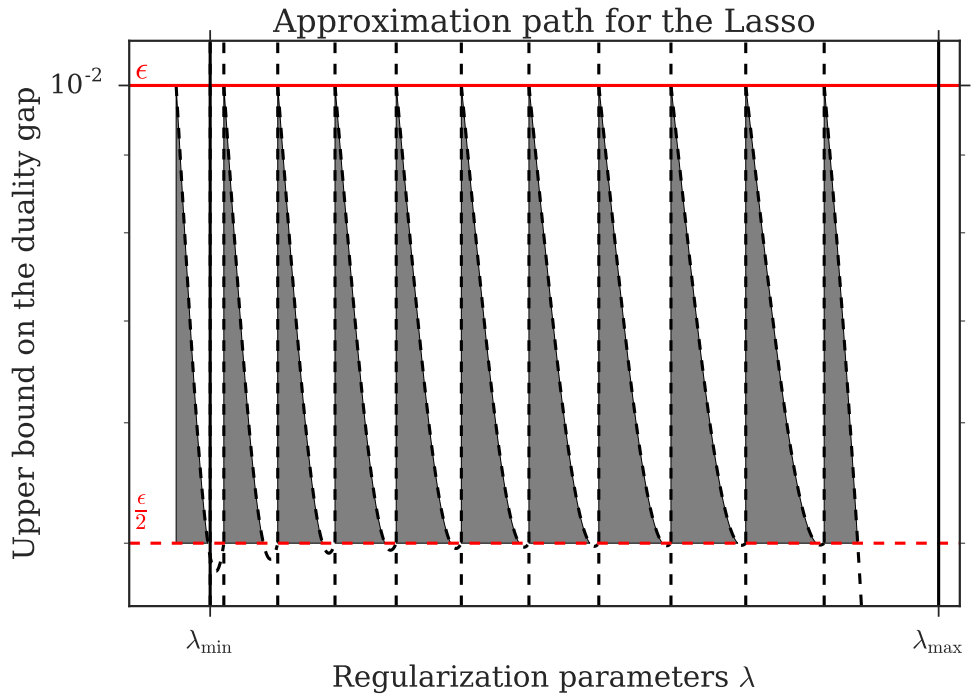


Figure 3.1 – Illustration of ϵ-path for the Lasso at accuracy $\epsilon = 10^{-2}$. This corresponds to the *adaptive unilateral sampling* where at each step t, the primal objective is optimized up to accuracy $\mathrm{Gap}_t = \epsilon_{\mathrm{o}} = \epsilon/2$ to satisfy the hypotheses of Proposition 29.

Adaptive Bilateral

Often, we can make a larger step by combining the information given by the left and right step sizes. Indeed let us assume that we explore the parameter range from $\lambda_{\max}$ to $\lambda_{\min}$. Starting from a parameter λ_t, we define the next step, given by Proposition 29, $\lambda_t^\ell := \lambda_t(1-\rho_t^\ell)$. Then it exists $\lambda_{t'} \leqslant \lambda_t^\ell$ such that $\lambda_{t'}^r := \lambda_{t'}(1+\rho_{t'}^r) = \lambda_t^\ell$. Thus a larger step can be done by using $\lambda_{t'} = \lambda_t \times \frac{1-\rho_t^\ell}{1+\rho_{t'}^r}$ instead of λ_t^ℓ. However $\rho_{t'}^r$ depends on the (approximated) solution $\beta^{(\lambda_{t'})}$ that we do not know before optimizing the problem at parameter $\lambda_{t'}$ when computing sequentially the grid points in decreasing order *i.e.* $\lambda_{t'} \leqslant \lambda_t$. We overcome this issue by (upper) bounding all the constant in $Q_{t',\mathcal{V}_{f^*}}(\rho)$ that depend on the solution $\beta^{(\lambda_{t'})}$, by constants involving only information available when once $\beta^{(\lambda_t)}$ has been approximated. For it, we need the following technical lemma, valid on the class of uniformly convex functions, that provides suitable bounds for deriving *Bilateral* (and later *Uniform*) approximation paths.

We first define the following quantities

$$\widetilde{R}_t := \mathcal{V}_f^{*-1}\left(f(X\beta^{(\lambda_t)}) + \frac{2\epsilon_{\mathrm{o}}}{\rho_t^\ell(\epsilon)}\right) \ , \tag{3.17}$$

$$\widetilde{\Delta}_t := \widetilde{R}_t \times \mathcal{U}_f^{-1}(\epsilon_{\mathrm{o}}) \ . \tag{3.18}$$

The next lemma shows how the terms Δ_t and $\|z_t\|$ can be directly bounded by quantity independent of t. Note that using the dual vector in Equation (3.16), we have

$$\|z_t\| = \|-\lambda_t\theta^{(\lambda_t)}\| \leqslant \|\nabla f(X\beta^{(\lambda_t)})\| \ .$$

Lemma 13. *Assuming f is $\mathcal{U}_f$-uniformly convex, we have*

$$\|\nabla f(X\beta^{(\lambda_{t'})})\| \leqslant \widetilde{R}_t \text{ and } \Delta_{t'} \leqslant \widetilde{\Delta}_t \ .$$

Proof. given in the supplementary material. □

Combining Lemma 13 and Lemma 12, we obtain

$$\mathrm{Gap}_\lambda(\beta^{(\lambda_{t'})}, \theta^{(\lambda_{t'})}) \leqslant Q_{t',\mathcal{V}_{f^*}}(\rho) \leqslant \widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho) \tag{3.19}$$

where $\rho = 1-\lambda/\lambda_{t'}$ and the mapping $\rho \mapsto \widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho)$ is independent of the approximated solution $(\beta^{(\lambda_{t'})}, \theta^{(\lambda_{t'})})$ at parameter $\lambda_{t'}$ and is defined as

$$\widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho) := \epsilon_{\mathrm{o}} + \rho\cdot(\widetilde{\Delta}_t - \epsilon_{\mathrm{o}}) + \mathcal{V}_{f^*}\left(|\rho|\cdot\widetilde{R}_t\right).$$

Hence we obtain the following approximation path with larger intermediate step:

$$\rho_t^{(b)}(\epsilon) = \frac{\rho_t^\ell(\epsilon) + \tilde{\rho}_t^r(\epsilon)}{1+\tilde{\rho}_t^r(\epsilon)}, \tag{3.20}$$

where $\rho_t^\ell(\epsilon)$ is defined in Proposition 29 and $\tilde{\rho}_t^r(\epsilon)$ is the largest non negative ρ such that $\widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho) \leqslant \epsilon$.

Proposition 32 (Bilateral Approximation Path)**.** *Assume that f is $\mathcal{U}_f$-uniformly convex and $\mathcal{V}_f$-uniformly smooth and let $0 < \epsilon_{\mathrm{o}} < \epsilon$ and $\mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant \epsilon_{\mathrm{o}}$ for all t. Define the grid $\Lambda^{(b)} := \{\lambda_0, \ldots, \lambda_{T-1}\}$ by*

$$\lambda_0 = \lambda_{\max}, \quad \lambda_{t+1} = \lambda_t \times (1-\rho_t^{(b)}(\epsilon)) \ .$$

Then the solution set $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda^{(b)}\}$ is an ϵ-path for problem (3.1)*.*

Uniform Unilateral and Bilateral

Given the initial information from the primal/dual vectors $(\beta^{(\lambda_{\max})}, \theta^{(\lambda_{\max})})$ at parameter $\lambda_{\max} = \lambda_0$, we can build a uniform grid that guarantee an ϵ-approximation before solving any optimization problem. Indeed, by using the same reasoning, we can build $Q_{\text{unif}}(\cdot)$ such that for $\rho = 1 - \lambda/\lambda_t$, we have $\text{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant Q_{\text{unif}}(\rho) := \widetilde{Q}_{0,\mathcal{V}_{f*}}(\rho)$.

We denote the uniform step as follow

$$\rho_{\text{unif}}(\epsilon) := \begin{cases} \rho^{\ell}_{\text{unif}}(\epsilon) & \text{for } \textit{Unilateral} \text{ path} \\ \frac{\rho^{\ell}_{\text{unif}}(\epsilon) + \tilde{\rho}^{r}_{\text{unif}}(\epsilon)}{1 + \tilde{\rho}^{r}_{\text{unif}}(\epsilon)} & \text{for } \textit{Bilateral} \text{ path.} \end{cases} \tag{3.21}$$

Proposition 33 (Uniform Approximation Path)**.** *Assume that f is $\mathcal{U}_f$-uniformly convex and $\mathcal{V}_f$-uniformly smooth and let $0 < \epsilon_{\text{o}} < \epsilon$ and $\text{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant \epsilon_{\text{o}}$, for all t. Define the grid $\Lambda^{(\text{unif})} = \{\lambda_0, \dots, \lambda_{T-1}\}$ by*

$$\lambda_0 = \lambda_{\max}, \quad \lambda_{t+1} = \lambda_t(1 - \rho_{\text{unif}}(\epsilon)) \ .$$

Then the solution set $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda^{(\text{unif})}\}$ is an ϵ-path for problem (3.1) *with at most T_ϵ grid points where*

$$T_\epsilon = \left\lfloor \frac{\log(\lambda_{\max}/\lambda_{\min})}{\log(1 - \rho_{\text{unif}}(\epsilon))} \right\rfloor + 2 \ .$$

Remark 12. *Since the uniform grid depends only on the initially known value $\beta^{\lambda_{\max}}$ at $\lambda_0 = \lambda_{\max}$, it can be computed before solving the optimization problem at any parameter. This allows the valuable advantage of parallelizing the computations over the grid of parameters.*

3.1.3 Concerns about Previous Methods

Previous algorithms for maintaining an ϵ-approximation of the solution along the regularization path that have been considered in the literature to help the calibration of the hyperparameters (Clarkson, 2010; Giesen et al., 2010) have the quality of being able to apply to a very large number of problems. Indeed, they have been developed under restrictive assumptions in which they are optimal with a complexity of $O(1/\epsilon)$. Nevertheless, data fitting functions arising in machine learning have sometimes nicer regularities that must be exploited upstream to obtain more scalable algorithms. This is all the more striking in the Lasso example where a better complexity was obtained in (Mairal and Yu, 2012) and at that time, it was the only algorithm to enjoy complexity in $O(1/\sqrt{\epsilon})$. But, what is so special about the Lasso and what can be said about others formulations?

First, we would like to emphasize the fact that without specific assumptions on the parameter, the ϵ-path complexity can be arbitrarily large even for "simple" functions. Indeed, let us consider the function

$$[0, 1] \times \mathbb{R} \ni (\lambda, \beta) \mapsto P_\lambda(\beta) = \sqrt[d]{|\beta - \lambda|} - \lambda^d \tag{3.22}$$

which is minimal at $\hat{\beta}^{(\lambda)} = \lambda$ and we have $P_\lambda(\hat{\beta}^{(\lambda)}) = -\lambda^d$. Hence for any β, the interval $I_\beta = \{\lambda \in [0, 1] : \mathcal{E}_\lambda(\beta) = \sqrt[d]{|\beta - \lambda|} \leqslant \epsilon\}$ has a length $2\epsilon^d$ independent of β. Hence this function has an ϵ-path complexity of $1/(2\epsilon^d)$ which gives an exponential lower bound for approximation path. Fortunately the complexity can be simplified if we assume more structure, (Giesen et al., 2012) proposed to consider functions such that:

$$\begin{cases} \beta \mapsto P_\lambda(\beta) \text{ are bounded from below for any } \lambda \text{ in an interval } I, \\ \lambda \mapsto P_\cdot(\beta) \text{ is concave for any } \beta \text{ in some set } D. \end{cases} \tag{3.23}$$

Within this framework, they show a lower bound of the ϵ-path complexity in $O(1/\sqrt{\epsilon})$ and also an structural upper bound in $O(1/\sqrt{\epsilon})$ showing the tightness of their bounds and analysis. They also

Algorithm 5 ϵ_v-Path for Validation Set

Input: $f, \Omega, \epsilon_{\mathrm{v}}, [\lambda_{\min}, \lambda_{\max}]$
Compute $\epsilon_{\mathrm{p}} = \xi(\epsilon_{\mathrm{v}}, \mu, X')$ according to Proposition 36
Set $\Lambda_{\mathrm{val}} = \texttt{Training\_path}\,(f, \Omega, \epsilon_{\mathrm{p}}, [\lambda_{\min}, \lambda_{\max}])$
Return: Λ_{val}

propose a generic algorithm capable to compute an ϵ-path that achieves this complexity building on a polynomial lower bound of the objective function.

We can notice two major concerns: the lower bound may be too pessimistic for the framework (3.23) (see Section 3.1.4) and we can find important machine learning examples where a reasonable polynomial lower bound on the objective function is hardly available. For instance, let us consider the ℓ_1 regularized logistic regression, in this case, the dual loss f^* is not smooth since the loss function f is not strongly convex. However, one can overcome this issue by restricting on any compact set as in (Dünner et al., 2016). Let us consider the one dimensional toy example where $\beta \in \mathbb{R}$, $X = \mathrm{Id}$ and $y = -1$, $f(X\beta) = \log(1 + \exp(\beta))$. We have, $\nabla^2 f(\beta) = \exp(\beta)/(1 + \exp(\beta))^2$. Then for Problem (3.1), since $P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant P_\lambda(0)$, we have $|\hat{\beta}^{(\lambda)}| \in [0, \log(2)/\lambda]$ and the smoothness constant of the dual can be estimated as $\nu_{f^*} = (1 + \exp(\log(2)/\lambda))^2$ at each step. This unfortunately leads to an infeasible algorithm with tiny step size since for $\lambda = \lambda_{\max}/10$ we already have $\nu_{f^*} \approx \exp(100)$ in Proposition 31. Moreover, note that the dual function is not polynomial thus algorithm previously proposed in (Giesen et al., 2012) do not handle the logistic regression case. Yet, as we will see in the next paragraph, we can efficiently build an ϵ-path with $O(1/\sqrt{\epsilon})$ complexity thanks to Generalized Self-Concordance bounds.

3.1.4 Complexity Analysis and Link with the Regularity of the Loss

Lower bounds. From our analysis, the lower bound on the duality gap in Lemma 12 tells us how close the proposed step in Proposition 29 is from the best possible step one can achieve for smooth loss functions. Indeed, at the optimal solution, we have $\mathrm{Gap}_t = \Delta_t = 0$. Thus the largest possible step — starting at λ_t and moving in decreasing order — is given by the smallest λ between $\lambda_{\min}$ and λ_t such that $\mathcal{U}_{f^*}\left(-\lambda_t \hat{\theta}^{(\lambda_t)} \times (1 - \frac{\lambda}{\lambda_t})\right) > \epsilon$. Hence, *any* algorithm for computing ϵ-path for $\mathcal{U}_{f^*}$-uniformly convex dual loss, have a complexity of order $O(1/\mathcal{U}_{f^*}^{-1}(\epsilon))$.

Our framework has the noticeable advantage to naturally adapt to the regularity of the loss function and do not require specific algebra for each function as it was done previously in the literature.

Upper bounds. We denote T_ϵ the cardinality of the grid returned by Algorithm 4. Let $(\rho_t)_{t \in [0:T_\epsilon - 1]}$ be the set of step size needed to cover the interval $[\lambda_{\min}, \lambda_{\max}]$. Using $\rho_t = 1 - \frac{\lambda_{t+1}}{\lambda_t}$, we have

$$\log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) = \log\left(\prod_{t=0}^{T_\epsilon - 1} \frac{\lambda_t}{\lambda_{t+1}}\right) = \sum_{t=0}^{T_\epsilon - 1} \log\left(\frac{1}{1 - \rho_t}\right) \quad .$$

Hence, denoting $\rho_{\min}(\epsilon) = \min_{t \in [0:T_\epsilon - 1]} \rho_t$, we have

$$T_\epsilon \times \rho_{\min}(\epsilon) \leqslant \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \quad . \tag{3.24}$$

Moreover, to simplify our analysis we will suppose that at each step λ_t, we have solved the optimization problem with two measures of accuracy $\mathrm{Gap}_t \leqslant \epsilon_{\mathrm{o}}$ and $\Delta_t \leqslant \epsilon_{\mathrm{o}}$ for $\epsilon_{\mathrm{o}} < \epsilon$. Also, we assume that we explore the parameter range in decreasing order. Then we recall from Lemma 12

that $\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant Q_{t,\mathcal{V}_{f^*}}(\rho)$ which is smaller than ϵ as soon as $\mathcal{V}_{f^*,z_t}(-z_t \cdot \rho) \leqslant \epsilon - \epsilon_{\mathrm{o}}$. Since $\rho_{\min}(\epsilon) = \min_{t\in[0:T_\epsilon-1]} \rho_t = \min_{t\in[0:T_\epsilon-1]} \sup\{\rho \ : \ Q_{t,\mathcal{V}_{f^*}}(\rho) \leqslant \epsilon\}$, then

$$\rho_{\min}(\epsilon) \geqslant \min_{t\in[0:T_\epsilon-1]} \sup\{\rho \ : \ \mathcal{V}_{f^*,z_t}(-z_t \cdot \rho) \leqslant \epsilon - \epsilon_{\mathrm{o}}\} \ . \tag{3.25}$$

Hence the complexity of the path is bounded as follows.

Proposition 34 (Complexity for Uniformly Convex Functions)**.** *If f and f^* are uniformly convex and uniformly smooth, then*

$$T_\epsilon \leqslant \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \times \frac{\mathcal{V}_f^{*-1}\left(f(X\beta^{(\lambda_0)}) + \frac{2\epsilon_{\mathrm{o}}}{\rho_0^\ell(\epsilon)}\right)}{\mathcal{V}_{f^*}^{-1}(\epsilon - \epsilon_{\mathrm{o}})} \ .$$

Proof. In the uniformly convex case, $\mathcal{V}_{f^*,z_t}(-z_t \cdot \rho) = \mathcal{V}_{f^*}(\rho \left\|z_t\right\|)$, hence we can deduce from Equation (3.24) and (3.25) that

$$T_\epsilon \leqslant \frac{1}{\rho_{\min}(\epsilon)} \times \log\big(\frac{\lambda_{\max}}{\lambda_{\min}}\big) \leqslant \log\big(\frac{\lambda_{\max}}{\lambda_{\min}}\big) \times \frac{\max_{t\in[0:N_\epsilon-1]} \left\|z_t\right\|}{\mathcal{V}_{f^*}^{-1}(\epsilon - \epsilon_{\mathbf{o}})} \ ,$$

so we just need to uniformly bound $\|z_s\|$. This bound follows from (3.17) and (3.16). □

Note that the initial step size $\rho_0^\ell(\epsilon) \leqslant \mathcal{V}_{f^*}^{-1}(\epsilon - \epsilon_{\mathbf{o}})/\left\|z_0\right\|$. Hence for loss function f such that $\mathcal{V}_{f^*}(\cdot) = \nu_{f^*} \left\|\cdot\right\|^d / d$, the complexity of the ϵ-path corresponds to $T_\epsilon \in O(1/\sqrt[d]{\epsilon})$.

For Generalized Self-Concordant functions, we show explicitly the complexity only for the logistic regression.

Proposition 35 (Complexity for Logistic Regression)**.** *If $f(z) = \sum_{i=1}^n \log(1 + e^{z_i}) - y_i z_i$, then there exists $B_{f^*,\lambda_0} > 0$ and $B'_{f^*,\lambda_0} > 0$ such that*

$$T_\epsilon \leqslant \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \max\left(\frac{B_{f^*,\lambda_0}}{\sqrt{\epsilon - \epsilon_{\mathbf{o}}}}, \frac{1}{B'_{f^*,\lambda_0}}\right).$$

Proof. We recall that f^* is generalized self-concordant with $\nu = 4$. The function $w_\nu(\cdot)$ is increasing and $w_\nu(0) = 1/2$, hence there exists a positive constant a_ν such that $w_\nu(\tau) \leqslant 1$ for $\tau \in [0, a_\nu]$ (in fact $a_\nu = 1$ for the logistic regression). Thus, provided $\rho d_\nu(z_t) \leqslant a_\nu$, we can derive the bound $\mathcal{V}_{f^*}(-z_t \times \rho) \leqslant \rho^2 \left\|z_t\right\|_{z_t}^2$.

Like in the uniformly convex case, in order to get the complexity of the ϵ-path, we also need a uniform bound on $\left\|z_t\right\|_{z_t}$.

By taking (3.5) on f^* with $x \leftarrow z_t$ and $z \leftarrow 0$, we obtain

$$\begin{aligned}\mathcal{U}_{f^*,z_t}(-z_t) &\leqslant f^*(0) - f^*(z_t) - \langle \nabla f^*(z_t), -z_t\rangle = f(\nabla f^*(z_t)) = f(X\beta^{(\lambda_t)}) - \Delta_t \\ &\leqslant f(X\beta^{(\lambda_t)}) + \epsilon_{\mathbf{o}} \leqslant f(X\beta^{(\lambda_0)}) + \frac{2\epsilon_{\mathbf{o}}}{\rho_0^\ell(\epsilon)} + \epsilon_{\mathbf{o}}\end{aligned}$$

where we have used the equality case of Fenchel-Young inequality and $f^*(0) = -\inf f = 0$.

Since $\nabla^2 f^*(z) = \mathrm{diag}(h_1(z_1), \cdots, h_n(z_n))$ where $h_i(z_i) = 1/(z_i(1-z_i))$ for all $i \in [n]$, we have $\|z\|_z^2 = \sum_{i=1}^n z_i^2 h_i(z_i)$. Whence we deduce that $\|z\|_2 \to 0 \Rightarrow \|z\|_z \to 0$. This gives that if $\|z\|_z \to +\infty$, then $\|z\|_2$ must be lower bounded. Then, as $\mathcal{U}_{f^*,z}(-z) = w_4\left(-\frac{\|z\|_z^2}{\|z\|_2}\right)\|z\|_z^2 \to +\infty$ when $\|z\|_z \to \infty$, we conclude that $\left\|z_t\right\|_{z_t}$ must be upper bounded by a quantity depending only on f^* and on $f(X\beta^{(\lambda_0)}) + \frac{2\epsilon_{\mathbf{o}}}{\rho_0^\ell} + \epsilon_{\mathbf{o}}$. Let us denote this bound B_{f^*,λ_0}. Likewise, we can show that $d_\nu(z_t)$ is upper bounded by a constant B'_{f^*,λ_0}. □

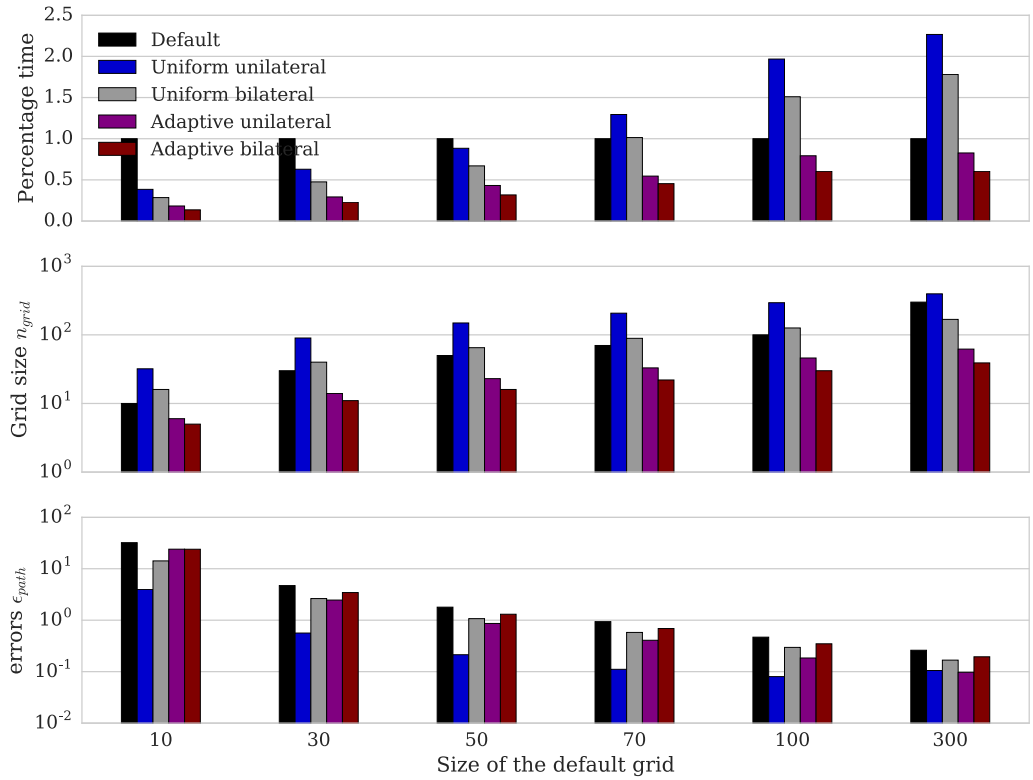


Figure 3.2 – ℓ_1 least-squares regression on climate data set NCEP/NCAR Reanalysis with $n = 814$ observations and $p = 73577$.

3.2 Validation Path and Approximation of the Best Hyperparameter

Considering the validation data (X', y') and loss $\mathcal{L}$, we define the validation error of the estimate β as

$$E_v(\beta) = \mathcal{L}(y', X'\beta) \ . \tag{3.26}$$

The objective is to solve the *bi-level* optimization problem

$$\begin{aligned} \underset{\lambda \in [\lambda_{\min}, \lambda_{\max}]}{\arg\min} \quad & E_v(\hat{\beta}^{(\lambda)}) = \mathcal{L}(y', X'\hat{\beta}^{(\lambda)}) \\ \text{s.t. } & \hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, f(X\beta) + \lambda\Omega(\beta) \ . \end{aligned}$$

Recent works address this problem, by using gradient-based algorithms see for instance (Pedregosa, 2016), have shown promising results in computational time and scalability w.r.t. multiple hyperparameters. However, they require assumptions such as smoothness of the validation function E_v and Non-singular Hessian of the inner optimization problem at optimal values which are difficult/impossible to check in practice since they rely on the exact knowledge of the optimal solutions and fail to hold for root mean square error and indicator loss. Moreover, they can only guarantee convergence to stationary point.

Here we show that with a safe and simple exploration of the parameter space, our algorithm has a global convergence property. Indeed, for any fixed tolerance ϵ_v, Algorithm 4 returns a solution with a validation error no larger than ϵ_v to the smallest possible error.

The following conditions, on the validation loss and on the inner optimization objective, are

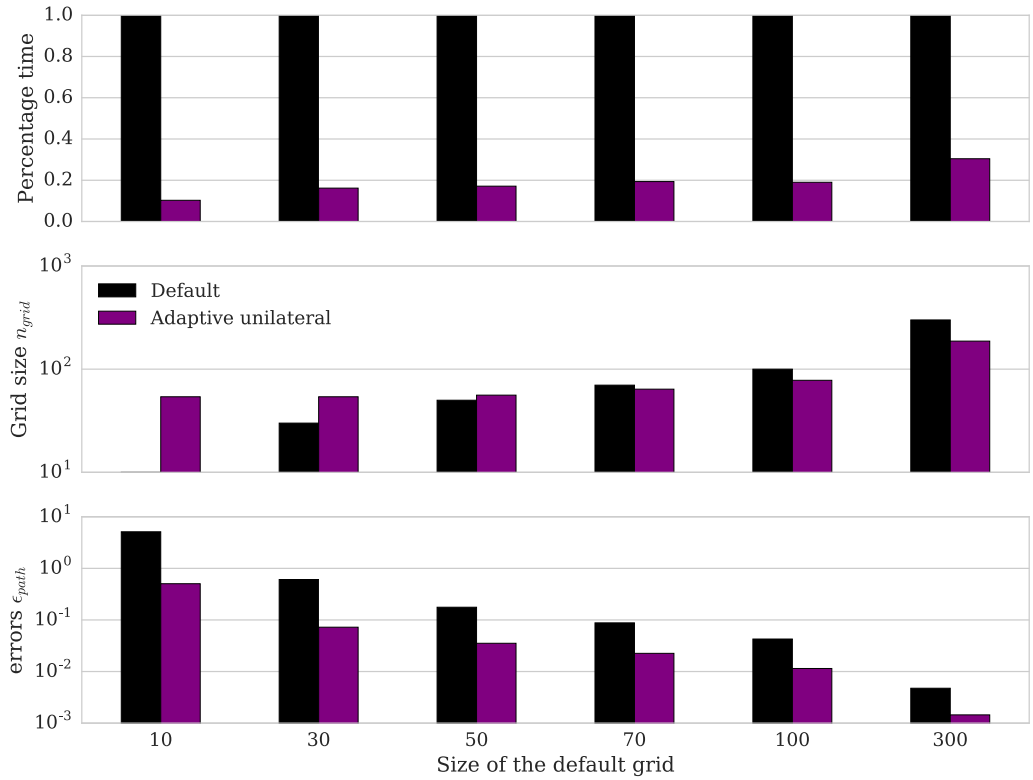


Figure 3.3 – ℓ_1 logistic regression on leukemia data set with $n = 72$ observations and $p = 7129$ features.

Figure 3.4 – Computation of the approximation path at the same error than the default grid.

assumed through the section:

$$\textbf{(A1):} \quad |\mathcal{L}(a,b) - \mathcal{L}(a,c)| \leqslant \mathcal{L}(b,c), \tag{3.27}$$

$$\textbf{(A2):} \quad \beta \mapsto P_\lambda(\beta) \text{ is } \mu\text{-strongly convex.} \tag{3.28}$$

The assumption on the loss function is verified for norms (regression case) and indicator function (classification). Indeed, for any norm $\mathcal{L}(a,b) = \|a-b\|$, we have from the triangle inequality $\big|\, \|a-b\| - \|a-c\| \,\big| \leqslant \|b-c\|$. For the indicator function, since $|\mathbf{1}_{ab<0} - \mathbf{1}_{ac<0}| \leqslant \mathbf{1}_{bc<0}$, we have $\big|\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{a_i b_i<0} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{a_i c_i<0}\big| \leqslant \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{b_i c_i<0}$.

Definition 18. *Given a primal solution $\hat{\beta}^{(\lambda)}$ at the regularization parameter parameter λ, we define the gap on the validation error between two parameter λ and λ_t as*

$$\Delta E_v(\lambda_t, \lambda) := \big|E_v(\hat{\beta}^{(\lambda)}) - E_v(\beta^{(\lambda_t)})\big|. \tag{3.29}$$

Suppose we have fixed a tolerance ϵ_v on the gap on validation error *i.e.* $\Delta E_v(\lambda_t, \lambda) \leqslant \epsilon_v$. Based on Inequality (3.27) in assumption **(A1)**, if there is a region $\mathcal{R}_\lambda$ that contains the optimal solution $\hat{\beta}^{(\lambda)}$ at parameter λ, then we have

$$\Delta E_v(\lambda_t, \lambda) \leqslant \mathcal{L}(X'\hat{\beta}^{(\lambda)}, X'\beta^{(\lambda_t)}) \leqslant \max_{\beta \in \mathcal{R}_\lambda} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}).$$

A simple strategy consists in choosing $\mathcal{R}_\lambda$ as a ball. Indeed, under the assumption **(A2)**, we have

Lemma 14 (Gap Safe Region)**.** *Assuming that $P_\lambda(\beta)$ is μ-strongly convex, the primal optimal solution $\hat{\beta}^{(\lambda)}$ belongs to the euclidean ball $\mathcal{B}(\beta^{(\lambda_t)}, r_{\lambda,\mu})$ where*

$$r_{\lambda,\mu} := r_{\lambda,\mu}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) = \sqrt{\frac{2}{\mu}\,\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})}\ . \tag{3.30}$$

Such a ball based on the duality gap radius was recently used to accelerate sparse optimization algorithm by iteratively identifying, with guarantee, the sparsity structure of the optimal solutions. Such strategy are known in machine learning as *safe screening rules* (El Ghaoui et al., 2012; Fercoq et al., 2015; Shibagaki et al., 2016; Ndiaye et al., 2017b).

Since the radius of the ball depends explicitly on the duality gap, we can sequentially track a range of parameters for which the gap on the validation error remains below a prescribed tolerance by controlling the optimization error measured with the duality gap. Hence we define

$$\xi(\epsilon_v, \mu, X') := \begin{cases} \frac{\mu}{2} \times \left(\frac{\epsilon_v}{\|X'\|}\right)^2 & \text{(regression case),} \\ \frac{\mu}{2} \times \left(\frac{x'^\top_{(\lfloor n\epsilon_v \rfloor+1)}\beta^{(\lambda_t)}}{\|x'_{(\lfloor n\epsilon_v \rfloor+1)}\|}\right)^2 & \text{(classification case).} \end{cases}$$

Proposition 36 (Grid for a prescribed validation error)**.** *Suppose that we have solved problem* (3.1) *for a parameter λ_t up to accuracy $\mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant \xi(\epsilon_v, \mu, X')$, then we have $\Delta E_v(\lambda_t, \lambda) \leqslant \epsilon_v$ for all*

$$\lambda \in \lambda_t \times \left[1 - \rho_t^\ell(\xi(\epsilon_v, \mu, X')),\ 1 + \rho_t^r(\xi(\epsilon_v, \mu, X'))\right]$$

where $\rho_t^\ell(\epsilon)$ and $\rho_t^r(\epsilon)$ for $\epsilon > 0$ are defined in Proposition 29.

Remark 13 (Stopping condition for the training step)**.** *Considering the current regularization parameter $\lambda = \lambda_t$, we have $\Delta E_v(\lambda_t, \lambda_t) = \left|E_v(\hat{\beta}^{(\lambda_t)}) - E_v(\beta^{(\lambda_t)})\right| \leqslant \epsilon_v$ provided that duality gap satisfies $\mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leqslant \xi(\epsilon_v, \mu, X')$. This gives us a stopping criterion for solving problems on the training set relative to the desired accuracy ϵ_v on the validation set (X', y').*

Proof. In case where the loss function is a norm, we have:

$$\max_{\beta \in \mathcal{B}(\beta^{(\lambda_t)}, r)} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}) = \max_{\beta \in \mathcal{B}(\beta^{(\lambda_t)}, r)} \|X'(\beta - \beta^{(\lambda_t)})\| \leqslant r_{\lambda,\mu}\|X'\|$$

where $r_{\lambda,\mu}$ is defined in Equation (3.30). Hence by using the bounds on the duality gap in Lemma 12, we can ensure $\Delta E_v(\lambda_t, \lambda) \leqslant \epsilon_v$ for all $\rho = 1 - \lambda/\lambda_t$ such that $Q_{t,\nu_{f^*}}(\rho) \leqslant \frac{\mu\epsilon_v^2}{2\|X'\|^2}$.

For the indicator loss function, using the inequality $-2ab \leqslant (a-b)^2 - b^2$ for $a = x'^\top_i\beta$ and $b = x'^\top_i\beta^{(\lambda_0)}$ and $|x'^\top_i(\beta - \beta^{(\lambda_0)})| \leqslant r\|x'_i\|$ for all $\beta \in \mathcal{B}(\beta^{(\lambda_0)}, r)$ we have:

$$-2(x'^\top_i\beta)(x'^\top_i\beta^{(\lambda_0)}) \leqslant (r\|x'_i\|)^2 - (x'^\top_i\beta^{(\lambda_0)})^2.$$

Hence we obtain the following upper bound

$$\max_{\beta \in \mathcal{B}(\beta^{(\lambda_0)}, r)} \mathcal{L}(X'\beta, X'\beta^{(\lambda_0)}) = \max_{\beta \in \mathcal{B}(\beta^{(\lambda_0)}, r)} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(x'^\top_i\beta^{(\lambda_0)})(x'^\top_i\beta)<0} \leqslant \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{|x'^\top_i\beta^{(\lambda_0)}| \leqslant r\|x'_i\|}.$$

By using the bound on the duality gap, we can ensure $\Delta E_v(\lambda_0, \lambda) \leqslant \epsilon_v$ for all λ such that:

$$\#\left\{i \in [n] : \xi_i := \frac{\mu}{2}\left(\frac{x'^\top_i\beta^{(\lambda_0)}}{\|x_i\|}\right)^2 \leqslant Q_{t,\nu_{f^*}}(1 - \lambda/\lambda_t)\right\} \leqslant \lfloor n\epsilon_v \rfloor.$$

By denoting $\left(\xi_{(i)}\right)_{i\in[n]}$ the (increasing) ordered sequence, we need the inequality to be true for at most the $\lfloor n\epsilon_v \rfloor$ first values *i.e.* we choose λ such that:

$$Q_{t,\nu_{f^*}}\left(1-\frac{\lambda}{\lambda_t}\right) < \frac{\mu}{2}\left(\frac{x'^{\top}_{(\lfloor n\epsilon_v\rfloor+1)}\beta^{(\lambda_0)}}{\left\|x'_{(\lfloor n\epsilon_v\rfloor+1)}\right\|}\right)^2 .$$

□

The Algorithm 5 outputs a discrete set of parameters Λ_{val} such that $\{\beta^{(\lambda_t)} \text{ for } \lambda_t \in \Lambda_{\text{val}}\}$ is an ϵ_v-approximation path for the validation objective function. As a direct consequence, for all λ in $[\lambda_{\min}, \lambda_{\max}]$, it exists $\lambda_t \in \Lambda_{\text{val}}$ outputted by Algorithm 5, such that

$$E_v(\beta^{(\lambda_t)}) - \epsilon_v \leqslant E_v(\hat{\beta}^{(\lambda)}).$$

By taking successively, the minimum over all λ_t in the grid on the left hand side, and the minimum over all λ in the parameter range on the right hand side, we obtain

Corollary 1. *If the set $\{\beta^{(\lambda_t)} \text{ for } \lambda_t \in \Lambda_{\text{val}}\}$ is an ϵ_v-path for the validation function, then*

$$\min_{\lambda_t\in\Lambda_{\text{val}}} E_v(\beta^{(\lambda_t)}) - \min_{\lambda\in[\lambda_{\min},\lambda_{\max}]} E_v(\hat{\beta}^{(\lambda)}) \leqslant \epsilon_v. \tag{3.31}$$

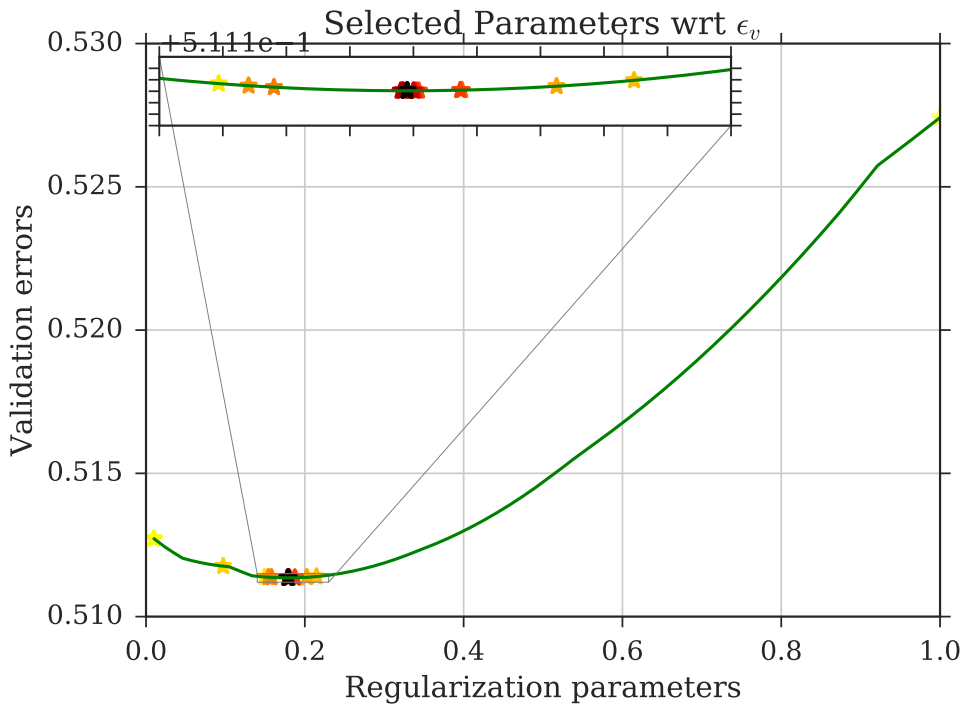


Figure 3.5 – Selecting the optimal hyperparameter for ℓ_1 Elastic Net regression at different accuracy ϵ_v and for Diabetes data set with $n = 442$ observations and $p = 10$. We illustrate the parameter selected by our algorithm at different precision levels. The color map ranges from yellow (low precision) to dark red (high precision)

3.3 Support Path for Sparse Regularization

In Chapter 2, we have provided a general framework for identifying active structure in convex optimization problems and have introduced the Gap Safe Rules which allows to eliminate more

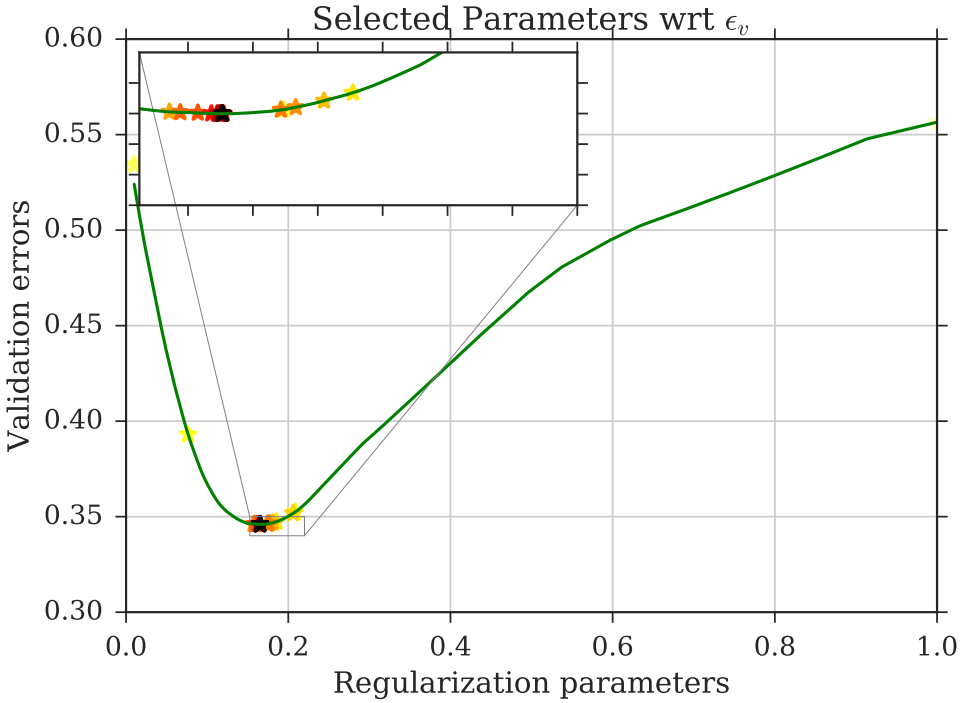


Figure 3.6 – Selecting the optimal hyperparameter for ℓ_1 Elastic Net regression at different accuracy ϵ_v for Synthetic data set $n = 500$ observations and $p = 5000$. We illustrate the parameter selected by our algorithm at different precision levels. The color map ranges from yellow (low precision) to dark red (high precision)

variables than previous methods. In this section, we study how to follow the variations of the active set *w.r.t.* to the regularization parameter λ. For simplicity, we restrict the discussions to the case where Ω is the ℓ_1 norm. Extensions to the general case taking into account the svm or other hierarchical penalties should not pose difficulties.

Following the results in Section 2.2.3, the sequential screening rule for ℓ_1 norm reads

$$\forall j \in [p] : \ |X_j^\top \theta^{(\lambda_t)}| + \|X_j\| \, r_t < 1 \Longrightarrow \hat{\beta}_j^{(\lambda)} = 0 \ . \tag{3.32}$$

where r_t is the Gap Safe Sphere radius sequentially defined by

$$r_t := \sqrt{\frac{2}{\gamma \lambda^2} \operatorname{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})} \ .$$

We recall from Definition 10, that the active sets for ℓ_1 norm are given by

$$\hat{\mathcal{A}}^{(\lambda_t)} := \left\{ j \in [p] : \ |X_j^\top \hat{\theta}^{(\lambda_t)}| \geqslant 1 \right\} \ , \tag{3.33}$$

$$\mathcal{A}^{(\lambda_t)} := \left\{ j \in [p] : \ |X_j^\top \theta^{(\lambda_t)}| + \|X_j\| \, r_t \geqslant 1 \right\} \ . \tag{3.34}$$

Leveraging our bounds on the duality gap Lemma 12, we can lower and upper bound the sequential gap screening radius as

$$\sqrt{\frac{2}{\gamma \lambda^2} Q_{t, \mathcal{U}_{f^*}} \left(1 - \frac{\lambda}{\lambda_t} \right)} \leqslant \sqrt{\frac{2}{\gamma \lambda^2} \operatorname{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})} \leqslant \sqrt{\frac{2}{\gamma \lambda^2} Q_{t, \mathcal{V}_{f^*}} \left(1 - \frac{\lambda}{\lambda_t} \right)} \ . \tag{3.35}$$

This allows to explicitly control the size of the active set with the sparsity inducing parameter λ. Indeed, for a non active feature j *i.e.* $j \in \bar{\mathcal{A}}^{(\lambda_t)}$, the bound (3.35) guarantee that it remains non active at regularization parameter λ *i.e.* $j \in \bar{\mathcal{A}}^{(\lambda)}$ for all λ in $[\lambda_t^{j,\text{in}}, \lambda_t]$ where

$$\lambda_t^{j,\text{in}} = \inf\left\{\lambda \leqslant \lambda_t : \ |X_j^\top \theta^{(\lambda_t)}| + \|X_j\| \sqrt{\frac{2}{\gamma\lambda^2} Q_{t,\mathcal{V}_{f^*}}\left(1 - \frac{\lambda}{\lambda_t}\right)} < 1\right\} \ . \tag{3.36}$$

Similarly, if $j \in \mathcal{A}^{(\lambda_t)}$, then $j \in \mathcal{A}^{(\lambda)}$ for all λ in $[\lambda_t^{j,\text{out}}, \lambda_t]$ where

$$\lambda_t^{j,\text{out}} = \inf\left\{\lambda \leqslant \lambda_t : \ |X_j^\top \theta^{(\lambda_t)}| + \|X_j\| \sqrt{\frac{2}{\gamma\lambda^2} Q_{t,\mathcal{U}_{f^*}}\left(1 - \frac{\lambda}{\lambda_t}\right)} \geqslant 1\right\} \ . \tag{3.37}$$

This allows to mimic the Lars algorithm (Efron et al., 2004) and check when

- a variable in the active set $\mathcal{A}^{(\lambda_t)}$ should leave $\mathcal{A}^{(\lambda_t)}$ and enters in the zero set:

$$\lambda_t^{\text{out}} = \min_{j \in \mathcal{A}^{(\lambda_t)}} \lambda_t^{j,\text{out}} \ , \tag{3.38}$$

- a variable in the non active set $\bar{\mathcal{A}}^{(\lambda_t)}$ should leave $\bar{\mathcal{A}}^{(\lambda_t)}$ and becomes active:

$$\lambda_t^{\text{in}} = \min_{j \in \bar{\mathcal{A}}^{(\lambda_t)}} \lambda_t^{j,\text{out}} \ . \tag{3.39}$$

Hence, given at step t, the next moment where the active set changes is given by

$$\lambda_{t+1} = \min\{\lambda_t^{\text{in}}, \lambda_t^{\text{out}}\} \ . \tag{3.40}$$

Remark 14. *So given $\mathcal{A}^{(\lambda_t)}$ of size s_t, we can find a smaller parameter λ_{t+1} with a prescribed size of the active set $s_{t+1} = |\mathcal{A}^{(\lambda_{t+1})}|$. Indeed, we have to choose $s_{t+1} - s_t$ new variable by sorting (in decreasing order) the $\lambda_t^{(j),\text{in}}$. If one variable from the active set becomes non active, we select $s_{t+1} - s_t + 1$ instead, and we reiterate.*

In the following we show that the variables stay in the safe active set forever once they hit it. Thus, the size of the sequential safe active set is constant between two kink.

Lemma 15. *For the Lasso problem, the mapping $\lambda \longmapsto \#\mathcal{A}^{(\lambda_t)}$ is piecewise decreasing.*

Proof. Since for any λ_t, $\text{Gap}_\lambda(\hat{\beta}^{(\lambda_t)}, \hat{\theta}^{(\lambda_t)}) = Q_{t,\mathcal{V}_{f^*}}\left(1 - \frac{\lambda}{\lambda_t}\right) = \frac{1}{2}\|y - X\hat{\beta}^{(\lambda_t)}\|^2 \left(1 - \frac{\lambda}{\lambda_t}\right)^2$, if $j \in \mathcal{A}^{(\lambda_t)}$, then the next step when j leaves the active set is given by

$$\begin{aligned}
\lambda_t^{j,\text{out}} &= \inf\left\{\lambda \leqslant \lambda_t : \ |X_j^\top \hat{\theta}^{(\lambda_t)}| + \|X_j\| \sqrt{\frac{2}{\lambda^2}\frac{1}{2}\|y - X\hat{\beta}^{(\lambda_t)}\|^2 \left(1 - \frac{\lambda}{\lambda_t}\right)^2} \geqslant 1\right\}, \\
&= \inf\left\{\lambda \leqslant \lambda_t : \ \frac{1}{\lambda}\left(1 - \frac{\lambda}{\lambda_t}\right) \geqslant \frac{1 - |X_j^\top \hat{\theta}^{(\lambda_t)}|}{\|X_j\| \, \|y - X\hat{\beta}^{(\lambda_t)}\|}\right\} = 0.
\end{aligned}$$

□

For the lasso, the joining time of a non active feature j is explicitly given by

$$\begin{aligned}
\lambda_t^{j,\text{in}} &= \inf\left\{\lambda \leqslant \lambda_t : \ |X_j^\top \hat{\theta}^{((\lambda_t))}| + \|X_j\| \sqrt{\frac{2}{\lambda^2}\frac{1}{2}\|y - X\hat{\beta}^{(\lambda_t)}\|^2 \left(1 - \frac{\lambda}{\lambda_t}\right)^2} < 1\right\}, \\
&= \inf\left\{\lambda \leqslant \lambda_t : \ \frac{1}{\lambda}\left(1 - \frac{\lambda}{\lambda_t}\right) < \frac{1 - |X_j^\top \hat{\theta}^{((\lambda_t))}|}{\|X_j\| \, \|y - X\hat{\beta}^{(\lambda_t)}\|}\right\}, \\
&= \frac{\lambda_t \|y - X\hat{\beta}^{(\lambda_t)}\|}{\|y - X\hat{\beta}^{(\lambda_t)}\| + \lambda_t \frac{1 - |X_j^\top \hat{\theta}^{((\lambda_t))}|}{\|X_j\|}} \ .
\end{aligned}$$

Hence from λ_t, the next time the active set changes corresponds to the next joining time given by $\lambda_{t+1} = \min_{j \in \bar{\mathcal{A}}^{(\lambda_t)}} \lambda_t^{j,\text{in}}$. This means that we can compute a safe support path for the Lasso since for any $\lambda \in [\lambda_{t+1}, \lambda_t]$, we have $\text{supp}(\hat{\beta}^{(\lambda)}) \subset \mathcal{A}^{(\lambda_t)}$.

3.4 Iteration Complexity of Pathwise Optimization

When solving the primal problem (3.1) at a given parameter λ_t, we denote $\beta_k^{(\lambda_t)}$ the vector obtained after k iterations and $\theta_k^{(\lambda_t)}$ its associated dual vector.

Lemma 16. *Assume that Ω is strongly convex and that we use a linearly convergent algorithm initialized with $\beta^{(\lambda_{t-1})}$ and let $\theta^{(\lambda_{t-1})}$ be its associated dual feasible pair. For some $\kappa > 0$, it holds*

$$C_{f,\Omega,X} \operatorname{Gap}_{\lambda_t}(\beta_{k+1}^{(\lambda_t)}, \theta_{k+1}^{(\lambda_t)}) \leqslant (1-\kappa)^k \operatorname{Gap}_{\lambda_t}(\beta^{(\lambda_{t-1})}, \theta^{(\lambda_{t-1})}) \ .$$

Proof. From Equation (2.19), we have

$$\begin{aligned} C_{f,\Omega,X} \operatorname{Gap}_{\lambda_t}(\beta_{k+1}^{(\lambda_t)}, \theta_{k+1}^{(\lambda_t)}) &\leqslant (1-\kappa)^k [P_{\lambda_t}(\beta_0^{(\lambda_t)}) - P_{\lambda_t}(\hat{\beta}^{(\lambda)})] \\ &\leqslant (1-\kappa)^k \operatorname{Gap}_{\lambda_t}(\beta_0^{(\lambda_t)}, \theta_0^{(\lambda_t)}) \ . \end{aligned}$$

Hence the conclusion since $\beta^{(\lambda_{t-1})}, \theta^{(\lambda_{t-1})}$ are used as initialization. □

Using the bound on the warm start initialization error Lemma 12 and Lemma 16, we deduce

$$C_{f,\Omega,X} \operatorname{Gap}_{\lambda_t}(\beta_{k+1}^{(\lambda_t)}, \theta_{k+1}^{(\lambda_t)}) \leqslant (1-\kappa)^k \left[\frac{\lambda_t}{\lambda_{t-1}} \operatorname{Gap}_{t-1} + \mathcal{V}_{f*}\left(-z_t(1 - \frac{\lambda_t}{\lambda_{t-1}}) \right) \right]$$

where $\operatorname{Gap}_{t-1} = \operatorname{Gap}_{\lambda_{t-1}}(\beta^{(\lambda_{t-1})}, \theta^{(\lambda_{t-1})})$. Denoting $W(\rho, \delta) := (1-\rho)\delta + \mathcal{V}_{f*}(-z_t\rho)$, we have $\operatorname{Gap}_{\lambda_t}(\beta_{k+1}^{(\lambda_t)}, \theta_{k+1}^{(\lambda_t)}) \leqslant \epsilon$ as soon as $k \geqslant \frac{1}{\kappa} \log \left(C_{f,\Omega,X} \frac{W(\frac{\lambda_t}{\lambda_{t-1}}, \operatorname{Gap}_{t-1})}{\epsilon} \right)$ which means that the iteration complexity at time t can be controlled by the optimization precision at time $t-1$ and the ratio between λ_t and λ_{t-1}.

Proposition 37. *Assume that Ω is strongly convex and that we use a linearly convergent algorithm initialized with $\beta^{(\lambda_{t-1})}$ as a warm start initialization at parameter λ_t. Then the duality gap at the last step $\operatorname{Gap}_{\lambda_T}(\beta_{k+1}^{(\lambda_T)}, \theta_{k+1}^{(\lambda_T)})$ is smaller than ϵ after at most K_ϵ iterations where*

$$K_\epsilon \leqslant \sum_{t=1}^{T} \frac{1}{\kappa_t} \log \left(C_{f,\Omega,X} \times \frac{W(\frac{\lambda_t}{\lambda_{t-1}}, \operatorname{Gap}_{t-1})}{\operatorname{Gap}_t} \right) \ , \tag{3.41}$$

with $W(\rho, \delta) := (1-\rho)\delta + \mathcal{V}_{f*}(-z_t\rho)$ and $C_{f,\Omega,X} := (\sigma_X \nu_f + \mu_\Omega)/\mu_\Omega$ for some $\kappa_t > 0$.

3.5 Numerical Experiments

To illustrate the behavior of our method, we compare the computational times and number of grid point needed to achieve a prescribed error ϵ on the duality gap for any regularization parameter λ on a given range $[\lambda_{\min}, \lambda_{\max}]$. More precisely, given the default grid, commonly used in `sklearn` (Pedregosa et al., 2011) and `glmnet` (Friedman et al., 2010b) *i.e.* $\lambda_t = \lambda_{\max} 10^{-\delta t/(T-1)}$ ($\delta = 3$ in our experiments), we report the times and numbers of grid point needed to achieve a smaller approximation error (measured thanks to Proposition 30) than the default grid.

Our experiments were conducted with the real and synthetic databases {Diabetes, Leukemia, synthetic (*make_regression*)} available in `sklearn` and the climate data NCEP/NCAR Reanalysis (Kalnay et al., 1996).

Results in Figure 3.4 show that we build a smaller grid thanks to the greater steps we take in each iteration. This often results in a significant gain in computational time. The convergence of our algorithm is illustrated numerically in Figure 3.5 and Figure 3.6 for different value of validation error ϵ_v.

Conclusion and Perspectives

We have introduced a general framework of approximation of the regularization paths by exploiting the optimality certificates provided by the duality gap. Our approach allows to manage a larger class of learning problem and extends easily to the approximation of the validation path for both classification and regression. Consequently, we have proposed a hyperparameter selection algorithm with a guarantee of global convergence towards the best hyperparameter of the empirical risk on the validation data.

Although providing strong optimality properties, improvements in the overall computation time can be obtained. In fact, the algorithms we have introduced are based on a sequential exploration of the hyperparameter space, which forces us to launch optimization algorithms on parameters that are not necessarily promising. To avoid this issue, we plan to mix our strategy with bandit like algorithm (Li et al., 2016a) that dynamically allows more computational resources to most promising hyperparameter. This can lead to a significant speed up while preserving our guarantees. Note also that our error bounds depends on the duality gap which has to be small when using such a dynamic strategy. In this case, we can rely on the extrapolation of the residual in (Massias et al., 2018b) to accelerate the convergence of the duality gap toward zero. Hence given a grid of parameter, our guess is that we will be able to screen-out faster irrelevant hyperparameters in terms of prediction performance.

3.6 Appendix

Proofs for Bounds on the Duality Gap

Lemma 17 (Fenchel-Young inequalities)**.** *Let f be a continuously differentiable function. For all x, x^*, we have*

$$f(x) + f^*(x^*) \geqslant \langle x^*, x \rangle, \tag{3.42}$$

with equalities if and only if $x^* = \nabla f(x)$ (or equivalently $x = \nabla f^*(x^*)$). Moreover, if f is $\mathcal{U}_{f,x}$-convex (resp. $\mathcal{V}_{f,x}$-smooth) Inequality (3.43) *(resp. Inequality* (3.44)*) holds true:*

$$f(x) + f^*(x^*) \geqslant \langle x^*, x \rangle + \mathcal{U}_{f,x}(x - \nabla f^*(x^*)), \tag{3.43}$$
$$f(x) + f^*(x^*) \leqslant \langle x^*, x \rangle + \mathcal{V}_{f,x}(x - \nabla f^*(x^*)). \tag{3.44}$$

Proof. We have from the $\mathcal{U}_{f,x}$-convexity and the equality $f(z) + f^*(\nabla f(z)) = \langle \nabla f(z), z \rangle$

$$-f^*(\nabla f(z)) + \langle \nabla f(z), x \rangle + \mathcal{U}_{f,x}(x - z) = f(z) + \langle \nabla f(z), x - z \rangle + \mathcal{U}_{f,x}(x - z) \leqslant f(x).$$

We conclude by applying the inequality at $z = \nabla f^*(x^*)$ and remark that $\nabla f(z) = x^*$. The same proof holds for the upper bound (3.44). □

Applying Fenchel-Young Inequalities (3.43) and (3.44) give the following bounds.

Lemma 18. *We assume that $-\lambda\theta \in \mathrm{Dom}(f^*)$ and $X^\top\theta \in \mathrm{Dom}(\Omega^*)$. Then, the Inequality* (3.45) *(resp.* (3.45)*) provided that f is $\mathcal{U}_f$-convex (resp. $\mathcal{V}_f$-smooth).*

$$\lambda\widetilde{\Omega}(\beta, \theta) + \mathcal{U}_f(X\beta - \nabla f^*(-\lambda\theta)) \leqslant \mathrm{Gap}_\lambda(\beta, \theta) \tag{3.45}$$
$$\lambda\widetilde{\Omega}(\beta, \theta) + \mathcal{V}_f(X\beta - \nabla f^*(-\lambda\theta)) \geqslant \mathrm{Gap}_\lambda(\beta, \theta) \tag{3.46}$$

where $\widetilde{\Omega}(\beta, \theta) = \Omega(\beta) + \Omega^*(X^\top\theta) + \langle \beta, -X^\top\theta \rangle$.

Proof. We apply the Fenchel-Young inequality (3.43) to obtain

$$\begin{aligned}
&\mathrm{Gap}_\lambda(\beta, \theta) = f(X\beta) + f^*(-\lambda\theta) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \\
&\geqslant \langle X\beta, -\lambda\theta \rangle + \mathcal{U}_f(X\beta - \nabla f^*(-\lambda\theta)) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \\
&= \mathcal{U}_f(X\beta - \nabla f^*(-\lambda\theta)) + \lambda\left(\Omega(\beta) + \Omega^*(X^\top\theta) + \langle \beta, -X^\top\theta \rangle\right).
\end{aligned}$$

The same technique applies for the upper bound with the Fenchel-Young inequality (3.44) □

Remark 15. *From the Fenchel-Young inequality* (3.42)*, we have $\Omega(\beta) + \Omega^*(X^\top\theta) \geqslant \langle \beta, X^\top\theta \rangle$, so the lower bound is always non negative.*

Lemma 19. *For $x \in \mathrm{Dom}(f)$, if f is $\mathcal{V}_{f,x}$-smooth, then writing $\mathcal{V}^*_{f,x} = (\mathcal{V}_{f,x})^*$ for the Fenchel-Legendre transform, one has*

$$\mathcal{V}^*_{f,x}(-\nabla f(x)) \leqslant f(x) - \inf_z f(z).$$

Proof. From the smoothness of f, we have

$$\inf_z f(z) \leqslant \inf_z \left(f(x) + \langle \nabla f(x), z - x \rangle + \mathcal{V}_{f,x}(z - x)\right) = f(x) - (\mathcal{V}_{f,x})^*(-\nabla f(x)).$$

□

Lemma 20. *Let $\beta^{(\lambda_t)}$ (resp. $\beta^{(\lambda_{t'})}$) be an ϵ-solution at parameter λ_t (resp. $\lambda_{t'}$), then we have*

$$\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right)\left(f(X\beta^{(\lambda_{t'})}) - f(X\beta^{(\lambda_t)})\right) \leqslant \mathrm{Gap}_{t'} + \frac{\lambda_{t'}}{\lambda_t}\mathrm{Gap}_t\,.$$

where $\mathrm{Gap}_s := \mathrm{Gap}_{\lambda_s}(\beta^{(\lambda_s)}, \theta^{(\lambda_s)})$ for $s \in \{t, t'\}$. Hence the mapping $\lambda \mapsto f(X\hat{\beta}^{(\lambda)})$ is non-increasing.

Proof. Since $\hat{\beta}^{(\lambda)}$ is optimal at parameter λ, we have:

$$f(X\beta^{(\lambda)}) + \lambda\Omega(\beta^{(\lambda)}) - \epsilon \leqslant f(X\hat{\beta}^{(\lambda)}) + \lambda\Omega(\hat{\beta}^{(\lambda)}) \leqslant f(X\beta^{(\lambda_t)}) + \lambda\Omega(\beta^{(\lambda_t)})\ .$$

Moreover,

$$\begin{aligned}
f(X\beta^{(\lambda_t)}) + \lambda\Omega(\beta^{(\lambda_t)}) &= \frac{\lambda}{\lambda_t}\left(f(X\beta^{(\lambda_t)}) + \lambda_t\Omega(\beta^{(\lambda_t)})\right) + \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\beta^{(\lambda_t)}) \\
&\leqslant \frac{\lambda}{\lambda_t}\left(f(X\hat{\beta}^{(\lambda_t)}) + \lambda_t\Omega(\hat{\beta}^{(\lambda_t)}) + \epsilon\right) + \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\beta^{(\lambda_t)}) \\
&\leqslant \frac{\lambda}{\lambda_t}\left(f(X\beta^{(\lambda)}) + \lambda_t\Omega(\beta^{(\lambda)}) + \epsilon\right) + \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\beta^{(\lambda_t)})\ .
\end{aligned}$$

The last inequality comes from the optimality of $\hat{\beta}^{(\lambda_t)}$ at parameter λ_t. Hence

$$f(X\beta^{(\lambda)}) + \lambda\Omega(\beta^{(\lambda)}) - \epsilon \leqslant \frac{\lambda}{\lambda_t}\left(f(X\beta^{(\lambda)}) + \lambda_t\Omega(\beta^{(\lambda)}) + \epsilon\right) + \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\beta^{(\lambda_t)}).$$

At optimality, $\epsilon = 0$ and we can deduce that $\left(1 - \frac{\lambda}{\lambda_t}\right) f(X\hat{\beta}^{(\lambda)}) \leqslant \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\hat{\beta}^{(\lambda_t)})$, hence the second result. □

We can furthermore bound the norm of the gradient of the loss when the parameter λ varies. A direct application of Lemma 19 and Lemma 20 yields:

Lemma 21. *Assume that f is ν_f-smooth and let $\beta^{(\lambda_{t'})}$ (resp. $\beta^{(\lambda_t)}$) be an ϵ-solution at parameter $\lambda_{t'}$ (resp. λ_t). Then for $\delta_\epsilon(\lambda_{t'}, \lambda_t) := \frac{\lambda_t + \lambda_{t'}}{\lambda_t - \lambda_{t'}}\,\epsilon$, we have*

$$\mathcal{V}_f^*(-\nabla f(X\beta^{(\lambda_{t'})})) \leqslant f(X\beta^{(\lambda_t)}) + \delta_\epsilon(\lambda_{t'}, \lambda_t). \tag{3.47}$$

At optimality $\epsilon = 0$ and so $\delta_\epsilon(\lambda_{t'}, \lambda_t) = 0$ and we have

$$\mathcal{V}_f^*(-\nabla f(X\hat{\beta}^{(\lambda_{t'})})) \leqslant f(X\hat{\beta}^{(\lambda_t)}). \tag{3.48}$$

Lemma 22. *Assuming f is $\mathcal{U}_f$-uniformly convex, we have $\|\nabla f(X\beta^{(\lambda_t)})\| \leqslant \widetilde{R}_t$ and $\Delta_{t'} \leqslant \widetilde{\Delta}_t$.*

Proof. Since f is convex, we have

$$\begin{aligned}
\Delta_t := f(X\beta^{(\lambda_t)}) - f(\nabla f^*(-\lambda_t\theta^{(\lambda_t)})) &\leqslant -\langle \nabla f(X\beta^{(\lambda_t)}), \nabla f^*(-\lambda_t\theta^{(\lambda_t)}) - X\beta^{(\lambda_t)}\rangle \\
&\leqslant \|\nabla f(X\beta^{(\lambda_t)})\|_* \|\nabla f^*(-\lambda_t\theta^{(\lambda_t)}) - X\beta^{(\lambda_t)}\| \\
&\leqslant \|\nabla f(X\beta^{(\lambda_t)})\| \times \mathcal{U}_f^{-1}(\mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}))
\end{aligned}$$

where the two last inequalities comes from Holder inequality and Lemma 18. □

We can also easily obtain the following guarantee on the affine interpolation

Proposition 38. *Considering primal dual pairs* $\left(\beta^{(\lambda_s)}, \theta^{(\lambda_s)}\right)$ *such that* $\mathrm{Gap}_{\lambda_s}(\beta^{(\lambda_s)}, \theta^{(\lambda_s)}) \leqslant \epsilon_o$ *for* $i \in \{t, t'\}$. *For any* $\alpha \in [0,1]$, *we define* $\lambda := (1-\alpha)\lambda_t + \alpha\lambda_{t'}$, $\beta^{(\lambda)} := (1-\alpha)\beta^{(\lambda_t)} + \alpha\beta^{(\lambda_{t'})}$ *and* $\theta^{(\lambda)} := (1-\alpha)\theta^{(\lambda_t)} + \alpha\theta^{(\lambda_{t'})}$. *For any* $\epsilon \geqslant \epsilon_o$, $\alpha := \frac{\lambda - \lambda_{t'}}{\lambda_t - \lambda_{t'}}$ *and* $\lambda_{t'} \in \lambda_t \times \left[\frac{1}{1+\rho}, 1\right]$ *we have* $\mathrm{Gap}_\lambda(\beta^{(\lambda)}, \theta^{(\lambda)}) \leqslant \epsilon$ *where*

$$\rho = \frac{\sqrt{2\nu \max(R_t, R_{t'})(\epsilon - \epsilon_o) + (\Delta_t - \epsilon_o)^2} - (\Delta_t - \epsilon_o)}{\nu \max(R_t, R_{t'})}. \tag{3.49}$$

Proof. From the convexity of the duality gap, we have:

$$\begin{aligned}
\mathrm{Gap}_\lambda(\beta^{(\lambda)}, \theta^{(\lambda)}) &\leqslant (1-\alpha)\,\mathrm{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) + \alpha\,\mathrm{Gap}_\lambda(\beta^{(\lambda_{t'})}, \theta^{(\lambda_{t'})}) \\
&\leqslant (1-\alpha)\left[\frac{\lambda}{\lambda_t}\mathrm{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) + \Delta_t\left(1 - \frac{\lambda}{\lambda_t}\right) + R_t\left(1 - \frac{\lambda}{\lambda_t}\right)^2\right] \\
&\quad + \alpha\left[\frac{\lambda}{\lambda_{t'}}\mathrm{Gap}_{\lambda_{t'}}(\beta^{(\lambda_{t'})}, \theta^{(\lambda_{t'})}) + \tilde{R}_{t'}\left(1 - \frac{\lambda}{\lambda_{t'}}\right)_+ + R_{t'}\left(1 - \frac{\lambda}{\lambda_{t'}}\right)^2\right].
\end{aligned}$$

Since $\lambda_t \geqslant \lambda \geqslant \lambda_{t'}$, we have $\frac{\lambda}{\lambda_t} \leqslant \frac{\lambda_t}{\lambda_{t'}}$, $1 - \frac{\lambda}{\lambda_t} \leqslant 1 - \frac{\lambda_{t'}}{\lambda_t}$ and $1 - \frac{\lambda}{\lambda_{t'}} \leqslant 1 - \frac{\lambda_{t'}}{\lambda_t}$. Hence

$$\begin{aligned}
\mathrm{Gap}_\lambda(\beta^{(\lambda)}, \theta^{(\lambda)}) \leqslant& (1-\alpha)\left[\frac{\lambda_t}{\lambda_{t'}}\epsilon_o + \Delta_t\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right) + R_t\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right)^2\right] \\
&+ \alpha\left[\frac{\lambda_t}{\lambda_{t'}}\epsilon_o + R_{t'}\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right)^2\right] \\
=& \frac{\lambda_t}{\lambda_{t'}}\epsilon_o + \left(1 - \frac{\lambda_{t'}}{\lambda_t}\right)^2\left[(1-\alpha)R_t + \alpha R_{t'}\right] + \alpha\Delta_t\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right).
\end{aligned}$$

Using a uniform bound independent of $\alpha \in [0,1]$, we have:

$$\mathrm{Gap}_\lambda(\beta^{(\lambda)}, \theta^{(\lambda)}) \leqslant \frac{\lambda_t}{\lambda_{t'}}\epsilon_o + \max(R_t, R_{t'})\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right)^2 + \Delta_t\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right).$$

Since $\lambda_{t'} \leqslant \lambda_t$, we have $(1-\lambda_{t'}/\lambda_t)^2 \leqslant (1-\lambda_t/\lambda_{t'})^2$ and $(1-\lambda_{t'}/\lambda_t) \leqslant -(1-\lambda_t/\lambda_{t'})$. So we can simplify the bound as $\mathrm{Gap}_\lambda(\beta^{(\lambda)}, \theta^{(\lambda)}) \leqslant \epsilon$ as soon as $(\lambda_t/\lambda_{t'})\epsilon_o + \max(R_t, R_{t'})\,(1 - \lambda_t/\lambda_{t'})^2 - \Delta_t\,(1 - \lambda_t/\lambda_{t'}) \leqslant \epsilon$. Hence we obtain the result by solving the quadratic inequality in $x = (1 - \lambda_t/\lambda_{t'}) \leqslant 0$. □

Chapter 4

Join Optimization for Concomitant Location-Scale Estimations

In the context of high dimensional regression where the number of features is greater than the number of observations, standard least squares need some regularization to both avoid overfitting and ease the interpretation of discriminant features. The Lasso (Chen and Donoho, 1995; Tibshirani, 1996) use the ℓ_1 norm as a sparsity inducing regularization and is one of the most popular methods for variable selection. It is defined as

$$\hat{\beta}_{\mathrm{L}}^{(\lambda)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad . \tag{4.1}$$

Though this estimator is well understood theoretically, the choice of the tuning parameter λ still raises critical questions in practice as well as in theory. The statistical guarantees of the Lasso (Bühlmann and van de Geer, 2011, Chapter 6) rely on choosing the tuning parameter proportional to the noise level. Indeed, under the linear model $y = X\beta^\star + \sigma_\star \varepsilon$ where $\varepsilon \sim \mathcal{N}(0,1)$ and denoting $S_\star = \mathrm{supp}(\beta^\star)$, $s_\star$ the cardinality of $S_\star$ and assuming that it exists a (compatibility) constant $\phi_\star > 0$ such that $\frac{1}{n}\|X\beta\|_2^2 \geqslant \frac{\phi_\star}{s_\star}\|\beta_{S_\star}\|_1^2$ for all β satisfying $\|\beta_{-S_\star}\|_1 \leqslant 3\|\beta_{S_\star}\|_1$, we have (with high probability):

$$\lambda \approx \sigma_\star \sqrt{n \log p} \text{ implies } \frac{1}{n}\|X\hat{\beta}_{\mathrm{L}}^{(\lambda)} - X\beta^\star\|_2^2 \leqslant \frac{\sigma_\star^2}{\phi_\star}\frac{s_\star \log p}{n} \quad . \tag{4.2}$$

Unfortunately, the quantity $\sigma_\star$ is usually *unknown* to practitioners. Beside, the noise level is of practical interest since it is also required in the computation of model selection criterion depending on the likelihood such as `AIC` (Akaike, 1974), `BIC` (Schwarz, 1978), `SURE` (Stein, 1981) or in the construction of confidence sets.

A way to estimate both the regression coefficients and the noise level is to estimate them simultaneously *e.g.* by computing the maximum likelihood at y which leads to the loss function

$$(\beta, \sigma) \to \log(\sigma) + \frac{1}{2\sigma^2}\|y - X\beta\|^2 \quad .$$

Unfortunately, it fails to be jointly convex. Also, when $y = X\beta$ and σ tends to zero *i.e.* approaching the boundary of the parameter space, the objective function tends to $-\infty$ making the statistical analysis and the global optimization problem difficult.

In this chapter, we first recall three different strategies used in the literature to alleviate the dependence on the unknown noise $\sigma_\star$ of the underlying statistical model.

Perspective transformation. A way to perform such a joint estimation with a convex formulation was proposed in the robust statistic theory (Huber, 1964) popularized in (Huber, 1981)

particularly in the context of location-scale estimation and (Huber and Dutter, 1974) proposed an alternating minimization to get the corresponding estimators. It relies on the joint convexity of the perspective of a convex function where the noise level plays the role of a dilation parameter. Later, Owen (2007) extended it to handle sparsity inducing penalty in high dimensional setting. It was then thoroughly analyzed in (Sun and Zhang, 2012), under the name Scaled-Lasso. In this chapter, we coin all these estimators *"Concomitant"* following the terminology proposed by Huber.

Pivotal estimator. While investigating estimator pivotal *w.r.t.* the noise level, (Belloni et al., 2011) proposed to solve the following convex program: modify the standard Lasso by removing the square in the data fitting term. Thus, they termed their estimator the Square-root Lasso; see also (Chrétien and Darses, 2011). Under a standard design assumption it is proved that the Square-root Lasso reaches optimal rates (4.2) for sparse regression, with the additional benefit that the regularization parameter is independent of the noise level. A second approach leading to this very formulation, was proposed by Xu et al. (2010) to account for adversarial corruption in the design matrix. Interestingly their robust construction led exactly to the Square-root Lasso formulation.

Re-parameterization. An important remark is that the maximization of the likelihood over the canonical parameters of a distribution from the exponential family is a convex problem. Hence, we can recover a jointly convex formulation through a change of variable. This strategy was used by Städler et al. (2010) for estimating the parameter of Mixture Regression Models and also recently in (Yu and Bien, 2017) under the name of Natural Lasso. Interestingly, the perspective formulation above was mentioned in (Antoniadis, 2010), in a response to Städler et al. (2010) providing another convex alternative for joint estimation.

Among the solutions to compute the Concomitant Lasso, two roads have been pursued so far. On the one hand, considering the Scaled-Lasso formulation, (Sun and Zhang, 2010, 2012) have proposed an iterative procedure that alternates Lasso steps and noise estimation steps, the later leading to rescaling the tuning parameter iteratively. On the other hand, considering the Square-root Lasso formulation, Belloni et al. (2011) have leaned on second order cone programming solvers, *e.g.* TFOCS (Becker et al., 2011). Despite the appealing properties listed above, among which the superiority of the theoretical results is the most striking, no consensus for an efficient solver has yet emerged for the Concomitant Lasso.

Our contribution aims at providing a more numerically stable formulation, called the Smoothed Concomitant Lasso. This variant also allows to obtain a fast solver: we first adapt a coordinate descent algorithm to the smooth version of the original problem, see (Nesterov, 2005; Beck and Teboulle, 2012)). Then, we apply the safe rules strategies and the active warm start developed in Chapter 2. This leads to important acceleration in practice both on real and simulated data. Overall, our method presents the same computational cost as for the Lasso, but enjoys the nice features mentioned earlier in terms of statistical properties and is less sensitive to the smoothing parameter. The Concomitant Lasso also has a matrix formulation more suitable to multivariate settings (van de Geer and Stucky, 2016) and a smooth version (Massias et al., 2018a) allows for similar computational gains.

The contents of this chapter are based on our published paper

Authors: E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, J. Salmon.

- "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". *Journal of Physics: Conference Series 904 (1), 012006*, 2017.

Notation. For a set $S \subset [p]$, we denote by $P_{X,S} = X_S \left(X_S^\top X_S\right)^+ X_S^\top$ the projection operator onto Span$\{X_j : j \in S\}$, where A^+ represents the Moore-Penrose pseudo-inverse. We note $\mathrm{tr(X)}$ the trace of matrix X and $\widehat{\Sigma} = X^\top X/n$ the normalized Gram matrix of X.

4.1 Concomitant Lasso

Let us first introduce the Concomitant Lasso estimator following the formulation proposed in (Huber, 1981, Chapter 7).

$$\min_{\beta \in \mathbb{R}^p, \sigma > 0} \frac{1}{n} \sum_{i=1}^{n} \left[\rho \left(\frac{y_i - x_i^\top \beta}{\sigma} \right) + a \right] \sigma \ , \tag{4.3}$$

where ρ is convex and vanishes at 0, $a > 0$. Interestingly, the objective function in formulation (4.3) is jointly convex in β and σ (see Proposition 39 below). A simple choice is the quadratic loss $\rho(z) = z^2/2$. For robustness purpose, (Huber, 1981, Eq. (7.14)) suggested the function ρ to be chosen so that it balances between the ℓ_2^2 loss for small error and the ℓ_1 loss for large error *i.e.* $\rho = \mathcal{H}_s$ where s is a non negative scale and $\mathcal{H}_s$ originally defined in (Huber, 1964) as

$$\mathcal{H}_s(z) = \begin{cases} \frac{z^2}{2} & \text{if } |z| \leqslant s, \\ s|z| - \frac{s^2}{2} & \text{if } |z| > s \ . \end{cases} \tag{4.4}$$

A similar point of view can be taken when it comes to defining a regularization function allowing at the same time to have the robustness of Ridge ℓ_2^2 but also the sparsity of the Lasso ℓ_1. As a result, we can consider a reversed version of Huber's function, called *Berhu*

$$\mathcal{B}_t(z) = \begin{cases} |z| & \text{if } |z| \leqslant t, \\ \frac{z^2}{2t} + \frac{t}{2} & \text{if } |z| > t \ . \end{cases} \tag{4.5}$$

Taking into account the dependence in the scale estimates as in (4.3), (Owen, 2007) propose the following concomitant estimator of location and scale as a *robust hybrid of Lasso and Ridge regression*:

$$\min_{(\beta, \sigma, \tau) \in \mathbb{R}^p \times \mathbb{R}_{++} \times \mathbb{R}_{++}} \frac{n\sigma}{2} + \sum_{i=1}^{n} \mathcal{H}_s \left(\frac{y_i - X_{i,:}\beta}{\sigma} \right) \sigma + n\lambda \left(\frac{p\tau}{2} + \sum_{j=1}^{p} \mathcal{B}_t \left(\frac{\beta_j}{\tau} \right) \tau \right) \ . \tag{4.6}$$

For s large enough, $\mathcal{H}_s$ becomes $|\cdot|^2/2$, and for t large enough $\mathcal{B}_t$ yields $|\cdot|$. In such a case, the optimization over the variable τ disappears from the formulation since $|\cdot|$ is 1-homogeneous.

This new estimator simultaneously brings together many desirable properties namely robustness to outliers, the ability to select the most relevant explanatory variables while being equivariant *w.r.t.* shift and scale transformation. However, there are still important challenges to overcome as suggested by Owen's conclusion:

> *«It remains to investigate the accuracy of the method for prediction and coefficient estimation. There is also a need for an automatic means of choosing λ. Both of these tasks must however wait on the development of faster algorithms for computing the hybrid traces. »*

For the sake of simplicity, we will first consider the case where $\mathcal{H}_s$ (resp. $\mathcal{B}_t$) reduce to the ℓ_2^2 loss (resp. ℓ_1 regularization). We call Concomitant Lasso (Owen, 2007; Antoniadis, 2010) the solution of the following optimization problem:

Definition 19. *For $\lambda > 0$, the Concomitant Lasso estimator $\hat{\beta}^{(\lambda)}$ is defined as a solution of the primal optimization problem*

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \underbrace{\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}_{P_\lambda(\beta, \sigma)} \ . \tag{4.7}$$

The statistical analysis was performed in (Sun and Zhang, 2012) where this estimator was called Scaled Lasso. Here we take an optimization point of view and propose a faster algorithm for solving (4.7).

Link with the Perspective of Convex Function. Given a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ as the function $\operatorname{persp}_f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that

$$\operatorname{persp}_f(\sigma, r) = \begin{cases} \sigma f\left(\frac{r}{\sigma}\right), & \text{if } \sigma > 0, \\ +\infty, & \text{if } \sigma \leqslant 0. \end{cases} \tag{4.8}$$

Proposition 39. *If f is convex, then $\operatorname{persp}_f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is also (jointly) convex.*

Proof. It suffices to show that $\operatorname{epi} \operatorname{persp}_f$ (see Equation (1.11)) is a convex set. By definition, we have

$$\begin{aligned} \operatorname{epi} \operatorname{persp}_f &= \{(\sigma, r, t) \in \mathbb{R}_*^+ \times \mathbb{R}^n \times \mathbb{R} : \quad \operatorname{persp}_f(\sigma, r) \leqslant t\} \\ &= \left\{(\sigma, r, t) \in \mathbb{R}_*^+ \times \mathbb{R}^n \times \mathbb{R} : \quad f\left(\frac{r}{\sigma}\right) \leqslant \frac{r}{\sigma}\right\} \\ &= \left\{\sigma \times (1, r', t') \in \mathbb{R}_*^+ \times (\mathbb{R}^n \times \mathbb{R}) : \quad f(r') \leqslant t'\right\} = \mathbb{R}_*^+ \times \operatorname{epi} f. \end{aligned}$$

Whence $\operatorname{epi} \operatorname{persp}_f$ is a Cartesian product of convex sets. □

Taking $f(r) = \frac{1}{2n}\|r\|_2^2 + \frac{1}{2}$, we have $\frac{1}{2n\sigma}\|y - X\beta\|^2 + \frac{\sigma}{2} = \operatorname{persp}_f^{**}(\sigma, y - X\beta)$ whence the convexity of problem (4.7) comes from Proposition 39.

However, the Concomitant Lasso estimator is ill-defined. Indeed, the set over which we optimize is not closed and the optimization problem may have no solution. Also, the perspective is not lower semi-continuous in general. However, lower semi-continuity is a very desirable property; together with the fact that the function is infinite at infinity, which guarantees the existence of minimizers (Peypouquet, 2015, Proposition 2.19). We circumvent this difficulty by considering instead the Fenchel biconjugate of the objective function which is always lower semi-continuous (Bauschke and Combettes, 2011, proposition 13.32). One can show (Bauschke and Combettes, 2011, Example 13.8) that the Fenchel conjugate of persp_f is

$$\operatorname{persp}_f^*(\nu, \theta) = \begin{cases} 0, & \text{if } \nu + f^*(\theta) \leqslant 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

and

$$\operatorname{persp}_f^{**}(\sigma, r) = \begin{cases} \sigma f^{**}\left(\frac{r}{\sigma}\right), & \text{if } \sigma > 0, \\ \sup\limits_{\theta \in \operatorname{dom} f^*} \langle \theta, r \rangle, & \text{if } \sigma = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

In our case, $f(r) = \frac{1}{2n}\|r\|_2^2 + \frac{1}{2}$ and so $f^{**} = f$ and $\operatorname{dom} f^* = \mathbb{R}^n$. Hence, we get

$$\operatorname{persp}_f^{**}(\sigma, r) = \begin{cases} \frac{1}{2n\sigma}\|r\|_2^2 + \frac{\sigma}{2}, & \text{if } \sigma > 0, \\ 0, & \text{if } \sigma = 0 \text{ and } r = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Taking this lower semi-continuous function leads to a well defined Concomitant Lasso estimator thanks to the following formulation

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}}{\arg\min} \ \operatorname{persp}_f^{**}(\sigma, y - X\beta) + \lambda \|\beta\|_1 . \tag{4.9}$$

The only difference with the original one is that we take $\hat{\sigma}^{(\lambda)} = 0$ if $y - X\hat{\beta}^{(\lambda)} = 0$. The actual objective function accepts $\sigma = 0$ as soon as $y = X\beta$. In the rest of this chapter, we will write (4.7) instead of the minimization of the biconjugate (4.9) as a slight abuse of notation. We refer to (Combettes, 2016) for a recent analysis of perspective functions.

4.1.1 Different Approaches and Points of View

As mentioned in the introduction, different independent approaches have led to equivalent formulations of the Concomitant Lasso. Among those we know, there is:

Link with the Square-root Lasso. Independently, another approach to overcome dependency of the Lasso estimator (4.1) was investigated in (Belloni et al., 2011) through the Square-root Lasso formulation:

$$\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{\sqrt{n}} \|y - X\beta\|_2 + \lambda \|\beta\|_1 \quad . \tag{4.10}$$

They show that the estimator $\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}}$ is pivotal *w.r.t.* to the standard deviation $\sigma_\star$ and does not require its estimate. Interestingly, this estimator is strongly related to the Concomitant Lasso. Recalling a basic relation in optimization (Hiriart-Urruty and Lemaréchal, 2012, Chapter 0): the decoupling also called transitivity of infima.

Proposition 40 (Decoupling of infima)**.** *Let P be a real valued function defined on $Z = Z_1 \times Z_2$. The couple of vector $(z_1^\star, z_2^\star)$ minimizes P over Z if and only if $z_2^\star$ minimizes $P(z_1^\star, \cdot)$ over Z_2 and $z_1^\star$ minimizes the function $\tilde{P} : z_1 \mapsto \inf_{z_2} P(z_1, z_2)$ over Z_1. Whence*

$$\inf_{(z_1,z_2)\in Z} P(z_1, z_2) = \inf_{z_1 \in Z_1} \inf_{z_2 \in Z_2} P(z_1, z_2) = \inf_{z_2 \in Z_2} \inf_{z_1 \in Z_1} P(z_1, z_2) \quad . \tag{4.11}$$

Applying Proposition 40, the map $\sigma \mapsto \frac{1}{2n\sigma} \|y - X\beta\|_2^2 + \frac{\sigma}{2}$ is minimized by $\sigma(\beta) = \frac{\|y - X\beta\|_2}{\sqrt{n}}$, whence $\min_{\beta \in \mathbb{R}^p, \sigma \geqslant 0} P_\lambda(\beta, \sigma) = \min_{\beta \in \mathbb{R}^p} \frac{1}{2n\sigma(\beta)} \|y - X\beta\|_2^2 + \frac{\sigma(\beta)}{2} + \lambda \|\beta\|_1$. Then we conclude that $\left(\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}}, \hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}}\right)$ with $\hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}} = \frac{1}{\sqrt{n}} \|y - X\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}}\|_2$, is a solution of the Concomitant Lasso (4.7) for all $\lambda > 0$. This shows that the Concomitant Lasso is a variational formulation of the Square-root Lasso.

Connection with the Lasso Path. The Lasso estimator (4.1) and the Concomitant Lasso (4.7) are strongly related, they share the same solution path up to rescaling of the regularization parameter. Indeed, by denoting temporarily $(\hat{\beta}^{(\lambda)}_{CL}, \widehat{\sigma}^{(\lambda)}_{\text{CL}})$ a solution of (4.7) and applying the Fermat's rule, we have:

$$\begin{aligned} X^\top (y - X\hat{\beta}^{(\lambda\sigma)}_{\text{L}}) &\in n\lambda\sigma \operatorname{sign}(\hat{\beta}^{(\lambda\sigma)}_{\text{L}}), \ \forall \sigma > 0, \\ X^\top (y - X\hat{\beta}^{(\lambda)}_{\text{CL}}) &\in n\lambda\widehat{\sigma}^{(\lambda)}_{\text{CL}} \operatorname{sign}(\hat{\beta}^{(\lambda)}_{\text{CL}}) \quad . \end{aligned}$$

Hence $(\hat{\beta}^{(\lambda\widehat{\sigma}^{(\lambda)}_{\text{CL}})}_{\text{L}}, \widehat{\sigma}^{(\lambda)}_{\text{CL}})$ is optimal for problem (4.7). It also gives the connection between the solution path $\hat{\beta}^{(\lambda\widehat{\sigma}^{(\lambda)}_{\text{CL}})}_{\text{L}} = \hat{\beta}^{(\lambda)}_{CL}$. This relation was exploited in (Sun and Zhang, 2012) in order to compute solutions of the Concomitant Lasso by alternating minimization (see Section 4.6) which turn out to be very efficient if one has access to the full Lasso path (by using for instance Lars homotopy algorithm (Efron et al., 2004)). However, as shown in (Mairal and Yu, 2012; Gärtner et al., 2012), the complexity of the full regularization path can be exponential in number of features p.

Robustness to Disturbance of the Features. Studying the robustness properties of the Lasso in the presence of a design matrix corrupted with a bounded disturbance, (Xu et al., 2010) have shown the following result.

Proposition 41 (Xu et al. (2010, Theorem 1))**.** *Given the uncertainty set* $\mathcal{U} := \{(\delta_1, \cdots, \delta_p) : \|\delta_j\|_2 \leqslant \lambda, \forall j \in [p]\}$*, the set of minimizers of the Square-root Lasso coincides with the minimizers of the following robust optimization problem*

$$\min_{\beta \in \mathbb{R}^p} \max_{\Delta X \in \mathcal{U}} \|y - (X + \Delta X)\beta\|_2 \quad . \tag{4.12}$$

Thus, we have an explicit connection between the Square-root Lasso which is pivotal *w.r.t.* to the noise level and a robust optimization problem whose solutions ensure protection against disturbances of the matrix of features. This property is therefore transmitted naturally to the Lasso estimators along the entire path of regularization. Therefore, the addition of parsimonious regularization can be re-interpreted as the exclusion of variables contaminated by a malicious disturbance. This highlight a fundamental correspondence between robustness and sparsity.

4.1.2 Critical Parameters for the Concomitant Lasso

Since it is difficult to get the right regularization parameter in advance, a principled way to tune Lasso-type programs is to perform a cross-validation procedure over a pre-set finite grid of parameters. This leads to a data-driven choice of regularizer requiring the computation of many estimators, one for each λ value. Usually, a geometrical grid $\lambda_t = \lambda_{\max}^{\mathrm{L}} 10^{-\delta(t-1)/(T-1)}, t \in [T]$ is used, for instance it is the default grid in `scikit-learn` (Pedregosa et al., 2011) and `glmnet` (Friedman et al., 2007), with $\delta = 3$. For the Concomitant Lasso, we now show that this method presents some numerical drawbacks. Let us first investigate the Fenchel dual formulation and the solutions for extreme values of λ.

Proposition 42. *Denoting* $\Delta_{X,\lambda} = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leqslant 1, \lambda\sqrt{n}\|\theta\| \leqslant 1\right\}$*, the dual formulation of the Concomitant Lasso reads*

$$\hat{\theta}^{(\lambda)} \in \operatorname*{arg\,max}_{\theta \in \Delta_{X,\lambda}} \underbrace{\langle y, \lambda\theta \rangle}_{D_\lambda(\theta)}. \tag{4.13}$$

For an optimal primal vector $\hat{\beta}^{(\lambda)}$*,* $\widehat{\sigma}^{(\lambda)} = \|y - X\hat{\beta}^{(\lambda)}\|/\sqrt{n}$*. Moreover, the Fermat's rule reads*

$$y = n\lambda\widehat{\sigma}^{(\lambda)}\hat{\theta}^{(\lambda)} + X\hat{\beta}^{(\lambda)}, \tag{4.14}$$

$$X^\top(y - X\hat{\beta}^{(\lambda)}) \in n\lambda\widehat{\sigma}^{(\lambda)}\partial\|\cdot\|_1(\hat{\beta}^{(\lambda)}). \tag{4.15}$$

As for the Lasso, the null vector is optimal for the Concomitant Lasso problem as soon as the regularization parameter becomes too large, as detailed in the next proposition.

Proposition 43. *We have* $\hat{\beta}^{(\lambda)} = 0$ *for all* $\lambda \geqslant \lambda_{\max} := \frac{\|X^\top y\|_\infty}{\|y\|\sqrt{n}}$.

Proof. The Fermat's rule states:

$$(0, \widehat{\sigma}^{(\lambda)}) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p, \sigma > 0} P_\lambda(\beta, \sigma) \iff 0 \in \frac{\{-X^\top y\}}{n\widehat{\sigma}^{(\lambda)}} + \lambda\mathcal{B}_\infty \iff \frac{1}{n\widehat{\sigma}^{(\lambda)}} \left\|X^\top y\right\|_\infty \leqslant \lambda.$$

Thus, the critical parameter is given by $\lambda_{\max} = \|X^\top y\|_\infty/(n\widehat{\sigma}^{(\lambda)})$, so the result follows noticing that when $\hat{\beta}^{(\lambda)} = 0$ one has $\widehat{\sigma}^{(\lambda)} = \|y\|/\sqrt{n} > 0$ (remind that we assumed $y \neq 0$, since otherwise $(0, 0)$ would be a solution for any λ). □

However, for the Concomitant Lasso, there is another extreme. Indeed, there exists a critical parameter $\lambda_{\min}$ such that the Concomitant Lasso is equivalent to the Basis Pursuit (Chen and Donoho, 1995) for all $\lambda \leqslant \lambda_{\min}$ and gives an estimate $\widehat{\sigma}^{(\lambda)} = 0$. We recall that the Basis Pursuit and its dual are given by

$$\hat{\beta}^{\mathrm{BP}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p : y = X\beta} \|\beta\|_1, \qquad \hat{\theta}^{\mathrm{BP}} \in \operatorname*{arg\,max}_{\theta \in \mathbb{R}^n : \|X^\top\theta\|_\infty \leqslant 1} \langle y, \theta \rangle. \tag{4.16}$$

Proposition 44. *For any $\lambda \leqslant \lambda_{\min} := 1/(\|\hat{\theta}^{\mathrm{BP}}\|\sqrt{n})$, $(\hat{\beta}^{\mathrm{BP}}, 0)$ is optimal for P_λ and $\hat{\theta}^{\mathrm{BP}}$ is optimal for D_λ.*

Proof. By strong duality in the Basis Pursuit problem $\|\hat{\beta}^{\mathrm{BP}}\|_1 = \langle y, \hat{\theta}^{\mathrm{BP}} \rangle$. Now, $(\hat{\beta}^{\mathrm{BP}}, 0)$ is admissible for P_λ (see formulation (4.9)) and $\hat{\theta}^{\mathrm{BP}}$ is admissible for D_λ as soon as $\lambda \leqslant \lambda_{\min} := 1/(\|\hat{\theta}^{\mathrm{BP}}\|\sqrt{n})$. One can check for $\lambda \leqslant \lambda_{\min}$ that

$$P_\lambda(\hat{\beta}^{\mathrm{BP}}, 0) = \lambda \|\hat{\beta}^{\mathrm{BP}}\|_1 = \lambda \langle y, \hat{\theta}^{\mathrm{BP}} \rangle = D_\lambda(\hat{\theta}^{\mathrm{BP}}).$$

We conclude that $(\hat{\beta}^{\mathrm{BP}}, 0)$ is optimal for the primal and $\hat{\theta}^{\mathrm{BP}}$ is optimal for the dual.

□

We can guarantee the existence of minimizers to the Concomitant Lasso, even if $\widehat{\sigma}^{(\lambda)} = 0$, but the problem becomes more and more ill-conditioned for smaller and smaller $\widehat{\sigma}^{(\lambda)}$. The previous proposition shows that for too small λ's, a Basis Pursuit solution will always be found, though numerically this might be challenging to get. Indeed, when λ approaches $\lambda_{\min}$, a coordinate descent algorithm (similar to the one described in Algorithm 6) encounters trouble to perform dual gap computations. This is because we estimate the dual variable by a ratio having both denominator and numerator of the order of σ, which is problematic when $\sigma \to 0$, see Eq. (4.27).

A solution could be to pre-compute $\lambda_{\min}$ to prevent the user from requesting computation involving λ's too close from the critical value. Nevertheless, solving the Basis Pursuit problem first, to obtain $\lambda_{\min}$, is not realistic. For instance, the split Bregman algorithm (Goldstein and Osher, 2009) involves a sequence of Lasso problems to solve. In homotopy approaches (*i.e.* when computing a path of λ's) that we consider, the most difficult problem to solve are the one associated with λ close to 0. Hence attacking the problem by first solving the hardest case will slow down the whole process, as one would not benefit from warm start computations.

To avoid these issues, we propose a slight modification of the objective function by adding a constraint on σ. We refer to this method as the Smoothed Concomitant Lasso following the terminology introduced by Nesterov (2005), see also (Beck and Teboulle, 2012) for more on smoothing in optimization.

4.1.3 Smoothed Concomitant Lasso

We now introduce our Smoothed Concomitant Lasso, by adding a noise level limit σ_0, aimed at avoiding numerical instabilities for too small λ values.

Definition 20. *For $\lambda > 0$ and $\sigma_0 > 0$, the Smoothed Concomitant Lasso estimator $\widehat{\beta}^{(\lambda,\sigma_0)}$ and its associated noise level estimate $\widehat{\sigma}^{(\lambda,\sigma_0)}$ are defined as solutions of the primal optimization problem*

$$(\widehat{\beta}^{(\lambda,\sigma_0)}, \widehat{\sigma}^{(\lambda,\sigma_0)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}}{\arg\min} \underbrace{\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 + \iota_{[\sigma_0, +\infty[}(\sigma)}_{P_{\lambda,\sigma_0}(\beta,\sigma)}. \tag{4.17}$$

Remark 16. *Another simple smoothing consists in adding a term of the form $\epsilon_0/(2n\sigma)$ for $\epsilon_0 > 0$ which leads to*

$$\min_{\beta \in \mathbb{R}^p, \sigma > 0} \frac{\|y - X\beta\|_2^2 + \epsilon_0}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 = \min_{\beta \in \mathbb{R}^p} \sqrt{\|y - X\beta\|_2^2 + \epsilon_0} + \lambda \|\beta\|_1 .$$

However the obtained estimator will not satisfies the scale equivariance property, though leading to a twice continuously differentiable objective, allowing second order approaches such as proximal newton algorithm (Lee et al., 2014).

In the same way that we have expressed the link between Concomitant Lasso and Square-root Lasso, we find that the Smoothed Concomitant Lasso is a variational formulation of the (shifted) Huber criterion applied to ℓ_2. For $\sigma_0 > 0$, let H_{σ_0} be the function of $u \in \mathbb{R}$ define as

$$H_{\sigma_0}(u) = \begin{cases} \frac{u^2}{2\sigma_0} + \frac{\sigma_0}{2} & \text{if } |u| \leqslant \sigma_0, \\ |u| & \text{otherwise.} \end{cases} \tag{4.18}$$

4.1.4 Detour on Smoothing Techniques for Non-smooth Optimization

To solve non-smooth optimization problems, algorithms based on first order information like the subgradient method have convergence rates in $O(1/\epsilon^2)$ which is considered to be *slow* for large scale problems. An adequate way to improve the rate is to finely exploit the structure of the functions involved and to solve a smooth representation of the problem while maintaining optimality guarantees on the initial problem. This strategy has been used successfully in (Nesterov, 2005) and leads to an improved rate of $O(1/\epsilon)$ for "max"-type functions. Here we recall the smoothing techniques following the unified presentation in (Beck and Teboulle, 2012).

A key idea is that a proper, closed and convex function P is $1/\nu$-smooth if and only if its conjugate P^* is ν-strongly convex (see Proposition 3). Hence to find a smooth approximation of a non-smooth function P, it suffices to add a strongly convex regularization on its conjugates.

For a proper, closed and convex function P, we recall that the Fenchel conjugation is an involution *i.e.*

$$P(z) = P^{**}(z) := \sup_{z^* \in \mathbb{R}^d} \{\langle z^*, z\rangle - P^*(z^*)\} \ .$$

Definition 21 (Inf-conv Smoothing)**.** *Let P be a proper, closed and convex function, $\nu > 0$ and let w be a differentiable convex function with $1/\nu$-Lipschitz continuous gradient. For any $\mu > 0$, we define $P^*_\mu(z^*) := P^*(z^*) + \mu w^*(z^*)$. Its conjugate $P_\mu(z) := \sup_{z^* \in \mathbb{R}^d}\{\langle z^*, z\rangle - P^*_\mu(z^*)\}$ is called (inf-conv) μ-smooth approximation of P.*

Proposition 45 (Beck and Teboulle (2012, Theorem 4.1))**.** *The function P_μ defined inDefinition 21 is proper, closed, convex and differentiable with $1/(\nu\mu)$-Lipschitz continuous gradient. Moreover,*

$$P_\mu(z) = P^{**}_\mu(z) = (P^* + (\mu w)^*)^*(z) = (P \,\square\, w_\mu)(z) \ , \tag{4.19}$$

where $w_\mu(\cdot) := \mu w(\frac{\cdot}{\mu})$ is the dilation of w. Whence

$$P_\mu(z) := \inf_{u \in \mathbb{R}^d} \left\{ P(u) + \mu w\left(\frac{z-u}{\mu}\right)\right\} \ . \tag{4.20}$$

Under the lighting of this smoothing tools, similarly to the connection between Square-root Lasso and Concomitant Lasso, we have the connection between a Smoothed Square-root Lasso and the Smoothed Concomitant Lasso.

Smoothing of the Euclidean Norm. For $P(z) = \|z\|_2$, $w(z) = \|z\|_2^2/2$ and $\mu = \sigma_0\sqrt{n}$, we have $P_{\sigma_0}(z) = H_{\sigma_0\sqrt{n}}(\|z\|_2) - \frac{\sigma_0\sqrt{n}}{2}$. Whence the smoothed counterpart of the Square-root Lasso reads:

$$\min_{\beta \in \mathbb{R}^p} H_{\sigma_0\sqrt{n}}(\|y - X\beta\|_2) - \frac{\sigma_0\sqrt{n}}{2} + \lambda \|\beta\|_1 \ .$$

Remark 17 (Smoothing of the ℓ_1 norm)**.** *For $P(z) = \sum_{i=1}^n |z_i|$, $w(z) = \|z\|_2^2/2$ and $\mu = \sigma_0$, we have $P_{\sigma_0}(z) = \sum_{i=1}^n H_{\sigma_0}(y_i - x_i^\top\beta) - \frac{\sigma_0 n}{2}$. Whence the smoothed counterpart of the Least Absolute Deviation (LAD) with Lasso penalty reads: (Huber, 1964)*

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n H_{\sigma_0}(y_i - x_i^\top\beta) - \frac{\sigma_0 n}{2} + \lambda \|\beta\|_1 \ .$$

Note that, without the ℓ_1 penalty term, this leads to the re-weighted iterative least squares algorithm to compute the LAD regression (Schlossmacher, 1973; Lange et al., 2000), see also the variational surrogate in (Mairal, 2015).

The above smoothing techniques improve the conditioning of non-smooth convex optimization problems and make it possible to use algorithms with better convergences speeds. This, combining with the high performance of (proximal) coordinate descent algorithms (Nesterov, 2012) allows us to tackle efficiently large scale problems.

Concurrently to our work, (Li et al., 2016b) adopted the smoothing of the euclidean norm point of view combined with an Iterative Soft-thresholding algorithms. However, their strategy has two noticeable drawbacks:

— Due to full gradient update, each iteration of their algorithm is expensive and it may take a large amount of time to converge in large scale setting.
— If the smoothing parameter σ_0 is small, their algorithm will make a tiny gradient step which further slows down the progression towards the optimum.

Taking the variational point of view along with a coordinate descent algorithm, safe removal of inactive variables and a new active warm start method in a homotopy continuation framework, we present a faster algorithm with less dependence on the smoothing parameter for solving concomitant estimation problem.

Critical parameter of the Smoothed Concomitant Lasso

Proposition 46. *The dual formulation of the Smoothed Concomitant Lasso reads*

$$\widehat{\theta}^{(\lambda,\sigma_0)} = \underset{\theta \in \Delta_{X,\lambda}}{\arg\max} \underbrace{\langle y, \lambda\theta \rangle + \sigma_0 \left(\frac{1}{2} - \frac{\lambda^2 n}{2} \|\theta\|^2 \right)}_{D_{\lambda,\sigma_0}(\theta)}, \tag{4.21}$$

for $\Delta_{X,\lambda} = \left\{ \theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leqslant 1, \|\theta\| \leqslant 1/(\lambda\sqrt{n}) \right\}$. Associated to an optimal primal vector $\widehat{\beta}^{(\lambda,\sigma_0)}$, we must have $\widehat{\sigma}^{(\lambda,\sigma_0)} = \sigma_0 \vee (\|y - X\widehat{\beta}^{(\lambda,\sigma_0)}\|/\sqrt{n})$.

We also have the link-equation between primal and dual solutions:

$$n\lambda\widehat{\sigma}^{(\lambda,\sigma_0)}\widehat{\theta}^{(\lambda,\sigma_0)} + X\widehat{\beta}^{(\lambda,\sigma_0)} = y, \tag{4.22}$$

$$X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)}) \in n\lambda\widehat{\sigma}^{(\lambda,\sigma_0)} \partial \|\cdot\|_1(\widehat{\beta}^{(\lambda,\sigma_0)}). \tag{4.23}$$

Remark 18. *The dual problem* (4.21) *also reads*

$$\widehat{\theta}^{(\lambda,\sigma_0)} = \underset{\theta \in \Delta_{X,\lambda}}{\arg\max} \frac{1}{2} \left\| \frac{y}{\sigma_0 n} \right\|_2^2 - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda\sigma_0 n} \right\|_2^2 = \Pi_{\Delta_{X,\lambda}} \left(\frac{y}{\lambda\sigma_0 n} \right) \quad .$$

Since $\Delta_{X,\lambda}$ is convex and closed, the solution $\widehat{\theta}^{(\lambda,\sigma_0)}$ is unique.

A similar reasoning to Proposition 43 gives the following critical parameter.

Proposition 47. *We have $\widehat{\beta}^{(\lambda,\sigma_0)} = 0$, for all $\lambda \geqslant \lambda_{\max} := \frac{\|X^\top y\|_\infty}{n(\sigma_0 \vee (\|y\|/\sqrt{n}))}$.*

Contrarily to the Concomitant Lasso, the parameter corresponding to zero estimates of the noise for the Smoothed Concomitant Lasso is $\lambda_{\min} = 0$. Indeed, fromRemark 16 we know that $\widehat{\theta}^{(\lambda,\sigma_0)}$ belongs to the boundary of the feasible set hence it is non zero. This combined with the link equation (4.22) and $\widehat{\sigma}^{(\lambda,\sigma_0)} \geqslant \sigma_0 > 0$, we have $y = X\widehat{\beta}^{(\lambda,\sigma_0)}$ if and only if $\lambda = 0$.

Choice of the Smoothing Parameter

In practice, the choice of σ_0 can be motivated as follows:

— Suppose we have prior information on the minimal noise level expected in the data. Then we can set σ_0 as this bound. Indeed, if $\widehat{\sigma}^{(\lambda,\sigma_0)} > \sigma_0$, then the constraint $\sigma \geqslant \sigma_0$ is not active and the optimal solution to Problem (4.17) is equal to the optimal solution to Problem (4.7). The Smoothed Concomitant Lasso estimator will only be different from the Concomitant Lasso estimator when the prediction given by the Concomitant Lasso violates the a priori information.

— Without prior information we can consider a given accuracy ϵ, and set $\sigma_0 = \epsilon$. Then, the theory of smoothing (Nesterov, 2005; Beck and Teboulle, 2012) tells us that any $\epsilon/2$-solution to Problem (4.17) is an ϵ-solution to Problem (4.7). Thus we obtain the same solutions, but as an additional benefit we have a control on the conditioning of the problem.

— If departing slightly from the Concomitant Lasso estimator is not too big of an issue, one can also use an arbitrary proportion of the initial estimation of the noise variance *i.e.* $\sigma_0 = \|y\|/\sqrt{n} \times 10^{-\alpha}$. This was our choice in practice, and we have set $\alpha = 2$. Indeed, taking a large enough value for σ_0 leads to less numerical issues.

Duality gap and link with the Lasso

From the optimality condition in (4.14) and (4.15), one can remark that if $\widehat{\beta}^{(\lambda,\sigma_0)}$ is a solution of the Smoothed Concomitant Lasso, then it is also a solution of the Lasso with regularization parameter $\lambda\widehat{\sigma}^{(\lambda,\sigma_0)}$. The following proposition estimates the quality (in term of duality gap) of a primal-dual vector in the Lasso path compared to Concomitant Lasso path. We recall the Lasso problem and its dual:

$$\hat{\beta}^{\lambda}_{\mathrm{L}} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \underbrace{\frac{1}{2n}\|y - X\beta\|^2 + \lambda\|\beta\|_1}_{P^{\mathrm{L}}_{\lambda}(\beta)} \ ,$$

$$\hat{\theta}^{\lambda}_{\mathrm{L}} = \operatorname*{arg\,max}_{\theta\in\mathbb{R}^n : \|X^\top\theta\|_\infty \leqslant 1} \underbrace{\frac{1}{2n}\|y\|^2 - \frac{1}{2n}\|y - \lambda n\theta\|^2}_{D^{\mathrm{L}}_{\lambda}(\theta)} .$$

Hence, defining the duality gap of the Lasso $G^{\mathrm{L}}_{\lambda}(\beta,\theta) = P^{\mathrm{L}}_{\lambda}(\beta) - D^{\mathrm{L}}_{\lambda}(\theta)$, and the duality gap of the Smoothed Concomitant Lasso $G_{\lambda,\sigma_0}(\beta,\theta,\sigma) = P_{\lambda,\sigma_0}(\beta,\sigma) - D_{\lambda,\sigma_0}(\theta)$, we have

Proposition 48. $\forall \beta \in \mathbb{R}^p, \theta \in \Delta_{X,\lambda}, \sigma \geqslant \sigma_0$, $G^{\mathrm{L}}_{\sigma\lambda}(\beta,\theta) \leqslant \sigma G_{\lambda,\sigma_0}(\beta,\sigma,\theta)$.

Proof. Since $\sigma - \sigma_0 \geqslant 0$ and $\lambda\sqrt{n}\|\theta\| \leqslant 1$, we have

$$\frac{\sigma\lambda^2 n}{2}\|\theta\|^2 = \frac{\sigma - \sigma_0}{2}\lambda^2 n\|\theta\|^2 + \frac{\sigma_0\lambda^2 n}{2}\|\theta\|^2 \leqslant \frac{\sigma - \sigma_0}{2} + \frac{\sigma_0\lambda^2 n}{2}\|\theta\|^2.$$

$$\begin{aligned}
G^{\mathrm{L}}_{\sigma\lambda}(\beta,\theta) &= P^{\mathrm{L}}_{\sigma\lambda}(\beta) - D^{\mathrm{L}}_{\sigma\lambda}(\theta) \\
&= \frac{1}{2n}\|y - X\beta\|^2 + \sigma\lambda\|\beta\|_1 - \frac{1}{2n}\|y\|^2 + \frac{1}{2n}\|y - \sigma\lambda n\theta\|^2 \\
&= \frac{1}{2n}\|y - X\beta\|^2 + \sigma\lambda\|\beta\|_1 - \sigma\lambda\langle y,\theta\rangle + \frac{\sigma^2\lambda^2 n}{2}\|\theta\|^2 \\
&\leqslant \sigma\left(\frac{1}{2n\sigma}\|y - X\beta\|^2 + \lambda\|\beta\|_1 - \lambda\langle y,\theta\rangle + \frac{\sigma}{2} - \sigma_0(\frac{1}{2} - \frac{\lambda^2 n}{2}\|\theta\|^2)\right) \\
&= \sigma\left(P_{\lambda,\sigma_0}(\beta,\sigma) - D_{\lambda,\sigma_0}(\theta)\right) = \sigma G_{\lambda,\sigma_0}(\beta,\theta).
\end{aligned}$$

□

Hence, as $\forall \lambda,\ \widehat{\sigma}^{(\lambda)} \leqslant \|y\|/\sqrt{n}$, if the duality gap for the Smoothed Concomitant Lasso is small, so is the duality gap for the Lasso with the corresponding regularization parameter.

Extension to Multi-Task Lasso

The strategies adopted so far to obtain pivotal estimator *w.r.t.* to the noise level can be extended to matrix formulation with observations $Y = X\mathrm{B}^\star + \Sigma^\star E$ where E is a standard multivariate Gaussian noise, by considering the matrix formulation with the trace norm as a loss function (van de Geer and Stucky, 2016)

$$\min_{\mathrm{B}\in\mathbb{R}^{p\times q}} \|Y - X\mathrm{B}\|_* + \lambda\Omega(\mathrm{B}) \ . \tag{4.24}$$

Following the variational Formulation for the trace norm (Argyriou et al., 2008), (Bach et al., 2012, Chapter 5.2), (van de Geer and Stucky, 2016, lemma 1) we have

$$\|\mathrm{B}\|_* = \frac{1}{2}\inf_{\Sigma>0} \mathrm{tr}(\mathrm{B}^\top\Sigma^{-1}\mathrm{B} + \Sigma) \ ,$$

which implies that the optimization problem (4.24) is equivalent to the Concomitant estimation

$$\min_{\mathrm{B}\in\mathbb{R}^{p\times q},\Sigma>0} \frac{1}{2}\mathrm{tr}((Y - X\mathrm{B})^\top\Sigma^{-1}(Y - X\mathrm{B}) + \Sigma) + \lambda\Omega(\mathrm{B}) \ .$$

Similarly to the Smoothed Concomitant Lasso, the non-smoothness in the loss may causes several numerical issues in the optimization process. Instead, a regularized version of the trace norm was considered in (Massias et al., 2018a)

$$\min_{\mathrm{B}\in\mathbb{R}^{p\times q},\Sigma\succeq\Sigma_0} \frac{1}{2}\mathrm{tr}((Y - X\mathrm{B})^\top\Sigma^{-1}(Y - X\mathrm{B}) + \Sigma) + \lambda\Omega(\mathrm{B}) \ .$$

A connection with a perspective transform can also be considered if we rely on the framework for perspective of matrix convex function in (Effros, 2009; Ebadian et al., 2011).

4.2 Faster Algorithm for Concomitant Lasso

Safe Screening Rules

In order to achieve greater computational efficiency, we propose new safe screening rules (using the terminology introduced in the seminal work El Ghaoui et al. (2012)) for our problem and we compare their performance. The principle underlying safe screening rules is as follows: one can discard inactive features from the optimization problem, thanks to the sub-differential inclusion (4.15) and to a safe region $\mathcal{R}$ such that $\widehat{\theta}^{(\lambda,\sigma_0)} \in \mathcal{R}$:

$$\max_{\theta\in\mathcal{R}} |X_j^\top\theta| < 1 \Rightarrow |X_j^\top\widehat{\theta}^{(\lambda,\sigma_0)}| < 1 \Rightarrow \widehat{\beta}_j^{(\lambda,\sigma_0)} = 0. \tag{4.25}$$

Since the dual objective of the Smoothed Concomitant Lasso is $\lambda^2\sigma_0 n$-strongly concave, we can provide a dynamic and converging SAFE sphere region $\mathcal{R}$, following the methodology introduced in (Ndiaye et al., 2015).

Proposition 49 (Gap Safe rule)**.** *For all $(\beta, \sigma, \theta) \in \mathbb{R}^p \times \mathbb{R}_+ \times \Delta_{X,\lambda}$, then for*

$$r = \sqrt{2G_{\lambda,\sigma_0}(\beta, \sigma, \theta)/(\lambda^2\sigma_0 n)},$$

we have $\widehat{\theta}^{(\lambda,\sigma_0)} \in \mathcal{B}(\theta, r)$. Thus, we have the following safe sphere screening rule

$$|X_j^\top\theta| + r\,\|X_j\| < 1 \Longrightarrow \widehat{\beta}_j^{(\lambda,\sigma_0)} = 0. \tag{4.26}$$

Another test, valid when $\sigma_0 = 0$, can be derived if we assume upper/lower bounds: to eliminate feature j, it is enough to check whether

$$\max_\theta \{|X_j^\top \theta| \,:\, \lambda\sqrt{n}\|\theta\| \leqslant 1, \quad \underline{\eta} \leqslant D_\lambda(\theta) \leqslant \overline{\eta}\} < 1.$$

In our implementation, we use the primal and the dual objective as a natural bound on the problem since $\underline{\eta} = D_\lambda(\theta_k) \leqslant D_\lambda(\widehat{\theta}^{(\lambda,\sigma_0)}) \leqslant P_{\lambda,\sigma_0}(\beta_k, \sigma_k) = \overline{\eta}$.

Proposition 50 (Bound Safe rule)**.** *Assume that, for a given $\lambda > 0$, we have an upper bound $\overline{\eta} \in (0, +\infty]$, and a lower bound $\underline{\eta} \in (0, +\infty]$ over the Smoothed Concomitant Lasso problem* (4.17)*. Denote by $x_j = X_j/\|X_j\|$ and $y' = y/\|y\|$ two unit vectors, and by $\underline{\gamma} = (\underline{\eta} - \sigma_0/2)\sqrt{n}/\|y\|$ and $\overline{\gamma} = \overline{\eta}\sqrt{n}/\|y\|$. Then if one of the three following conditions is met*

— $|X_j^\top y'| > \overline{\gamma}$ *and* $\overline{\gamma}|X_j^\top y'| + \sqrt{1-\overline{\gamma}^2}\sqrt{1-(X_j^\top y')^2} < \lambda\sqrt{n}/\|X_j\|$,

— $\underline{\gamma} \leqslant |X_j^\top y'| \leqslant \overline{\gamma}$ *and* $1 < \lambda\sqrt{n}/\|X_j\|$,

— $|X_j^\top y'| < \underline{\gamma}$ *and* $\underline{\gamma}|X_j^\top y'| + \sqrt{1-\underline{\gamma}^2}\sqrt{1-(X_j^\top y')^2} < \lambda\sqrt{n}/\|X_j\|$,

the j-th feature can be discarded i.e. $\widehat{\beta}_j^{(\lambda,\sigma_0)} = 0$.

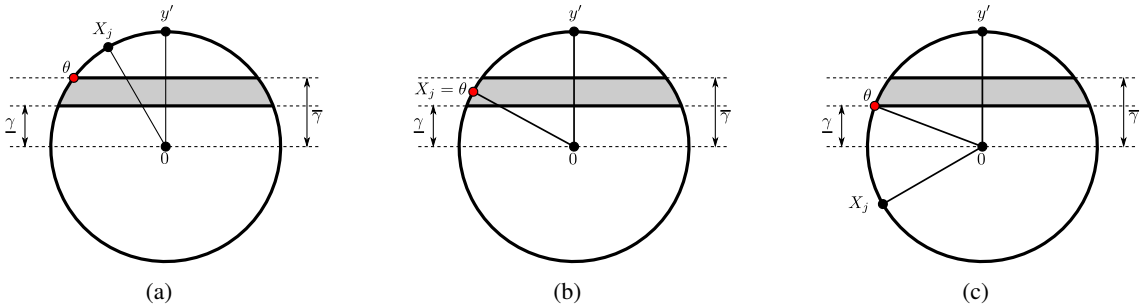


Figure 4.1 – *Screening out the jth feature X_j.* The gray region is the intersection between the ball $\{\theta \in \mathbb{R}^n : \|\theta\| \leqslant 1\}$ and the set $\{\theta \in \mathbb{R}^n : \underline{\gamma} \leqslant \theta^\top y' \leqslant \overline{\gamma}\}$, for the three possible regimes. The point θ displayed above is the one achieving $\max\{|\theta^\top X_j| : \|\theta\| \leqslant 1, \quad \underline{\gamma} \leqslant {y'}^\top \theta \leqslant \overline{\gamma}\}$, where X_j is any feature we aim at screening out.

First, note that there are two privileged directions in the optimization problem at stake: X_j and y. Then, we can see that the dual feasible set is constrained to be inside the Euclidean unit ball, and by strong duality we know that $\underline{\gamma} \leqslant \theta^\top y' \leqslant \overline{\gamma}$. Hence, to screen out X_j, we aim at solving

$$\max\{|\theta^\top x| \quad : \quad \|\theta\| \leqslant 1, \quad \underline{\gamma} \leqslant {y'}^\top \theta \leqslant \overline{\gamma}\},$$

for any $x = X_j$ (not yet dis-activated by a previous safe screening). As we can see on Figure 4.1, there are three regimes depending on the position of X_j relative to the bound constraints. Technical details of the derivation can be found in Section 4.6.

Smoothed Concomitant Lasso algorithm (SC)

We first present the inner loop of our main algorithm, *i.e.* the implementation of coordinate descent for the Smoothed Concomitant Lasso. In Algorithm 6, we denote by $\mathcal{A}$ the active set, *i.e.* the set of coordinates that we have not screened out. For safe screening rules, this set is guaranteed to contain the support of the optimal solution.

We now present in Algorithm 7, the fast solver we proposed for the Smoothed Concomitant Lasso, relying on the three following key features: coordinate descent, Gap Safe screening rules and improved warm start propositions.

Algorithm 6 CD4SCL – Coordinate Descent for the Smoothed Concomitant Lasso with Gap Safe screening

Input : $X, y, \epsilon, K, f^{\text{ce}}(= 10), \lambda, \sigma_0, \beta, \sigma$
$\mathcal{A} \leftarrow [p]$
for $k \in [K]$ **do**
  if $k \mod f^{\text{ce}} = 1$ **then**
    Compute θ thanks to (4.27)
    if $G_{\lambda,\sigma_0}(\beta, \sigma, \theta) = P_{\lambda_t,\sigma_0}(\beta, \sigma) - D_{\lambda_t,\sigma_0}(\theta) \leqslant \epsilon$. **then** // Stopping criterion
      break
    Update $\mathcal{A}$ thanks to Proposition 49 // Screening test
  for $j \in \mathcal{A}$ **do** // Loop over coordinates
    $\beta_j \leftarrow \mathrm{ST}_{\frac{n\sigma\lambda_t}{\|x_j\|^2}}\left(\beta_j - \frac{x_j^\top (X\beta - y)}{\|x_j\|^2}\right)$ // Soft-thresholding step
    $\sigma \leftarrow \sigma_0 \vee \frac{\|y - X\beta\|}{\sqrt{n}}$ // Noise estimation step
Output: β, σ, $\mathcal{A}$

Coordinate Descent

The algorithm we consider to compute the Smoothed Concomitant Lasso is coordinate descent, an efficient way to solve Lasso-type problem (even for multiple values of parameters) (Friedman et al., 2007). Previous attempts mainly focused on iteratively alternating Lasso steps along with noise level estimation (Sun and Zhang, 2012) [1], or conic programming (Becker et al., 2011). In (Li et al., 2016b), written concurrently to this work, the authors consider ISTA, a first order method using full gradient information at each iteration.

Here we provide a simple and efficient coordinate descent approach, *cf.* Algorithm 6. Our primal objective P_{λ,σ_0} can be written as the sum of a convex differentiable function $f(\beta, \sigma) = \|y - X\beta\|^2/(2n\sigma) + \sigma/2$ and of a separable function $g(\beta, \sigma) = \lambda\|\beta\|_1 + \iota_{[\sigma_0,+\infty[}(\sigma)$. Moreover, for $\sigma \geqslant \sigma_0 > 0$, the gradient of f is Lipschitz continuous. Hence, we know that the coordinate descent method converges to a minimizer of our problem (Yun, 2014). We choose to update the variable σ every other iteration because this can be done at a negligible cost.

Remark 19 (Larger (Coordinate) Gradient Step)**.** *For $\sigma_0 \leqslant \sigma$, the function $f(r, \sigma) = \|r\|_2^2/(2\sigma) + \sigma/2$, involved in the variational formulation, is $(1/\sigma)$-Lipschitz continuous gradient. The Huber criterion (smoothed variant of the loss of the Square-root Lasso) H_{σ_0} is $(1/\sigma_0)$-Lipschitz continuous gradient. A (coordinate) gradient based algorithm directly launched on this loss function will be slowed down when σ_0 is (too) small. However, it will always make larger step in the variational formulation with $f(r, \sigma)$ than the Huber loss H_{σ_0} specially when σ is much larger than σ_0. Hence the algorithm* (7) *enjoys the local smoothness and then is less sensitive to the parameter σ_0.*

Our stopping criterion is based on the duality gap defined by $G_{\lambda,\sigma_0}(\beta, \sigma, \theta) = P_{\lambda,\sigma_0}(\beta, \sigma) - D_{\lambda,\sigma_0}(\theta)$. This requires the computation of a dual feasible point, that, provided a primal vector β, can be obtained as follows

$$\theta = \frac{y - X\beta}{\lambda n\sigma_0 \vee \|X^\top(y - X\beta)\|_\infty \vee \lambda\sqrt{n}\,\|y - X\beta\|}. \tag{4.27}$$

This choice of dual point is motivated by the following convergence result.

Proposition 51. *Let $(\beta_k)_{k\in\mathbb{N}}$ be a sequence that converges to $\widehat{\beta}^{(\lambda,\sigma_0)}$. Then $(\theta_k)_{k\in\mathbb{N}}$ built thanks to* (4.27) *converges to $\widehat{\theta}^{(\lambda,\sigma_0)}$. Hence the sequence of dual gap $(G_{\lambda,\sigma_0}(\beta_k, \sigma_k, \theta_k))_{k\in\mathbb{N}}$ converges to zero.*

1. a description of their algorithm is given in Section 4.6 for completeness

Active warm start

In Algorithm 7, the first occurrence of CD4SCL is a warming step aimed at improving the current primal point at a low cost. For Gap Safe, we disable it by setting $K_0 = 0$. For the experiments with the Active warm start, we have set $K_0 = K = 5000$ and $\epsilon_0 = \epsilon$.

Concerning the parameter f^{ce} it governs how often we perform the dual gap evaluation. Due to the complexity of this step, we do not recommend to do this step every pass over the features, but rather compute this quantity less often, every f^{ce} passes. In practice we have fixed its value to $f^{\text{ce}} = 10$ for all our experiments.

Algorithm 7 Coordinate Descent for the Smoothed Concomitant Lasso with the Active warm start screening

Input : $X, y, \epsilon, \epsilon_0, K, K_0, f^{\text{ce}}, (\lambda_t)_{t\in[T-1]}, \sigma_0$
$\lambda_0 = \lambda_{\max} = \|X^\top y\|_\infty / (\|y\|\sqrt{n}), \quad \beta^{\lambda_0} = 0, \quad \sigma^{\lambda_0} = \|y\|/\sqrt{n}$
$\mathcal{A} \leftarrow [p]$
for $t \in [T-1]$ **do**
  $\beta, \sigma \leftarrow \beta^{\lambda_{t-1}}, \sigma^{\lambda_{t-1}}$ (previous ϵ-solution) // Get previous ϵ-solution
  $\beta, \sigma, \_ \leftarrow \text{CD4SCL}(X_{\mathcal{A}}, y, \epsilon_0, K_0, f^{\text{ce}}, \lambda_t, \sigma_0, \beta, \sigma)$ // Active warm start step
  $\beta, \sigma, \mathcal{A} \leftarrow \text{CD4SCL}(X \ , y, \epsilon \ , K \ , f^{\text{ce}}, \lambda_t, \sigma_0, \beta, \sigma)$ // Standard loop
  $\beta^{\lambda_t}, \sigma^{\lambda_t} \leftarrow \beta, \sigma$
Output: $(\beta^{\lambda_t})_{t\in[T-1]}$, $(\sigma^{\lambda_t})_{t\in[T-1]}$

4.3 Re-parameterization of Exponential Family

By studying finite mixture of regressions model in a context where the number of covariates are larger than the sample size, Städler et al. (2010) proposed to use a ℓ_1-penalized maximum likelihood estimator in order to obtain a join estimation of the mean parameter and noise level. Unfortunately, a direct formulation leads to the following non-convex optimization problem

$$\min_{\beta\in\mathbb{R}^p, \sigma^2} \log(\sigma) + \frac{1}{2n\sigma^2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \ .$$

Moreover, the corresponding estimators fail to be equivariant under scaling of the responses vector y. To overcome the aforementioned drawbacks, they propose a scaling of the regularization parameter with σ to achieves equivariance and an adequate re-parameterization to obtain convexity.

Scaled Regularization. As the estimator of the noise is indirectly affected by the amount of regularization λ, it is beneficial to take it into account as follows

$$\min_{\beta\in\mathbb{R}^p, \sigma^2} \log(\sigma) + \frac{1}{2n\sigma^2} \|y - X\beta\|_2^2 + \lambda \frac{\|\beta\|_1}{\sigma} \ . \tag{4.28}$$

It is now easy to see that the obtained estimator is equivariant under scaling of y but the optimization problem still non jointly convex (indeed it is not convex in σ).

Convex Re-parameterization. By simply taking the change of variable

$$\theta_j = \frac{\beta_j}{\sigma}, \qquad \rho = \frac{1}{\sigma} \ , \tag{4.29}$$

the formulation (4.28) reads

$$\min_{\phi, \rho>0} -\log(\rho) + \frac{1}{2n} \|\rho y - X\phi\|_2^2 + \lambda \|\phi\|_1 \tag{4.30}$$

which is jointly convex but it worth noting that it does not preserve the equivariance at all.

This convexity obtained by a simple change of variable is strongly related to the fact that Maximum Likelihood Estimator (MLE) with a distribution for the exponential family naturally leads to convex optimization problem. We recall from (Brown, 1986) the classical definition and convexity properties of exponential model.

Definition 22 (Exponential Family)**.** *Let ν be a σ-finite measure and $\lambda(\theta) = \int \mathrm{e}^{\theta y}\nu(dy)$ its Laplace transform and $N := \{\theta : \lambda(\theta) < +\infty\}$. For $P(\theta) = \log(\lambda(\theta))$, we define*

$$p_\theta(y) = \exp(\langle \theta, y\rangle - P(\theta)) \ . \tag{4.31}$$

The family of density $\{p_\theta : \theta \in \Theta \subset N\}$ is called (standard) exponential family.

Proposition 52. *N is a convex set and P is a convex function on N. Furthermore, P is lower semi-continuous on $\mathbb{R}^d$ and continuous on the interior of N.*

Proof. Let $\alpha \in [0, 1]$, we have:

$$\begin{aligned}
\lambda(\alpha\theta_1 + (1-\alpha)\theta_2) &= \int \mathrm{e}^{(\alpha\theta_1+(1-\alpha)\theta_2)y}\nu(dy) = \int \left(\mathrm{e}^{\theta_1 y}\right)^{\alpha}\left(\mathrm{e}^{\theta_2 y}\right)^{1-\alpha}\nu(dy) \\
&\leqslant \left(\int \mathrm{e}^{\theta_1 y}\nu(dy)\right)^{\alpha}\left(\int \mathrm{e}^{\theta_2 y}\nu(dy)\right)^{1-\alpha} \quad \text{by Holder's inequality} \\
&= \lambda(\theta_1)^{\alpha}\lambda(\theta_2)^{1-\alpha}.
\end{aligned}$$

Then θ_1 and θ_2 belong to N implies $\alpha\theta_1 + (1-\alpha)\theta_2$ also belongs to N. Thus N is a convex set and the convexity of P follows:

$$\begin{aligned}
P(\alpha\theta_1 + (1-\alpha)\theta_2) = \log(\lambda(\alpha\theta_1 + (1-\alpha)\theta_2)) &\leqslant \alpha\log(\lambda(\theta_1)) + (1-\alpha)\log(\lambda(\theta_2)) \\
&= \alpha P(\theta_1) + (1-\alpha)P(\theta_2).
\end{aligned}$$

□

Hence for a random variable y with gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, we have

$$p_{(\mu,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \exp\left\{\langle(\theta_1,\theta_2),(y,y^2)\rangle - P(\theta_1,\theta_2)\right\} \tag{4.32}$$

where

$$\theta = (\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right) \text{ and } P(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta_2}\right). \tag{4.33}$$

Hence considering a regression setting where $\mu = X\beta$, and considering the change of variable

$$\phi_j = \frac{\beta_j}{\sigma^2}, \qquad \rho = \frac{1}{\sigma^2} \ , \tag{4.34}$$

the ℓ_1-penalized MLE in (4.32) leads to the Natural Lasso proposed in (Yu and Bien, 2017)

$$\min_{\phi\in\mathbb{R}^p,\rho>0} -\frac{1}{2}\log(\rho) + \rho\frac{\|y\|_2^2}{2n} - \frac{1}{n}y^\top X\phi + \frac{\|X\phi\|_2^2}{2n\rho} + \lambda\|\phi\|_1 \ . \tag{4.35}$$

The obtained estimator is jointly convex but fail to be equivariant. Written in a factorized form, we see better the similarity with the estimator proposed by Städler et al. (2010):

$$\min_{\phi,\rho>0} -\log(\sqrt{\rho}) + \frac{1}{2n}\left\|\sqrt{\rho}y - X\frac{\phi}{\sqrt{\rho}}\right\|_2^2 + \lambda\|\phi\|_1 . \tag{4.36}$$

$\hat{\sigma}_{OR}$	$\hat{\sigma}_{\mathcal{M}-CV}$	$\hat{\sigma}_{\mathcal{M}-LS}$	$\hat{\sigma}_i$	$\widehat{\sigma}_{D2}$
$\frac{\|y-P_{X,S^\star}y\|}{\sqrt{n-\|S^\star\|}}$	$\frac{\|y-X\hat{\beta}^{\lambda_{cv}}_{\mathcal{M}}\|}{\sqrt{n-\|\hat{S}^{\lambda_{cv}}_{\mathcal{M}}\|}}$	$\frac{\|y-P_{X,\hat{S}_{\mathcal{M}}}y\|}{\sqrt{n-\|\hat{S}_{\mathcal{M}}\|}}$	$\frac{\|y^{(i')}-P_{X^{(i')},\hat{S}_i}y^{(i')}\|}{\sqrt{n/2-\|\hat{S}_i\|}}$	$\frac{\left(1+\frac{p\widehat{m}_1^2}{(n+1)\widehat{m}_2}\right)\|y\|^2}{n} - \frac{\widehat{m}_1\|X^\top y\|^2}{\sqrt{n(n+1)\widehat{m}_2}}$

Table 4.1 – The estimator $\hat{\beta}_{\mathcal{M}}$ are obtained by a method $\mathcal{M}$ and $\mathcal{M}-LS$ is its least square refitting. We note $S^{\odot} = \{j \in [p],\, \beta_j^{\odot} \neq 0\}$, $D_i = (y^{(i)}, X^{(i)})_{i\in[2]}$ is a split in two parts of the observations, and $\hat{S}_i$ the support selected after a cross-validation on the part D_i. The RCV estimator is $\hat{\sigma}_{RCV} = ((\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2)^{1/2}$, and $\widehat{m}_1 = \mathrm{tr}(\widehat{\Sigma})/p$ and $\widehat{m}_2 = \mathrm{tr}(\widehat{\Sigma}^2)/p - (\mathrm{tr}(\widehat{\Sigma}))^2/(pn)$.

To obtain an equivariant estimator, (Yu and Bien, 2017) use a perspective function of the squared ℓ_1 norm as a penalty term and propose the Organic Lasso

$$\min_{\phi,\rho>0} -\log(\sqrt{\rho}) + \frac{1}{2n}\left\|\sqrt{\rho}y - X\frac{\phi}{\sqrt{\rho}}\right\|_2^2 + \lambda\left\|\frac{\phi}{\sqrt{\rho}}\right\|_1^2 \tag{4.37}$$

Furthermore, they show that the estimator obtained with the Organic Lasso formulation is a minimizer of the ℓ_1^2 penalized least square

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|y - X\beta\|^2 + \lambda\|\beta\|_1^2 \quad .$$

Remark 20 (Homogeneity and Equivariance)**.** *The functions ρ and Ω are positively homogeneous with the same degree d if and only of for any $t \in \mathbb{R}$, we have*

$$\min_{\beta\in\mathbb{R}^p} \ell(ty - X(t\beta)) + \lambda\Omega(t\beta) = |t|^d \min_{\beta\in\mathbb{R}^p} \ell(y - X\beta) + \lambda\Omega(\beta) \quad .$$

Whence the Square-root Lasso and Organic Lasso are equivariant under scaling transformation of the observations y and the Lasso is not. Basically, any couple of loss and regularizer (ℓ, Ω) with same degree of homogeneity will produce an equivariant estimator.

Extension to Multivariate Setting

The re-parameterization procedure also works in matrix setting with Multivariate Gaussian Distribution where the density is expressed as

$$\begin{aligned} p_{\mu,\Sigma}(y) &= \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{(y-\mu)^\top\Sigma^{-1}(x-\mu)}{2}\right) \\ &= \exp\left\{\langle(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}), (y, yy^\top)\rangle - P\left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right)\right\} \quad , \end{aligned}$$

where

$$P\left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right) = -\frac{1}{4}\langle(-\frac{1}{2}\Sigma^{-1})^{-1}\Sigma^{-1}\mu, \Sigma^{-1}\mu\rangle + \frac{1}{2}\log\left(\frac{(-\pi)^d}{\det(-\frac{1}{2}\Sigma^{-1})}\right) \quad .$$

In this case, the canonical parameter is $\Theta = (\Theta_1, \Theta_2) = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$ and the cumulant function $P(\Theta)$ is convex. Hence following the same route as (Yu and Bien, 2017), a multivariate Natural (and Organic) Lasso can be obtained.

4.4 Numerical Experiments

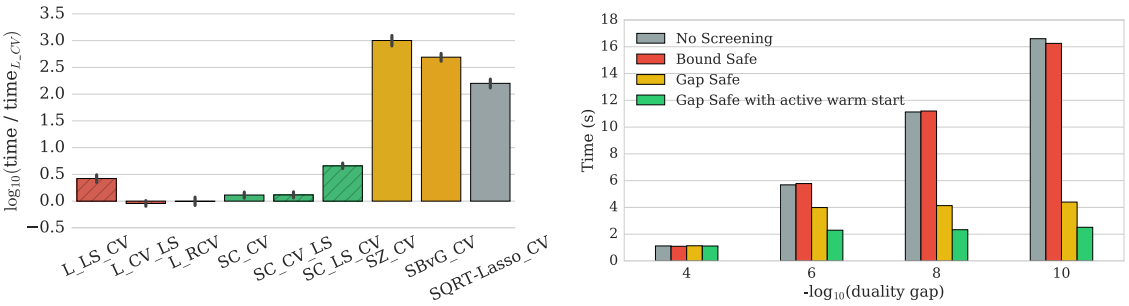
We compare the estimation performance and computation times of standard deviation estimators which are presently the state-of-the-art in high dimensional settings. We refer to (Reid et al.,

2013) for a recent comparison. In our simulations we use the common setup: $y = X\beta^\star + \sigma\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_n)$ and $X \in \mathbb{R}^{n\times p}$ follows a multivariate normal distribution with covariance $\Sigma = (\rho^{|i-j|})_{i,j\in[p]}$. We define $\beta^\star = \alpha\beta$ where the coordinates of β are drawn from a standard Laplace distribution and we randomly set $s\%$ of them to zero. The scalar α is chosen in order to satisfy a prescribed signal to noise ratio denoted snr: $\alpha = \sqrt{\text{snr} \times \sigma^2/\beta^\top\Sigma\beta}$. We note $S^\star = \{j \in [p],\, \beta^\star_j \neq 0\}$.

The procedures we have compared are summarized in Table 4.1. Namely, our reference is the oracle estimator (OR) $\hat{\sigma}_{OR}$, the cross-validated estimator (CV) $\hat{\sigma}_{\mathcal{M}-CV}$ whith a parameter λ_{cv} chosen by 5-fold cross-validation, the least-square refitting estimator (LS) $\hat{\sigma}_{\mathcal{M}-LS}$, the refitted cross-validation (RCV) $\hat{\sigma}_{RCV}$ and $\hat{\sigma}_{D2}$ the estimator introduced in (Dicker, 2014).

We run all the following algorithms over the non-increasing sequence $\lambda_t = \lambda_{\max} 10^{-\delta\frac{t-1}{T-1}}$ for t in $[T]$ with the default value $\delta = 2$ and $T = 100$. The regularization grid for the joint estimations (Scaled-Lasso, with solver from (Sun and Zhang, 2012) (SZ), Smoothed Concomitant Lasso (SC), Square-root Lasso (Belloni et al., 2011) (SQRT-Lasso) and the estimator introduced in (Städler et al., 2010) (SBvG)) begins at $\lambda_{\max}$ given in Proposition 47. We set Smoothed Concomitant Lasso with the default value $\sigma_0 = \|y\|/\sqrt{n} \times 10^{-2}$. As explained in Section 4.1.3 this choice improves numerical efficiency at the cost of departing slightly from the Concomitant Lasso estimator in the low noise regime. The grid for the Lasso (L) estimators begins with $\lambda^{\mathrm{L}}_{\max} = \|X^\top y\|_\infty/n$. The Lasso with the universal parameter $\lambda = \sqrt{2\log(p)/n}$ is denoted (L_U) and SZ refers to Concomitant Lasso with the quantile regularization described in (Sun and Zhang, 2013) in Fig. 4.3(a).

For each method, 50 replications are computed from the model aforementioned.



(a) Times to run simulations using synthetic dataset.

(b) Time to reach convergence using Leukemia dataset.

Figure 4.2 – Comparisons of the computational times using different estimation method (time presented relative to the mean time of the Lasso). (b): speed up using screening rules for the Smoothed Concomitant Lasso w.r.t. to duality gap and for $(\lambda_t)_{t\in[100]}$. The dimensions of Leukemia dataset are $(n = 72, p = 7129)$.

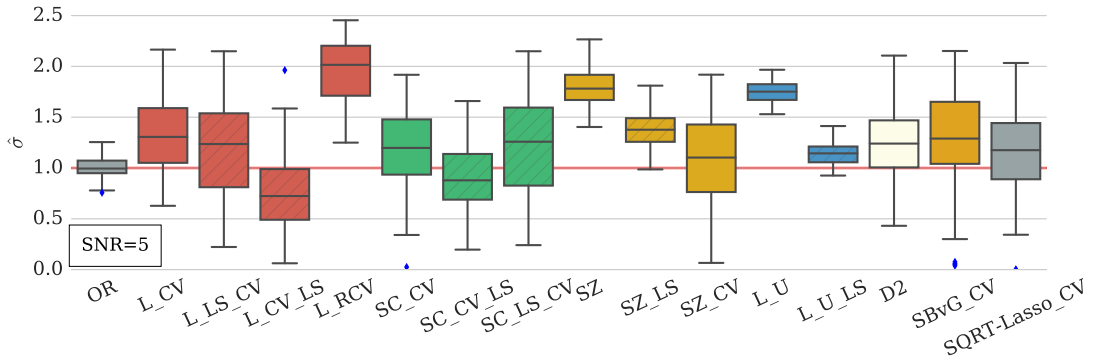
Computational Performance

Figure 4.2(a) presents on the Leukemia dataset the computation times observed for the different CV methods. The Smoothed Concomitant Lasso is based on the coordinate descent algorithm described in Algorithm 7, written in Python and Cython to generate low level C code, offering high performance. When a Lasso solver is needed, we have used the one from `scikit-learn`, that is coded similarly. For SZ_CV, computations are quite heavy as one uses the alternating algorithm proposed in (Sun and Zhang, 2012). Depending on the regularization parameter (for instance when one approaches $\lambda_{\min}$) the SZ_CV method is quite intractable and the algorithm faces the numerical issues mentioned earlier. The generic solver used for SBvG and SQRT-Lasso, is the `CVXPY` package (Diamond and Boyd, 2016), explaining why these methods are two orders of magnitude slower than a Lasso. This is in contrast to our solver that reaches similar computing time w.r.t. an efficient Lasso solver, with the additional benefit of jointly estimating the coefficients and the

standard deviation of the noise.

Figure 4.2(b) shows the benefit one can obtain thanks to the safe screening rules introduced above. The Bound safe rule on the Smoothed Concomitant Lasso problem does not show significant acceleration w.r.t. the Gap Safe rule. Indeed, the Gap Safe rule greatly benefits from the convergence of the dual vector, leading to smaller and smaller safe sphere as the iterations proceeds (Fercoq et al., 2015; Ndiaye et al., 2015). Another nice feature for the Gap Safe rules relies on a new warm start strategy when computing the full grid $(\lambda_t)_{t\in T}$. For a new λ, one first performs the optimization over the safe active set (*i.e.* the non discarded variables) from the previous λ. This *active warm start* strategy improves the warm start by providing a better primal vector. It helps achieving solutions with great precision at lower cost (up to $8\times$ speed-up on the Leukemia dataset).

Performance of Standard Deviation Estimators



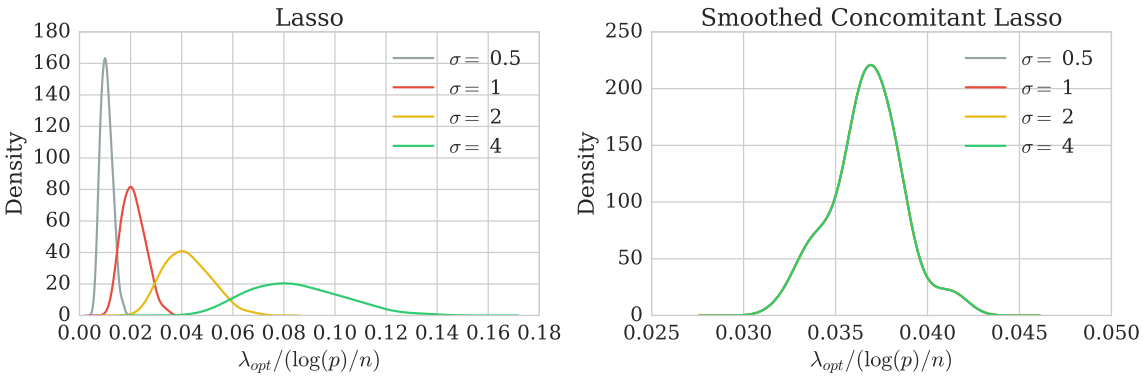
(a) Estimated performance using synthetic dataset.

Figure 4.3 – Comparison of quality of different estimators of the noise σ normalized to 1. The synthetic datasets are generated with the settings ($n = 100, p = 500, \rho = 0.6, \text{snr} = 5, s = 0.9$, 50 replications)

As noted earlier in (Fan et al., 2012), spurious correlations can strongly affect sparse regression and usually lead to large biases. This makes the standard deviation estimation very challenging and affects the cross-validation estimator based on the Lasso as they usually underestimate the standard deviation. The phenomenon is amplified when one uses least squares refitting on the cross-validated Lasso, as noticed in (Reid et al., 2013). Here we show an example where refitting cross-validation degrades the estimation. Results are presented as boxplots in Figure 4.3(a) (see Section 4.6 for additional settings).

In our experiments, we observe that SC and SZ are very efficient in high sparsity settings with low correlations, correcting for the positive bias of the estimator estimator from (Städler et al., 2010) (SBvG). In (Reid et al., 2013), it was also argued that the cross-validation estimator based on Lasso is more stable and performs better when the sparsity decreases and when the snr increases. We would like to emphasize that this is not the case when one performs a cross-validation procedure on the Concomitant Lasso. Here, we show that the latter achieves performances of the same order than the Lasso. It is worth noting that our method is consistently good over the whole experiments we conducted especially when applying least squares refitting.

Another appealing good property of the Smoothed Concomitant Lasso compared to the Lasso is the invariance of the optimal $\lambda_{opt} := \arg\min_{\lambda\in(\lambda_t)_{t\in[T]}} \|X\widehat{\beta}^{(\lambda,\sigma_0)} - X\beta^\star\|_2$ w.r.t. different levels of noise. We show on Figure 4.4(a) a kernel density plot of its distribution on synthetic data with different values of σ. A similar experiment was performed in (Li et al., 2016b) leading to the same conclusion with an optimal λ chosen by a train/test procedure.



(a) Estimated distribution of the optimal λ_{opt}.

Figure 4.4 – Comparisons of the distribution of optimal regularizer λ_{opt} under different levels of noise.

4.5 Conclusion and Perspectives

We have explored the joint estimation of the coefficients and noise level for ℓ_1 regularized regression. We have corrected some numerical drawbacks of the Concomitant Lasso estimator by proposing a slightly smoother formulation, leading to the Smoothed Concomitant Lasso. A fast algorithm, relying both on coordinate descent and on safe screening rules with improved warm start was investigated, and it was shown to achieve the same numerical efficiency than for the Lasso while also estimating the noise level. It could be interesting in future research to extend our work to more general data-fitting terms (Owen, 2007) and to combine sketching techniques as in (Pham and Ghaoui, 2015).

The linear model considered in this chapter is a special case of the Generalized Linear Model (McCullagh and Nelder, 1989) which turn out to be a special case of Dispersion Model (Jorgensen, 1997). In this later case, one assume that the observation y has a probability distribution of the form

$$p(y, \mu, \sigma^2) = a(y, \sigma^2) \exp\left(-\frac{\ell(y, \mu)}{\sigma^2}\right),$$

where respectively μ and $\sigma^2 > 0$ are *position* and *dispersion* parameters, a is a nonnegative function and ℓ a loss function. It allows for a general joint modeling of the mean and dispersion parameter. A future work could be to extent our approach and propose faster algorithm for learning such a model while associated with a variable selection regularizer (Antoniadis et al., 2016).

Recently, Combettes and Müller (2018) have proposed a general framework for representing convex M-estimators with concomitant scale as perspective functions and show that they can be solved with proximal splitting algorithms. It should be interesting to study the coordinate descent (with variational formulation) counterpart of such algorithms.

4.6 Appendix

Scaled-Lasso Algorithm (SZ)

We describe in Algorithm 8 the algorithm proposed by Sun and Zhang (2012) to compute the Scaled-Lasso and refer to it as SZ.

In our experiments we have used with the default parameters of the associated R packages `scalreg-package`, and a choice of λ's following the quantile oriented method described in (Sun and Zhang, 2013).

Note that contrary to our approach the stopping criterion is only based on checking the absence of consecutive increments on the noise level, whereas we consider dual gap evaluations as a more principled way.

Concerning the Lasso steps, as for other Lasso computations in our experiments, we have used the Lasso solver from `scikit-learn` with a dual gap tolerance of 10^{-4} and the other parameters set to their default values.

Algorithm 8 Scaled-Lasso algorithm (Sun and Zhang, 2012) for a fixed λ value

Input : $X, y, \epsilon(= 10^{-4}), K = 100, \lambda, \sigma_{\text{old}}(= 5), \sigma_{\text{new}}(= 0.1)$
$k = 0$
while $|\sigma_{\text{old}} - \sigma_{\text{new}}| > \epsilon$ *and* $k < K$ **do**
  $k \leftarrow k + 1$
  $\sigma_{\text{old}} = \sigma_{\text{new}}$
  $\lambda^L \leftarrow \lambda \sigma_{\text{old}}$
  $\beta \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|^2 + \lambda^L \|\beta\|_1$ `// Lasso step with parameter` λ^L
  $\sigma_{\text{new}} = \|y - X\beta\| / \sqrt{n}$ `// Noise estimation step`
Output: $(\beta, \sigma_{\text{new}})$

Additional Experiments

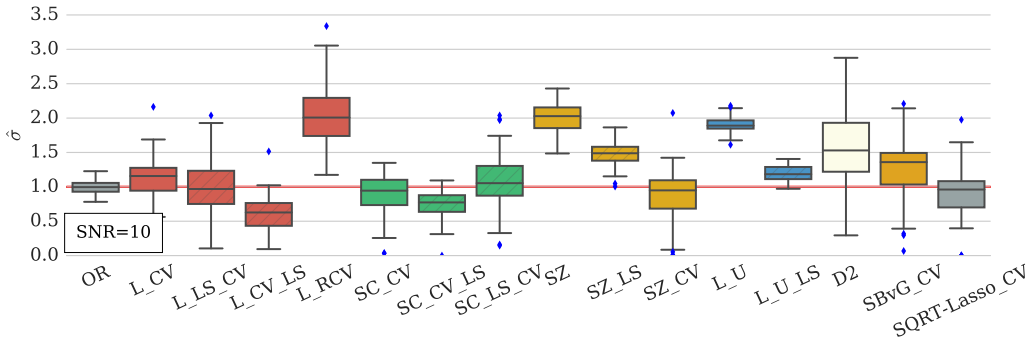
In this section, we present some extensive benchmarks with synthetic datasets with less sparse signal than in Section 4.4. The main observation, presented in Fig. 4.5, is that Smoothed Concomitant Lasso with cross-validation is stable w.r.t. various settings and provides similar performance to other Lasso variants investigated.

For each setting, we compare the mean running time for 50 simulations. The results are displayed in Figure 4.6. The computational time of our algorithm is in the same order of magnitude as the computational time for the Lasso.

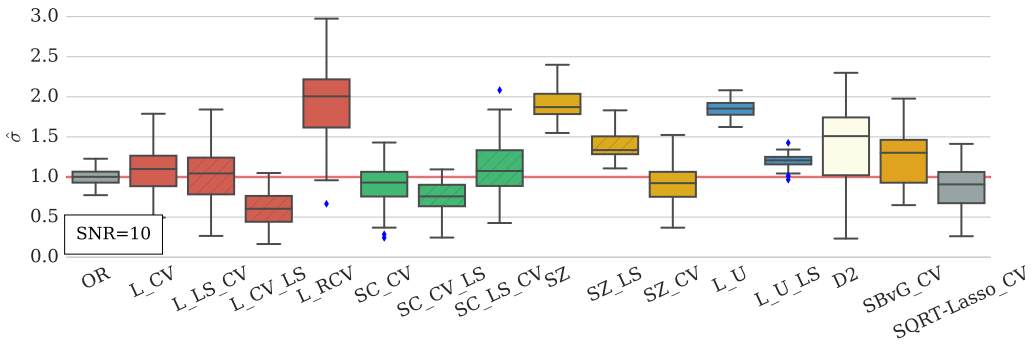
Perspective of a Convex Function

The concomitant scale estimator introduced by Huber (1981, Ch. 7.7 and 7.8) (see also (Owen, 2007; Antoniadis, 2010)), is related to the perspective of a function defined for a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ as the function $\text{persp}_f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that

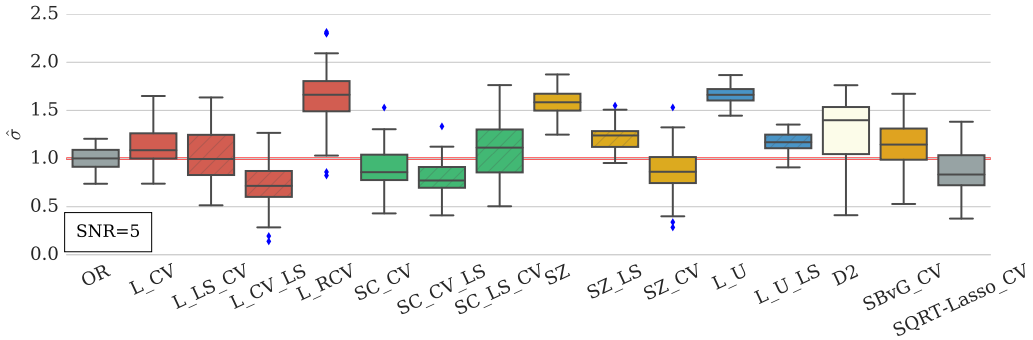
$$\text{persp}_f(r, \sigma) = \begin{cases} \sigma f\left(\frac{r}{\sigma}\right), & \text{if } \sigma > 0, \\ +\infty, & \text{if } \sigma \leqslant 0. \end{cases}$$



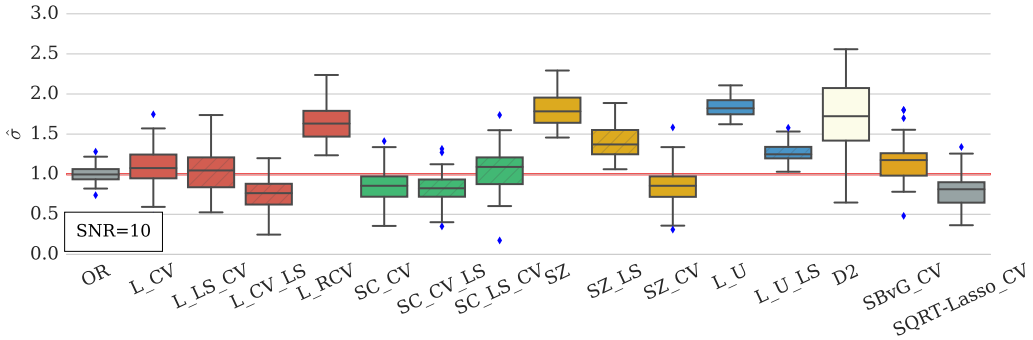
(a) $(n = 100, p = 200, \rho = 0, snr = 10, s = 0.8)$



(b) $(n = 100, p = 200, \rho = 0.2, snr = 10, s = 0.8)$

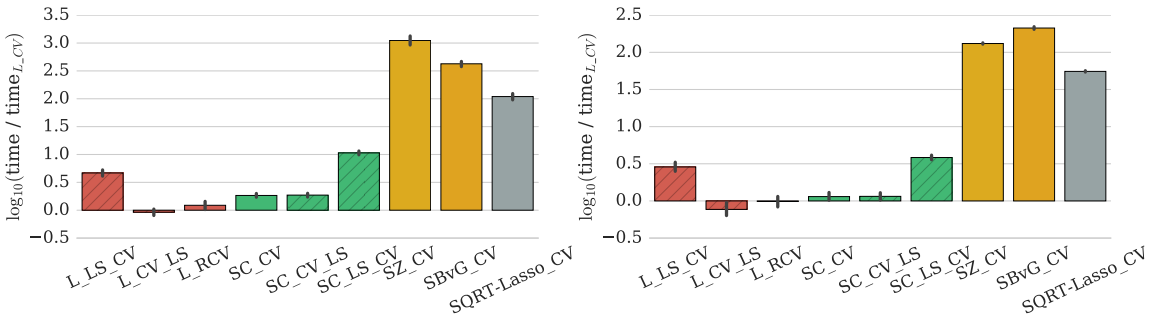


(c) $(n = 100, p = 200, \rho = 0.6, snr = 5, s = 0.8)$

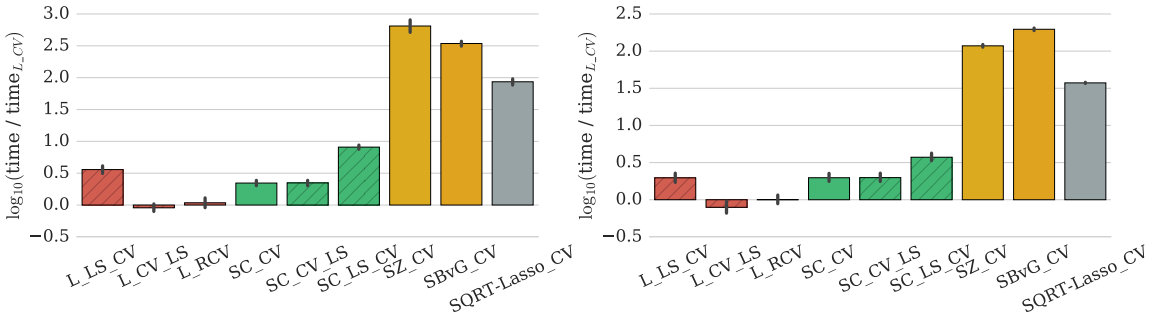


(d) $(n = 100, p = 200, \rho = 0.8, snr = 10, s = 0.8)$

Figure 4.5 – Estimated performance on synthetic dataset for different parameters.



(a) $(n = 100, p = 200, \rho = 0, snr = 10, s = 0.8)$ (b) $(n = 100, p = 200, \rho = 0.2, snr = 10, s = 0.8)$



(c) $(n = 100, p = 200, \rho = 0.6, snr = 5, s = 0.8)$ (d) $(n = 100, p = 200, \rho = 0.8, snr = 10, s = 0.8)$

Figure 4.6 – Computational time for 50 simulations on synthetic dataset.

This function is not lower semi-continuous in general. However, lower semi-continuity is a very desirable property. Together with the fact that the function is infinite at infinity, this guarantees the existence of minimizers (Peypouquet, 2015, Theorem 2.19). Hence we consider instead its biconjugate, which is always lower semi-continuous (Bauschke and Combettes, 2011, Theorem 13.32). One can show (Bauschke and Combettes, 2011, Example 13.8) that the Fenchel conjugate of persp_f is

$$\mathrm{persp}_f^*(\theta, \nu) = \begin{cases} 0, & \text{if } \nu + f^*(\theta) \leqslant 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Hence a direct calculation shows that

Proposition 53.

$$\mathrm{persp}_f^{**}(r, \sigma) = \begin{cases} \sigma f^{**}\left(\frac{r}{\sigma}\right), & \textit{if } \sigma > 0, \\ \sup\limits_{\theta \in \mathrm{dom} f^*} \langle \theta, r \rangle, & \textit{if } \sigma = 0, \\ +\infty, & \textit{otherwise.} \end{cases}$$

Proof. Let us define $g = \mathrm{persp}_f^*$ for simplicity.

First case: $\sigma > 0$.

$$\mathrm{persp}_f^{**}(r, \sigma) = \sup_{\theta \in \mathbb{R}^n, \nu \in \mathbb{R}} \langle \theta, r \rangle + \sigma\nu - g(\theta, \nu) = \sup_{\theta \in \mathbb{R}^n, \nu \in \mathbb{R}} \{\langle \theta, r \rangle + \sigma\nu : \nu + f^*(\theta) \leqslant 0\}$$

As $\sigma > 0$, for a given β, one should take ν the largest possible, hence $\nu = -f^*(\theta)$.

$$\mathrm{persp}_f^{**}(r, \sigma) = \sup_{\theta \in \mathbb{R}^n} \langle \theta, r \rangle - \sigma f^*(\theta) = \sigma \sup_{\theta \in \mathbb{R}^n} \langle \theta, r/\sigma \rangle - f^*(\theta) = \sigma f^{**}(r/\sigma)$$

Second case: $\sigma = 0$.

$$\mathrm{persp}_f^{**}(r, 0) = \sup_{\theta \in \mathbb{R}^n, \nu \in \mathbb{R}} \langle \theta, r \rangle - g(\theta, \nu) = \sup_{\theta \in \mathbb{R}^n, \nu \in \mathbb{R}} \{\langle \theta, r \rangle : \nu + f^*(\theta) \leqslant 0\}.$$

As ν has no influence on the value of the objective, we can choose it as small as we want and so the only requirement on θ is that it should belong to the domain of f^*. We get

$$\mathrm{persp}_f^{**}(r, 0) = \sup_{\theta \in \mathrm{dom} f^*} \langle \theta, r \rangle$$

Third case: $\sigma < 0$. If $\sigma < 0$, we can let ν go to $-\infty$ in the formula of $\mathrm{persp}_f^{**}(r, \sigma)$ which leads to $\mathrm{persp}_f^{**}(r, \sigma) = +\infty$. □

Dual of the Smoothed Concomitant Lasso

Proposition 54. *For $\lambda > 0$ and $\sigma_0 > 0$, the Smoothed Concomitant Lasso estimator $\widehat{\beta}^{(\lambda, \sigma_0)}$ and its associated noise level estimate $\widehat{\sigma}^{(\lambda, \sigma_0)}$ are defined as solutions of the primal optimization problem*

$$(\widehat{\beta}^{(\lambda, \sigma_0)}, \widehat{\sigma}^{(\lambda, \sigma_0)}) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p, \sigma \geqslant \sigma_0} \frac{1}{2n\sigma} \|y - X\beta\|^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 , \tag{4.38}$$

With $\Delta_{X,\lambda} = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leqslant 1, \|\theta\| \leqslant 1/(\lambda\sqrt{n})\right\}$, the dual formulation of the Smoothed Concomitant Lasso reads

$$\widehat{\theta}^{(\lambda, \sigma_0)} = \operatorname*{arg\,max}_{\theta \in \Delta_{X,\lambda}} \underbrace{\langle y, \lambda\theta \rangle + \sigma_0 \left(\frac{1}{2} - \frac{\lambda^2 n}{2} \|\theta\|^2 \right)}_{D_{\lambda, \sigma_0}(\theta)} . \tag{4.39}$$

For an optimal primal vector $\widehat{\beta}^{(\lambda, \sigma_0)}$, we must have $\widehat{\sigma}^{(\lambda, \sigma_0)} = \sigma_0 \vee (\|y - X\widehat{\beta}^{(\lambda, \sigma_0)}\|/\sqrt{n})$. We also have the link-equation between primal and dual solutions: $y = n\lambda\widehat{\sigma}^{(\lambda, \sigma_0)}\widehat{\theta}^{(\lambda, \sigma_0)} + X\widehat{\beta}^{(\lambda, \sigma_0)}$.

Proof.

$$\begin{aligned}
&\min_{\beta \in \mathbb{R}^p, \sigma \geqslant \sigma_0} \frac{1}{2n\sigma} \|y - X\beta\|^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 \\
&= \min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n, \sigma \geqslant \sigma_0} \frac{1}{2n\sigma} \|y - z\|^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 \text{ s.t. } z = X\beta \\
&= \min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n, \sigma \geqslant \sigma_0} \max_{\theta \in \mathbb{R}^n} \underbrace{\frac{1}{2n\sigma} \|y - z\|^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 + \lambda\theta^\top (z - X\beta)}_{\mathcal{L}(\beta, \sigma, \theta, z)} \\
&= \max_{\theta \in \mathbb{R}^n} \min_{\sigma \geqslant \sigma_0} \frac{\sigma}{2} - \max_{z \in \mathbb{R}^n} \left\{ \langle -\lambda\theta, z \rangle - \frac{1}{2n\sigma} \|y - z\|^2 \right\} - \lambda \max_{\beta \in \mathbb{R}^p} \langle X^\top \theta, \beta \rangle - \|\beta\|_1 \\
&= \max_{\theta \in \mathbb{R}^n} \min_{\sigma \geqslant \sigma_0} \frac{\sigma}{2} - \frac{\lambda^2 n\sigma}{2} \|\theta\|^2 + \langle \lambda\theta, y \rangle - \iota_{\mathcal{B}_\infty}(X^\top \theta).
\end{aligned}$$

The fourth line is true because the Slater's condition is met, hence we can permute $\min$ and $\max$ thanks to strong duality. Finally we obtain the dual problem since

$$\min_{\sigma \geqslant \sigma_0} \sigma \left(\frac{1}{2} - \frac{\lambda^2 n}{2} \|\theta\|^2 \right) = \begin{cases} \sigma_0 \left(\frac{1}{2} - \frac{\lambda^2 n}{2} \|\theta\|^2 \right), & \text{if } \frac{1}{2} - \frac{\lambda^2 n}{2} \|\theta\|^2 \geqslant 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

□

Let us use the same Lagragian notation as above, and denote

$$\left(\widehat{\beta}^{(\lambda,\sigma_0)}, \widehat{\sigma}^{(\lambda,\sigma_0)}, \widehat{\theta}^{(\lambda,\sigma_0)}, \hat{z}^{(\lambda,\sigma_0)}\right) \in \underset{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n, \sigma\geqslant\sigma_0}{\arg\min} \max_{\theta\in\mathbb{R}^n} \mathcal{L}(\beta,\sigma,\theta,z).$$

The primal-dual link equation follows directly from the Fermat's rule:

$$\begin{aligned}
&\frac{\partial \mathcal{L}(\widehat{\beta}^{(\lambda,\sigma_0)}, \widehat{\sigma}^{(\lambda,\sigma_0)}, \cdot, \hat{z}^{(\lambda,\sigma_0)})}{\partial \theta}(\widehat{\theta}^{(\lambda,\sigma_0)}) = \hat{z}^{(\lambda,\sigma_0)} - X\widehat{\beta}^{(\lambda,\sigma_0)} = 0,\\
&\frac{\partial \mathcal{L}(\widehat{\beta}^{(\lambda,\sigma_0)}, \widehat{\sigma}^{(\lambda,\sigma_0)}, \widehat{\theta}^{(\lambda,\sigma_0)}, \cdot)}{\partial z}(\hat{z}^{(\lambda,\sigma_0)}) = -\frac{1}{n\widehat{\sigma}^{(\lambda,\sigma_0)}}(y - \hat{z}^{(\lambda,\sigma_0)}) + \lambda\widehat{\theta}^{(\lambda,\sigma_0)} = 0.
\end{aligned}$$

Convergence of the Dual Vectors

Proposition 55. *Let $(\beta_k)_{k\in\mathbb{N}}$ be a sequence that converges to $\widehat{\beta}^{(\lambda,\sigma_0)}$. Then $(\theta_k)_{k\in\mathbb{N}}$ built from $\theta_k = (y - X\beta_k)/((\lambda n\sigma_0) \vee \|X^\top(y - X\beta_k)\|_\infty \vee (\lambda\sqrt{n}\|y - X\beta_k\|))$ converges to $\widehat{\theta}^{(\lambda,\sigma_0)}$. Hence the sequence of dual gap $(G_{\lambda,\sigma_0}(\beta_k,\sigma_k,\theta_k))_{k\in\mathbb{N}}$ converges to zero.*

Proof. Let $\alpha_k = (\lambda n\sigma_0) \vee (\|X^\top(y - X\beta_k)\|_\infty) \vee (\lambda\sqrt{n}\|y - X\beta_k\|)$, then we have:

$$\begin{aligned}
\left\|\theta_k - \widehat{\theta}^{(\lambda,\sigma_0)}\right\| &= \left\|\frac{1}{\alpha_k}(y - X\beta_k) - \frac{1}{\lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}}(y - X\widehat{\beta}^{(\lambda,\sigma_0)})\right\|\\
&= \left\|\left(\frac{1}{\alpha_k} - \frac{1}{\lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}}\right)(y - X\beta_k) - \frac{(X\widehat{\beta}^{(\lambda,\sigma_0)} - X\beta_k)}{\lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}}\right\|\\
&\leqslant \left|\frac{1}{\alpha_k} - \frac{1}{\lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}}\right| \|y - X\beta_k\| + \left\|\frac{X\widehat{\beta}^{(\lambda,\sigma_0)} - X\beta_k}{\lambda}\right\|.
\end{aligned}$$

If $\beta_k \to \widehat{\beta}^{(\lambda,\sigma_0)}$, then the second term in the last display converges to zero, and for the first term, we show below that $\alpha_k \to \alpha := (\lambda n\sigma_0) \vee (\|X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)})\|_\infty) \vee (\lambda\sqrt{n}\|y - X\widehat{\beta}^{(\lambda,\sigma_0)}\|)$. Recall that from Fermat's rule, we have $y - X\widehat{\beta}^{(\lambda,\sigma_0)} = \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}\widehat{\theta}^{(\lambda,\sigma_0)}$ and $X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)}) \in \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}\partial\|\cdot\|_1(\widehat{\beta}^{(\lambda,\sigma_0)})$, leading to one of the three following situations:

— if $\widehat{\sigma}^{(\lambda,\sigma_0)} > \sigma_0$, then $\|X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)})\|_\infty \leqslant \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)} = \lambda\sqrt{n}\|y - X\widehat{\beta}^{(\lambda,\sigma_0)}\| = \alpha$.

— If $\widehat{\sigma}^{(\lambda,\sigma_0)} = \sigma_0$ and $\widehat{\beta}^{(\lambda,\sigma_0)} \neq 0$, we have $X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)}) = \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}\hat{v}$ where $\hat{v} \in \partial\|\cdot\|_1(\widehat{\beta}^{(\lambda,\sigma_0)})$. Since $\widehat{\beta}^{(\lambda,\sigma_0)} \neq 0$, there exists a coordinate j such that $\widehat{\beta}_j^{(\lambda,\sigma_0)} \neq 0$ and so $|\hat{v}_j| = 1$ which implies that $\|\hat{v}\|_\infty = 1$. Hence $\|X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)})\|_\infty = \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}$. Moreover, $\|y - X\widehat{\beta}^{(\lambda,\sigma_0)}\| = \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}\|\widehat{\theta}^{(\lambda,\sigma_0)}\| \leqslant \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}/(\lambda\sqrt{n})$ since $\widehat{\theta}^{(\lambda,\sigma_0)} \in \Delta_{X,\lambda}$. Hence, $\lambda\sqrt{n}\|y - X\widehat{\beta}^{(\lambda,\sigma_0)}\| \leqslant \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)} = \|X^\top(y - X\widehat{\beta}^{(\lambda,\sigma_0)})\|_\infty = \alpha$.

— If $\widehat{\sigma}^{(\lambda,\sigma_0)} = \sigma_0$ and $\widehat{\beta}^{(\lambda,\sigma_0)} = 0$, then $y = \lambda n\sigma_0\widehat{\theta}^{(\lambda,\sigma_0)}$, $\lambda\sqrt{n}\|y\| \leqslant \lambda n\sigma_0$ since $\widehat{\theta}^{(\lambda,\sigma_0)} \in \Delta_{X,\lambda}$, and $\|X^\top y\|_\infty \leqslant \lambda n\sigma_0$. Hence $\alpha = \lambda n\sigma_0$.
Finally, we have shown that in all cases, $(\alpha_k)_{k\in\mathbb{N}}$ converges to $\alpha = \lambda n\widehat{\sigma}^{(\lambda,\sigma_0)}$, so the first term also converges to zero. □

Safe Screening Rules

Proposition 56. *For all $(\beta,\sigma,\theta) \in \mathbb{R}^p \times \mathbb{R}_+ \times \Delta_{X,\lambda}$, then for*

$$r = \sqrt{\frac{2G_{\lambda,\sigma_0}(\beta,\sigma,\theta)}{\lambda^2\sigma_0 n}},$$

we have $\widehat{\theta}^{(\lambda,\sigma_0)} \in \mathcal{B}(\theta, r)$. *Thus, we have the following safe sphere screening rule*

$$|X_j^\top \theta| + r \|X_j\| < 1 \Longrightarrow \widehat{\beta}_j^{(\lambda,\sigma_0)} = 0. \tag{4.40}$$

Proof. The proof follows (Ndiaye et al., 2015). We give it for the sake of completeness. By weak duality, for all $\beta \in \mathbb{R}^p$, $D_{\lambda,\sigma_0}(\theta) \leqslant P_{\lambda,\sigma_0}(\beta, \sigma)$. Then, note that the dual objective function of the Smoothed Concomitant Lasso is $\lambda^2\sigma_0 n$-strongly concave. This implies that:

$$\forall (\theta, \theta') \in \Delta_{X,\lambda} \times \Delta_{X,\lambda}, \quad D_{\lambda,\sigma_0}(\theta) \leqslant D_{\lambda,\sigma_0}(\theta') + \nabla D_{\lambda,\sigma_0}(\theta')^\top (\theta - \theta') - \frac{\lambda^2\sigma_0 n}{2} \left\|\theta - \theta'\right\|^2.$$

Moreover, since $\widehat{\theta}^{(\lambda,\sigma_0)}$ maximizes the concave function D_{λ,σ_0}, the following inequality holds true:

$$\forall\, \theta \in \Delta_{X,\lambda}, \quad \nabla D_\lambda(\widehat{\theta}^{(\lambda,\sigma_0)})^\top (\theta - \widehat{\theta}^{(\lambda,\sigma_0)}) \leqslant 0.$$

Hence, we have for all $\theta \in \Delta_{X,\lambda}$ and $\beta \in \mathbb{R}^p$:

$$\begin{aligned}\frac{\lambda^2\sigma_0 n}{2} \left\|\theta - \widehat{\theta}^{(\lambda,\sigma_0)}\right\|^2 &\leqslant D_{\lambda,\sigma_0}(\widehat{\theta}^{(\lambda,\sigma_0)}) - D_{\lambda,\sigma_0}(\theta) \\ &\leqslant P_\lambda(\beta, \sigma) - D_{\lambda,\sigma_0}(\theta) = G_{\lambda,\sigma_0}(\beta, \sigma, \theta).\end{aligned}$$

Furthermore,

$$\max_{\bar{\theta} \in \mathcal{B}(\theta,r)} |X_j^\top \bar{\theta}| \leqslant |X_j^\top \theta| + \max_{\bar{\theta} \in \mathcal{B}(\theta,r)} |X_j^\top (\bar{\theta} - \theta)| \leqslant |X_j^\top \theta| + \|X_j\| \max_{\bar{\theta} \in \mathcal{B}(\theta,r)} \left\|\bar{\theta} - \theta\right\| = |X_j^\top \theta| + r \|X_j\|.$$

Hence $\max_{\bar{\theta} \in \mathcal{B}(\theta,r)} |X_j^\top \bar{\theta}| = |X_j^\top \theta| + r \|X_j\|$ since the vector $\bar{\bar{\theta}} := \theta + X_j \frac{r}{\|X_j\|}$ is feasible and attains the bound. □

In this section we derive the Bound Safe screening rules of Proposition 50. First, we need two technical lemmas.

Lemma 23. *Let* y' *and* x *be two unit vectors, and consider* $0 \leqslant \underline{\gamma} \leqslant \overline{\gamma} \leqslant 1$. *The optimal value of*

$$\max\{\theta^\top x \quad : \quad \|\theta\| \leqslant 1, \quad \underline{\gamma} \leqslant y'^\top \theta \leqslant \overline{\gamma}\},$$

is given by

$$\begin{cases} \overline{\gamma} x^\top y' + \sqrt{1 - \overline{\gamma}^2}\sqrt{1 - (x^\top y')^2}, & \text{if } x^\top y' > \overline{\gamma}, \\ 1, & \text{if } \underline{\gamma} \leqslant x^\top y' \leqslant \overline{\gamma}, \\ \underline{\gamma} x^\top y' + \sqrt{1 - \underline{\gamma}^2}\sqrt{1 - (x^\top y')^2}, & \text{if } x^\top y' < \underline{\gamma}. \end{cases}$$

Proof. First remark that x and y are two privileged directions in the optimization problem at stake. Indeed, if θ has a nonzero component in a direction orthogonal to both x and y', then, because of the constraint $\|\theta\| = 1$, this reduces the freedom in $\mathrm{Span}(x, y')$ while giving no progress in the objective and the linear constraints. Hence, from now on we can restrict ourselves to the plane $\mathrm{Span}(x, y)$.

We denote by $\angle(w, z) \in \mathbb{R}/2\pi\mathbb{Z}$ the directed angle between unitary vectors w and z. We recall that $\cos(\angle(w, z)) = w^\top z$, so we can narrow down our analysis to the three following cases:

1. Assume that $x^\top y' > \overline{\gamma}$. Then the optimal θ is such that (see Figure (4.1).(a)) $\|\theta\| = 1$, $\theta^\top y' = \overline{\gamma}$ and x is "between" θ and y', which implies that $\sin(\angle(\theta, y'))\sin(\angle(y', x)) < 0$. Hence,

$$\begin{aligned}\theta^\top x &= \cos(\angle(\theta, x)) = \cos\big(\angle(\theta, y') + \angle(y', x)\big) \\ &= \cos\big(\angle(\theta, y')\big)\cos\big(\angle(y', x)\big) - \sin\big(\angle(\theta, y')\big)\sin\big(\angle(y', x)\big) \\ &= \theta^\top y' . y'^\top x + \big|\sin(\angle(\theta, y'))\sin(\angle(y', x))\big| \\ &= \overline{\gamma} y'^\top x + \sqrt{1 - \overline{\gamma}^2}\sqrt{1 - (y'^\top x)^2}.\end{aligned}$$

2. Assume that $\underline{\gamma} \leqslant x^\top y' \leqslant \overline{\gamma}$, then $\theta = x$ is admissible, and the maximum is 1 (see Figure (4.1).(b)).
3. Assume that $-1 \leqslant x^\top y' < \underline{\gamma}$ (see Figure (4.1).(c)), then the optimal θ is such that $\|\theta\| = 1$, $\theta^\top y' = \underline{\gamma}$ and θ is "between" x and y', which implies that $\sin(\angle(\theta, y'))\sin(\angle(y', x)) < 0$. Hence, elementary trigonometry gives

$$\theta^\top x = \cos\big(\angle(\theta, y') + \angle(y', x)\big) = \underline{\gamma} x^\top y' + \sqrt{1-\underline{\gamma}^2}\sqrt{1-(x^\top y')^2}.$$

□

Lemma 24. *Let y' and x be two unit vectors, and consider $0 \leqslant \underline{\gamma} \leqslant \overline{\gamma} \leqslant 1$. The optimal value of*

$$\max\{|\theta^\top x| \quad : \quad \|\theta\| \leqslant 1, \quad \underline{\gamma} \leqslant {y'}^\top \theta \leqslant \overline{\gamma}\},$$

is given by

$$\begin{cases} \overline{\gamma}|x^\top y'| + \sqrt{1-\overline{\gamma}^2}\sqrt{1-(x^\top y')^2}, & \text{if } |x^\top y'| > \overline{\gamma}, \\ 1, & \text{if } \underline{\gamma} \leqslant |x^\top y'| \leqslant \overline{\gamma}, \\ \underline{\gamma}|x^\top y'| + \sqrt{1-\underline{\gamma}^2}\sqrt{1-(x^\top y')^2}, & \text{if } |x^\top y'| < \underline{\gamma}. \end{cases}$$

Proof. We need to compute

$$\max\{|\theta^\top x| \quad : \quad \|\theta\| \leqslant 1, \quad \underline{\gamma} \leqslant {y'}^\top \theta \leqslant \overline{\gamma}\}.$$

We apply Lemma 23 with $x \leftarrow x$ and $x \leftarrow -x$. We get five cases and for each the value is a maximum between two choices. In fact, one of the two choices is always dominated by the other one. We just present one case for conciseness.

Suppose that $x^\top y' > \overline{\gamma}$ (and thus $-x^\top y' < \underline{\gamma}$ since $\underline{\gamma} \geqslant 0$). Then the optimal θ satisfies

$$\left|\theta^\top x\right| = \left(\overline{\gamma} x^\top y' + \sqrt{1-\overline{\gamma}^2}\sqrt{1-(x^\top y')^2}\right) \vee \left(-\underline{\gamma} x^\top y' + \sqrt{1-\underline{\gamma}^2}\sqrt{1-(x^\top y')^2}\right).$$

We now remark the equivalence

$$\begin{aligned} & -\underline{\gamma} x^\top y' + \sqrt{1-\underline{\gamma}^2}\sqrt{1-(x^\top y')^2} \leqslant \overline{\gamma} x^\top y' + \sqrt{1-\overline{\gamma}^2}\sqrt{1-(x^\top y')^2} \\ \Leftrightarrow & \left(\sqrt{1-\underline{\gamma}^2} - \sqrt{1-\overline{\gamma}^2}\right)\sqrt{1-(x^\top y')^2} - (\overline{\gamma} + \underline{\gamma}) x^\top y' \leqslant 0. \end{aligned}$$

This function is decreasing in $x^\top y'$ so

$$\begin{aligned} & \left(\sqrt{1-\underline{\gamma}^2} - \sqrt{1-\overline{\gamma}^2}\right)\sqrt{1-(x^\top y')^2} - (\overline{\gamma} + \underline{\gamma}) x^\top y' \\ & \qquad \leqslant \left(\sqrt{1-\underline{\gamma}^2} - \sqrt{1-\overline{\gamma}^2}\right)\sqrt{1-\overline{\gamma}^2} - (\overline{\gamma} + \underline{\gamma})\overline{\gamma} \\ & \qquad = \sqrt{1-\underline{\gamma}^2}\sqrt{1-\overline{\gamma}^2} - 1 + \overline{\gamma}^2 - \overline{\gamma}^2 - \underline{\gamma}\overline{\gamma} \leqslant 0. \end{aligned}$$

Thus, the second term in the maximum is never selected and we can simplify the expression. The other cases can be handled similarly. □

Proposition 57. *Assume that, for a given $\lambda > 0$, we have an upper bound $\overline{\eta} \in (0, +\infty]$, and a lower bound $\underline{\eta} \in (0, +\infty]$ over the Smoothed Concomitant Lasso problem* (4.17)*. Denote by $x_j = X_j/\|X_j\|$ and $y' = y/\|y\|$ two unit vectors, and by $\underline{\gamma} = (\underline{\eta} - \sigma_0/2)\sqrt{n}/\|y\|$ and $\overline{\gamma} = \overline{\eta}\sqrt{n}/\|y\|$. Then if one of the three following conditions is met*

— $|x_j^\top y'| > \overline{\gamma}$ *and* $\overline{\gamma}|x_j^\top y'| + \sqrt{1-\overline{\gamma}^2}\sqrt{1-(x_j^\top y')^2} < \lambda\sqrt{n}/\|X_j\|$.

— $\underline{\gamma} \leqslant |x_j^\top y'| \leqslant \overline{\gamma}$ *and* $1 < \lambda\sqrt{n}/\|X_j\|$.
— $|x_j^\top y'| < \underline{\gamma}$ *and* $\underline{\gamma}|x_j^\top y'| + \sqrt{1-\underline{\gamma}^2}\sqrt{1-(x_j^\top y')^2} < \lambda\sqrt{n}/\|X_j\|$.

then the j-th feature can be discarded i.e. $\widehat{\beta}_j^{(\lambda,\sigma_0)} = 0$.

Proof. If $\underline{\eta} \leqslant D_\lambda(\widehat{\theta}^{(\lambda,\sigma_0)})\rangle \leqslant \overline{\eta}$ and

$$\max\{|X_j^\top \theta| \,:\, \lambda\sqrt{n}\,\|\theta\| \leqslant 1, \quad \underline{\eta} \leqslant D_\lambda(\theta)\rangle \leqslant \overline{\eta}\} < 1, \tag{4.41}$$

then the j-th feature can be discarded (see Eq. (4.25)).

For the standard Concomitant Lasso formulation, $D_\lambda(\theta) = \langle y, \lambda\theta\rangle$ and Lemma 24 can be directly applied to get a safe screening rule from (4.41). To treat the Smoothed Concomitant Lasso (4.17), we check that if $\underline{\eta} \leqslant D_{\lambda,\sigma_0}(\theta) \leqslant \overline{\eta}$ then $\underline{\eta} - \sigma_0/2 \leqslant \langle y, \lambda\theta\rangle \leqslant \overline{\eta}$. Thus, we obtain a new screening test

$$\max\{|X_j^\top \theta| \,:\, \lambda\sqrt{n}\,\|\theta\| \leqslant 1, \quad \underline{\eta} - \sigma_0/2 \leqslant \langle y, \lambda\theta\rangle \leqslant \overline{\eta}\} < 1. \tag{4.42}$$

To leverage Lemma 24 we reformulate the test as

$$\max\left\{\left|\frac{\lambda\sqrt{n}}{\|X_j\|}X_j^\top\theta\right| \quad : \quad \sqrt{n}\lambda\,\|\theta\| \leqslant 1, \quad \frac{(\underline{\eta}-\sigma_0/2)\sqrt{n}}{\|y\|} \leqslant \left\langle \frac{y}{\|y\|}, \sqrt{n}\lambda\theta \right\rangle \leqslant \frac{\overline{\eta}\sqrt{n}}{\|y\|}\right\} < \frac{\sqrt{n}\lambda}{\|X_j\|}.$$

Denoting by $x_j' = X_j/\|X_j\|$ and $y' = y/\|y\|$ two unit vectors, and by $\underline{\gamma} = (\underline{\eta} - \sigma_0/2)\sqrt{n}/\|y\|$ and $\overline{\gamma} = \overline{\eta}\sqrt{n}/\|y\|$, the test (4.42) now reads

$$\max\left\{\left|\theta^\top x_j'\right| \quad : \quad \|\theta\| \leqslant 1, \quad \underline{\gamma} \leqslant \left\langle y', \theta\right\rangle \leqslant \overline{\gamma}\right\} < \frac{\sqrt{n}\lambda}{\|X_j\|}.$$

Lemma 24 concludes the proof. □

Chapitre 5

Abrégé des Contributions de la Thèse

Le traitement automatique des données est devenu omniprésent dans les technologies actuelles. Il a d'importantes applications dans de nombreuses disciplines scientifiques telles que la médecine, la biologie ou la météorologie mais aussi dans des outils numériques tels que la traduction de texte, la publicité ciblée ou la détection de spam. Il s'inscrit au coeur des problématiques de traitement du signal, de théorie de l'information et des statistiques où l'on cherche à comprendre et résumer les informations essentielles dans les données collectées. Ces dernières sont souvent accompagnées par des informations à priori sur leurs structures, qui peuvent ensuite être utilisées pour établir des modèles et algorithmes prédictifs. Cependant, ces algorithmes dépendent fortement des méthodes d'optimisation mathématique et il est devenu crucial d'avoir des outils qui restent efficaces lorsque la taille des bases de données augmente. Nous étudions des méthodes de réduction de complexité dans certains problèmes d'optimisation découlant de l'apprentissage statistique avec un principal fil conducteur :économiser du temps de calcul en exploitant des certificats optimalité et des structures de régularité spécifiques aux problèmes.

Dans ce chapitre, nous résumons les contributions de la thèse en présentant les principales idées et résultats développés dans ce manuscrit.

5.1 Optimization Convexe en Apprentissage Statistique

Nous suivons une formalisation classique des tâches d'apprentissage statistique comme dans (Hastie et al., 2009; Shalev-Shwartz and Ben-David, 2014). Soit $\mathcal{X}$ (resp. $\mathcal{Y}$) un ensemble de vecteurs d'entrée (resp. sortie) et X (resp. Y) une variable aléatoire évaluée dans $\mathcal{X}$ (resp. $\mathcal{Y}$). Nous appelons *tâche d'apprentissage* l'identification d'une application $h : \mathcal{X} \mapsto \mathcal{Y}$ qui explique la relation entre l'entrée X et la sortie Y.

Considérant une fonction de *perte* ℓ telle que $\ell(y, y) = 0$, $\ell(y, y') \geqslant 0$, nous voulons apprendre une fonction de prédiction h minimisant l'erreur de prédiction $\ell(h(X), Y)$ en moyenne. Pour simplifier, nous considérons que h sera recherché sur une famille de fonctions paramétrée (prédéfinie) $\mathcal{H} := \{h(\cdot, \beta) : \beta \in \mathbb{R}^p\}$ qui code les connaissances à priori que nous avons sur les données. Ensuite, la tâche d'apprentissage peut être écrite sous la forme du problème d'optimisation suivant :

$$\min_{\beta \in \mathbb{R}^p} R(\beta) := \mathbb{E}[\ell(h(X, \beta), Y)] \ . \tag{5.1}$$

Puisque l'espérance est prise sous la distribution de probabilité conjointe $\mathbb{P}_{X,Y}$ des variables X, Y qui est supposé être *inconnu*, h ne peut pas être appris directement de cette manière. On devrait plutôt apprendre en considérant un échantillon de données. Nous supposons que les observations sur notre ensemble de données $\{(x_i, y_i)\}_{i \in [n]}$ sont indépendantes et identiquement distri-

buées. De ce fait, par la loi des grands nombres, la loi empirique $\frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)}$, où δ représente les masses de Dirac, approximent la vraie distribution $\mathbb{P}_{X,Y}$ si le nombre d'observations n est suffisamment grand. D'où le paradigme *Minimisation du Risque Empirique* (MRE) :

$$\min_{\beta\in\mathbb{R}^p} R_n(\beta) := \frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i,\beta),y_i) \ . \tag{5.2}$$

Une instanciation populaire des formulations (5.1), (5.2) est l'outil fondamental en Statistique, connu sous le nom de *Estimation du Maximum de Vraisemblance (EMV)*. Nous nous référons à (Van der Vaart, 1998) pour une description complète et (Stigler, 2007) pour l'histoire passionnée de cette méthode. Fait intéressant, le EMV de la famille exponentielle conduit naturellement à un problème d'optimisation convexe (Brown, 1986).

Definition 23 (Famille Exponentielle)**.** *Soit ν une mesure σ-finie et $\lambda(\theta) = \int\exp(\theta y)\nu(dy)$ sa transformée de Laplace de domaine $N = \{\theta\in\mathbb{R}^n : \lambda(\theta) < +\infty\}$. Pour $P(\theta) = \log(\lambda(\theta))$, soit*

$$p_\theta(y) = \exp\left(\langle\theta, y\rangle - P(\theta)\right) \ . \tag{5.3}$$

Soit Θ un sous-ensemble convexe de N, la famille de densité $\{p_\theta : \theta\in\Theta\}$ est appelé famille exponentielle (standard).

La convexité du problème d'optimisation de EMV d'une famille exponentielle découle de la convexité de la transformée de Log-Laplace P, nous rappelons la preuve dans le Chapitre 4.

Theorem 5 (Brown (1986, Theorem 1.13))**.** *L'ensemble N est convexe et P est convexe sur N. De plus, P est semi-continue inférieurement sur $\mathbb{R}^n$ et continue sur l'intérieur de N.*

En inférence statistique, il est courant de supposer que la distribution des observations est paramétrée par un certain $\theta_0\in\Theta$ inconnu. L'objectif est d'approximer et de fournir des informations sur le paramètre du modèle, à défaut de le trouver exactement, à partir de variables aléatoires distribuées sous cette loi. Une méthode inférentielle classique est le EMV. Pour une variable y dans le support convexe de ν et Θ un sous-ensemble convexe de N, la fonction $\Theta\ni\theta\mapsto p_\theta(y)$ s'appelle *vraisemblance* au point y. En supposant que le paramètre θ_0 appartient à Θ et que y est une variable aléatoire de loi P_{θ_0}, l'estimateur du maximum de vraisemblance est défini par

$$\hat\theta(y) = \arg\max_{\theta\in\Theta} p_\theta(y) = \arg\min_{\theta\in\Theta} -\log(p_\theta(y)) = \arg\min_{\theta\in\Theta} P(\theta) - \langle\theta, y\rangle$$

which is a convex optimization problem with a loss function $\ell(\theta, y) := P(\theta) - \langle\theta, y\rangle$. Ainsi, l'EMV pour des observations indépendantes et identiquement distribuées $y = (y_1, \cdots, y_n)$ s'écrit

$$\hat\theta(y) \in \arg\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,\theta) \ . \tag{5.4}$$

Dans un contexte d'apprentissage, un exemple important est la généralisation de la régression conduisant à la famille de *Modèle Linéaire Généralisé* (MLG) (McCullagh and Nelder, 1989) où le modèle statistique contient une partie déterministe donnée par une combinaison linéaire des covariables $\eta = X\beta$ et la partie aléatoire donnée par $\mu = \mathbb{E}[Y]$ où Y est supposée appartenir à une famille exponentielle, sont liées par $h(\eta) = \mu$. Selon la distribution supposée des observations, nous obtenons l'estimateur des moindres carrés et la regression logistique comme exemples canoniques pour les tâches de régression et de classification.

Moindre Carrés. Soit un échantillon d'observations $y = (y_1, \cdots, y_n)$ indépendantes et de même loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, *i.e.*

$$\begin{aligned} p_{(\mu,\sigma^2)}(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\} \\ &= \exp\left\{ \langle (\theta_1, \theta_2), (y, y^2) \rangle - P(\theta_1, \theta_2) \right\} \end{aligned}$$

où $\theta = (\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ et $P(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(-\frac{\pi}{\theta_2})$. En supposant que σ^2 (donc θ_2) est *connu*, l'EMV (5.4) pour un modèle gaussien de moyenne $\mu = X\beta$ est donné par :

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \frac{1}{2\sigma^2} (y_i - x_i^\top \beta)^2 \ . \tag{5.5}$$

Régression Logistique. Soit une variable binaire $y \in \{0, 1\}$ suivant une loi de Bernoulli de paramètre μ *i.e.*

$$p_\mu(y) = \mu^y (1-\mu)^{1-y} = \exp\left\{ \langle \theta, y \rangle - P(\theta) \right\} \ ,$$

où $\theta = \log(\frac{\mu}{1-\mu})$ et $P(\theta) = \log(\frac{1}{1+\mathrm{e}^\theta})$. Dans un contexte de regression où la partie deterministe est $\theta = X\beta$, l'EMV (5.4) pour un échantillon de variables indépendant et identiquement distribuée $y = (y_1, \cdots, y_n)$, est donné par :

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left(1 + \exp(x_i^\top \beta)\right) - y_i x_i^\top \beta \ .$$

Cependant, le cadre de la minimisation du risque empirique n'est pas limitée aux modèles statistiques d'estimation basée sur la vraisemblance. D'autres paradigmes d'apprentissage, comme le perceptron et les machines à vecteur support, fournissent un bon prédicteur sans hypothèses sur la distribution sous-jacente des données.

Dans de nombreux cas, il s'avère trés difficile de résoudre le problème de minimisation du risque empirique. La plus part du temps, il n'existe pas de formulation explite des solutions. Et en toute généralité, nous ne savons pas résoudre exactement les problèmes d'optimisation ! Voir (Nesterov, 2004, Chapitre 1). Nous allons donc nous contenté d'approximation à un certain niveau de précision ϵ. Dans une configuration à grande échelle, la dimensionnalité dans les problème d'optimisation (5.2) peut être si importante que les algorithmes nécessitant une évaluation des quantités reposant sur le jeu de données complet deviennent impraticables. Une tendance populaire dans l'optimisation pour l'apprentissage statistique est de revenir à des méthodes simples développées avec des ressources de calcul limitées et popularisées dans les 50 (voir (Bottou et al., 2016) pour une revue récente). Par conséquent, les algorithmes qui fournissent des calculs rapides et peu coûteux avec des informations limitées ont été privilégiés. Nous pouvons citer par exemple l'optimisation incrémentale incluant la descente stochastique du gradient (Robbins and Monro, 1951), l'algorithme Frank-Wolfe (Frank and Wolfe, 1956), la descente par bloc de coordonnée (Warga, 1963) et les méthodes d'ensemble actives.

5.2 Régularisation Structurée et Choix d'Hyperparamètre

L'ajout d'un terme de régularisation apparaît naturellement en apprentissage statistique pour améliorer la stabilité numérique et éviter des phénomènes de sur-apprentissage. En effet, résoudre le problème de minimisation du risque empirique (5.2) n'est souvent pas suffisant pour trouver de bons prédicteurs, car le problème a tendance à être mal conditionné dans des contextes de

grandes dimensions. Cette régularisation encode une connaissance supplémentaire sur la structure des données. Par exemple, il peut être utilisé pour imposer le choix de modèles plus simples et peut être formulé comme suit :

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i, \beta), y_i) + \lambda \Omega(\beta) \ , \tag{5.6}$$

où Ω est la fonction de régularisation qui pénalise les solutions complexes et $\lambda > 0$ contrôle le niveau de biais inductif. Il est généralement lié au principe de *simplicité* de G. Ockham (14ème siècle) ou (Wrinch and Jeffreys, 1921). Le terme de régularisation équilibre la minimisation du risque empirique et la simplicité structurelle du modèle à travers l'hyperparamètre λ. Il est essentiel de trouver l'équilibre optimal pour obtenir une bonne prédiction sur des ensembles de données inédits :les petits λ conduisent à des modèles complexes qui risquent de sur-apprendre sur les données d'entrainement tandis que les grands λ conduisent à des modèles simplistes avec une puissance de prédiction médiocre. Une approche courante pour sélectionner un "bon" paramètre consiste à utiliser la *validation croisée*. Essentiellement, cette méthode évite d'entrainer et d'évaluer la performance d'un estimateur sur les mêmes données. Il a été introduit dans (Larson, 1931) ; voir (Arlot and Celisse, 2010) pour une description plus complète. Pour simplifier, nous traitons ici la version simplifiée qui consiste à scinder les données $\{(x_i, y_i)\}_{i \in [n]}$ en deux parties $(X_{\text{train}}, y_{\text{train}})$ et $(X_{\text{test}}, y_{\text{test}})$ et considérons Λ comme un ensemble d'hyperparamètres discret. Avec une fonction de perte de validation $\mathcal{L}$ qui mesure l'erreur de prédiction sur l'ensemble de tests, la validation croisée correspond à la réalisation des deux étapes suivantes :

1. résoudre le problème (1.6) avec les données $(X_{\text{train}}, y_{\text{train}})$ pour tout $\lambda \in \Lambda$,
2. choisir le $\lambda \in \Lambda$ qui minimise l'erreur de validation $\mathcal{L}(h(X_{\text{test}}, \hat{\beta}^{(\lambda)}), y_{\text{test}})$.

Une grille standard considérée dans la littérature est $\lambda_t = \lambda_{\max} 10^{-\delta t/(T-1)}$ avec un petit δ ($\delta = 10^{-2}$ ou 10^{-3}), voir par exemple (Bühlmann and van de Geer, 2011) [2.12.1] ou le paquet `glmnet` (Friedman et al., 2010b) et `scikit-learn` (Pedregosa et al., 2011). Choisir δ est un défi du point de vue statistique (les performances ont tendance à diminuer à mesure que δ devient proche de zéro, en raison du surapprentissage) et du point de vue de l'optimisation, la complexité de calcul tend à augmenter pour les petits λ, les itérés dans le primal étant denses et le problème à résoudre de plus en plus mal posé. Il est de coutume de commencer par un assez grand paramètre de régularisation $\lambda_0 = \lambda_{\max}$ puis d'effectuer séquentiellement le calcul de $\hat{\beta}^{(\lambda_t)}$ après celui de $\hat{\beta}^{(\lambda_{t-1})}$. Souvent, elle conduit à calculer les modèles dans l'ordre croissant de complexité : cela permet une accélération importante en profitant de l'initialisation du démarrage à chaud.

Selon le contexte, plusieurs fonctions de régularisation Ω ont été introduites pour prendre en compte la régularité à priori des estimateurs. Les exemples utilisés dans nos expériences sont :

Régularisation Ridge ou de Tikhonov. La fonction de régularisation $\Omega(\beta) = \|\beta\|_2^2/2$ a été introduite dans (Tikhonov, 1943) pour améliorer la stabilité des problèmes inverses, et en Statistiques (Hoerl, 1962; Hoerl and Kennard, 1970) pour réduire l'erreur quadratique moyenne de l'estimateur de moindres carrés classique lorsque la matrice de design est de rang déficient. En l'apprentissage statistique, il est souvent considéré comme un stabilisateur de l'algorithme d'apprentissage, en ce sens que la prédiction ne change pas beaucoup lorsque les données d'entrée sont légèrement perturbées. Par conséquent, l'erreur d'apprentissage reste proche de l'erreur de test, ce qui empêche l'algorithme de surapprendre sur les données d'entrainement (Shalev-Shwartz and Ben-David, 2014, Chapitre 13.2).

Bien que fondamental, la prévention du phénomène de surapprentissage n'est pas suffisante dans de nombreuses applications. Souvent, il faut également avoir une bonne représentation des

données et fournir des modèles de prédiction interprétables. Il est donc crucial de pouvoir sélectionner les variables explicatives les plus pertinentes, ce qui a motivé l'introduction de méthodes de régularisation parcimonieuse.

Régularisation Parcimonieuse de Type Lasso. La régularisation $\Omega(\beta) = \|\beta\|_1$ a été introduite dans (Chen and Donoho, 1995; Tibshirani, 1996) en traitement du signal et en Statistiques. Elle suit les méthodes classiques de sélection variables explicatives dans la régression multiple (Efroymson, 1960) pour la régression adapdative ou (Breiman, 1995) pour la sélection avec la non-négative garrote. La pénalité ℓ_1 norm a l'avantage de pouvoir sélectionner des variables de manière continue et sa formulation convexe permet d'utiliser un algorithme itératif rapide.

Plus tard, plusieurs extensions ont été proposées, notamment par Zou and Hastie (2005) pour la régularisation Elastic Net $\Omega(\beta) = \alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 /2$ qui fait une interpolation entre le Ridge et le Lasso, par Hebiri and van de Geer (2011) pour le Lasso lissé où $\Omega(\beta) = \alpha \|\beta\|_1 + \gamma \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2$, ou pour des régularisations de groupe hiérarchiques plus complexes (Friedman et al., 2010a; Sprechmann et al., 2011). Une enquête fournissant une théorie unifiée pour les normes induisant une faible densité structurée convexe a récemment été proposée dans (Obozinski and Bach, 2016). Notez que la parcimonie peut également être incorporée dans le terme d'ajustement au données. C'est le cas de la perte charnière (hinge loss) qui, d'ailleurs peut également être utilisée comme critère de sélection des variables (Guyon et al., 2002; Rakotomamonjy, 2003).

En utilisant de telle régularisation, la performance en généralisation, des estimateurs obtenus en minimisant le risque empirique, est alors fortement liée aux capacités de réglage du paramètre de régularisation λ. Cela nécessite souvent le calcul du chemin complet des solutions dans le cadre des méthodes d'homotopie sur une plage (souvent un ensemble discret) d'hyperparamètres Λ. En effet, il est généralement impossible de calculer le chemin complet dans un ensemble continu si on a pas accés aux solutions exacte dans l'Equation (5.6). Cependant, pour les problèmes impliquant une perte quadratique par morceaux et des régularisations linéaires par morceaux, le chemin des solutions $\{\hat{\beta}^{(\lambda)}, \lambda \in \Lambda\}$ est continu et linéaire par morceaux (Rosset and Zhu, 2007). Cette linéarité par morceaux et très spécifique et permet de calculer exactement la totalité du chemin de la solution. Ce type de propriété a été redécouvert plusieurs fois dans la littérature. Par exemple, dans (Markowitz, 1952) pour la sélection de portefeuille, (Osborne, 1992) pour les problèmes de régression quantile, (Osborne et al., 2000a) pour Lasso, (Efron et al., 2004; Park and Hastie, 2007) pour le modèle linéaire généralisé avec une régularisation avec la norme ℓ_1.

Outre la régularité générale des fonctions en jeu, l'exploitation explicite de la structure des fonctions permet de concevoir des algorithmes d'optimisation plus rapides. L'une des principales contributions de cette thèse est de proposer des accélérations supplémentaires en économisant une quantité considérable de calculs effectués le long des itérations. Ici, nous ne considérerons que les problèmes d'optimisation convexe *i.e.* les fonctions dans (5.6) où la classe d'hypothèses $\mathcal{H}$ et la fonction de perte ℓ sont supposées être toutes deux convexes. Nous avons vu qu'une telle formulation convexe inclut déjà une grande classe de tâches d'apprentissage statistique telles que l'estimation du maximum de vraisemblance pour la famille des distribution exponentielle, mais également des formulations résultant du paradigme des machine à vecteurs de support (SVM).

5.3 Publications et Résumé des Chapitres

Publications. Les contributions de la thèse ont fait l'objet de publications et de présentations dans des conférences et journaux d'apprentissage statistique :

Auteurs : E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon.
1− "Gap Safe Screening Rules for Sparse Multi-task and Multi-class Models".
Advances in Neural Information Processing Systems, 811-819, 2015.

2− "Gap Safe Screening Rules for Sparse-Group Lasso".
Advances in Neural Information Processing Systems, 388-396, 2016.

3− "Gap Safe Screening Rules for Sparsity Enforcing Penalties".
The Journal of Machine Learning Research 18 (1), 4671-4703, 2017.

Auteurs : E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, J. Salmon.
4− "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression".
Journal of Physics : Conference Series 904 (1), 012006, 2017.

Auteurs : E. Ndiaye, T. Le, O. Fercoq, J. Salmon, I. Takeuchi.
5− Safe Grid Search with Optimal Complexity. *Soumis dans une revue*, 2018.

Nous présentons les résultats obtenus dans les différents chapitres de la thèse comme suit.

Notations. La variable d'optimisation est un vecteur $\beta = (\beta_1, \dots, \beta_p)^\top$ admettant une structure de groupe. Un groupe de fonctionnalités est un sous-ensemble $g \subset [p]$ et $|g|$ est sa cardinalité. L'ensemble de groupes est noté $\mathcal{G}$ et nous nous concentrons uniquement sur les groupes ne se chevauchant pas qui forment une partition de l'ensemble $[p]$. Nous désignons par β_g le vecteur dans $\mathbb{R}^{|g|}$, qui est la restriction de β aux indices de g. Nous utilisons également la notation $X_g \in \mathbb{R}^{n \times n_g}$ pour désigner la sous-matrice de X assemblée à partir des colonnes d'indices $j \in$ et X_j lorsque les groupes une seule fonctionnalité, *i.e.* quand $g = \{j\}$. Des notations similaires sont utilisées pour les observations et le groupe d'échantillons sera noté $\mathcal{I}$.

Nous considérons les problèmes de minimisation du risque empirique régularisés comme étant la somme d'un terme lisse (ajustement aux données) et d'un terme non lisse (pénalité sur la complexité de la solution, indirectement sa parcimonie), ou inversement. Utilisant la dualité de Fenchel, nous travaillons avec les formulations primaux/duaux suivantes :

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{I}} f_i(x_i^\top \beta) + \lambda \sum_{g \in \mathcal{G}} \Omega_g(\beta_g) =: P_\lambda(\beta) \qquad \text{(Primal)}, \tag{5.7}$$

$$\hat{\theta}^{(\lambda)} \in \arg\max_{\theta \in \mathbb{R}^n} -\sum_{i \in \mathcal{I}} f_i^*(-\lambda \theta_i) - \lambda \sum_{g \in \mathcal{G}} \Omega_g^*(X_g^\top \theta) =: D_\lambda(\theta) \qquad \text{(Dual)}. \tag{5.8}$$

De plus, nous avons les conditions d'optimalité reliant les solutions primale et duale :

$$\forall i \in \mathcal{I}, \quad -\lambda \hat{\theta}_i^{(\lambda)} \in \partial f_i(x_i^\top \hat{\beta}^{(\lambda)}) \Longleftrightarrow x_i^\top \hat{\beta}^{(\lambda)} \in \partial f_i^*(-\lambda \hat{\theta}_i^{(\lambda)}), \tag{5.9}$$

$$\forall g \in \mathcal{G}, \quad X_g^\top \hat{\theta}^{(\lambda)} \in \partial \Omega_g(\hat{\beta}_g^{(\lambda)}) \Longleftrightarrow \hat{\beta}_g^{(\lambda)} \in \partial \Omega_g^*(X_g^\top \hat{\theta}^{(\lambda)}). \tag{5.10}$$

Nous montrons comment exploiter une structure particulière des solutions (leur parcimonie) pour ignorer des variables sans importance lors du processus d'optimisation. Nous garantissons de ne pas les exclure à tord et par conséquent nous accélérons la résolution des problèmes 5.7 et 5.8. La raison sous-jacente est qu'il n'y a aucun avantage à effectuer des calculs sans valeur impliquant des caractéristiques ou des observations non influentes. Cette stratégie dite de *criblage sûre* suit le travail fondateur de El Ghaoui et al. (2012) et a rapidement conduit à une littérature de plus en plus florissante afin de l'appliquer à différentes instanciations du problème (1.6).

Nous proposons ici, un cadre unificateur qui met en évidence les structures sous-jacentes des fonctions convexes qui sont souvent implicitement exploitées pour établir ces règles de filtrage cans la formulation de la minimisation du risque empirique régularisée séparable et non différentiable. Notre méthode repose sur l'exploitation de conditions d'optimalité de premier ordre et des propriétés de séparation des sous-différentiels de fonctions convexes. Ainsi, nous généralisons les règles de criblage de variables précédemment connues dans la littérature. Il s'applique à une grande classe de tâches d'apprentissage supervisées telles que Lasso, Sparse-Group Lasso, Lasso multitâche, régression logistique binaire et multinomiale, machine à vecteurs de support (SVM), pour en nommer quelques-unes.

Nous présentons brièvement une règle générale d'indentification des groupes de variables ainsi que la construction des régions de sécurité permettant de garantir que l'on élimine que des variables non influentes.

Theorem 6. *Soit $\mathcal{R}^*$ un ensemble (dual) fermé et convexe qui contient la solution (duale) optimale $\hat{\theta}^{(\lambda)}$. Pour tout groupe g de variable dans $\mathcal{G}$, pour tout vecteur $\beta_g^\star$ tel que $\mathrm{int}\partial\Omega_g(\beta_g^\star)$ est non vide, nous obtenons les règles d'identification suivantes :*

$$\textbf{\textit{Règle de Criblage :}} \qquad \textit{Si } \Omega_g^\circ(X_g^\top \hat{\theta}^{(\lambda)}, \beta_g^\star) < 1 \textit{ alors } \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ . \tag{5.11}$$

$$\textbf{\textit{Règle de Criblage Sûre :}} \qquad \textit{Si } \max_{\theta \in \mathcal{R}^*} \Omega_g^\circ(X_g^\top \theta, \beta_g^\star) < 1 \textit{ alors } \hat{\beta}_g^{(\lambda)} = \beta_g^\star \ . \tag{5.12}$$

Le terme $\Omega_g^\circ(X_g^\top \theta, \beta_g^\star)$ quantifie une distance entre le vecteur $X_g^\top \theta$ et la frontière de l'ensemble $\partial\Omega_g(\beta_g^\star)$. Les règles de criblage permettent donc d'identifier les groupes de vecteur qui ne saturent pas les certificats d'optimalité. Dans le cas du Lasso, les coordonnées du vecteur éliminées correspondent aux emplacements j où la solution optimale $\hat{\beta}_j^{(\lambda)}$ est nulle. Ainsi, en se focalisant que sur les coordonnées non nulles, nous réduisons la dimensionnalité du problème et économisons des calculs.

Theorem 7 (Construction d'une Région de Sécurité dans le Dual)**.** *Nous supposons que f_i admet un gradient $1/\gamma$-Lipschitz. Pour tout $\beta \in \mathrm{dom}P_\lambda$ et $\theta \in \mathrm{dom}D_\lambda$, le saut de dualité est défini par* $\mathrm{Gap}_\lambda(\beta, \theta) := P_\lambda(\beta) - D_\lambda(\theta)$. *Nous avons*

$$\left\| \hat{\theta}^{(\lambda)} - \theta \right\| \leqslant \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\beta, \theta)}{\gamma \lambda^2}} \quad . \tag{5.13}$$

De ce fait, la boule $\mathcal{R}^* := \mathcal{B}(\theta, \sqrt{2\,\mathrm{Gap}_\lambda(\beta, \theta)/\gamma\lambda^2})$ est une region de sécurité i.e. *contient la solution dual optimal $\hat{\theta}^{(\lambda)}$ quelque soit $\beta \in \mathrm{dom}P_\lambda$ et $\theta \in \mathrm{dom}D_\lambda$. Nous proposons de construire un vecteur dual en mettant à l'echelle l'image de l'application du gradient*

$$\theta := \frac{-\nabla f(X\beta)}{\max(\lambda, \mathcal{S}^\circ_{\mathrm{dom}\Omega^*}(X^\top \nabla f(X\beta)))} \in \mathrm{dom}D_\lambda \quad . \tag{5.14}$$

En exploitant les informations fournies par les bornes dépendantes du saut de dualité, nous fournissons des résultats théoriques, tels que la complexité en itération de l'identification des ensembles actifs (optimaux), et concevons de nouveaux algorithmes rapides pour éliminer en toute sécurité davantage de variables que les règles qui ont été considérées auparavant, notamment pour les paramètres de régularisation de faible intensité. Notre approche peut s'adapter à n'importe quel algorithm itératif, mais convient particulièrement bien aux méthodes de descente par blocs de coordonnées. Nous introduisons également de nouvelles stratégies de démarrage à chaud qui ont montré nette amélioration des performances. Dans toutes nos expériences numériques, nous rapportons des accélérations significatives par rapport aux règles de criblage proposées précédemment dans la litterature sur tous les jeux de données testés.

Chapitre 3. **Optimisation Globale par Calcul du Chemin Complet de Régularisation et Sélection Optimale d'Hyperparamètre.**

Nous discutons de l'optimisation par approximation de chemin et de son application dans la sélection de modèle. Malgré la propriété intéressante des méthodes d'homotopie pour fournir une meilleure prédiction en termes de performances en généralisation, la sélection de l'intensité optimale de régularisation λ pour l'erreur de validation est parfois difficile, même pour un problème tel que Lasso, dans lequel nous pouvons trouver un algorithme qui calcule exactement la solution entière. Les algorithmes de calcul du chemin exact tel que le `Lars` (Efron et al., 2004) ou les méthodes basées sur un prédicteur-correcteur peuvent souffrir d'instabilités numériques dues à plusieurs inversions de matrice et à leur complexité, c'est-à-dire le nombre de segments linéaires dans le chemin, pouvant être exponentielle en la dimension du problème. Par exemple, la complexité pire cas pour le Lasso est exactement $(3^p + 1)/2$ (Mairal and Yu, 2012) et $O(2^n)$ pour le SVM (Gärtner et al., 2012). Ces complexités pire cas montrent que même si ces algorithmes peuvent être très efficaces et fournissent des solutions exactes (ou de très grande précision), ils peuvent être totalement impraticables en grande dimension.

Dans ce chapitre, nous revenons sur les techniques d'approximation du chemin de regularisation pour une tolérance prédéfinie ϵ, dans un cadre unifié et nous montrons que sa complexité est de $O(1/\sqrt[d]{\epsilon})$ pour les fonctions uniformément convexes d'ordre $d > 0$ et $O(1/\sqrt{\epsilon})$ pour les fonctions auto-concordantes généralisées. Cela inclut des exemples tels que la perte quadratique qui intervient dans l'estimateur des moindres carrés, mais aussi l'important exemple de perte logistique qui, à notre connaissance, n'a pas été efficacement traité par les travaux antérieurs. De plus, nous clarifions le lien entre la complexité de l'approximation du chemin de regularisation et la régularité de la fonction de perte considérée dans le cadre de la minimisation du risque empirique réguralisée.

Enfin, nous tirons parti de notre technique pour exposer des bornes plus précises sur l'erreur de validation et fournissons un algorithme pratique pour la sélection d'hyperparamètre avec de plus forte garantie. Plus précisément, étant donné le fractionnement des données d'apprentissage et de validation $(y_{\text{train}}, X_{\text{train}}, y_{\text{val}}, X_{\text{val}})$, nous formulons le problème du choix d'hyperparamètre comme une optimisation à deux niveaux

$$\begin{aligned} \underset{\lambda \in [\lambda_{\min}, \lambda_{\max}]}{\arg\min} \; & E_v(\hat{\beta}^{(\lambda)}) = \mathcal{L}(y_{\text{val}}, X_{\text{val}} \hat{\beta}^{(\lambda)}) \\ & \text{s.t. } \hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \ell(X_{\text{train}} \beta, y_{\text{train}}) + \lambda \Omega(\beta) \ . \end{aligned}$$

Pour une tolérance prescrite $\epsilon_v > 0$ de l'erreur de prédiction sur les données de validation, nous montrons comment concevoir une grille discrète du paramètre $\Lambda_{\text{val}}(\epsilon_v)$ incluse dans le segment $[\lambda_{\min}, \lambda_{\max}]$ tels que :

$$\min_{\lambda_t \in \Lambda_{\text{val}}(\epsilon_v)} E_v(\beta^{(\lambda_t)}) - \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} E_v(\hat{\beta}^{(\lambda)}) \leqslant \epsilon_v.$$

Par conséquent, notre approche consiste simplement en un algorithme d'exploration séquentielle sur une grille de paramètre. Elle fournit un schema numérique garantissant une convergence globale pour approximer l'hyperparamètre optimal à un niveau d'erreur de validation ϵ_v quelconque. Elle s'applique à une grande classe de tâches d'apprentissage statistique. Nous illustrons par des experiences numériques l'utilisation de notre algorithme en pratique sur des problèmes de regression et de classification.

L'Estimation du Maximum de Vraisemblance (EMV) est un paradigme d'apprentissage statistique classique et important, qui nécessite la spécification d'un bon modèle statistique. Par exemple, un modèle linéaire avec du bruit gaussien nécessite l'estimation des paramètres de position et de dispersion (μ, σ^2). Si σ est *connu* et que les observations $y = (y_1, \cdots, y_n)$ sont indépendantes et distribuées de manière identique, l'EMV conduit à l'estimation classique des moindres carrés (5.5) où l'influence de σ peut être ignoré dans la résolution du problème d'optimisatio. Cependant, si σ est *inconnu*, l'estimation seul de μ n'est pas suffisante pour caractériser la distribution P_{μ,σ^2}. Ce qui conduit à un modèle incomplet.

De plus, dans un contexte de grandes dimensions où le nombre d'observations est (très) inférieur au nombre de variable caractéristiques, les méthodes d'estimation fournissant des solutions parcimonieuses, telle que le Lasso, sont très populaires car elles permettent de sélectionner des variables importantes et facilitent l'interprétation des variables discriminantes.

Pour plus d'efficacité, ils s'appuient sur un paramètre de régularisation qui permet de calibrer l'ajustement aux données contre la parcimonie et doit être proportionnel au niveau de bruit σ (Bickel et al., 2009). Pourtant, ce dernier est souvent *inconnu* dans la pratique.

Une solution possible consiste à optimiser conjointement le paramètre de régression μ ainsi que le niveau de bruit σ. Une formulation directe de l'EMV dans le cas gaussien donne

$$\min_{\beta \in \mathbb{R}^p, \sigma > 0} \log(\sigma) + \frac{1}{2\sigma^2} \|y - X\beta\|^2 \quad .$$

Une telle formulation n'aboutit malheuresement pas à un problème conjointement convexe. En outre, lorsque $y = X\beta$ et σ tend vers 0 c'est-à-dire à l'approche des frontière de l'espace des paramètres, la fonction objective tend vers $-\infty$. Le fait qu'elle ne soit pas bornée inférieurement complique à la fois l'analyse statistique et l'optimisation globale.

Nous étudions différentes formulations convexes qui ont été considérées dans la littérature, à savoir (Huber, 1981; Owen, 2007) ainsi que des méthodes de re-paramétrage (Städler et al., 2010). Du point de vue de l'optimisation, nous illustrons les problèmes numériques de la formulation du Lasso concomitant et proposons une modification lisse que nous avons dénommé Smoothed Concomitant Lasso, visant à augmenter les stabilités numériques. Notre proposition s'appuie sur les techniques de lissage à la Nesterov (2005); Beck and Teboulle (2012) du problème initial. En utilisant les règles de criblage sûres sur des chemins de régularisation et le démarrage à chaud développés dans les chapitres précédants, nous proposons une implementation efficace et précise pour une estimation conjointe. Nous évaluons un coÃ»t de calcul similaire à celui du Lasso. Il s'agit d'une avancée significative par rapport aux méthodes précédentes basées sur des solveurs génériques de programmation conique ou de procédure itérative alternant les étapes de Lasso et les étapes d'estimation de bruit.

Pour assurer une reproductibilité de nos expériences numériques, les implémentations des algorithmes proposés dans cette thèse ainsi que leurs codes sources sont disponibles dans la page

`https://github.com/EugeneNdiaye`.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 1974.

E. L. Allgower and K. Georg. *Numerical continuation methods: an introduction*. Springer Science & Business Media, 2012.

A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 2010.

A. Antoniadis, I. Gijbels, S. Lambert-Lacroix, and J-M. Poggi. Joint estimation and variable selection for mean and dispersion in proper dispersion models. *Electronic Journal of Statistics*, 2016.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *NIPS*, 2006.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 2008.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 2010.

D. Azé and J-P. Penot. Uniformly convex and uniformly smooth convex functions. *Annales de la faculté des sciences de Toulouse*, 1995.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 2012.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2009.

A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 2012.

S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 2011.

S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith. Cython: The best of both worlds. *Computing in Science Engineering*, 2011.

V. Bellon, V. Stoven, and C-A. Azencott. Multitask feature selection with task descriptors. In *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, 2016.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 2009.

A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. A dynamic screening principle for the lasso. *EUSIPCO*, 2014.

A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Trans. Signal Process.*, 2015.

J. M. Borwein and H. Wolkowicz. Facial reduction for a cone-convex programming problem. *Journal of the Australian Mathematical Society*, 1981.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. *NIPS*, 2008.

A. L. Brearley, G. Mitra, and H. P. Williams. Analysis of mathematical programming problems prior to applying the simplex algorithm. *Mathematical programming*, 1975.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 1995.

L. D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics, 1986.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, 2011.

O. Burdakov. A new vector norm for nonlinear curve fitting and some other optimization problems. *33. Int. Wiss. Kolloq. Fortragsreihe "Mathematische Optimierung | Theorie und Anwendungen"*, 1988.

O. Burdakov and B. Merkulov. On a new norm for data fitting and optimization problems. *Linköping University, Linköping, Sweden, Tech. Rep. LiTH-MAT*, 2001.

S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly. Sparse group lasso: Consistency and climate applications. *SIAM International Conference on Data Mining*, 2012.

S. S. Chen and D. L. Donoho. Atomic decomposition by basis pursuit. *Tech. Report, Stanford University*, 1995.

S. Chrétien and S. Darses. Sparse recovery with unknown variance: a lasso-type approach. *IEEE Trans. Inf. Theory*, 2011.

K. L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, page 63, 2010.

P. L. Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued and Variational Analysis*, 2016.

P. L. Combettes and C. L. Müller. Perspective maximum likelihood-type estimation via proximal decomposition. *arXiv preprint arXiv:1805.06098*, 2018.

P. L. Combettes and J-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011.

L. Condat. Fast projection onto the simplex and the l1 ball. *Mathematical Programming*, 2016.

C. F. Dantas and R. Gribonval. Dynamic screening with approximate dictionaries. In *XXVIème colloque GRETSI*, 2017.

C. F. Dantas and R. Gribonval. Faster and still safe: combining screening techniques and structured dictionaries to accelerate the lasso. In *ICASSP 2018-IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res*, 2016.

L. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 2014.

D. Drusvyatskiy and H. Wolkowicz. The many faces of degeneracy in conic optimization. *Foundations and Trends in Optimization*, 2017.

C. Dünner, S. Forte, M. Takáč, and M. Jaggi. Primal-dual rates and certificates. *ICML*, 2016.

A. Ebadian, I. Nikoufar, and M. E. Gordji. Perspectives of matrix convex functions. *Proceedings of the National Academy of Sciences*, 2011.

E. G. Effros. A matrix convexity approach to some celebrated quantum inequalities. *Proceedings of the National Academy of Sciences*, 2009.

B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 2004.

M. A. Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 2012.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *JRSSB*, 2008.

J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Roy. Statist. Soc. Ser. B*, 2012.

R.-E Fan, K.-W Chang, C.-J Hsieh, X.-R Wang, and C.-J Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res*, 2008.

O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 2015.

O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. *ICML*, pages 333–342, 2015.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 1956.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 2007.

J. Friedman, T. Hastie, and Robert R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010a.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 2010b.

B. Gärtner, M. Jaggi, and C. Maria. An exponential lower bound on the complexity of regularization paths. *Journal of Computational Geometry*, 2012.

J. Giesen, M. Jaggi, and S. Laue. Approximating parameterized convex optimization problems. *European Symposium on Algorithms*, 2010.

J. Giesen, J. K. Müller, S. Laue, and S. Swiercy. Approximating concavely parameterized optimization problems. *NIPS*, 2012.

T. Goldstein and Stanley Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2009.

A. Gramfort, M. Kowalski, and M. Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Phys. Med. Biol.*, 2012.

A. Gramfort, D. Strohmeier, J. Haueisen, M.S. Hamalainen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *NeuroImage*, 2013.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 2002.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *J. Mach. Learn. Res*, 2004.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2009.

M. Hebiri and S. van de Geer. The smooth-lasso and other l1 + l2-penalized methods. *Electronic Journal of Statistics*, 2011.

J-B. Hiriart-Urruty. A note on the Legendre-Fenchel transform of convex composite functions. *Nonsmooth Mechanics and Analysis*, 2006.

J-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. II*. Springer-Verlag, 1993.

J-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer, 2012.

A. E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 1962.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970.

P. J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, 1964.

P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.

P. J. Huber and R. Dutter. Numerical solution of robust regression problems. In *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*, pages 165–172. Physica Verlag, Vienna, 1974.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 2004.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res*, 2011.

T. B. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. *ICML*, 2015.

T. B. Johnson and C. Guestrin. Unified methods for exploiting piecewise linear structure in convex optimization. *NIPS*, 2016.

B. Jorgensen. *The theory of dispersion models*. CRC Press, 1997.

A. Juditski and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 1996.

K. Koh, S-J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res*, 2007.

K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 2000.

S. C. Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 1931.

J. D. Lee, Y. Sun, and M. A. Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 2014.

S. Lee and E. P. Xing. Screening rules for overlapping group lasso. *preprint arXiv:1410.6880v1*, 2014.

S. Lee, J. Zhu, and E. P. Xing. Adaptive multi-task lasso: with application to eqtl detection. *NIPS*, 2010.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*, 2016a.

X. Li, J. Haupt, R. Arora, H. Liu, M. Hong, and T. Zhao. A first order free lunch for sqrt-lasso. *arXiv preprint arXiv:1605.07950*, 2016b.

J. Liang, J. Fadili, and G. Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 2017.

H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. *ICML*, 2009.

J. Liu, Z. Zhao, J. Wang, and J. Ye. Safe screening with variational inequalities and its application to lasso. *ICML*, 2014.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 2015.

J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. *ICML*, 2012.

H. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics (NRL)*, 1956.

Harry Markowitz. Portfolio selection. *The journal of finance*, 1952.

B. Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 1970.

M. Massias, A. Gramfort, and J. Salmon. From safe screening rules to working sets for faster lasso-type solvers. *arXiv preprint arXiv:1703.07285*, 2017.

M. Massias, O. Fercoq, A. Gramfort, and J. Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. *AISTATS*, 2018a.

M. Massias, A. Gramfort, and J. Salmon. Celer: a Fast Solver for the Lasso with Dual Extrapolation. In *ICML*, 2018b.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

Cs Mészáros and U. H. Suhl. Advanced preprocessing techniques for linear and quadratic programming. *OR Spectrum*, 2003.

C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of rn. *Journal of Optimization Theory and Applications*, 1986.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP safe screening rules for sparse multi-task and multi-class models. *NIPS*, 2015.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP safe screening rules for Sparse-Group Lasso. *NIPS*, 2016.

E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclere, and J. Salmon. Efficient smoothed concomitant Lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 2017a.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res*, 2017b.

Y. Nesterov. *Introductory lectures on convex optimization*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 2005.

Y. Nesterov. Gradient methods for minimizing composite objective function. *Citeseer*, 2007.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 2012.

J. Nutini, I. Laradji, and M. Schmidt. Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*, 2017.

G. Obozinski and F. Bach. A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. *HAL Id : hal-01412385, version 1*, 2016.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 2010.

K. Ogawa, Y. Suzuki, and I. Takeuchi. Safe screening of non-support vectors in pathwise svm computation. *ICML*, 2013.

J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Siam, 1970.

M. R. Osborne. An effective method for computing regression quantiles. *IMA Journal of Numerical Analysis*, 1992.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 2000a.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 2000b.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 2007.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 2014.

M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007.

F. Pedregosa. Hyperparameter optimization with approximate gradient. *ICML*, 2016.

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. *J. Mach. Learn. Res*, 2011.

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D-Y. Noh, J. R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 2010.

J. Peypouquet. *Convex optimization in normed spaces: theory, methods and examples*. Springer, 2015.

V. Pham and L. El Ghaoui. Robust sketching for multiple square-root LASSO problems. In *AISTATS*, 2015.

B. Playe, C-A. Azencott, and V. Stoven. Efficient multi-task chemogenomics for drug specificity prediction. *bioRxiv*, 2018.

A. Raj, J. Olbrich, B. Gärtner, B. Schölkopf, and M. Jaggi. Screening rules for convex problems. *arXiv preprint arXiv:1609.07478*, 2016.

A. Rakotomamonjy. Variable selection using svm-based criteria. *J. Mach. Learn. Res*, 2003.

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*, 2013.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 2014.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.

R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1997. Reprint of the 1970 original, Princeton Paperbacks.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 2007.

M. Sangnier, O. Fercoq, and F. d'Alché Buc. Data sparse nonparametric regression with ε-insensitive losses. In *Asian Conference on Machine Learning*, 2017.

E. J. Schlossmacher. An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association*, 1973.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 1978.

D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, 2016.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *ICML*, 2014.

A. Shibagaki, Y. Suzuki, M. Karasuyama, and I. Takeuchi. Regularization path of cross-validation error lower bounds. *NIPS*, 2015.

A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. *ICML*, 2016.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *J. Comput. Graph. Statist.*, 2013.

P. Sprechmann, I. Ramirez, G. Sapiro, and C. E. Yonina. Collaborative hierarchical sparse modeling. *Information Sciences and Systems (CISS)*, 2010.

P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *IEEE Trans. Signal Process.*, 2011.

N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 2010.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1981.

S. M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 2007.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: A recipe for newton-type methods. *arXiv preprint arXiv:1703.04599*, 2017.

T. Sun and C-H. Zhang. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 2010.

T. Sun and C-H. Zhang. Scaled sparse linear regression. *Biometrika*, 2012.

T. Sun and C-H. Zhang. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res*, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2012.

R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 2013.

A. N. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39:176–179, 1943.

S. van de Geer and B. Stucky. χ2-confidence sets in high-dimensional regression. *Statistical Analysis for High-Dimensional Data*, 2016.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, 1998.

J. Wang and J. Ye. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. *NIPS*, 2014.

J. Wang, J. Zhou, J. Liu, P. Wonka, and J. Ye. A safe screening rule for sparse logistic regression. *NIPS*, 2014.

J. Wang, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. *J. Mach. Learn. Res*, 2015.

J. Warga. Minimizing certain convex functions. *Journal of the Society for Industrial and Applied Mathematics*, 1963.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 2015.

D. Wrinch and H. Jeffreys. Xlii. on certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1921.

Z. J. Xiang, H. Xu, and P. J. Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. *NIPS*, 2011.

H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. *IEEE Trans. Inf. Theory*, 2010.

Q. Xu, S. J. Pan, H. Xue, and Q. Yang. Multitask learning for protein subcellular location prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2011.

T. Yoshida, I. Takeuchi, and M. Karasuyama. Safe triplet screening for distance metric learning. *arXiv preprint arXiv:1802.03923*, 2018.

G. Yu and J. Bien. Estimating the error variance in a high-dimensional linear model. *arXiv preprint arXiv:1712.02412*, 2017.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 2006.

S. Yun. On the iteration complexity of cyclic coordinate gradient descent methods. *SIAM J. Optim.*, 2014.

Y. Zeng and P. Breheny. The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in R. *arXiv preprint arXiv:1701.05936*, 2017.

D. Zhang, D. Shen, and Alzheimer's Disease Neuroimaging Initiative. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *Neuroimage*, 2012.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 2005.

Titre : Algorithmes d'Optimisation Sûrs pour la Sélection de Variables et le Réglage d'Hyperparamètre.

Mots Clefs : Optimisation Convexe, Parcimonie Structurée, Élimination Sûre de Variables, Ensemble Actif, Chemin de Régularisation, Estimation de Variance.

Résumé :
Le traitement massif et automatique des données requiert le développement de techniques de filtration des informations les plus importantes. Parmi ces méthodes, celles présentant des structures parcimonieuses se sont révélées idoines pour améliorer l'efficacité statistique et computationnelle des estimateurs, dans un contexte de grande dimension. Elles s'expriment souvent comme solution de la minimisation du risque empirique régularisé s'écrivant comme une somme d'un terme lisse qui mesure la qualité de l'ajustement aux données, et d'un terme non lisse qui pénalise les solutions complexes. Cependant, une telle manière d'inclure des informations a priori introduit de nombreuses difficultés numériques pour résoudre le problème d'optimisation sous-jacent et pour calibrer le niveau de régularisation. Ces problématiques ont été au coeur des questions que nous avons abordées dans cette thèse.

Une technique récente, appelée «Screening Rules», propose d'ignorer certaines variables pendant le processus d'optimisation en tirant bénéfice de la parcimonie attendue des solutions. Ces règles d'élimination sont dites sûres lorsqu'elles garantissent de ne pas rejeter les variables à tort. Nous proposons un cadre unifié pour identifier les structures importantes dans ces problèmes d'optimisation convexe, et introduisons les règles «Gap Safe Screening Rules». Elles permettent d'obtenir des gains considérables en temps de calcul grâce à la réduction de la dimension induite par cette méthode. De plus, elles s'incorporent facilement aux algorithmes itératifs et s'appliquent à un plus grand nombre de problèmes que les méthodes précédentes.

Pour trouver un bon compromis entre minimisation du risque et introduction d'un biais d'apprentissage, les algorithmes d'homotopie offrent la possibilité de tracer la courbe des solutions en fonction du paramètre de régularisation. Toutefois, ils présentent des instabilités numériques dues à plusieurs inversions de matrice, et sont souvent coûteux en grande dimension. Aussi, ils ont des complexités exponentielles en la dimension du modèle dans des cas défavorables. En autorisant des solutions approchées, une approximation de la courbe des solutions permet de contourner les inconvénients susmentionnés. Nous revisitons les techniques d'approximation des chemins de régularisation pour une tolérance prédéfinie, et nous analysons leur complexité en fonction de la régularité des fonctions de perte en jeu. Il s'ensuit une proposition d'algorithmes optimaux ainsi que diverses stratégies d'exploration de l'espace des paramètres. Ceci permet de proposer une méthode de calibration de la régularisation avec une garantie de convergence globale pour la minimisation du risque empirique sur les données de validation.

Le Lasso, un des estimateurs parcimonieux les plus célèbres et les plus étudiés, repose sur une théorie statistique qui suggère de choisir la régularisation en fonction de la variance des observations. Ceci est difficilement utilisable en pratique car la variance du modèle est une quantité souvent inconnue. Dans de tels cas, il est possible d'optimiser conjointement les coefficients de régression et le niveau de bruit. Ces estimations concomitantes, apparues dans la littérature sous les noms de Scaled Lasso, Square-Root Lasso, fournissent des résultats théoriques aussi satisfaisants que celui du Lasso tout en étant indépendants de la variance réelle. Bien que présentant des avancées théoriques et pratiques importantes, ces méthodes sont numériquement instables et les algorithmes actuellement disponibles sont coûteux en temps de calcul. Nous illustrons ces difficultés et nous proposons à la fois des modifications basées sur des techniques de lissage pour accroître la stabilité numérique de ces estimateurs, ainsi qu'un algorithme plus efficace pour les obtenir.

Title : Safe Optimization Algorithms for Variable Selection and Hyperparameter Tuning

Keys words : Convex Optimization, Structured Sparsity, Safe Screening Rules, Active set Regularization Path, Variance Estimation

Abstract :
Massive and automatic data processing requires the development of techniques able to filter the most important information. Among these methods, those with sparse structures have been shown to improve the statistical and computational efficiency of estimators in a context of large dimension. They can often be expressed as a solution of regularized empirical risk minimization, and generally lead to non differentiable optimization problems in the form of a sum of a smooth term, measuring the quality of the fit, and a non-smooth term, penalizing complex solutions. Although it has considerable advantages, such a way of including prior information, unfortunately introduces many numerical difficulties, both for solving the underlying optimization problem and to calibrate the level of regularization. Solving these issues has been at the heart of this thesis.

A recently introduced technique, called "Screening Rules", proposes to ignore some variables during the optimization process by benefiting from the expected sparsity of the solutions. These elimination rules are said to be safe when the procedure guarantees that no variable is wrongly rejected. In this work, we propose a unified framework for identifying important structures in these convex optimization problems and we introduce the "Gap Safe Screening Rules". They allows to obtain significant gains in computational time thanks to the dimensionality reduction induced by this method. In addition, they can be easily inserted into iterative algorithms and apply to a large number of problems.

To find a good compromise between minimizing risk and introducing a learning bias, (exact) homotopy continuation algorithms offer the possibility of tracking the curve of the solutions as a function of the regularization parameters. However, they exhibit numerical instabilities due to several matrix inversions and are often expensive in large dimension. Another weakness is that a worst-case analysis shows that they have exact complexities that are exponential in the dimension of the model parameter. Allowing approximated solutions makes possible to circumvent the aforementioned drawbacks by approximating the curve of the solutions. In this thesis, we revisit the approximation techniques of the regularization paths given a predefined tolerance and we propose an in-depth analysis of their complexity *w.r.t.* the regularity of the loss functions involved. Hence, we propose optimal algorithms as well as various strategies for exploring the parameters space. We also provide a calibration method (for the regularization parameter) that enjoys global convergence guarantees for the minimization of the empirical risk on the validation data.

Among sparse regularization methods, the Lasso is one of the most celebrated and studied. Its statistical theory suggests choosing the level of regularization according to the amount of variance in the observations, which is difficult to use in practice because the variance of the model is often an unknown quantity. In such case, it is possible to jointly optimize the regression parameter as well as the level of noise. These concomitant estimates, appeared in the literature under the names of Scaled Lasso or Square-Root Lasso, and provide theoretical results as sharp as that of the Lasso while being independent of the actual noise level of the observations. Although presenting important advances, these methods are numerically unstable and the currently available algorithms are expensive in computation time. We illustrate these difficulties and we propose modifications based on smoothing techniques to increase stability of these estimators as well as to introduce a faster algorithm.