# Structured Sparsity-Inducing Norms : Statistical and Algorithmic Properties with Applications to Neuroimaging

Rodolphe Jenatton



# THÈSE DE DOCTORAT
# DE L'ÉCOLE NORMALE SUPÉRIEURE DE CACHAN


présentée par **RODOLPHE JENATTON**


pour obtenir le grade de
**DOCTEUR DE L'ÉCOLE NORMALE SUPÉRIEURE DE CACHAN**

Domaine : **MATHÉMATIQUES APPLIQUÉES**

Sujet de la thèse :

**Normes Parcimonieuses Structurées : Propriétés Statistiques et Algorithmiques avec Applications à l'Imagerie Cérébrale**

—

**Structured Sparsity-Inducing Norms : Statistical and Algorithmic Properties with Applications to Neuroimaging**

---

Thèse présentée et soutenue à Cachan le 24 Novembre 2011 devant le jury composé de :

| | | |
|---|---|---|
| Jean-Yves AUDIBERT | Directeur de recherche, Ecole des Ponts ParisTech | Directeur de thèse |
| Francis BACH | Directeur de recherche, INRIA Paris-Rocquencourt | Directeur de thèse |
| Laurent EL GHAOUI | Professeur, University of California, Berkeley | Rapporteur |
| Rémi GRIBONVAL | Directeur de recherche, INRIA Rennes | Examinateur |
| Eric MOULINES | Professeur, Télécom Paris Tech | Examinateur |
| Guillaume OBOZINSKI | Chargé de recherche, INRIA Paris-Rocquencourt | Examinateur |
| Massimiliano PONTIL | Professeur, University College London | Rapporteur |
| Bertrand THIRION | Directeur de recherche, INRIA Saclay | Examinateur |

---


Thèse préparée au sein des équipes SIERRA et WILLOW du laboratoire d'informatique de l'École Normale Supérieure, Paris.
(INRIA/ENS/CNRS UMR 8548). 23 avenue d'Italie, 75214 Paris.

## RÉSUMÉ

De nombreux domaines issus de l'industrie et des sciences appliquées ont été, au cours des dernières années, les témoins d'une révolution numérique. Cette tendance s'est accompagnée d'une croissance continue du volume des données—vidéos, musiques et images, dont le traitement est devenu un véritable défi technique. Par exemple, il est aujourd'hui fréquent de prendre des centaines de photographies de plusieurs millions de pixels, la moindre application de méthodes du traitement de l'image devenant alors une opération difficile. Dans ce contexte, la parcimonie est apparue comme un concept central en apprentissage statistique et traitement du signal. Il est en effet naturel de représenter, analyser et exploiter les données disponibles à travers un nombre réduit de paramètres. Par exemple, on peut imaginer effectuer de la reconnaissance d'objets sur des images de hautes résolutions en n'utilisant qu'un petit sous-ensemble pertinent de pixels. Alors que les approches générales favorisant la parcimonie ont déjà été l'objet de nombreux travaux—débouchant sur d'élégantes fondations théoriques, des outils algorithmiques efficaces et plusieurs succès pratiques—cette thèse se concentre sur une forme particulière et plus récente de parcimonie, nommée *parcimonie structurée*.

Comme son nom l'indique, nous considérerons des situations où nous ne serons pas simplement intéréssés par la parcimonie, mais où nous aurons également à disposition des connaissances a priori nous renseignant sur certaines propriétés structurelles. En continuant d'exploiter l'exemple de la reconnaissance d'objets mentioné ci-dessus, nous savons que des pixels voisins sur une image ont tendance à partager des propriétés similaires, telles que la classe de l'objet à laquelle ils appartiennent. Ainsi, une approche encourageant la parcimonie devrait tirer partie de cette information spatiale.

L'objectif de cette thèse est de comprendre et analyser le concept de parcimonie structurée, en se basant sur des considérations statistiques, algorithmiques et appliquées. Nous commencerons par introduire une famille de normes structurées parcimonieuses dont les propriétés sont étudiées en détail. En particulier, nous montrerons à quel type d'information structurelle ces normes correspondent, et nous présenterons sous quelles conditions statistiques elles sont capables de produire une séléction consistente de variables. Nous étudierons ensuite l'apprentissage de dictionnaires parcimonieux et structurés, où nous exploiterons les normes introduites précédemment dans un cadre de factorisation de matrices. L'approche qui en résulte est fléxible et versatile, et nous montrerons que les éléments de dictionnaire appris exhibent une structure parcimonieuse adaptée à la classe de signaux considérée. Concernant l'optimisation, nous proposerons différents outils algorithmiques efficaces et capables de passer à l'échelle, tels que des stratégies à ensemble de variables actives ou encore des méthodes proximales. Grâce à ces outils algorithmiques, nous illustrerons sur de nombreuses applications issues de domaines variés, quand et pourquoi la parcimonie structurée peut être bénéfique. Ces illustrations contiennent par exemple, des tâches de restauration en traitement de l'image, la modélisation de documents textuels sous la forme d'une hiérarchie de thèmes, la prédiction de la taille d'objets à partir de signaux d'imagerie par résonance magnétique fonctionnelle, ou encore des problèmes de segmentation d'images en vision par ordinateur.

# INTRODUCTION ET CONTRIBUTIONS DE LA THÈSE

Une grande variété de problèmes en apprentissage statistique peuvent se résumer par l'apprentissage, à partir de données, d'un jeu de paramètres maximisant un critère préétabli, d'une manière soit supervisée ou non supervisée. De tels problèmes apparaissent par exemple lorsqu'on désire expliquer un ensemble de réponses à partir de certaines observations, ou lorsqu'on souhaite résumer, organiser, ou compresser un grand volume de données. Dans tous les cas, et de manière similaire à d'autres domaines des sciences appliquées, une solution simple, *parcimonieuse*, est souvent privilégiée face à d'autres alternatives plus "complexes". Ce biais en faveur de la parcimonie peut être justifié par deux arguments : On peut soit croire a priori que le phénomène étudié admet effectivement une solution parcimonieuse, ou alternativement, et sans cette connaissance a priori, on peut chercher une explication simple parce qu'elle se prête bien à l'interprétation, à la compréhension et conduit à des post-traitements moins coûteux.

En apprentissage statistique, on fait référence aux approches parcimonieuses pour désigner les maniéres de résoudre un problème en n'utilisant qu'un nombre limité de paramètres.

Le thème principal de cette thèse est la *parcimonie structurée*. Comme son nom l'indique, nous allons nous concentrer sur des configurations où non seulement la parcimonie est pertinente, mais aussi où des connaissances a priori suggèrent que la solution attendue exhibe certaines *structures* intéressantes.

Des connaissances a priori de nature structurelle peuvent se manifester sous différentes formes, et reflètent par exemple des propriétés ordinales, temporelles, ou encore spatiales du problème étudié. Cette importante idée est probablement plus simple à comprendre à travers une série d'exemples concrets.

Dans le contexte de la vision par ordinateur, la ségmentation consiste à partitioner des images en sous-parties pertientes (e.g., premier plan et arrière plan). Dans ce cas, il semble naturel de prendre en compte le fait que des pixels voisins dans l'image sont susceptibles de partager des attributs similaires. Les résultats de ségmentation sont effectivement améliorés grâce à cette information spatiale (e.g., voir Boykov et al., 2001). Pour l'étude de séries temporelles non-alignées d'expression de gênes, l'ordonnancement dans le temps des motifs de régulation (e.g., motifs montants et descendants) jouent un rôle majeur. Comme précédemment, modéliser cette information ordinale conduit à de meilleures performances (Tibau Puig et al., 2011). De même, pour le traitement du langage naturel, les données sont générallement représentées par des motifs qui, par exemple, peuvent désigner des pairs de mots contigus labélisés par leurs classes syntaxiques. De tels motifs peuvent être combinés ensemble pour alors former de nouveaux motifs plus complexes. Dans un cadre parcimonieux, il semble naturel de considérer d'abord les paramètres associés aux motifs élémentaires avant d'analyser des motifs plus sophistiqués. En pratique, cette approche hiérarchique se révèle être performante (Martins et al., 2011).

Toutes les illustrations mentionées ci-dessus constituent des exemples où la parcimonie structurées est bénéfique. Dans cette thèse, nous essayons de comprendre ce concept que nous étudions sous de divers angles, algorithmique, statistique et appliqué.

## Résumé des Contributions de la Thèse

Nous listons ci-dessous les contributions qui émergent de la thèse :

- **Chapter 2 :** Ce chapitre introduit et étudie précisemment le principal objet de cette thèse, à savoir, les normes encourageant la parcimonie structurée. Ces dernières sont définies comme combinaisons linéaires de normes élémentaires construites à partir de groupes de variables. Nous considérons l'ensemble de tous les groupes possibles et charactérisons exactement quel type de connaissances a priori elles sont capables d'encoder. Avant de décrire comment on peut passer des groupes à l'ensemble des motifs parcimonieux induits, nous montrons qu'il est possible de construire automatiquement une unique norme minimale, adaptée à la famille de motifs parcimonieux désirés, comme par exemple, l'ensemble des rectangles sur une grille en deux dimensions. Par ailleurs, nous étudions sous quelles conditions statistiques, ces normes permettent d'estimer de manière consistente un motif parcimonieux structuré, aussi bien en faible qu'en haute dimensions. Enfin, nous introduisons un algorithmique efficace et rapide basé sur la construction d'un ensemble de variables actives.
- **Chapter 3 :** Cette seconde partie est dédiée à l'apprentissage de dictionnaires structurés, exploitant et étendant le schéma de régularisation introduit au Chapitre 2. Nous montrons comment on peut apprendre des dictionnaires dont les atomes ont des motifs parcimonieux structurés qui sont adaptés à la classe de signaux considérés. Pour cela, nous présentons une technique d'optimisation simple et efficace qui reposent sur des méthodes de descente "par bloque" disposant de mises à jour explicites. On applique finalement notre outil à un problème de reconnaissance de visages, où nous sommes capables d'apprendre des déscripteurs robustes aux occlusions.
- **Chapter 4 :** Ce chapitre s'intéresse à des méthodes algorithmiques efficaces pour résoudre des problèmes de parcimonie structurée où nous supposons que les variables possèdent une structure hiérarchique. Plus précisemment, nous considérons les méthodes proximales pour lesquelles nous montrons que l'opérateur proximal associé aux normes hiérarchiques peut être calculé exactement via une approche duale. Notre procédure a une complexité linéaire, ou presque linéaire, par rapport au nombre d'atomes, ce qui rend l'application des méthodes proximales accélérées rapides et efficaces. Nous illustrons notre approche par des applications variées, telles que le débruitage d'images naturelles, ou encore la représentation de documents textuels sous forme de hiérarchies de thématiques, établissant par la même occasion des liens avec les modèles probabilistiques utilisés dans ce même cadre.
- **Chapter 5 :** Ce quatrième chapitre considère deux applications de la parcimonie structurée à la neuroimagerie.

D'une part, nous nous intéressons à la prédiction de la taille d'objets à travers dix sujets, en nous basant sur des signaux d'imagerie par résonance magnétique fonctionelle (IRMf). Pour cela, nous introduisons une régularisation hiérarchique structurée construite à partir des données d'entrainement, via une procédure non supervisée exploitant des contraintes spatiales spécifiques. Cette construction permet de prendre en compte la structure spatiale multi-échelle des signaux IRMf, augmentant ainsi la robustesse aux variabilitées inter-sujets. Nous réalisons une comparaison expérimentale impliquant plusieurs algorithmes et formulations, et nous illustrons la capacité de notre approche à localiser des régions du cerveau dédiées au traitement des stimulis visuels.
D'autre part, nous introduisons un modèle génératif pour étudier des séries temporelles du cerveau au repos. Cette classe de signaux est modélisée grâce à l'apprentissage de dictionnaires structurés où les atomes que nous apprenons présentent des formes compactes et localisées en trois dimensions. De plus, notre approche se prête naturellement à la séléction de modèles et à une évaluation quantitative ; dans ce cadre, nous obtenons des améliorations par rapport aux techniques non structurées, mesurées par une vraisemblance sur des données de test.

- **Chapter 6 :** Le cinquième et dernier chapitre de la thèse se situe légèrement en marge de la thématique principale de parcimonie structurée. L'objectif de ce partie est de caractériser les minima locaux du problème (non convexe) de l'apprentissage de dictionnaires parcimonieux, utilisé par exemple dans des extensions structurées aux Chapitres 3 et 4. En particulier, nous considérons un modèle probabiliste de signaux parcimonieux, et nous montrons qu'avec forte probabilité, le problème de l'apprentissage de dictionnaires parcimonieux admet un minimum local suivant certaines courbes passant par le dictionnaire de référence ayant généré les signaux. Notre analyse couvre les cas des dictionnaires redondants et des signaux bruités, étendant ainsi les travaux précédants limités à une configuration sans bruit. L'étude que nous réalisons est non asymptotique et permet de mieux comprendre comment les grandeurs clé du problème, telles que la cohérence ou le bruit, peuvent varier avec la dimension des signaux, le nombre d'observations, le niveau de sparsité et le nombre d'atomes. Ce travail en cours constitue un premier pas vers la preuve plus complexe de l'existence d'un minimum local dans l'ensemble de la variété des dictionnaires normalisés.

ABSTRACT

Numerous fields of applied sciences and industries have been witnessing a process of digitisation over the past few years. This trend has come with a steady increase in the amount of available digital data whose processing was become a challenging task. For instance, it is nowadays common to take thousands of pictures of several millions of pixels, which makes any subsequent image-processing/computer-vision task a computationally demanding exercise.

In this context, parsimony—also known as *sparsity*—has emerged as a key concept in machine learning, statistics and signal processing. It is indeed appealing to represent, analyze, and exploit data through a reduced number of parameters, e.g., performing object recognition over high-resolution images based only on some relevant subsets of pixels. While general sparsity-inducing approaches have already been well-studied—with elegant theoretical foundations, efficient algorithmic tools and successful applications, this thesis focuses on a particular and more recent form of sparsity, referred to as *structured sparsity*.

As its name indicates, we shall consider situations where we are not only interested in sparsity, but where some structural prior knowledge is also available. Continuing the example of object recognition, we know that neighbouring pixels on images tend to share similar properties—e.g., the label of the object class to which they belong—so that sparsity-inducing approaches should take advantage of this spatial information.

The goal of this thesis is to understand and analyze the concept of structured sparsity, based on statistical, algorithmic and applied considerations. To begin with, we introduce a family of structured sparsity-inducing norms whose properties are closely studied. In particular, we show what type of structural prior knowledge they correspond to, and we present the statistical conditions under which these norms are capable of consistently performing structured variable selection. We then turn to the study of sparse structured dictionary learning, where we use the aforementioned norms within the framework of matrix factorization. The resulting approach is flexible and versatile, and it is shown to learn representations whose structured sparsity patterns are adapted to the considered class of signals. From an optimization viewpoint, we derive several efficient and scalable algorithmic tools, such as, working-set strategies and proximal-gradient techniques. With these methods in place, we illustrate on numerous real-world applications from various fields, when and why structured sparsity is useful. This includes, for instance, restoration tasks in image processing, the modelling of text documents as hierarchy of topics, the inter-subject prediction of sizes of objects from fMRI signals, and background-subtraction problems in computer vision.

# ACKNOWLEDGMENTS

It is commonly said that superlatives are inappropriate for scientific writing; in this section, I will have to violate this rule.

The second sentence of these acknowledgements is naturally dedicated to my advisors Jean-Yves Audibert and Francis Bach. I would like to deeply thank them for all their help, support, kindness and enthusiasm over the past three years. It is an invaluable privilege to have had the opportunity to work with them.

I would like to warmly thank Laurent El Ghaoui and Massimiliano Pontil for having reviewed my manuscript. I am also grateful to Rémi Gribonval, Eric Moulines, Guillaume Obozinski, and Bertrand Thirion for having accepted to be part of the jury.

Back in 2008 and while leaving the finance industry and arriving in the Willow (and later, Sierra) Project team, I could not expect such an exceptional environment. Jean Ponce along with Francis Bach have been succeeding in creating great teams with a unique atmosphere of work and with worldwide synergies. It is hard to describe how nice it is to be only surrounded by extremely friendly and smart colleagues, all at the cutting edge in their respective fields of research. I therefore want to thank (in alphabetical order), Sylvain Arlot (for all my annoying questions about concentration of measure), Louise Benoit, Y-Lan Boureau, Florent Couzinié-Devy, Vincent Delaitre, Olivier Duchenne, Cécile Espiègle, Loïc Février, Jan Van Gemert, Edouard Grave, Warith Harchaoui, Toby Hocking, Armand Joulin (for all my measure-theoretic questions that he always considered in a good mood), Ivan Laptev, Augustin Lefèvre (for the pleasure it was to share his office, and for his patience regarding my numerous questions), Nicolas Le Roux, Bryan Russell, Mark Schmidt, Josef Sivic, Mathieu Solnon, and Oliver Whyte.

I also have a special thank for my two good friends, Julien Mairal and Guillaume Obozinski, with whom I closely worked and collaborated. They have this rare ability to convey their passions for research through their charisma and enthusiasm. They have really been role models to look up to, and I am more than happy to see that they are becoming the next worldwide experts in their fields.

This thesis gave me the opportunity to meet and interact with researchers and students from various places, who all positively contributed to my experience. I notably would like to thank: Laurent El Gahoui, Alexandre Gramfort, Rémi Gribonval, Zaid Harchaoui, Laurent Jacob, Jinzhu Jia, Percy Liang, Vincent Michel, Jean Ponce, Bertrand Thirion, Gaël Varoquaux, and Bin Yu. Finally, I would like to thank my high-school math teacher, Jean-Luc Brischoux, who made me enjoy mathematics.

This three-year scientific challenge was made possible by the financial support of INRIA, which I gratefully thank.

Last but not least, I want to warmly thank my close relatives, my family, my parents, sisters, and especially Marie, for her love and her unfailing support over these three years.

# 1

# Introduction and Related Work

A large variety of problems in machine learning and computational statistics amount to learning, from some data, a set of parameters that maximizes a predefined criterion, in either a supervised or unsupervised fashion. Such problems for instance arise when we desire to relate some input observations to some output response, or when we simply wish to summarize, organize or compress large amounts of data. In any case, and similarly to other fields of applied sciences, a simple, *parsimonious* solution is often preferred over more "complex" ones. This bias towards parsimony can be underpined in two ways: First, we may a priori believe that the studied phenomenon has indeed a parsimonious solution, or second, and without this prior knowledge, we may seek a simple explanation because it lends itself well to understanding, interpretability and less processing.

In machine learning and statistics, parsimony is also known as *sparsity*, and, in a broad sense, we shall refer to sparsity-inducing approaches as ways of tackling problems by only resorting to a reduced number of parameters.

The central theme of this thesis is *structured sparsity*. As its name indicates, we shall concentrate on settings where not only sparsity is relevant, but also where prior knowledge suggests that the expected solution exhibits some *structure* of interest.

Structural prior knowledge can come up under various forms and may for instance reflect ordering, spatial or temporal properties of the problem at hand. This important idea is probably easier to understand through concrete examples.

In the context of computer vision, segmentation consists in partitioning images into meaningful parts (e.g., background versus foreground). In this case, it seems natural to take into account the fact that neighboring pixels on the image are likely to share similar attributes. Better segmentation results are indeed obtained thanks to this spatial prior knowledge (e.g., see Boykov et al., 2001, and references therein). While studying misaligned time series of gene expressions across different subjects, the temporal ordering of the regulation patterns (e.g., up- and down-regulations) plays a prominent part. Again, modeling this ordinal information leads to improved performance (Tibau Puig et al., 2011). In natural language processing, the features associated with the data are usually built from templates which may, for instance, represent pairs of contiguous words labeled by their part-of-speech tags. Such templates can be combined with each other, hence forming more complex features. In this context, it seems natural to first consider parameters relative to elementary templates before looking at more sophisticated ones; this hierarchical approach turns out to be powerful in practice (Martins et al., 2011).

All the aforementioned illustrations constitute examples where structured sparsity is beneficial. In this thesis, we try to understand this concept and we thoroughly study it from various viewpoints—algorithmic, statistical as well as applied.

The remainder of the introduction is organized as follows: We first summarize the contributions made by this thesis before introducing more formally the framework of structured sparsity and the key object in our analysis, namely structured sparsity-inducing norms. We then present some background material about dictionary learning and its variants, which are techniques used throughout the manuscript. Finally, since most chapters make heavy use of concepts from convex optimization and/or from statistical learning theory, we dedicate two sections of this introduction to present the necessary material.

## 1.1 Summary of the Contributions of the Thesis

We list the contributions that emerge from this thesis:

- **Chapter 2:** This chapter introduces and studies in details the core object of this thesis, namely, *structured sparsity-inducing norms*. These are defined as linear combinations of norms built over groups of variables. We consider all possible sets of groups and characterize exactly what type of prior knowledge can be encoded by considering families of overlapping groups of variables. Before describing how to go from groups to nonzero patterns, we show that it is possible to "reverse-engineer" a given set of nonzero patterns, i.e., to build the unique minimal set of groups that will generate these patterns. This allows the automatic design of sparsity-inducing norms, adapted to target sparsity patterns such as rectangles on a two-dimensional grid. We moreover study under which conditions we can obtain a consistent estimation of structured nonzero patterns, both in low- and high-dimensional settings. Finally, we introduce an efficient and scalable active-set algorithm.
- **Chapter 3:** This second part is dedicated to sparse structured dictionary-learning, leveraging and extending the regularization scheme introduced in Chapter 2. We show how we can learn dictionaries whose atoms have structured nonzero patterns adapted to the class of signals considered. To this end, we propose an efficient and simple optimization technique based on nested block-coordinate descents with closed-form updates. We eventually apply our unsupervised method to face recognition, where we learn localized features robust to occlusions.
- **Chapter 4:** This chapter focuses on designing efficient algorithmic tools to solve sparse decomposition problems where we assume some hierarchical structure among the variables. Specifically, we resort to proximal methods for which we show that the proximal operator associated with the hierarchical norm is computable exactly via a dual approach. Our procedure has a complexity linear, or close to linear, in the number of atoms, which makes accelerated proximal-gradient techniques highly scalable and efficient. We illustrate our method on various problems, such as the denoising of natural image patches and the repre-

sentation of text documents over hierarchies of topics, thus establishing interesting connections with probabilistic topic models.

- **Chapter 5:** This fourth section concentrates on two applications of structured sparsity to neuroimaging.
  On the one hand, we consider the prediction of sizes of objects across ten subjects based on data investigating object coding in high-level visual cortex. To this end, we introduce a sparse hierarchical structured regularization built in a data-driven fashion, from a spatially-constrained agglomerative clustering. This construction makes it possible to take into account the multi-scale spatial structure of fMRI data, thus resulting in more robustness to inter-subject variability. We conduct an experimental comparison of several algorithms and formulations and illustrate the ability of the proposed method to localize in space and in scale some brain regions involved in the processing of visual stimuli.
  On the other hand, we introduce a generative model to study brain resting-state time series. This class of signals is modeled by sparse structured dictionary learning where we learn atoms that exhibit three-dimensional localized clusters. Moreover, the proposed approach provides a natural framework for model selection and quantitative evaluation, where we obtain improvements over unstructured methods, as measured by some likelihood on held-out data.
- **Chapter 6:** The fifth and last chapter of this thesis is slightly next to the scope of the main theme developed so far, namely, structured sparsity. In this part, we aim at characterizing the local minima of sparse coding, also known as sparse dictionary learning (see Chapters 3 and 4 for some structured extensions) which relies on a non-convex procedure whose local minima have not been fully analyzed yet. To this end, we consider a probabilistic model of sparse signals, and show that, with high probability, sparse coding admits a local minimum along some curves passing through the reference dictionary generating the signals.
  Our study takes into account the case of over-complete dictionaries and noisy signals, thus extending previous work limited to noiseless settings and/or under-complete dictionaries. The analysis we conduct is non-asymptotic and makes it possible to understand how the key quantities of the problem, such as the coherence or the level of noise, are allowed to scale with respect to the dimension of the signals, the number of atoms, the sparsity and the number of observations. This work in progress constitutes a first step towards the more involved proof of the existence of the local minimum over the entire manifold of normalized dictionaries.

## 1.2 Notation

Throughout the manuscript, we shall refer to vectors as bold lower case (possibly Greek) letters, and matrices by bold upper case ones. For any integer $j$ in the set $[\![1;p]\!] \triangleq \{1,\dots,p\}$, we denote the $j$-th coefficient of a $p$-dimensional vector $\mathbf{w} \in \mathbb{R}^p$ by $\mathbf{w}_j$. Similarly, for any matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$, we refer to the entry at the $i$-th row and $j$-th

column as $\mathbf{W}_{ij}$, for any $(i,j) \in [\![1;n]\!] \times [\![1;p]\!]$. We will also have to describe sub-vectors of $\mathbf{w} \in \mathbb{R}^p$, that is, for any $\mathrm{J} \subseteq [\![1;p]\!]$, we denote by $\mathbf{w}_\mathrm{J} \in \mathbb{R}^{|\mathrm{J}|}$ the vector formed by the entries of $\mathbf{w}$ indexed by J. Likewise, for any $\mathrm{I} \subseteq [\![1;n]\!]$, $\mathrm{J} \subseteq [\![1;p]\!]$, we denote by $\mathbf{W}_{\mathrm{IJ}} \in \mathbb{R}^{|\mathrm{I}|\times|\mathrm{J}|}$ the sub-matrix of $\mathbf{W}$ formed by the rows (respectively, the columns) indexed by I (respectively by J).

We extensively manipulate norms in this thesis. We thus define the $\ell_q$-norm for any vector $\mathbf{w} \in \mathbb{R}^p$ by

$$\|\mathbf{w}\|_q \triangleq \Big[\sum_{j=1}^{p} |\mathbf{w}_j|^q\Big]^{1/q} \text{ for } q \in [1,\infty), \quad \text{and} \quad \|\mathbf{w}\|_\infty \triangleq \max_{j\in[\![1;p]\!]} |\mathbf{w}_j|.$$

For $q \in (0,1)$, we extend the definition above to $\ell_q$ quasi-norms. In the same vein, for any matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$, we define the Frobenius norm of $\mathbf{W}$ by

$$\|\mathbf{W}\|_\mathrm{F} \triangleq \Big[\sum_{i=1}^{n}\sum_{j=1}^{p} \mathbf{W}_{ij}^2\Big]^{1/2}.$$

Also, we refer to the set of nonnegative real numbers as $\mathbb{R}_+ \triangleq \{t \in \mathbb{R};\, t \geq 0\}$. If need be, additional and more specific notation may be introduced at the beginning of some chapters.

## 1.3 Variable Selection and Sparsity-Inducing Regularizations

In a broad sense, a significant fraction of the work presented in this thesis is about *variable selection*, also known as *feature selection*. Throughout this manuscript, we should understand "variable" and "feature" as being a descriptor used to represent the data, for instance, the intensity of a pixel taken from an image or the frequency of a word in a document. Nowadays, data are not only becoming abundant in many scientific and industrial fields (e.g., the web and finance industries), but they are also being made available through richer and more complex representations (e.g., the ever increasing resolution of images).

In this context, variable selection is an essential tool that has primarily three purposes (Guyon and Elisseeff, 2003): (1) "summarizing" the description of the data to make it more *interpretable* and understandable, (2) obtaining a more compact and effective representation—e.g., for compression, (3) and finally, concentrating the *predictive power* of the different features when prediction accuracy matters.

In this thesis, we are interested in these three aspects, laying the emphasis on the first and last points. More concretely, variable selection shall amount to finding a subset of relevant features among a total of $p$ variables, that is, learning a *sparse* vector of parameters $\mathbf{w}$ in $\mathbb{R}^p$ whose set of nonzero coefficients characterizes the corresponding set of selected features. Note that the formulations of the underlying learning problems will be made more formal in the subsequent sections. There exist numerous ways of

addressing variable selection, such as, for instance, univariate statistical tests and greedy forward/backward search strategies to name only a few (Guyon and Elisseeff, 2003). In the remainder of this thesis, we shall focus instead on the concept of *regularization*, and more precisely, *sparsity-inducing regularization*. Let us introduce more formally this concept.

### 1.3.1 Sparsity-Inducing Regularization

In machine learning, statistics and signal processing, we usually learn a vector of parameters $\mathbf{w}$ in $\mathbb{R}^p$ by minimizing a convex function $f : \mathbb{R}^p \to \mathbb{R}_+$ that measures how well $\mathbf{w}$ fits some data. In most of the cases encountered in this thesis, we shall assume the function $f$ to be smooth—typically, differentiable with Lipschitz-continuous gradient. The choice of this function is driven by the application at hand, and $f$ generally corresponds to either a data-fitting term, or an empirical risk, i.e., the average of a loss function over a training set of data-points (e.g., see Shawe-Taylor and Cristianini (2004) for a thorough description of loss functions).

To express our a priori assumption that the learned vector $\mathbf{w}$ should be sparse in order to perform variable selection, we also consider a regularization term $\Omega : \mathbb{R}^p \to \mathbb{R}_+$, so that our formulation becomes

$$\min_{\mathbf{w}\in\mathcal{W}} [f(\mathbf{w}) + \lambda\Omega(\mathbf{w})]. \tag{1.1}$$

The scalar $\lambda \geq 0$ is known as the regularization parameter and it controls the effect of $\Omega$, while $\mathcal{W} \subseteq \mathbb{R}^p$ is a convex set that can possibly encode further properties of the problem, such as the non-negativity of the coefficients of $\mathbf{w}$. To promote sparse solutions, $\Omega$ should intuitively penalize vectors $\mathbf{w}$ that have many nonzero coefficients. Thus, a natural candidate to consider is the $\ell_0$ pseudo-norm, that is,

$$\|\mathbf{w}\|_0 \triangleq |\{j \in [\![1; p]\!];\ \mathbf{w}_j \neq 0\}|.$$

Indeed, by definition, $\|\mathbf{w}\|_0$ records the number of nonzero components of the vector $\mathbf{w}$, so that when it is used in (1.1), the less sparse a vector, the more heavily penalized. However, this regularizer does not lend itself well to optimization ($\mathbf{w} \mapsto \|\mathbf{w}\|_0$ is not even continuous), and generally leads to combinatorial and intractable problems (e.g., see Natarajan (1995) in the context of least-squares regression). This computational challenge prompts the need for surrogates (or relaxations) of the $\ell_0$ pseudo-norm that would not only preserve the desired sparsity-inducing properties, but would also be amenable to optimization.

### 1.3.2 Sparsity-Inducing Norms and the Example of the $\ell_1$-Norm

In this section, we present one way of reformulating $\ell_0$-based problems via sparsity-inducing norms[1]. In particular, we first focus on the $\ell_1$-norm which is the most popular

1. A norm $\|.\|$ on a real-valued vector space $\mathcal{V}$ is a function $\mathcal{V} \to \mathbb{R}_+$ such that (positive homogeneity) for any $(a, \mathbf{v}) \in \mathbb{R} \times \mathcal{V}$, we have $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$, (triangular inequality) for any $\mathbf{u}, \mathbf{v} \in \mathcal{V}$, it holds $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$, and (separability of points) $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in \mathcal{V}$.

example of such reformulations, and which makes it possible to understand the key properties common to more general sparsity-inducing norms.

Regularizing by the $\ell_1$-norm has been a topic of intensive research over the last decade. This line of work has witnessed the development of nice theoretical frameworks (Tibshirani, 1996; Chen et al., 1998; Mallat, 1999; Tropp, 2004, 2006; Zhao and Yu, 2006; Zou, 2006; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009) and the emergence of many efficient algorithmic tools (Efron et al., 2004; Nesterov, 2007; Friedman et al., 2007; Wu and Lange, 2008; Beck and Teboulle, 2009; Wright et al., 2009; Needell and Tropp, 2009; Yuan et al., 2010). Moreover, various applications were considered, for instance, in compressed sensing (Candes and Tao, 2005), for the structure estimation of graphical models (Meinshausen and Bühlmann, 2006) and for several reconstruction tasks involving natural images (e.g., see Mairal, 2010, for a review).

Within the context of least-squares regression, $\ell_1$-norm regularization is known as Lasso (Tibshirani, 1996) in statistics and basis pursuit in signal processing (Chen et al., 1998). Formulation (1.1) thus reduces to

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1\right] \tag{1.2}$$

in the Lasso case, and becomes

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \left[\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\right] \tag{1.3}$$

for basis pursuit. Though similar from an optimization viewpoint, we have explicitly written down the two formulations (1.2) and (1.3) to point out the conceptual differences there exist in statistics and signal processing settings. On the one hand, we use $\mathbf{X} \in \mathbb{R}^{n\times p}$ to denote a set of $n$ observations described by $p$ variables, while $\mathbf{y} \in \mathbb{R}^n$ represents the corresponding set of $n$ targets that we try to predict. For instance, $\mathbf{y}$ may have discrete entries in the context of classification. On the other hand, and for basis pursuit, we consider a $m$-dimensional signal $\mathbf{x} \in \mathbb{R}^m$ that we express as a linear combination of $p$ dictionary elements composing the dictionary $\mathbf{D} \triangleq [\mathbf{d}^1, \ldots, \mathbf{d}^p] \in \mathbb{R}^{m\times p}$. While the matrix $\mathbf{X}$ is assumed fixed and given beforehand, we shall see in Section 1.4 that the dictionary may correspond to either some pre-defined basis (e.g., see Mallat, 1999, for wavelet basis) or to some *learned* representations (Olshausen and Field, 1996). To avoid confusion, we will try to respect as much as possible this notation throughout the manuscript.

Since the $\ell_1$-norm is convex (as any norm), it is possible to design tractable algorithms to solve (1.1) in this case. We shall review some of these methods in Section 1.5. Nonetheless, we have not justified yet why the $\ell_1$-norm is supposed to have the same sparsity-inducing behavior as the $\ell_0$ pseudo-norm. While a formal answer will be provided in Section 1.5 based on adapted optimization tools, we first give "intuitive" arguments.

### One-Dimensional Solution to Lasso: Soft-Thresholding Operator

In order to understand why the $\ell_1$-regularization promotes sparse solutions, it is convenient to consider the Lasso formulation (1.2) when the data matrix $\mathbf{X}$ is orthogonal, in which case the minimization decouples into $p$ independent one-dimensional problems of the form

$$\min_{w\in\mathbb{R}} \Big[\frac{1}{2}(u-w)^2 + \lambda|w|\Big] \quad \text{for some real number } u \in \mathbb{R}. \tag{1.4}$$

Although the function $w \mapsto \phi_{u,\lambda}(w) \triangleq \frac{1}{2}(u-w)^2 + \lambda|w|$ is convex, its minimization is not straightforward since it is non-smooth, with the presence of the absolute value. If the minimum does not happen at zero, we can take the derivative of $\phi_{u,\lambda}$ and obtain a closed-form solution for the minimizer. On the other hand, if the function $\phi_{u,\lambda}$ has its smallest value at zero, its left/right derivative should be positive to guarantee first order optimality conditions. These facts can be summarized in the following statement, that is

$$\arg\min_{w\in\mathbb{R}} \Big[\frac{1}{2}(u-w)^2 + \lambda|w|\Big] = \operatorname{sign}(u)\max\{0, |u|-\lambda\}.$$

As a matter of fact, the minimizer of (1.4) is well-known and corresponds to the *soft-thresholding* operator introduced by Donoho and Johnstone (1995). We shall discuss at greater length in Section 1.5 why this operator is important and how it constitutes a building block of efficient algorithms. It is interesting to see how the solution $\hat{w}$ varies with respect to the regularization parameter. As long as the absolute value $|u|$ is smaller than $\lambda$, the solution $\hat{w}$ remains equal to zero, and sparsity is indeed promoted. Then, when $|u|$ goes beyond the threshold $\lambda$, the minimizer becomes equal to a "shifted" version of $u$.

To understand how the $\ell_1$-norm mimics the effect of the $\ell_0$ pseudo-norm, we consider as well the following problem

$$\min_{w\in\mathbb{R}} \Big[\frac{1}{2}(u-w)^2 + \lambda \mathbb{1}_{\{w\neq 0\}}\Big] \quad , \tag{1.5}$$

where the indicator function $\mathbb{1}_{\{w\neq 0\}}$ is equal to one if $w$ is nonzero, and zero otherwise. Again, problem (1.5) can be proved to have a closed-form minimizer whose expression is given by $\hat{w} = u.\mathbb{1}_{\{|u|\geq\sqrt{2\lambda}\}}$. We display on Figure 1.1 the profiles of the minimizers for both (1.4) and (1.5).

In order to gain more insight into the sparsity-inducing behavior of the $\ell_1$-norm, we now consider some intuitive geometrical argument.

### Geometrical Intuition Through the $\ell_1$-Norm Ball

Although we consider in (1.1) a regularized formulation, we could equivalently focus on a *constrained* problem, that is,

$$\min_{\mathbf{w}\in\mathbb{R}^p} f(\mathbf{w}) \quad \text{such that} \quad \Omega(\mathbf{w}) \leq \mu, \tag{1.6}$$

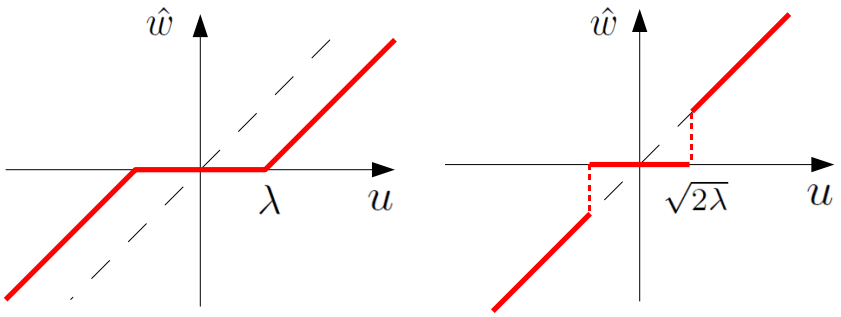



Figure 1.1: Comparison between soft- and hard-thresholding, respectively on the left and right figures.

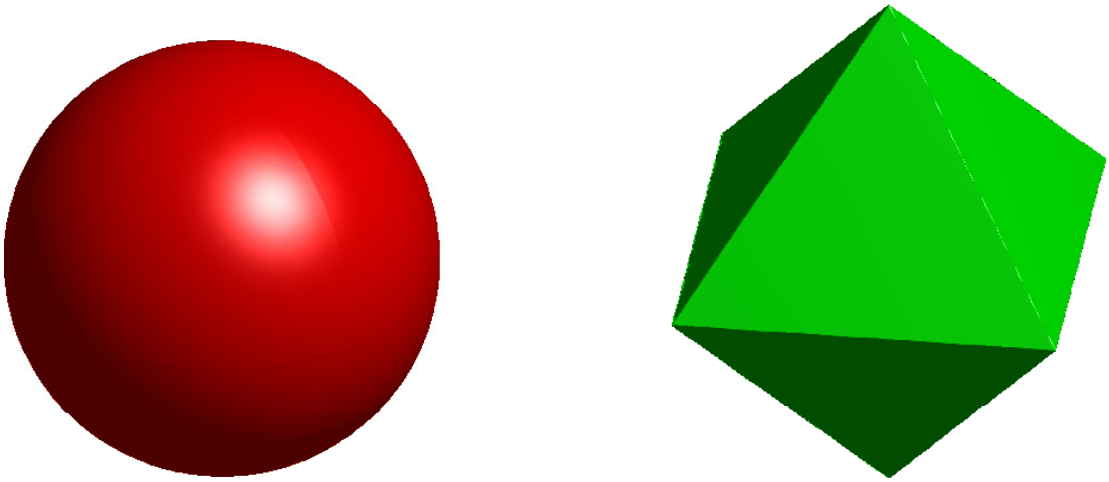
for some $\mu \in \mathbb{R}_+$, and where we have assumed for simplicity that $\mathcal{W} = \mathbb{R}^p$. The set of solutions of (1.6) parametrized by $\mu$ is the same as that of (1.1), as described by some value of $\lambda_\mu$ depending on $\mu$ (e.g., see Section 3.2 in Borwein and Lewis, 2006). At optimality, the opposite of the gradient of $f$ evaluated at any solution $\hat{\mathbf{w}}$ of (1.6) is known to belong to the normal cone to $\mathcal{B} = \{\mathbf{w} \in \mathbb{R}^p;\ \Omega(\mathbf{w}) \leq \mu\}$ at $\hat{\mathbf{w}}$ (Borwein and Lewis, 2006). In other words, for sufficiently small values of $\mu$, i.e., so that the constraint is active, the level set of $f$ for the value $f(\hat{\mathbf{w}})$ is tangent to $\mathcal{B}$.

As a consequence, the geometry of the ball $\mathcal{B}$ is directly related to the properties of the solutions $\hat{\mathbf{w}}$. If $\Omega$ is taken to be the $\ell_2$-norm, then the resulting ball $\mathcal{B}$ is the standard, isotropic, "round" ball that does not favor any specific direction of the space. On the other hand, when $\Omega$ is the $\ell_1$-norm, $\mathcal{B}$ corresponds to a diamond-shaped pattern in two dimensions, and to a pyramid in three dimensions. In particular, $\mathcal{B}$ is anisotropic and exhibits some singular points due to the non-smoothness of $\Omega$. Moreover, these singular points are located along the axis of $\mathbb{R}^p$, so that if the level set of $f$ happens to be tangent at one of those points, sparse solutions are obtained. In order to better understand why such constraints favour solutions located at vertices or some degenerate parts of the domain boundaries, we refer the interested reader to Barvinok (1995); Pataki (1998) for related results in the context of cone programming and low-rank matrices. We display on Figure 1.2 the balls $\mathcal{B}$ for both the $\ell_1$- and $\ell_2$-norms.

After having introduced the $\ell_1$-norm, we now turn to more sophisticated sparsity-inducing norms capable of encoding additional information about the data and the problem at hand.

### 1.3.3 Structured Sparsity-Inducing Norms

We have so far focused on the $\ell_1$-norm as a computationally-tractable surrogate for the $\ell_0$-pseudo-norm. The primary goal of these regularizers is to penalize dense vectors of parameters, that is, those whose *number* of nonzero coefficients is large. In other words, these regularization schemes only care about *cardinality*: they treat each variable individually and they are blind to potential relationships that may exist between the features. A simple way of justifying the latter property is to see that for any permutation



(a) $\ell_2$-norm ball (b) $\ell_1$-norm ball

Figure 1.2: Comparison between the $\ell_2$-norm and $\ell_1$-norm balls in three dimensions, respectively on the left and right figures. The $\ell_1$-norm ball presents some singular points located along the axis of $\mathbb{R}^3$.

matrix $\mathbf{P} \in \{0,1\}^{p \times p}$, we have the following invariance

$$\|\mathbf{P}\mathbf{w}\|_1 = \|\mathbf{w}\|_1 \quad \text{and} \quad \|\mathbf{P}\mathbf{w}\|_0 = \|\mathbf{w}\|_0.$$

One objective of this thesis is to come up with sparsity-inducing norms capable of encoding some additional *structure* about the variables. We shall assume this structural information available and known a priori. In addition, we will loosely speak about *structure*, without providing with a formal definition of what we mean by this term; instead, our statements will be motivated and illustrated by concrete examples.

### Sparsity-Inducing Norms with Non-Overlapping Groups of Variables

Let us introduce a first extension of the $\ell_1$-norm. Assume that we consider two-dimensional data-points of the form $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]$, described by two *categorical* variables taking values in the set $\{c_1, c_2, c_3\}$. One way of encoding these unordered categories consists of introducing an augmented space of dummy variables, so that a data-point $\mathbf{x}$ can represented as

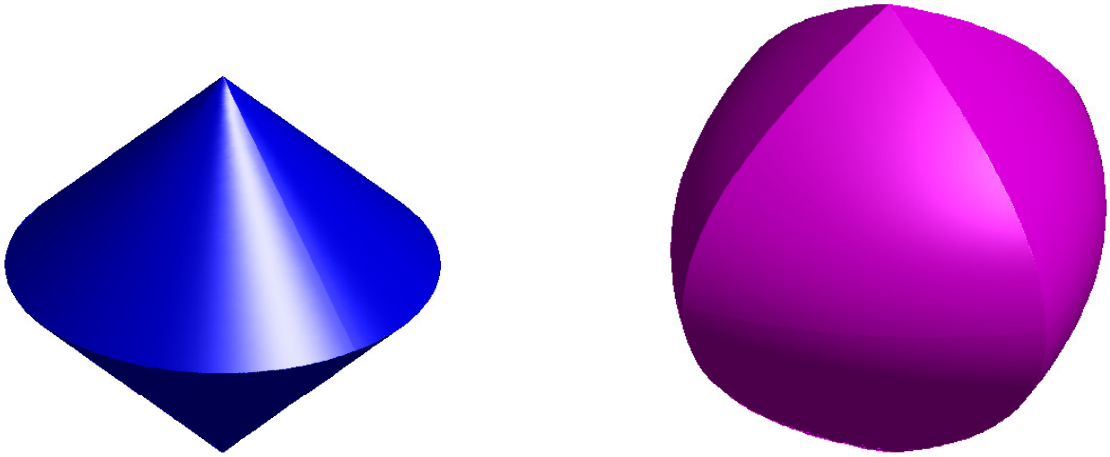
$$\mathbf{x}_{\text{aug}} = [\mathbf{x}_1^{c_1}, \mathbf{x}_1^{c_2}, \mathbf{x}_1^{c_3}, \mathbf{x}_2^{c_1}, \mathbf{x}_2^{c_2}, \mathbf{x}_2^{c_3}] \in \{0,1\}^6, \text{ with } \mathbf{x}_k^{c_l} \triangleq \begin{cases} 1 & \text{if } \mathbf{x}_k = c_l, \\ 0 & \text{otherwise.} \end{cases}$$

Now, further assume that we want to perform variable selection in a problem involving a set of such data-points. The $\ell_1$-regularization is not appropriate in this case since each dummy variable is going to be penalized independently, and as a result, the coding

scheme as triplets is likely to be lost. It is therefore appealing to keep the triplet-structure while performing variable selection. This setting is actually an example where *group sparsity* can be beneficial. Let us consider the set $\mathcal{G} = \{\{1,2,3\},\{4,5,6\}\}$ of subsets of $[\![1;6]\!]$, and introduce the norm

$$\Omega(\mathbf{w}) = \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_2 = \sqrt{\mathbf{w}_1^2 + \mathbf{w}_2^2 + \mathbf{w}_3^2} + \sqrt{\mathbf{w}_4^2 + \mathbf{w}_5^2 + \mathbf{w}_6^2}.$$

Intuitively, $\Omega$ acts as an $\ell_1$-norm on the vector $\{\|\mathbf{w}_g\|_2\}_{g\in\mathcal{G}}$. Regularizing by $\Omega$ therefore causes some $\|\mathbf{w}_g\|_2$ (and equivalently $\mathbf{w}_g$) to be zeroed out for some $g$ in $\mathcal{G}$. Conversely, within each *groups* of variables $g$ in $\mathcal{G}$, the $\ell_2$-norm does not induce sparsity. As a result, the norm $\Omega$ promotes sparsity at the level of the groups of variables $g$ in $\mathcal{G}$. Back to our example, this choice of $\Omega$ makes it possible to preserve the triplet-structure of the coding scheme while performing variable selection. We illustrate the properties of such a norm on Figure 1.3-(a) where we display the unit ball associated with $\Omega$, which is to be contrasted with the $\ell_1$-norm ball represented on Figure 1.2.



(a) $\ell_1/\ell_2$-norm ball without overlaps: $\Omega(\mathbf{w}) = \|\mathbf{w}_{\{1,2\}}\|_2 + |\mathbf{w}_3|$

(b) $\ell_1/\ell_2$-norm ball with overlaps: $\Omega(\mathbf{w}) = \|\mathbf{w}_{\{1,2,3\}}\|_2 + |\mathbf{w}_1| + |\mathbf{w}_2|$

Figure 1.3: Comparison between two mixed $\ell_1/\ell_2$-norm balls in three dimensions, without and with overlapping groups of variables, respectively on the left and right figures. The singular points appearing on these balls describe the sparsity-inducing behavior of the underlying norms $\Omega$.

We have introduced here a specific instance of *structured sparsity*—sometimes referred to as *group sparsity*—where structure is understood as non-overlapping block of variables. More generally, if $\mathcal{G}$ denotes a partition of $[\![1;p]\!]$, we can extend the definition of $\Omega$ as

$$\Omega(\mathbf{w}) = \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q \quad \text{for any } q \in (1,\infty]. \tag{1.7}$$

Thus defined, $\Omega$ is usually referred to as a mixed $\ell_1/\ell_q$-norm, and in practice, popular choices for $q$ are $\{2, \infty\}$. In the context of least-squares regression, this regularization is known as group Lasso (Turlach et al., 2005; Yuan and Lin, 2006). It has be shown to improve the prediction performance and/or interpretability of the learned models when the block structure is relevant (Roth and Fischer, 2008; Stojnic et al., 2009; Lounici et al., 2009; Huang and Zhang, 2010). Moreover, applications of this regularization scheme arose in the context of multi-task learning (Obozinski et al., 2009; Quattoni et al., 2009; Liu et al., 2009) to account for features shared across tasks, and multiple kernel learning (Bach, 2008b) for the selection of different kernels.

Specifying in advance the exact partition of the set features is in many situations too strong a requirement. We now consider a more general family of norms that notably reduces this constraint and makes it possible to encode richer structures.

### Sparsity-Inducing Norms with Overlapping Groups of Variables

Sparsity-inducing norms with *overlapping* groups of variables constitutes an essential building block of this thesis and one of its contributions. Chapter 2 is dedicated to a thorough and formal analysis of the properties of these norms; we give in this section a rather informal overview of these objects and present in which practical circumstances they might prove to be interesting.

Starting from the definition of $\Omega$ in Eq. (1.7), it is natural to study what happens when the set of groups $\mathcal{G}$ is allowed to contain elements that overlap. In fact, and as shown in Chapter 2, the sparsity-inducing behavior of $\Omega$ remains the same. As a result, when regularizing by $\Omega$, some entire groups of variables $g$ in $\mathcal{G}$ are set to zero. We illustrate the properties of such a norm on Figure 1.3-(b) where we display the unit ball associated with $\Omega$. While the resulting patterns of nonzero variables—also referred to as *supports*, or *nonzero patterns*—were obvious in the non-overlapping case, it is interesting to understand here the relationship that ties together the set of groups $\mathcal{G}$ and its associated set of possible nonzero patterns. Let us denote by $\mathcal{P}$ the latter set.

Under mild conditions, it can be proved (see Chapter 2) that given any *intersection-closed*[2] family of patterns $\mathcal{P}$ of variables, such as all the rectangles on a two-dimensional grid of variables, it is possible to build an ad-hoc set of groups $\mathcal{G}$—and hence, a regularization norm $\Omega$—that enforces the support of the solutions of (1.1) to belong to $\mathcal{P}$. Moreover, the converse is also true, meaning that given any norm of the form (1.7), we can characterize the set of possible nonzero patterns that the solutions of (1.1) can have.

These properties have important practical implications and make it possible to *design* norms that are adapted to the structure of the problem at hand; we give below interesting examples:

**One-dimensional Sequence.** Given $p$ variables organized in a sequence, if we want to select only contiguous nonzero patterns, we represent on Figure 1.4 the set of groups

2. A finite set $\mathcal{A}$ is said to be intersection-closed, if for any two elements $a_1, a_2 \in \mathcal{A}$, we have $a_1 \cap a_2 \in \mathcal{A}$. In words, this means that the the set $\mathcal{A}$ is stable with respect to the intersection.

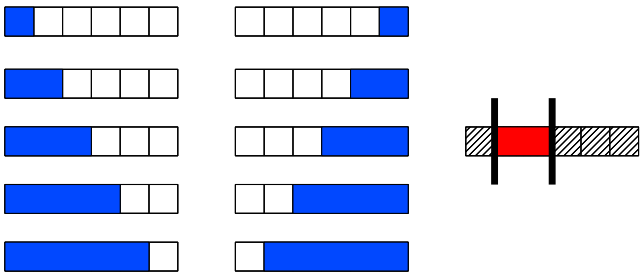


Figure 1.4: (Left) The set of blue groups to penalize in order to select contiguous patterns in a sequence. (Right) In red, an example of such a nonzero pattern with its corresponding zero pattern (hatched area).

$\mathcal{G}$ to consider. In this case, we have $|\mathcal{G}| = O(p)$. Imposing the contiguity of the nonzero patterns is for instance relevant in the context of time series, or for the diagnosis of tumors, based on the profiles of arrayCGH (Rapaport et al., 2008). Indeed, because of the specific spatial organization of bacterial artificial chromosomes along the genome, the set of discriminative features is expected to have specific contiguous patterns. Note that the sparsifying effect we obtain here is to be contrasted with that of a penalty based on total-variation (Rudin et al., 1992) plus the $\ell_1$-norm, where contiguous patterns would be encouraged but without the guarantee of selecting a *single* contiguous sequence. Moreover, the profiles of the learned coefficients will be piecewise constant in this case, a property which may not be desirable.

**Two-dimensional Grid.** In the same way, assume now the $p$ variables are organized on a two-dimensional grid. If we want the possible nonzero patterns $\mathcal{P}$ to be the set of all rectangles on this grid, the appropriate groups $\mathcal{G}$ to consider can be shown (see Chapter 2) to be those represented on Figure (1.5). In this setting, we have $|\mathcal{G}| = O(\sqrt{p})$.

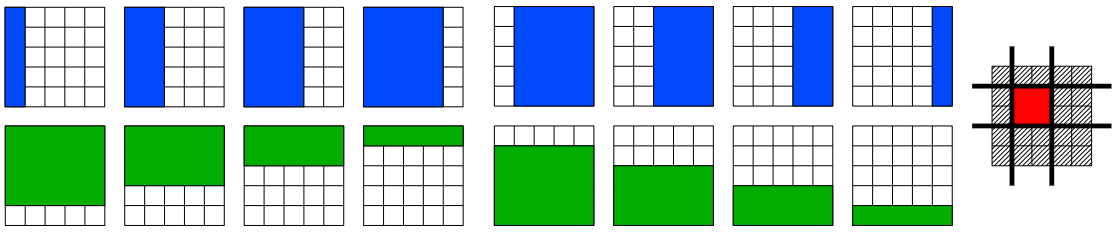


Figure 1.5: Vertical and horizontal groups: (Left) the set of blue and green groups to penalize in order to select rectangles. (Right) In red, an example of nonzero pattern recovered in this setting, with its corresponding zero pattern (hatched area).

Sparsity-inducing regularizations built upon such group structures have resulted in good performances for background subtraction (Cevher et al., 2008; Baraniuk et al., 2010;

Huang et al., 2009; Mairal et al., 2010b), topographic dictionary learning (Kavukcuoglu et al., 2009; Mairal et al., 2011), wavelet-based denoising (Rao et al., 2011), and for face recognition with corruption by occlusions (Jenatton et al., 2010b).

**Hierarchical Structure.** A third interesting example assumes that the variables have a hierarchical structure. Specifically, we consider that the $p$ variables correspond to the nodes of tree $\mathcal{T}$ (or a forest of trees). Moreover, we assume that we want to select the variables according to a certain order: a feature can be selected only if all its ancestors in $\mathcal{T}$ are already selected. This hierarchical rule can be shown to lead to the family of groups displayed on Figure 1.6.

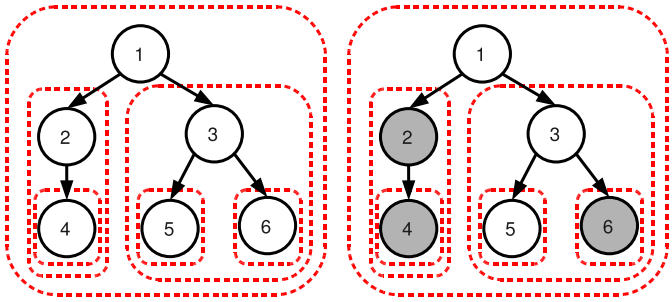



Figure 1.6: Left: example of a tree-structured set of groups $\mathcal{G}$ (dashed contours in red), corresponding to a tree $\mathcal{T}$ with $p = 6$ nodes represented by black circles. Right: example of a sparsity pattern induced by the tree-structured norm corresponding to $\mathcal{G}$: the groups $\{2, 4\}, \{4\}$ and $\{6\}$ are set to zero, so that the corresponding nodes (in gray) that form subtrees of $\mathcal{T}$ are removed. The remaining nonzero variables $\{1, 3, 5\}$ form a rooted and connected subtree of $\mathcal{T}$. This sparsity pattern obeys the following equivalent rules: (i) if a node is selected, the same goes for all its ancestors. (ii) if a node is not selected, then its descendant are not selected.

Chapter 4 largely discusses the properties of the norm $\Omega$ in this case, and presents efficient ways of optimizing the corresponding problem (1.1). This penalty was first used in Zhao et al. (2009); since then, this type of groups has led to numerous applications, for instance, wavelet-based denoising (Zhao et al., 2009; Baraniuk et al., 2010; Huang et al., 2009; Jenatton et al., 2011c), hierarchical dictionary learning for both topic modeling and image restoration (Jenatton et al., 2010a, 2011c), log-linear models for the selection of potential orders (Schmidt and Murphy, 2010), bioinformatics, to exploit the tree structure of gene networks for multi-task regression (Kim and Xing, 2010), and multi-scale mining of fMRI data for the prediction of some cognitive task (Jenatton et al., 2011b).

**Extensions.** The possible choices for the sets of groups $\mathcal{G}$ are not limited to the aforementioned examples. More complicated topologies can be considered, for instance, three-

dimensional spaces discretized in cubes or spherical volumes discretized in slices. An application to neuroimaging in Chapter 5 pursues this idea.

Before introducing more material in this introduction, we next review alternative approaches to structured sparsity that have emerged in the literature.

### Related Approaches to Structured Sparsity

We classify these parallel approaches into three categories:

**Convex Formulations.** As mentioned in the previous section, the family of norms defined in (1.7) is adapted to *intersection-closed* sets of nonzero patterns. However, some applications exhibit structures that can be more naturally modelled by *union-closed* families of supports. This idea was developed by Jacob et al. (2009) who, given a set of groups $\mathcal{G}$, introduced the following norm

$$\Omega_{\text{union}}(\mathbf{w}) \triangleq \min_{\boldsymbol{\xi} \in \mathbb{R}^{p \times |\mathcal{G}|}} \sum_{g \in \mathcal{G}} \|\boldsymbol{\xi}^g\|_2 \quad \text{such that} \quad \begin{cases} \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g = \mathbf{w}, \\ \forall g \in \mathcal{G},\ \boldsymbol{\xi}^g_j = 0 \text{ if } j \notin g. \end{cases} \tag{1.8}$$

As it will be discussed at length below, there also are non-convex formulations adapted to union-closed families of supports. We now turn to another convex way of inducing structured sparsity. Unlike the approaches previously presented, the work from Micchelli et al. (2010) does not rely on a set of groups $\mathcal{G}$. Instead, for a given nonempty convex cone $\mathcal{C} \subseteq \mathbb{R}^p$, the following norm is considered

$$\Omega_{\mathcal{C}}(\mathbf{w}) = \inf_{\mathbf{c} \in \mathcal{C}} \frac{1}{2} \sum_{j=1}^{p} \Big[ \frac{\mathbf{w}_j^2}{\mathbf{c}_j} + \mathbf{c}_j \Big].$$

The authors from Micchelli et al. (2010) show that $\Omega_{\mathcal{C}}$ penalizes less the vectors $\mathbf{w}$ which satisfy $\{|\mathbf{w}_j|\}_{j \in [\![1;p]\!]} \in \mathcal{C}$. As a result, and for judicious choices of $\mathcal{C}$, the norm $\Omega_{\mathcal{C}}$ promotes some form of structure, as encoded by $\mathcal{C}$. Examples of such structures are contiguous, or ordered, coefficients on a sequence.

**Submodular Formulations.** Another approach for structured sparsity draws connections with submodular analysis (Bach, 2010a). Starting from non-decreasing, submodular [3] set-functions $F$ of the supports of the parameter vector $\mathbf{w}$—i.e., $\mathbf{w} \mapsto F(\{j \in [\![1;p]\!];\ \mathbf{w}_j \neq 0\})$—structured sparsity-inducing norms can be built by considering the convex envelope of $F$ on the unit $\ell_\infty$-norm ball. By selecting the appropriate set-function $F$, similar structures to those described above can be obtained. This idea was further extended to symmetric, submodular set-functions of the level sets of $\mathbf{w}$, that is, $\mathbf{w} \mapsto \max_{\nu \in \mathbb{R}} F(\{j \in [\![1;p]\!];\ \mathbf{w}_j \geq \nu\})$, thus leading to different types of structures (Bach, 2010b).

3. Let $S$ be a finite set. A function $F : 2^S \to \mathbb{R}$ is said to be submodular if for any subset $A, B \subseteq S$, we have the inequality $F(A \cap B) + F(A \cup B) \leq F(A) + F(B)$; see Bach (2010a) and references therein.

**Non-convex Formulations.** We end this review of structured sparsity approaches by listing non-convex formulations. In the same flavor as the norm (1.8), Huang et al. (2009) considered the penalty

$$\psi(\mathbf{w}) \triangleq \min_{\mathcal{H}\subseteq\mathcal{G}} \sum_{g\in\mathcal{H}} \omega_g, \text{ such that } \{j \in [\![1;p]\!];\ \mathbf{w}_j \neq 0\} \subseteq \bigcup_{g\in\mathcal{H}} g,$$

where $\mathcal{G}$ is a given set of groups, and $\{\omega_g\}_{g\in\mathcal{G}}$ is a set of positive weights which defines a *coding length*. In other words, the penalty $\psi$ measures from an information-theoretic viewpoint, "how much it costs" to represent $\mathbf{w}$. A related approach was considered in Haupt and Nowak (2006). Finally, in the context of compressed sensing, the work of Baraniuk et al. (2010) also focuses on union-closed families of supports. In this case, however, the formulation does not bring into play a penalty derived from information-theoretic considerations.

## 1.4 Dictionary Learning with Structured Sparsity-Inducing Penalties

So far, we not have not discussed how we should choose the representations of the available signals and data. For instance, in the basis pursuit formulation (1.3), we have implicitly assumed that we were given a fixed dictionary $\mathbf{D}$—e.g., a basis of wavelets (Mallat, 1999)—over which it was sensible to decompose the signal $\mathbf{x}$. This section is dedicated to the introduction of matrix-factorization and dictionary-learning techniques which make it possible to *learn* a representation *adapted* to the considered class of signals.

### 1.4.1 Background Material on Matrix Factorization and Dictionary Learning

Throughout this introduction to dictionary learning, we take the notation and the viewpoint from signal processing, following the basis-pursuit formulation (1.3). Let us consider a set of $n$ signals $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{m\times n}$ described by $m$ features, typically $n$ natural image patches composed of $m$ pixels. The goal of the methods we next present consists of expressing each of these signals as linear combinations of *dictionary elements*, also known as *atoms*, taken from a *learned* dictionary.

Concretely, let us denote by $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p] \in \mathbb{R}^{m\times p}$ the dictionary and introduce $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n] \in \mathbb{R}^{p\times n}$ the matrix containing the decompositions, also referred to as *codes*, for each of the $n$ signals. With this notation in place, we try to simultaneously learn $(\mathbf{A}, \mathbf{D})$ in order to obtain

$$\mathbf{X} \approx \mathbf{D}\mathbf{A},$$

as measured by some data-fitting term. Most of the work in this thesis focuses on the square loss function [4], but some fields of applications benefit from other data-fitting

4. More precisely, since we manipulate matrices, we measure how close $\mathbf{D}\mathbf{A}$ is from $\mathbf{X}$ via the Frobenius norm, that is, $\frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2$.

terms, e.g., in audio processing (Févotte et al., 2009; Lefèvre et al., 2011a). Stated in this way, the problem of finding some interesting pair $(\mathbf{A}, \mathbf{D})$ is still very general; it is therefore appealing to reduce the space of candidate pairs $(\mathbf{A}, \mathbf{D})$ by further adding some constraints on $\mathbf{A}$ and/or $\mathbf{D}$. Those constraints should reflect some expected properties of the problem at hand. We list below some interesting and well-known examples of such formulations:

### Nonnegative Matrix-Factorization

This type of factorization is well-tailored for nonnegative signals. Nonnegative matrix-factorization (NMF) (Lee and Seung, 1999) has notably become popular thanks to its application to face images. In this context, it was observed to retrieve sets of variables that are partly localized on the face and capture some features or parts of the face which seem intuitively meaningful given our a priori (for a more detailed discussion and comparisons to structured approaches, see Chapter 3). More formally, the formulation of NMF reads

$$\min_{\mathbf{A}\in\mathbb{R}_+^{p\times n},\ \mathbf{D}\in\mathbb{R}_+^{m\times p}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2.$$

Thus, the coefficients of both $\mathbf{A}$ and $\mathbf{D}$ are forced to be nonnegative. This standard formulation was extended in many ways, see, e.g., Hoyer (2004); Cai et al. (2010).

### K-means Clustering

Another example of matrix-factorization problem is K-means (Lloyd, 1982). The objective of K-means is to find clusters and cluster centers, also referred to as centroids, starting from a set of unlabeled data-points. Once the desired number of clusters is chosen, say $p \in \mathbb{N}$, the formulation of K-means is given by

$$\min_{\substack{\mathbf{A}\in\{0,1\}^{p\times n}\\ \mathbf{D}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2, \quad \text{such that for any } i\in[\![1;n]\!],\ \|\boldsymbol{\alpha}^i\|_1 = 1.$$

The additional constraints on the columns of $\mathbf{A}$ enforce that one signal belongs to a single cluster out of the $p$ possibles ones.

The previous list of examples is of course not exhaustive, and could be further completed, e.g., by principal component analysis (Jolliffe, 1986) which is probably one of the most popular tools for data analysis and unsupervised dimensionality reduction. Moreover, and as it shall be discussed at length in Chapter 4, several probabilistic topic models (Blei et al., 2003) and other stochastic block models (Airoldi et al., 2008) can be viewed as matrix-factorization problems as well (Buntine, 2002). For further discussions, we refer the interested readers to some reviews about matrix factorization and dictionary learning, e.g., Singh and Gordon (2008); Mairal (2010); Tosic and Frossard (2011).

We now turn to the formulation of dictionary learning we adopt throughout this manuscript.

### 1.4.2 Dictionary-Learning with (Structured) Sparsity-Inducing Norms

Let us introduce two convex sets $\mathcal{A} \subseteq \mathbb{R}^{p\times n}$ and $\mathcal{D} \subseteq \mathbb{R}^{m\times p}$. We shall now focus on the following formulation

$$\min_{\mathbf{A}\in\mathcal{A},\ \mathbf{D}\in\mathcal{D}} \Big[\frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda\Omega(\mathbf{A})\Big], \tag{1.9}$$

where $\Omega$ is a sparsity-inducing regularizer applied to the matrix $\mathbf{A}$, which will usually decompose into a sum of vector-based regularizers for each of the columns (or rows) of $\mathbf{A}$. By introducing $\Omega$ in the cost function, we create a unbalance that can cause the coefficients of the matrix $\mathbf{A}$ to end up with small magnitudes; we therefore impose the set $\mathcal{D}$ to be bounded to avoid any degenerated solution. As a side comment, we know from Bach et al. (2008) that problem (1.9) has equivalent formulations, and we can go from regularized to constrained objective functions, and conversely. However, some of the resulting formulations are more suitable for optimization than others, and we shall mostly consider the form of problem (1.9).

**Optimization.** In terms of optimization, the presence of the product $\mathbf{DA}$ in problem (1.9) implies that there is unfortunately no *joint* convexity in the pair $(\mathbf{A},\mathbf{D})$. However, when one of the matrices is kept fixed, the problem is convex with respect to the other one; this property can be exploited within an alternative optimization scheme which leads to good results in practice and has become a commonly-used procedure (Olshausen and Field, 1997; Aharon et al., 2006; Lee et al., 2007; Mairal et al., 2009b). Details about the computations of these two steps—i.e., the optimization over $\mathbf{A}$ for $\mathbf{D}$ fixed, and vice-versa—are discussed in Chapter 3 and 4. Furthermore, an efficient online optimization method was recently designed for dictionary learning (Mairal et al., 2010a), making it possible to handle extremely large sets of signals.

**Sparse Dictionary Learning.** A popular instance of (1.9) is sparse coding (Olshausen and Field, 1996, 1997), where we not only want each signal to be well explained, but we also seek sparse decompositions over the dictionary; this leads to

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times n},\ \mathbf{D}\in\mathcal{D}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda\sum_{i=1}^{n}\|\boldsymbol{\alpha}^i\|_1, \tag{1.10}$$

where $\mathcal{D}$ is typically taken to be the set of dictionaries whose atoms have their $\ell_2$-norm less than, or equal to one. Sparse coding has received a great deal of attention over the past few years and has led to state-of-the-art performances in computer vision (Boureau et al., 2010), as well as in various image-processing tasks, e.g., see Peyré (2009); Mairal (2010) and numerous references therein. It is worth noting that sparse dictionary learning can alternatively be formulated as convex (Bach et al., 2008; Bradley and Bagnell, 2009b), submodular (Krause and Cevher, 2010), and non-parametric Bayesian (Zhou et al., 2009) problems.

Related to sparse coding is the problem of sparse principal component analysis (Jolliffe et al., 2003; Zou et al., 2006; Moghaddam et al., 2006; Zass and Shashua, 2007; d'Aspremont et al., 2008; Mackey, 2009; Witten et al., 2009). In this setting, the decompositions of the signals are not encouraged to be sparse anymore, but instead, the dictionary elements are forced to involve only a few variables (i.e., a small fraction of the $m$ features). For further details and discussions about the underlying optimization problems, we refer the interested readers to the review from Zhang et al. (2011).

**Sparse Structured Dictionary Learning.** The framework of dictionary learning lends itself well to structured sparsity. Indeed, by having access to the factorization $\mathbf{DA}$, it possible to encode prior knowledge in various ways: through the features (via the columns of $\mathbf{D}$), over the latent variables (through the columns of $\mathbf{A}$ and the rows of $\mathbf{D}$), and also across the different signals (this time, thanks to the rows of $\mathbf{A}$).

Moreover, there is an important subtlety which is worth being exposed. In the traditional basis pursuit setting, as presented Eq. (1.3), the dictionary is assumed fixed, and we therefore need strong prior information to be able to organize the atoms according to the considered regularizers. Although such situations do exist, e.g., see Chapter 4 with some wavelet base that naturally exhibits a hierarchical structure, they remain rather uncommon. Of course, in some cases, we can still apply an adapted pre-processing step to make sure the data exactly match the structure encoded by the regularization, e.g., see Chapter 5 where a hierarchical clustering of the features is performed beforehand.

On the contrary, in the context of dictionary learning, since the dictionary elements are learned, we can argue that the atoms *will have to match well* the prior that is imposed by the regularization. In other words, combining structured regularization with dictionary learning has precisely the advantage that the dictionary elements will *self-organize* and *self-adapt* to match the prior.

Sparse structured dictionary learning has been successfully applied to various modalities, such as, for instance, misaligned gene-expression time series (Tibau Puig et al., 2011), hierarchical topic modeling (Jenatton et al., 2010a, 2011c), the design of topographic dictionaries (Kavukcuoglu et al., 2009; Mairal et al., 2011) and localized features for face recognition (Jenatton et al., 2010b), the denoising of natural image patches (Jenatton et al., 2010a, 2011c), and speaker/instrument identification (Sprechmann et al., 2010a) as well as source separation (Lefèvre et al., 2011b).

## 1.5 Some Elements of Convex Analysis and Convex Optimization for Sparse Methods

Some parts of this section are built upon the material developed in the following book chapter:

F. Bach, R. Jenatton, J. Mairal and G. Obozinski. Convex Optimization with Sparsity-Inducing Norms. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, 2011.

The goal of this section is twofold. On the one hand, we shall recall results from convex analysis that are important to study the properties of structured sparsity-inducing norms. Most of these results are well-known in the literature, and we refer the readers interested in a more extensive coverage of this topic to classical books, such as, e.g., Borwein and Lewis (2006); Boyd and Vandenberghe (2004); Bertsekas (1999). On the other hand, we will focus on optimization techniques that are well-suited to address problem (1.1) in the context of structured sparsity-inducing norms.

### 1.5.1 Background Material of Convex Analysis

We start this section by introducing the concept of subgradient which generalizes the notion of derivative for non-smooth functions, and which is essential to describe optimality conditions. As a reminder, a function $h : \mathbb{R}^p \to \mathbb{R}$ is said to be convex if and only if (1) $\mathrm{dom}(h) \triangleq \{\mathbf{w} \in \mathbb{R}^p;\ h(\mathbf{w}) < +\infty\}$ is a convex set, and (2) for any $(\mathbf{v}, \mathbf{w}, t) \in \mathrm{dom}(h) \times \mathrm{dom}(h) \times [0,1]$, we have

$$h(t\mathbf{v} + (1-t)\mathbf{w}) \leq t\, h(\mathbf{v}) + (1-t)h(\mathbf{w}).$$

Moreover, if the domain $\mathrm{dom}(h)$ is nonempty, $h$ is said to be proper. In particular, this definition allows convex functions to take the value $+\infty$, such as the indicator function of a convex set.

#### Subgradients and Optimality conditions

Given a convex function $h : \mathbb{R}^p \to \mathbb{R}$ and a vector $\mathbf{w}$ in $\mathbb{R}^p$, let us define the set of subgradients—or simply, the *subdifferential*—of $h$ at $\mathbf{w}$ as

$$\partial h(\mathbf{w}) \triangleq \{\mathbf{z} \in \mathbb{R}^p;\ h(\mathbf{w}) + \mathbf{z}^\top(\mathbf{v} - \mathbf{w}) \leq h(\mathbf{v}) \text{ for any vector } \mathbf{v} \in \mathbb{R}^p\}. \tag{1.11}$$

The elements of $\partial h(\mathbf{w})$ are called the *subgradients* of $h$ at $\mathbf{w}$. For convex functions and for all points that lie in the interior of the domain, the subdifferential at this point is nonempty. We present interesting examples of subdifferentials that will prove important in our analysis:

- For a convex and differentiable function $f$: In this case, the subdifferential is a singleton that reduces to the gradient of $f$, that is, $\partial f(\mathbf{w}) = \{\nabla f(\mathbf{w})\}$.
- For a norm $\Omega$: A little calculation shows in this case the following equality

$$\partial\Omega(\mathbf{w}) = \begin{cases} \{\mathbf{z} \in \mathbb{R}^p;\ \Omega^*(\mathbf{z}) \leq 1\} \text{ if } \mathbf{w} = 0, \\ \{\mathbf{z} \in \mathbb{R}^p;\ \Omega^*(\mathbf{z}) \leq 1 \text{ and } \mathbf{z}^\top \mathbf{w} = \Omega(\mathbf{w})\} \text{ otherwise.} \end{cases} \tag{1.12}$$

  The previous equation brings into play an object central to this thesis, namely the *dual norm* of $\Omega$, referred to as $\Omega^*$ and which is defined by

$$\Omega^*(\mathbf{z}) \triangleq \max_{\mathbf{w} \in \mathbb{R}^p} \mathbf{z}^\top \mathbf{w} \ \text{ such that } \ \Omega(\mathbf{w}) \leq 1. \tag{1.13}$$

  The dual norm notably arises in the analysis of estimation bounds (Negahban et al., 2009), in the definition of proper stopping criteria while minimizing convex functions, and in the design of working-set strategies (Bach et al., 2011), as will be shown in Section 1.5.2. Moreover, the dual norm of $\Omega^*$ is $\Omega$ itself, and as a consequence, the formula above holds also if the roles of $\Omega$ and $\Omega^*$ are exchanged. It is easy to show that in the case of an $\ell_q$-norm, for $q \in [1; +\infty]$, the dual norm is the $\ell_{q'}$-norm, with $q'$ in $[1; +\infty]$ such that $\frac{1}{q} + \frac{1}{q'} = 1$. In particular, the $\ell_1$- and $\ell_\infty$-norms are dual to each other, and the $\ell_2$-norm is self-dual (dual to itself).

The previous list is not exhaustive, and other examples of useful subdifferentials include the indicator function of a convex set $\mathcal{C}$, whose set of subgradients is given by the normal cone to $\mathcal{C}$. We now turn to the main application of subgradients in this thesis, that is, the characterization of optimality conditions for non-smooth optimization problems:

**Proposition 1** (Subgradients at Optimality)
*For any proper convex function $h : \mathbb{R}^p \to \mathbb{R}$, a point $\mathbf{w}$ in $\mathbb{R}^p$ is a global minimum of $h$ if and only if the condition $\mathbf{0} \in \partial h(\mathbf{w})$ holds.*

Note that the concept of subdifferential is mainly useful for non-smooth functions. If $h$ is differentiable at $\mathbf{w}$, the set $\partial h(\mathbf{w})$ is indeed the singleton $\{\nabla h(\mathbf{w})\}$, and the condition $\mathbf{0} \in \partial h(\mathbf{w})$ reduces to the classical first-order optimality condition $\nabla h(\mathbf{w}) = 0$.

With these technical results in place, we can characterize the optimality conditions of problem (1.1) when the data-fitting term is assumed to be smooth and $\Omega$ is taken to be a norm. Putting together the pieces with Eq. (1.12) and Proposition 1, and applying standard convex calculus rules (Rockafellar, 1997), we have that $\mathbf{w} \in \mathbb{R}^p$ is optimal if and only if

$$-\frac{1}{\lambda} \nabla f(\mathbf{w}) \in \partial\Omega(\mathbf{w}).$$

As a consequence, the vector $\mathbf{0}$ is solution if and only if $\Omega^*\big(\nabla f(\mathbf{0})\big) \leq \lambda$. The general optimality condition above can be specified to the $\ell_1$-norm. Recalling the definition of the subdifferential of a norm in Eq. (1.12) along with the fact that the $\ell_\infty$- and $\ell_1$-norms are dual to each other, we obtain that

$$\mathbf{w} \in \arg\min_{\mathbf{v} \in \mathbb{R}^p} \Big[ f(\mathbf{v}) + \lambda \|\mathbf{v}\|_1 \Big] \Leftrightarrow \begin{cases} \big|\big[\nabla f(\mathbf{w})\big]_j\big| \leq \lambda, & \text{for all } j \in [\![1; p]\!] \text{ with } \mathbf{w}_j = 0, \\ \big[\nabla f(\mathbf{w})\big]_j = \operatorname{sign}(\mathbf{w}_j)\lambda, & \text{otherwise.} \end{cases}$$

These optimality conditions provide with a formal justification to the sparsity-inducing property of the $\ell_1$-norm: depending on the strength of the regularization parameter $\lambda$, the directions along which the variation of the data-fitting term is not large enough are kept to zero, and sparsity is indeed promoted. In the case of the square loss function, note that we get back the traditional optimality conditions for Lasso and basis pursuit (Fuchs, 2005; Wainwright, 2009).

We now introduce additional material to derive duality gaps and monitor the progress of optimization algorithms.

### Fenchel Conjugate and Duality Gaps

Let us denote by $f^*$ the Fenchel conjugate of $f$ (Rockafellar, 1997), defined by

$$f^*(\mathbf{z}) \triangleq \sup_{\mathbf{w}\in\mathbb{R}^p} [\mathbf{z}^\top \mathbf{w} - f(\mathbf{w})].$$

Regardless of the nature of the function $f$, the conjugate is always convex. Moreover, under mild conditions (convexity and closedness of the level sets of $f$), the functions $f$ and its *biconjuguate* $f^{**}$ coincide. For many functions, the conjugate is available in closed form (Borwein and Lewis, 2006), and for the square loss function, we notably have that $t \mapsto \frac{1}{2}t^2$ is self-conjugate.

In the context of this thesis, it is notably useful to specify the expression of the conjugate of a norm. Perhaps surprisingly and misleadingly, [5] the conjugate of a norm is not equal to its dual norm, but corresponds instead to the indicator function of the unit ball of its dual norm. More formally, we have the result (e.g., see Example 3.26 in Boyd and Vandenberghe, 2004):

**Proposition 2** (Fenchel conjugate of a norm)
*Let $\Omega$ be a norm on $\mathbb{R}^p$. The following equality holds for any $\mathbf{z} \in \mathbb{R}^p$*

$$\sup_{\mathbf{w}\in\mathbb{R}^p} [\mathbf{z}^\top \mathbf{w} - \Omega(\mathbf{w})] = \begin{cases} 0 & \text{if } \Omega^*(\mathbf{z}) \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

Fenchel conjugates are particularly useful to derive *duality gaps*. As a brief reminder, the duality gap of a minimization problem is defined as the difference between the primal and dual objective functions, evaluated for a feasible pair of primal/dual variables (Boyd and Vandenberghe, 2004, see Section 5.5). This gap serves as a certificate of (sub)optimality: if it is equal to zero, then the optimum is reached, and provided that strong duality holds, the converse is true as well (Boyd and Vandenberghe, 2004, see Section 5.5). Dual problems and feasible dual variables can be obtained through the following result (Borwein and Lewis, 2006, Theorem 3.3.5):

**Proposition 3** (Fenchel duality)
*If $f^*$ and $\Omega^*$ are respectively the Fenchel conjugate of a convex and differentiable function*

5. For convenience, we "overload" the notation of the superscript $^*$ and refer to the dual norm of $\Omega$ as $\Omega^*$ even though it is not equal to the conjugate of $\Omega$.

*$f$ and the dual norm of $\Omega$, then we have*

$$\max_{\mathbf{z}\in\mathbb{R}^p:\,\Omega^*(\mathbf{z})\leq\lambda} -f^*(\mathbf{z}) \quad \leq \quad \min_{\mathbf{w}\in\mathbb{R}^p} f(\mathbf{w}) + \lambda\Omega(\mathbf{w}).$$

*Moreover, equality holds as soon as the domain of $f$ has nonempty interior.*

Hence, for any feasible $\mathbf{w}, \mathbf{z}$ in $\mathbb{R}^p$, we can compute the difference between the value of the primal objective function $f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$ and the dual objective function $-f^*(\mathbf{z})$, which results in the duality gap:

$$f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) + f^*(\mathbf{z}).$$

Proposition 3 shows that it is always positive, and that it vanishes at optimality, where we then speak about *strong duality*. Duality gaps are important in convex optimization because they provide an upper bound on the difference between the current value of an objective function and the optimal value, which makes it possible to set proper stopping criteria for iterative optimization algorithms. Given a current iterate $\mathbf{w}$, computing a duality gap requires choosing a "good" candidate for $\mathbf{z}$ (and in particular a feasible one). Given that at optimality, $\mathbf{z}^\star = \nabla f(\mathbf{w}^\star)$ is the unique solution to the dual problem, a natural choice of dual variable is $\mathbf{z} = \min\left(1, \frac{\lambda}{\Omega^*(\nabla f(\mathbf{w}))}\right)\nabla f(\mathbf{w})$, which reduces to $\mathbf{z}^\star$ at the optimum and therefore yields a zero duality gap at optimality.

The next section provides with an alternative tool to deal with non-smooth convex formulations, especially adapted to the structure (1.7) of $\Omega$.

### Conic Duality and Sparsity-Inducing Norms

Because of the structure of the regularizers we study in this thesis—namely linear combinations of (generally non-smooth) norms—conic duality (see Boyd and Vandenberghe, 2004, and references therein) appears as a powerful and appealing tool. In fact, conic duality makes it possible to derive representations of (1.7) [6] that may be more amenable to optimization (see, e.g., Schmidt and Murphy, 2010; Jenatton et al., 2010a) and/or better suited for theoretical analysis (see, for instance, Bach, 2008b; Jenatton et al., 2010a).

We review here simple properties of cones and give an example of reformulation of (1.7) that is particularly useful when dealing with overlapping groups of variables (Jenatton et al., 2010a; Mairal et al., 2011).

Let us consider $\mathcal{C} \subseteq \mathbb{R}^p$ a cone, that is, a set such that if $\mathbf{c} \in \mathcal{C}$, then $t\mathbf{c} \in \mathcal{C}$ for any nonnegative scalar $t$. If the cone $\mathcal{C}$ is further assumed to be *proper*, [7] it induces a partial ordering $\preccurlyeq_{\mathcal{C}}$ on $\mathbb{R}^p$, so that

$$\mathbf{a} \preccurlyeq_{\mathcal{C}} \mathbf{b} \Leftrightarrow \mathbf{b} - \mathbf{a} \in \mathcal{C}.$$

6. In fact, using conic duality in place of classical Fenchel duality is more natural in our context. Both approaches can lead to the same dual formulations, but conic duality is usually more straightforward. For these reasons, we have made the decision to present both frameworks.

7. A cone is said to be proper if it is convex, closed, has no empty interior and does not contain any line (Boyd and Vandenberghe, 2004).

A well-known example of such ordering is the componentwise inequality between $p$-dimensional vectors built from the proper cone $\mathcal{C} = \mathbb{R}^p_+$.

Based on the cone $\mathcal{C}$, a dual counterpart can be defined in the following way:

$$\mathcal{C}^* \triangleq \{\mathbf{a} \in \mathbb{R}^p;\ \mathbf{a}^\top \mathbf{b} \geq 0 \text{ for all } \mathbf{b} \in \mathcal{C}\}.$$

Moreover, it can be shown that if $\mathcal{C}$ is proper, the same goes for the dual cone $\mathcal{C}^*$ (Boyd and Vandenberghe, 2004), which notably implies that a dual ordering $\preccurlyeq_{\mathcal{C}^*}$ can also be defined, similarly to $\preccurlyeq_{\mathcal{C}}$. Interesting properties tie these two orderings together, among which

$$[\mathbf{a} \preccurlyeq_{\mathcal{C}} \mathbf{b}] \Leftrightarrow [\mathbf{c}^\top \mathbf{a} \leq \mathbf{c}^\top \mathbf{b} \text{ for all } \mathbf{c} \succcurlyeq_{\mathcal{C}^*} \mathbf{0}]. \tag{1.14}$$

Relationship (1.14) is central to extend Lagrangian duality and Karush-Kuhn-Tucker optimality conditions to convex programs where inequality constraints are specified via generalized conic orderings, also known as *generalized inequalities*.

In the context of this thesis, it is natural to consider the following proper cone

$$\mathcal{C} \triangleq \{(\mathbf{a}, t) \in \mathbb{R}^{p+1};\ \|\mathbf{a}\| \leq t\}, \tag{1.15}$$

where $\|.\|$ denotes any norm on $\mathbb{R}^p$. Interestingly, the dual cone of $\mathcal{C}$ has a simple expression that brings into play the dual norm of $\|.\|$ (see Boyd and Vandenberghe, 2004, Example 2.25), namely

$$\mathcal{C}^* \triangleq \{(\boldsymbol{\alpha}, \tau) \in \mathbb{R}^{p+1};\ \|\boldsymbol{\alpha}\|^* \leq \tau\}. \tag{1.16}$$

For instance, when the underlying norm is chosen to be the $\ell_2$-norm, we speak about second-order cone, or "ice-cream" cone (Boyd and Vandenberghe, 2004). In addition, we have by definition of the conic orderings

$$[\|\mathbf{a}\| \leq t] \text{ if and only if } [(\mathbf{a}, t) \succcurlyeq_{\mathcal{C}} \mathbf{0}], \quad \text{and} \quad [\|\boldsymbol{\alpha}\|^* \leq \tau] \text{ if and only if } [(\boldsymbol{\alpha}, \tau) \succcurlyeq_{\mathcal{C}^*} \mathbf{0}].$$

The two previous relationships are the starting point to reformulate (1.1). If $\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|$, a first approach consists of rewriting (1.1) under (convex) conic constraints to bypass the non-smoothness of $\Omega$, that is,

$$\min_{\mathbf{w} \in \mathbb{R}^p, \mathbf{t} \in \mathbb{R}^{|\mathcal{G}|}} f(\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} \mathbf{t}^g, \text{ such that } \forall\ g \in \mathcal{G},\ (\mathbf{w}_g, \mathbf{t}^g) \succcurlyeq_{\mathcal{C}} \mathbf{0}. \tag{1.17}$$

Provided that $f$ is smooth and convex, the overall objective function is also smooth and convex, so that projected-gradient schemes (Schmidt and Murphy, 2010) can be applied. For instance, when $\|.\|$ is chosen to be the $\ell_2$-norm, projecting onto a single conic constraint can be performed efficiently with the following closed-form solution:

$$\text{Proj}_{\mathcal{C}}((\mathbf{a}, t)) = \begin{cases} (\mathbf{0}, 0) & \text{if } \|\mathbf{a}\|_2 \leq -t, \\ (\mathbf{a}, t) & \text{if } \|\mathbf{a}\|_2 \leq t, \\ \frac{1 + t/\|\mathbf{a}\|_2}{2}(\mathbf{a}, t) & \text{otherwise.} \end{cases}$$

When the set of conic constraints is separable, the situation is simple since it is sufficient to perform the individual projections sequentially. Moreover, when separability does not hold anymore, i.e., when the groups in $\mathcal{G}$ overlap, Dykstra's cyclic algorithm can be used instead (Schmidt and Murphy, 2010). Generalized inequalities can be further exploited to obtain dual formulations of (1.1) (Jenatton et al., 2010a; Mairal et al., 2011). In particular, we have the following result:

**Proposition 4** (Duality through generalized inequalities)
*Let $f$ be a proper convex function. Let us consider the maximization problem*

$$\max_{\boldsymbol{\xi} \in \mathbb{R}^{p \times |\mathcal{G}|}, \bar{\boldsymbol{\xi}} \in \mathbb{R}^p} -f^*(\bar{\boldsymbol{\xi}}) \quad \textit{such that} \quad \begin{cases} \forall\ g \in \mathcal{G}, \quad \boldsymbol{\xi}_j^g = 0 \ \textit{for}\ j \notin g, \\ \forall\ g \in \mathcal{G}, \quad (\boldsymbol{\xi}^g, \lambda) \succcurlyeq_{\mathcal{C}^*} \mathbf{0}, \\ \bar{\boldsymbol{\xi}}_j = \sum_{g \in \mathcal{G}, g \ni j} \boldsymbol{\xi}_j^g. \end{cases} \tag{1.18}$$

*Problems (1.1) and (1.18) are dual to each other and strong duality holds.*

The previous proposition turns out to be useful to derive efficient algorithms when groups in $\mathcal{G}$ are overlapping (Jenatton et al., 2010a; Mairal et al., 2011) and to make interesting connections between sparsity-inducing norms and optimization problems encountered in operations research (for more details, see Mairal et al., 2011).

We next present algorithms that are well adapted to solve problem (1.1) with structured sparsity-inducing norms.

### 1.5.2 Optimization Methods with Structured Sparsity-Inducing Norms

While the previous section was dedicated to the theoretical study of problem (1.1) with tools borrowed from convex analysis, we now present practical optimization procedures. In particular, we shall pay attention to the (if known) convergence-rate guarantees of the presented algorithms and their ability to properly handle sparsity. By the latter statement, we mean that the procedures should be able to output solutions with an identifiable set of nonzero entries, without requiring a somewhat arbitrary thresholding operation. Finally, Chapter 4 contains some speed comparisons involving the different techniques presented below.

#### Generic Approaches

The first class of methods we consider here is blind to the structure of problem (1.1). If we assume we can compute the gradient of $f$, and since we can always obtain a subgradient for $\Omega$ as defined in (1.7), we can then resort to subgradient descent to solve problem (1.1) (see, e.g., Bertsekas, 1999, and references therein). In this case, we know from Nesterov (2004) that after $k$ iterations, the cost function $f + \lambda\Omega$ will be $\varepsilon$-close to the optimum, with $\varepsilon = O(1/\sqrt{k})$. This guarantee is rather weak, and we shall see afterwards methods with faster rates. In addition, subgradient descent cannot typically handle sparsity, except when combined with other strategies, e.g., some truncation step (Langford et al., 2009).

Before reviewing other techniques, it is worth mentioning that if $f$ is chosen to be the square loss, and when $\Omega$ is a mixed $\ell_1/\ell_2$- or $\ell_1/\ell_\infty$-norm, then problem (1.1) can be cast as a second-order cone program (SOCP) and a quadratic program (QP) respectively, for which there are generic interior-point solvers that are known to be accurate but not highly scalable. As a brief reminder, a SOCP is defined as a convex optimization problem with a linear cost function, under second-order cone constraints (see Section 1.5.1) and linear equality constraints (Boyd and Vandenberghe, 2004), that is,

$$\min_{\mathbf{u}\in\mathbb{R}^p} \mathbf{u}_0^\top \mathbf{u} \quad \text{such that} \quad \begin{cases} \|\mathbf{A}_i\mathbf{u}+\mathbf{b}_i\|_2 \leq \mathbf{c}_i^\top \mathbf{u} + d_i, \text{ for } i \in [\![1;m]\!], \\ \mathbf{E}\mathbf{u} = \mathbf{f}, \end{cases}$$

where $\mathbf{u}_0 \in \mathbb{R}^p, \mathbf{A}_i \in \mathbb{R}^{n\times p}, \mathbf{b}_i \in \mathbb{R}^n, \mathbf{c}_i \in \mathbb{R}^p, d_i \in \mathbb{R}, \mathbf{E} \in \mathbb{R}^{q\times p}$ and $\mathbf{f} \in \mathbb{R}^q$ are fixed parameters defining the set of constraints.

### Reweighted-$\ell_2$ Approaches

This second class of methods deals with the non-smoothness of $\Omega$ by expressing the norm as the minimum over a set of smooth functions. At the cost of adding new variables (to describe the set of smooth functions), the problem becomes more amenable to optimization. In particular, reweighted-$\ell_2$ schemes consist of approximating the norm $\Omega$ by successive quadratic upper bounds (Argyriou et al., 2007; Rakotomamonjy et al., 2008; Jenatton et al., 2010b; Kim and Xing, 2010; Daubechies et al., 2010; Micchelli et al., 2010). Indeed, based on the inequality of arithmetic and geometric means $2\sqrt{ab} \leq a+b$, it can be shown that

$$\Omega(\mathbf{w}) = \min_{(\boldsymbol{\eta}_g)_{g\in\mathcal{G}}\in\mathbb{R}_+^{|\mathcal{G}|}} \frac{1}{2}\Big\{\sum_{g\in\mathcal{G}} \frac{\|\mathbf{w}_g\|_2^2}{\boldsymbol{\eta}_g} + \boldsymbol{\eta}_g\Big\} = \min_{\boldsymbol{\eta}\in\mathbb{R}_+^{|\mathcal{G}|}} \frac{1}{2}\Big\{\sum_{j=1}^p \Big[\sum_{\substack{g\in\mathcal{G}\\ g\ni j}} \frac{1}{\boldsymbol{\eta}_g}\Big]\mathbf{w}_j^2 + \sum_{g\in\mathcal{G}} \boldsymbol{\eta}_g\Big\}, \quad (1.19)$$

where we have assumed that $\Omega$ is a mixed $\ell_1/\ell_2$-norm. Interestingly, equality holds above if and only if $\boldsymbol{\eta}_g = \|\mathbf{w}_g\|_2$ for all $g$ in $\mathcal{G}$. Plugging the previous relationship into Eq. (1.1), the optimization can then be performed by alternating between the updates of $\mathbf{w}$ and the additional variables $(\boldsymbol{\eta}_g)_{g\in\mathcal{G}}$.[8] We shall see in Chapter 3 extensions of (1.19) to the case of more sophisticated norms. When the norm $\Omega$ is defined as a linear combination of $\ell_\infty$-norms, we are not aware of the existence of such variational formulations.

Unlike subgradient descent or proximal methods (see next section), reweighted-$\ell_2$ methods do not have simple non-asymptotic guarantees on their convergence rate (e.g., in Daubechies et al., 2010, local convergence rates are studied).

8. Note that such a scheme is interesting only if the optimization with respect to $\mathbf{w}$ is simple, which is typically the case with the square loss function (Bach et al., 2011). Moreover, for this alternating scheme to be provably convergent, the variables $(\boldsymbol{\eta}_g)_{g\in\mathcal{G}}$ have to be bounded away from zero, resulting in solutions whose entries may have small values, but not "true" zeros.

### Proximal Methods

Proximal methods constitute a class of first-order techniques well-adapted to deal with problem (1.1) (Nesterov, 2007; Beck and Teboulle, 2009; Combettes and Pesquet, 2010). In fact, and as opposed to the generic approaches presented above, proximal schemes will specifically take advantage of the structure of (1.1), namely, the sum of two convex terms where only one of these components is assumed smooth. Thus, we will typically assume that the data-fitting function $f$ is convex differentiable, with Lipschitz-continuous gradient. On the other hand, $\Omega$ is only asked to be convex, and problem (1.1) with structured sparsity-inducing norms therefore matches these requirements.

Proximal methods have become increasingly popular over the past few years, in both the signal processing (e.g., Becker et al., 2009; Wright et al., 2009; Combettes and Pesquet, 2010, and numerous references therein) and the machine learning communities (e.g., Bach et al., 2011, and references therein). In a broad sense, these methods can be seen as a natural extension of gradient-based techniques when the objective function to minimize has a non-smooth part. Proximal methods are iterative procedures. The simplest version of this class of methods linearizes at each iteration the function $f$ around the current estimate $\hat{\mathbf{w}}$, and this estimate is updated as the (unique by strong convexity) solution of the *proximal problem*, defined as follows:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[ f(\hat{\mathbf{w}}) + (\mathbf{w}-\hat{\mathbf{w}})^\top \nabla f(\hat{\mathbf{w}}) + \lambda\Omega(\mathbf{w}) + \frac{L}{2}\|\mathbf{w}-\hat{\mathbf{w}}\|_2^2 \right]. \tag{1.20}$$

A quadratic term is added in order to keep the update in a neighborhood where $f$ is close to its linear approximation, and $L>0$ is a parameter which is an upper bound on the Lipschitz constant of $\nabla f$.

Provided that we can solve efficiently the proximal problem of Eq. (1.20), this first iterative scheme constitutes a simple way of solving problem (1.1). It appears under various names in the literature: proximal-gradient techniques (Nesterov, 2007), forward-backward splitting methods (Combettes and Pesquet, 2010), and iterative shrinkage-thresholding algorithm (Beck and Teboulle, 2009). Furthermore, convergence rates of the function value $f+\lambda\Omega$ can be proved (Nesterov, 2007; Beck and Teboulle, 2009), and after $k$ iterations, the precision obtained is in the order of $O(1/k)$, which is to be contrasted with the subgradient case $O(1/\sqrt{k})$.

Interestingly, this first iterative scheme was extended to "accelerated" versions (Nesterov, 2007; Beck and Teboulle, 2009). In these extensions, the update is not taken to be exactly the result from (1.20); instead, it consists of solving the proximal problem applied to a well-chosen linear combination of the previous estimates. In this case, the function value $f+\lambda\Omega$ converges to the optimum with a rate of $O(1/k^2)$, where $k$ is the iteration number. From Nesterov (2004), we know that this rate is optimal within the class of first-order techniques; in other words, accelerated proximal-gradient methods behave as if there were no non-smooth component to handle.

Various proximal schemes can be found in the literature, depending on the considered setting (e.g., batch versus stochastic) and/or the assumptions made on the function $f$. For online/stochastic frameworks, we refer the interested readers to Duchi and

Singer (2009); Hu et al. (2009); Xiao (2010). Moreover, if we only assume that $f$ is convex, without satisfying smoothness properties, Douglas-Rachford splitting algorithm can be applied (Combettes and Pesquet, 2010). For additional work in this direction, see also Duchi and Singer (2009); Xiao (2010).

We have so far given an overview of proximal methods, without specifying how we precisely handle its main building block, namely the computation of the proximal problem, as defined in (1.20).

**Proximal Problem.** We start by equivalently rewriting problem (1.20) as

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\Big\|\mathbf{w} - (\hat{\mathbf{w}} - \frac{1}{L}\nabla f(\hat{\mathbf{w}}))\Big\|_2^2 + \frac{\lambda}{L}\Omega(\mathbf{w}).$$

Under this form, we can readily observe that when $\lambda = 0$, the solution of the proximal problem amounts to the standard gradient update rule. The problem above can be more generally viewed as an instance of the *proximal operator* (Moreau, 1962) associated with $\lambda\Omega$:

$$\mathrm{Prox}_{\lambda\Omega} : \mathbf{u} \in \mathbb{R}^p \mapsto \arg\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\Omega(\mathbf{v}).$$

This operator enjoys many attractive properties, e.g., continuity and nonexpansivity (Combettes and Wajs, 2006; Combettes and Pesquet, 2010). In particular, there is a relation that ties $\mathrm{Prox}_{\lambda\Omega}$ together with the Euclidean projection $\mathrm{Proj}_{\{\mathbf{v}\in\mathbb{R}^p;\ \Omega^*(\mathbf{v})\leq\lambda\}}$ which projects onto the ball $\{\mathbf{v}\in\mathbb{R}^p;\ \Omega^*(\mathbf{v})\leq\lambda\}$. Specifically, we have for any $\mathbf{u}\in\mathbb{R}^p$ (see Example 2.17 in Combettes and Wajs, 2006):

$$\mathrm{Prox}_{\lambda\Omega}(\mathbf{u}) = \mathbf{u} - \mathrm{Proj}_{\{\mathbf{v}\in\mathbb{R}^p;\ \Omega^*(\mathbf{v})\leq\lambda\}}(\mathbf{u}). \tag{1.21}$$

For many choices of regularizers $\Omega$, the proximal operator leads to a closed-form solution, which makes proximal methods especially efficient. If $\Omega$ is chosen to be the $\ell_1$-norm, we get back the soft-thresholding operator (1.4) applied elementwise. Similarly, if $\Omega$ is a mixed $\ell_1/\ell_2$-norm with its underlying set of groups $\mathcal{G}$ forming a partition of $[\![1;p]\!]$, we obtain the *group* soft-thresholding operator (Turlach et al., 2005):

$$\text{For any } \mathbf{u}\in\mathbb{R}^p,\ \text{for any } g\in\mathcal{G},\quad \Big[\mathrm{Prox}_{\lambda\Omega}(\mathbf{u})\Big]_g = \begin{cases} \mathbf{0} & \text{if } \|\mathbf{u}_g\|_2 \leq \lambda, \\ \frac{\|\mathbf{u}_g\|_2-\lambda}{\|\mathbf{u}_g\|_2}\mathbf{u}_g & \text{otherwise.}\end{cases}$$

In some cases, the direct computation of $\mathrm{Prox}_{\lambda\Omega}$ is not easy, while the projection $\mathrm{Proj}_{\{\mathbf{v}\in\mathbb{R}^p;\ \Omega^*(\mathbf{v})\leq\lambda\}}$ is; we can then exploit (1.21). This situation arises for instance with a mixed $\ell_1/\ell_\infty$-norm, where $\mathcal{G}$ still defines a partition of $[\![1;p]\!]$. We indeed have

$$\text{For any } \mathbf{u}\in\mathbb{R}^p,\ \text{for any } g\in\mathcal{G},\quad \Big[\mathrm{Prox}_{\lambda\Omega}(\mathbf{u})\Big]_g = \mathbf{u}_g - \mathrm{Proj}_{\{\mathbf{v}\in\mathbb{R}^{|g|};\ \|\mathbf{v}\|_1\leq\lambda\}}(\mathbf{u}_g),$$

and the Euclidean projection onto the $\ell_1$-norm ball can be performed in linear time (Brucker, 1984; Duchi et al., 2008).

As soon as groups in $\mathcal{G}$ overlap, the situation becomes difficult since no closed-forms are available, even by resorting to (1.21); Chapter 4 focuses on these settings and provides efficient algorithmic solutions. A related question of interest is whether it is possible to compute $\text{Prox}_{\lambda\Omega}$ only *approximately*, while still enjoying convergence, and possibly guaranteeing the same rates as in the approximation-free case (Schmidt et al., 2011b).

### Block-Coordinate Methods

The main idea of (block-) coordinate descent techniques is to solve problem (1.1) by sequentially optimizing with respect to one variable (or block of variables) while keeping the other ones fixed. This approach is appealing when the subproblem involving a single variable (or, a single block of variables) can be computed efficiently (either exactly or approximately). Moreover, since several cycles over the $p$ variables at play may be necessary, (block-) coordinate descent techniques can be naturally coupled to working-set strategies (see next section).

Coordinate descent is particularly suited for the Lasso (1.2) and basis pursuit (1.3) problems (Fu, 1998; Friedman et al., 2007; Wu and Lange, 2008). In this case, it can be shown that each subproblem amounts to solving a proximal operator associated with the $\ell_1$-norm, whose solution is available in closed-form, as defined by the soft-thresholding operator (1.4). Convergence in this setting is guaranteed by results from Tseng (2001). [9] In this context, the simplicity of application of coordinate descent lies in the fact that (1) the data-fitting is quadratic, and (2) the $\ell_1$-norm is *separable* with respect to the $p$ variables.

The situation is more complex when considering more general loss functions (for instance, a logistic loss function) and/or a broader family of regularizers (for example, separable mixed $\ell_1/\ell_q$-norms with $\mathcal{G}$ forming a partition of $[\![1;p]\!]$). We briefly mention the extensions from Tseng and Yun (2009); Wright (2010) whose algorithms consist in using local quadratic approximations of the data-fitting term. [10] In both papers, convex block-separable regularizers can be handled by solving a proximal problem within each block. Such a method was for instance applied in Meier et al. (2008) for logistic regression with an $\ell_1/\ell_2$-norm regularization.

For structured sparsity, we are interested in regularizers with *overlapping* groups of variables (see Section 1.3.3), in which case we are not aware of efficient block-coordinate methods. However, and as developed in Chapter 4, some dual representations of $\Omega$ can circumvent the issue of overlapping groups (see, e.g., Proposition 4, page 24).

9. Note that the result from Bertsekas (1999) could also be applied, after reformulating the regularization term in the form of *separable* convex constraints.

10. The authors from Tseng and Yun (2009) consider more general quadratic approximations (i.e., with non-diagonal Hessians) for which an inexact line-search can be conducted.

### Working-set Approaches

Working-set algorithms address optimization problems by solving an increasing sequence of small subproblems of (1.1). The working set, that we will denote by $\mathrm{J} \subseteq [\![1; p]\!]$, refers to the subset of variables involved in the optimization of these subproblems.

Working-set algorithms proceed as follows: after computing a solution to the problem restricted to the variables in J, global optimality is checked to determine whether the algorithm has to continue. If this is the case, new variables enter the working set J according to a strategy that has to be defined. Note that we only consider *forward* algorithms, i.e., where the working set grows monotonically. In other words, there are no *backward* steps where variables would be allowed to leave the set J. Provided this assumption is met, it is easy to see that these procedures stop in a finite number of iterations.

This class of algorithms takes advantage of sparsity from a computational point of view (Lee et al., 2007; Szafranski et al., 2007; Bach, 2008a; Roth and Fischer, 2008; Obozinski et al., 2009; Jenatton et al., 2011a; Schmidt and Murphy, 2010), since the subproblems that need to be solved are typically much smaller than the original one.

Working-set algorithms require three ingredients:

- **Inner-loop solver**: At each iteration of the working-set algorithm, problem (1.1) has to be solved on J, i.e., subject to the additional equality constraint that $\mathbf{w}_j = 0$ for all $j$ in $\mathrm{J}^c$:

$$\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \text{ such that } \mathbf{w}_{\mathrm{J}^c} = 0. \tag{1.22}$$

  The computation can be performed by any of the methods presented in this section. Working-set algorithms should therefore be viewed as "meta-algorithms". Since solutions for successive working sets are typically close to each other (except for the newly-active variables), the approach is efficient if the method chosen can use *warm-restarts*. Finally, even though problem (1.22) is formally defined in $\mathbb{R}^p$, in practice, adding the equality constraint on $\mathrm{J}^c$ amounts to manipulating vectors in $\mathbb{R}^{|\mathrm{J}|}$ (with hopefully $|\mathrm{J}| \ll p$).
- **Computing the optimality conditions**: Given a solution $\hat{\mathbf{w}}$ of problem (1.22), it is then necessary to check whether $\hat{\mathbf{w}}$ is also a solution for the original problem (1.1), without explicitly forcing the components in $\mathrm{J}^c$ to be equal to zero. This test relies on the duality gaps of problems (1.22) and (1.1). In particular, if $\hat{\mathbf{w}}$ is a solution of problem (1.22), it follows from Proposition 3 that

$$f(\hat{\mathbf{w}}) + \lambda \Omega(\hat{\mathbf{w}}) + f^*(\nabla f(\hat{\mathbf{w}})) = 0.$$

  In fact, the Lagrangian parameter associated with the equality constraint ensures the feasibility of the dual variable formed from the gradient of $f$ at $\hat{\mathbf{w}}$. In turn, this guarantees that the duality gap of problem (1.22) vanishes. The candidate $\hat{\mathbf{w}}$ is now a solution of the full problem (1.1), i.e., without the equality constraint, if and only if

$$\Omega^*(\nabla f(\hat{\mathbf{w}})) \leq \lambda. \tag{1.23}$$

Condition (1.23) points out that the dual norm $\Omega^*$ is a key quantity to monitor the progress of the working-set algorithm (Jenatton et al., 2011a). In simple settings, for instance when $\Omega$ is the $\ell_1$-norm, checking condition (1.23) can be easily computed since $\Omega^*$ is just the $\ell_\infty$-norm. In this case, condition (1.23) becomes

$$|[\nabla f(\hat{\mathbf{w}})]_j| \leq \lambda, \text{ for all } j \text{ in } \{1, \ldots, p\}.$$

Note that by using the optimality of problem (1.22), the components of the gradient of $f$ indexed by J are already guaranteed to be no greater than $\lambda$.
For more general sparsity-inducing norms with overlapping groups of variables, the dual norm $\Omega^*$ cannot be computed easily anymore, prompting the need for approximations and upper-bounds of $\Omega^*$ (Bach, 2008a; Jenatton et al., 2011a; Schmidt and Murphy, 2010).

- **Strategy for the growth of the working set**: If condition (1.23) is not satisfied for the current working set J, some inactive variables in $\mathrm{J}^c$ have to become active. This point raises the questions of *how many* and *how* these variables should be chosen.
  First, depending on the structure of $\Omega$, a *single* or a *group* of inactive variables have to be considered to enter the working set. Furthermore, one natural way to proceed is to look at the variables that violate condition (1.23) most. In the example of $\ell_1$-regularized least squares regression with normalized predictors, this strategy amounts to selecting the inactive variable that has the highest correlation with the current residual.

The working-set algorithms we have described so far aim at solving problem (1.1) for a fixed value of the regularization parameter $\lambda$. However, for specific types of loss and regularization functions (e.g., see Rosset and Zhu, 2007), the set of solutions of problem (1.1) can be obtained efficiently for all possible values of $\lambda$, which is exploited by homotopy algorithms (Osborne et al., 2000b) such as LARS (Efron et al., 2004).

### Extensions

Other methods could be considered for solving efficiently problem (1.1) when $\Omega$ is a structured sparsity-inducing norm. For instance, we may design quasi-Newton strategies (Schmidt et al., 2011a), or further develop augmented-Lagrangian techniques (Combettes and Pesquet, 2010; Boyd et al., 2011), as already pioneered by Mairal et al. (2011); Qin and Goldfarb (2011). In the same vein as augmented-Lagrangian techniques, there is a last extension which would be interesting to consider in the context of norms with overlapping groups of variables. Starting from the objective

$$\min_{\mathbf{w} \in \mathbb{R}^p} f(\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|,$$

for any norm $\|.\|$, we may introduce *duplicate variables* for each group $g \in \mathcal{G}$ with the constraint of being equal to $\mathbf{w}_g$, hence resulting in the new problem:

$$\min_{\mathbf{w} \in \mathbb{R}^p, \mathbf{u} \in \mathbb{R}^{s_\mathcal{G}}} f(\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{u}_g\|, \text{ such that } \mathbf{u} = \mathbf{G}\mathbf{w},$$

where $s_{\mathcal{G}} \triangleq \sum_{g\in\mathcal{G}} |g|$ and the matrix $\mathbf{G} \in \mathbb{R}^{s_{\mathcal{G}}\times p}$ encodes the group structure of $\mathcal{G}$. The nice feature of the problem above is to replace the norm with overlaps by one penalty without overlapping groups of variables. This is however achieved at the cost of (1) dealing with a larger problem (i.e., with $p+\sum_{g\in\mathcal{G}}|g|$ parameters), and (2) being able to properly handle (e.g., via projections) the new equality constraint. Noticing that $\mathbf{G}$ is full column-rank when $\mathcal{G}$ spans the set $[\![1;p]\!]$ and does not contain two identical groups, we may go one step further and try to solve the equivalent problem

$$\min_{\mathbf{u}\in\mathbb{R}^{s_{\mathcal{G}}}} f((\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{G}^\top\mathbf{u}) + \lambda\sum_{g\in\mathcal{G}}\|\mathbf{u}_g\|, \text{ such that } \mathbf{u}\in\text{span}(\mathbf{G}).$$

Interesting questions for future work would consist in studying to what extent do we need to keep the constraint $\mathbf{u}\in\text{span}(\mathbf{G})$ to go back to the domain of $\mathbf{w}$, and when the inverse of $\mathbf{G}^\top\mathbf{G}$ can be computed efficiently thanks to the structure of $\mathcal{G}$.

**Conclusions**

We briefly summarize this review of optimization techniques in the light of the difficulties related to the use of norms with overlapping groups of variables. The optimization involving $\Omega$ built from overlapping groups of variables is deemed challenging because the computations of its proximal operator and of its dual norm are both non-trivial sub-problems, with a computational cost typically in the same order as that of the full initial problem. In particular, we do not have closed-form solutions available anymore. As a result, the applications of schemes based on proximal operators (or projections) and/or tests of optimality via the dual norm (e.g., in working-set/homotopy strategies) become tricky, or at least, not straightforward. Moreover, and as explained above, block-coordinate descent methods do not apply directly to the case of overlapping groups of variables. Finally, reweighted-$\ell_2$ schemes are generally limited by the inversion of linear systems they have to repeatedly perform.

Our different simulations and experiments have shown that proximal methods appear so far as the most efficient and versatile approach. Coupling them with working-set strategies (see Section 2.4 in Chapter 2) is of course possible, but we sometimes observed that the bounds on the optimality conditions on which these strategies rely (see Propositions 7 and 8 in Chapter 2) might not be as tight as what we would need in practice.

This section concludes the introduction to the concepts from convex optimization which we require in the remainder of the thesis. We now review tools and notions to analyze some statistical properties of the solutions of problem (1.1).

## 1.6 Some Ingredients to Study Statistical Properties of Sparse Methods

Chapters 2 and 6 study some statistical properties of sparse structured linear models. This section aims at providing some background material which should be useful in this perspective.

For convenience, we adopt in the subsequent paragraphs the notations from the statistics/machine-learning setting, similarly to the Lasso problem in Eq. (1.2).

### 1.6.1 Several Criteria and Measures of Quality

From now on, let us concentrate on linear models. We assume that the output, or response, $y$ is generated from the linear combination of a $p$-dimensional observation $\mathbf{x} \in \mathbb{R}^p$ with a *true* model $\mathbf{w}^\star \in \mathbb{R}^p$. In addition, we shall assume some corruption from noise, in the form of, typically sub-Gaussian, centered random variable $\varepsilon$. In other words, we have

$$y = \mathbf{x}^\top \mathbf{w}^\star + \varepsilon. \tag{1.24}$$

Stated in this way, the only source of randomness corresponds to the noise, and we then speak about *fixed-design* setting. Random-design analysis (e.g, Wainwright, 2009, in the case of Lasso) can also be conducted by adding some probabilistic assumptions on the observations.

We will assume that the true model $\mathbf{w}^\star$ has a sparse, structured nonzero pattern $\mathrm{J}^\star = \{j \in [\![1; p]\!];\ \mathbf{w}^\star_j \neq 0\}$, with $|\mathrm{J}^\star| \ll p$. As an estimator for $\mathbf{w}^\star$, it is thus natural to take any solution of problem (1.1), with $\Omega$ being a well-chosen structured sparsity-inducing norm. Let us denote by $\hat{\mathbf{w}}$ one of these solutions.[11] There are now several criteria of interest (for more details and formal definitions, see, e.g., Liu, 2010):

- **Estimation performance:** For this first criterion, we are interested in how far $\hat{\mathbf{w}}$ is from the true model, that is, how large $\|\hat{\mathbf{w}} - \mathbf{w}^\star\|$ is for some norm $\|.\|$.
- **Prediction performance:** In this case, we might well be far from the true model $\mathbf{w}^\star$; instead, we want predictions to be accurate, that is, the quantity $\|\mathbf{X}(\hat{\mathbf{w}} - \mathbf{w}^\star)\|$ to be small for some norm $\|.\|$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ stands for a set of $n$ observations in $\mathbb{R}^p$ stacked row-wise.
- **Model-selection consistency:** This last criterion, also referred to as *support consistency* or *sparsistency*, focuses on recovering the true sparsity pattern $\mathrm{J}^\star$ of $\mathbf{w}^\star$. In other words, we want to have

  $$\Pr\Big(\{j \in [\![1; p]\!];\ \hat{\mathbf{w}}_j \neq 0\} = \mathrm{J}^\star\Big) \approx 1.$$

  The probability above has to be understood as being taken with respect to all sources of randomness in the problem.

In this thesis, we mostly study the third criterion. Furthermore, we ideally wish to obtain a *non-asymptotic* characterization of this criterion in order to better understand the roles and the scaling of $(n, p, |\mathrm{J}^\star|)$. Finally, even tough it has not been listed above, *stability* (e.g., of the estimated nonzero pattern) is another important measure. It has lately been the focus of a great deal of research, see, e.g., Bach (2008c); Meinshausen and Bühlmann (2010); Liu et al. (2010b).

We now describe an important structure of proof which is recurring in the study of sparse models where estimators result from a convex program.

11. The problem of identifiability (and uniqueness) is further discussed in Chapter 2. Usually, the assumptions made on the observations lead to a unique estimator $\hat{\mathbf{w}}$.

### 1.6.2 The Primal-Dual "Witness" Proof

In this section, we present a method to characterize model-selection consistency (i.e., the third criterion above) when the estimator is derived from a convex program. Our situation is exactly in this scope, since we consider any solution of problem (1.1) for some structured sparsity-inducing norm.

This scheme of proof appeared in many places, for instance, to study the structures of graphical models (Wainwright et al., 2007; Ravikumar et al., 2008; Jalali et al., 2011) and to prove exact support recovery of Lasso (Wainwright, 2009), group Lasso (Bach, 2008b; Nardi and Rinaldo, 2008) and other sparse estimators (Jenatton et al., 2011a).

We now describe the different steps of the primal-dual "witness" proof:

- **Construction of an "oracle" solution:** This first step consists of building a primal candidate starting from the knowledge of $\mathrm{J}^\star$. More precisely, we consider

$$\hat{\mathbf{w}}_{\text{oracle}} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^p} \Big[ f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) \Big], \quad \text{such that } \mathbf{w}_j = 0 \text{ for any } j \in [\mathrm{J}^\star]^c, \quad (1.25)$$

  where we have assumed that the solution $\hat{\mathbf{w}}_{\text{oracle}}$ is unique (typically through assumptions on the observations and choice of $\lambda$). By construction, we know that $\hat{\mathbf{w}}_{\text{oracle}}$ has the correct sparsity pattern; we will have to prove that $\hat{\mathbf{w}}_{\text{oracle}}$ is also solution to the full problem, without the "oracle" equality constraint. In the Lasso case, problem (1.25) is equivalent to

$$\min_{\mathbf{w}_{\mathrm{J}^\star}\in\mathbb{R}^{|\mathrm{J}^\star|}} \Big[ \frac{1}{2}\|\mathbf{y} - \mathbf{X}_{\mathrm{J}^\star}\mathbf{w}_{\mathrm{J}^\star}\|_2^2 + \lambda\|\mathbf{w}_{\mathrm{J}^\star}\|_1 \Big].$$

- **Defining a candidate for the dual variable:** Let us denote by $\hat{\mathbf{w}}$ a solution of problem (1.25) without the equality constraint. We know from Proposition 1 that we can characterize through a subgradient condition the optimality of $\hat{\mathbf{w}}$. We therefore need to construct a subgradient candidate $\mathbf{z}$ whose entries in $\mathrm{J}^\star$ are in agreement with the optimality of $\hat{\mathbf{w}}_{\text{oracle}}$.
  By separately treating $\mathbf{z}_{\mathrm{J}^\star}$ and $\mathbf{z}_{[\mathrm{J}^\star]^c}$, we implicitly assume that the regularizer $\Omega$ and its subdifferential can be decomposed in this way; this is indeed the case for our structured sparsity-inducing norms (see Chapter 2). Generally, we will be able to express (explicitly or implicitly) $\mathbf{z}_{[\mathrm{J}^\star]^c}$ with respect to $\mathbf{z}_{\mathrm{J}^\star}$.
- **Check the validity of the candidate:** The third step of the procedure checks whether $\hat{\mathbf{w}}_{\text{oracle}}$ is optimal for the full problem, using the subgradient $\mathbf{z}$ previously defined as a certificate. In particular, we need to guarantee that $\mathbf{z} \in \partial\lambda\Omega(\hat{\mathbf{w}}_{\text{oracle}})$.
- **Probabilistic control:** While the reasoning in the previous steps was *deterministic*, it remains to define and control the probabilistic events over which the construction and the condition $\mathbf{z} \in \partial\lambda\Omega(\hat{\mathbf{w}}_{\text{oracle}})$ hold with high probability.

This technique of proof has to be related to working-set algorithms (see Section 1.5.2) in that they share a similar construction and use the same tools from convex analysis.

### 1.6.3 Zoology of Conditions Imposed on the Design Matrix

In the statistical analysis of the sparse models we are interested in,[12] theoretical guarantees—for prediction, estimation or support recovery—require some assumptions on the correlations of the $p$ variables. Since Chapters 2 and 6 bring into play such assumptions, it is worth reviewing and presenting these concepts. For simplicity, we consider the $\ell_1$-norm setting, but the quantities we describe below can be extended to structured scenarios (Negahban et al., 2009).

In the case of noiseless linear models, i.e., Eq. (1.24) with $\varepsilon = 0$, it is interesting to study when the true vector $\mathbf{w}^\star$ is exactly recovered by a procedure based on $\ell_1$-norm minimization. For a sparsity level of $|\mathrm{J}^\star|$, a necessary and sufficient condition for exact recovery is known as the *null space property* with parameter $|\mathrm{J}^\star|$ (Donoho and Huo, 2001; Feuer and Nemirovski, 2003; Cohen et al., 2009). As soon as the data are corrupted by some noise, exact recovery becomes impossible and it is then sensible to focus on the estimation error $\|\hat{\mathbf{w}} - \mathbf{w}^\star\|_2$. We thus analyze under which sufficient condition of the null space property, this error can be proved to be small.

We review below several sufficient conditions, starting from the most restrictive one (but also the simplest one). We notably illustrate these conditions by their *sampling complexity* in the case of random Gaussian matrices with i.i.d. entries, that is, under which scaling of $(n, p, |\mathrm{J}^\star|)$ the conditions hold with high probability. In the subsequent, we use the notation $a \gtrsim b$ if there exists some constant $K > 0$ such that $a \geq Kb$ holds.[13] Moreover, we shall assume that the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ composed of $n$ observations stacked row-wise has its columns normalized with unit $\ell_2$-norm.

- **Elementwise incoherence condition:** This first sufficient condition is rather intuitive, and explicitly asks for a control of the off-diagonal terms of the covariance matrix $\mathbf{X}^\top \mathbf{X}$ (Donoho and Huo, 2001; Feuer and Nemirovski, 2003; Lounici, 2008), that is,

$$\max_{j,k \in [\![1;p]\!], j \neq k} \Big| \frac{1}{n} [\mathbf{X}^\top \mathbf{X}]_{jk} \Big| \precsim \frac{1}{|\mathrm{J}^\star|}.$$

  For random Gaussian matrices, this inequality is known to be satisfied with high probability for the following scaling $n \gtrsim |\mathrm{J}^\star|^2 \log(p)$ (e.g., see Candès and Plan, 2009).

- **Restricted isometry:** The introduction of this second condition is slightly more involved. For $s \in \mathbb{N}$, we introduce the $s$-restricted isometry constant $\delta_s > 0$ (Candes and Tao, 2005) as the smallest scalar such that

$$(1 - \delta_s)\|\mathbf{a}\|_2^2 \leq \|\mathbf{X}\mathbf{a}\|_2^2 \leq (1 + \delta_s)\|\mathbf{a}\|_2^2,$$

  hold for any vector $\mathbf{a} \in \mathbb{R}^p$ with $|\{j \in [\![1;p]\!];\ \mathbf{a}_j \neq 0\}| \leq s$. In words, the previous equation says that all sub-matrices involving at most $s$ columns of $\mathbf{X}$ behave like

12. We do not discuss the body of work that do not assume any computational limit, for instance, aggregated estimators (e.g., see Section 3.5 in Audibert, 2010, and references therein). In this case, no correlation-based assumptions are needed.

13. The usual notation is $a = \Omega(b)$, but we keep this symbol to represent the regularization term.

an isometry. It can be shown (Candes and Tao, 2005; Baraniuk et al., 2008) that if the following inequality is valid

$$\delta_s + \delta_{2s} + \delta_{3s} < 1,$$

then, the null space property with parameter $s$ is also satisfied. Moreover, for random Gaussian matrices, we obtain the scaling $n \gtrsim |\mathrm{J}^\star| \log(p/|\mathrm{J}^\star|)$ (Mendelson et al., 2008), where we can note the improvement with respect to the incoherence condition.

- **Restricted eigenvalue:** The last sufficient condition we mention is less restrictive than those presented above. It characterizes a lower bound of the spectrum of $\mathbf{X}$ over a specific subset of vectors. It was introduced in Bickel et al. (2009), and further studied in Van de Geer and Bühlmann (2009); Raskutti et al. (2010); Zhou (2009). Formally, there exist $(\gamma, \kappa)$ strictly positive scalars such that

$$\frac{1}{n}\|\mathbf{X}\mathbf{a}\|_2^2 \geq \gamma^2 \|\mathbf{a}\|_2^2 \quad \text{for any vector } \mathbf{a} \in \mathbb{R}^p \text{ with } \|\mathbf{a}_{\mathrm{J}^\star}\|_1 \leq \kappa \|\mathbf{a}_{[\mathrm{J}^\star]^c}\|_1.$$

The scaling for random Gaussian matrices was recently proved by Zhou (2009) [14] and leads to $n \gtrsim |\mathrm{J}^\star| \log(p/|\mathrm{J}^\star|)$. Again, there is an improvement upon the incoherence condition.

Before concluding this section, there is another correlation condition worth being discussed. The *irrepresentability condition* (Fuchs, 2005; Zhao and Yu, 2006; Tropp, 2006; Wainwright, 2009) is central when the criterion of interest is the model-selection consistency, that is, the ability of the estimator to retrieve the nonzero pattern of $\mathbf{w}^\star$. This is to be contrasted with the previous conditions which are primarily introduced to control the estimation error. Moreover, the irrepresentability condition is tight since it is proved to be necessary and sufficient (Zhao and Yu, 2006; Wainwright, 2009). Formally, it can be stated as follows:

$$\left\| [\mathbf{X}^\top \mathbf{X}]_{[\mathrm{J}^\star]^c \mathrm{J}^\star} \left( [\mathbf{X}^\top \mathbf{X}]_{\mathrm{J}^\star \mathrm{J}^\star} \right)^{-1} \mathrm{sign}(\mathbf{w}^\star)_{\mathrm{J}^\star} \right\|_\infty < 1.$$

As opposed to the previously-described criteria that depend only on $|\mathrm{J}^\star|$, the condition above explicitly brings into play the sign vector $\mathrm{sign}(\mathbf{w}^\star)$ and the set $\mathrm{J}^\star$. At the cost of being more restrictive, we can further maximize out over all possible sign vectors and supports of size $|\mathrm{J}^\star|$ (Tropp, 2006); the resulting criterion is shown to be weaker than the one based on restricted eigenvalue (Van de Geer and Bühlmann, 2009). Eventually, we discuss in Chapter 2 how the irrepresentability condition is modified to adapt to structured settings.

Thus far, we have not taken into account any computational aspect. While it is relevant to look at ensembles of random matrices in the context of compressed sensing, the situation is different in statistics/machine-learning where the design matrix $\mathbf{X}$ is fixed. It would therefore be interesting to be able to *compute* some tests of these sufficient

14. Simultaneously, the authors from Raskutti et al. (2010) came up with the weaker scaling that imposes $n \gtrsim |\mathrm{J}^\star| \log(p)$.

conditions: Some works recently investigated this direction of research, e.g., d'Aspremont et al. (2008); Juditski and Nemirovski (2010).

We refer the readers to Van de Geer and Bühlmann (2009) for a complete comparison of various assumptions made in the analysis of the Lasso.

# 2

# Understanding the Properties of Structured Sparsity-Inducing Norms

**Chapter abstract:** We consider the empirical risk minimization problem for linear supervised learning, with regularization by structured sparsity-inducing norms. These are defined as sums of Euclidean norms on certain subsets of variables, extending the usual $\ell_1$-norm and the group $\ell_1$-norm by allowing the subsets to overlap. This leads to a specific set of allowed nonzero patterns for the solutions of such problems. We first explore the relationship between the groups defining the norm and the resulting nonzero patterns, providing both forward and backward algorithms to go back and forth from groups to patterns. This allows the design of norms adapted to specific prior knowledge expressed in terms of nonzero patterns. We also present an efficient active set algorithm, and analyze the consistency of variable selection for least-squares linear regression in low and high-dimensional settings.

The material of this chapter is based on the following work:

R. Jenatton, J.-Yves Audibert, F. Bach. Structured Variable Selection with Sparsity-Inducing Norms. In *Journal of Machine Learning Research*, 12, 2777-2824. 2011

## 2.1 Introduction

Sparse linear models have emerged as a powerful framework to deal with various supervised estimation tasks, in machine learning as well as in statistics and signal processing. These models basically seek to predict an output by linearly combining only a small subset of the features describing the data. To simultaneously address this variable selection and the linear model estimation, $\ell_1$-norm regularization has become a popular tool, that benefits both from efficient algorithms (see, e.g., Efron et al., 2004; Lee et al., 2007; Beck and Teboulle, 2009; Yuan et al., 2010, and multiple references therein) and well-developed theory for generalization properties and variable selection consistency (Zhao and Yu, 2006; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009).

When regularizing by the $\ell_1$-norm, sparsity is yielded by treating each variable individually, regardless of its position in the input feature vector, so that existing relationships and structures between the variables (e.g., spatial, hierarchical or related to the

physics of the problem at hand) are merely disregarded. However, many practical situations could benefit from this type of prior knowledge, potentially both for interpretability purposes and for improved predictive performance.

For instance, in neuroimaging, one is interested in localizing areas in functional magnetic resonance imaging (fMRI) or magnetoencephalography (MEG) signals that are discriminative to distinguish between different brain states (Gramfort and Kowalski, 2009; Xiang et al., 2009, and references therein). More precisely, fMRI responses consist in voxels whose three-dimensional spatial arrangement respects the anatomy of the brain. The discriminative voxels are thus expected to have a specific localized spatial organization (Xiang et al., 2009), which is important for the subsequent identification task performed by neuroscientists. In this case, regularizing by a plain $\ell_1$-norm to deal with the ill-conditionedness of the problem (typically only a few fMRI responses described by tens of thousands of voxels) would ignore this spatial configuration, with a potential loss in interpretability and performance.

Similarly, in face recognition, robustness to occlusions can be increased by considering as features, sets of pixels that form small convex regions on the face images (Jenatton et al., 2010b). Again, a plain $\ell_1$-norm regularization fails to encode this specific spatial locality constraint (Jenatton et al., 2010b). The same rationale supports the use of *structured sparsity* for background subtraction tasks (Cevher et al., 2008; Huang et al., 2009; Mairal et al., 2010b). Still in computer vision, object and scene recognition generally seek to extract bounding boxes in either images (Harzallah et al., 2009) or videos (Dalal et al., 2006). These boxes concentrate the predictive power associated with the considered object/scene class, and have to be found by respecting the spatial arrangement of the pixels over the images. In videos, where series of frames are studied over time, the temporal coherence also has to be taken into account. An unstructured sparsity-inducing penalty that would disregard this spatial and temporal information is therefore not adapted to select such boxes.

Another example of the need for higher-order prior knowledge comes from bioinformatics. Indeed, for the diagnosis of tumors, the profiles of array-based comparative genomic hybridization (arrayCGH) can be used as inputs to feed a classifier (Rapaport et al., 2008). These profiles are characterized by plenty of variables, but only a few samples of such profiles are available, prompting the need for variable selection. Because of the specific spatial organization of bacterial artificial chromosomes along the genome, the set of discriminative features is expected to have specific contiguous patterns. Using this prior knowledge on top of a standard sparsity-inducing method leads to improvement in classification accuracy (Rapaport et al., 2008). In the context of multi-task regression, a genetic problem of interest is to find a mapping between a small subset of single nucleotide polymorphisms (SNP's) that have a phenotypic impact on a given family of genes (Kim and Xing, 2010). This target family of genes has its own structure, where some genes share common genetic characteristics, so that these genes can be embedded into a underlying hierarchy (Kim and Xing, 2010). Exploiting directly this hierarchical information in the regularization term outperforms the unstructured approach with a standard $\ell_1$-norm. Such hierarchical structures have been likewise useful in the context

of wavelet regression (Baraniuk et al., 2010; Zhao et al., 2009; Huang et al., 2009; Jenatton et al., 2011c), kernel-based non linear variable selection (Bach, 2008a) and for topic modeling (Jenatton et al., 2011c).

These real world examples motivate the need for the design of sparsity-inducing regularization schemes, capable of encoding more sophisticated prior knowledge about the expected sparsity patterns.

As mentioned above, the $\ell_1$-norm focuses only on *cardinality* and cannot easily specify side information about the patterns of nonzero coefficients ("nonzero patterns") induced in the solution, since they are all theoretically possible. Group $\ell_1$-norms (Yuan and Lin, 2006; Roth and Fischer, 2008; Huang and Zhang, 2010) consider a partition of all variables into a certain number of subsets and penalize the sum of the Euclidean norms of each one, leading to selection of groups rather than individual variables. Moreover, recent works have considered overlapping but nested groups in constrained situations such as trees and directed acyclic graphs (Zhao et al., 2009; Bach, 2008a; Kim and Xing, 2010; Jenatton et al., 2010a, 2011c; Schmidt and Murphy, 2010).

In this chapter, we consider all possible sets of groups and characterize exactly what type of prior knowledge can be encoded by considering sums of norms of overlapping groups of variables. Before describing how to go from groups to nonzero patterns (or equivalently zero patterns), we show that it is possible to "reverse-engineer" a given set of nonzero patterns, i.e., to build the unique minimal set of groups that will generate these patterns. This allows the automatic design of sparsity-inducing norms, adapted to target sparsity patterns. We give in Section 2.3 some interesting examples of such designs in specific geometric and structured configurations, which covers the type of prior knowledge available in the real world applications described previously.

As will be shown in Section 2.3, for each set of groups, a notion of hull of a nonzero pattern may be naturally defined. For example, in the particular case of the two-dimensional planar grid considered in this chapter, this hull is exactly the axis-aligned bounding box or the regular convex hull. We show that, in our framework, the allowed nonzero patterns are exactly those equal to their hull, and that the hull of the relevant variables is consistently estimated under certain conditions, both in low and high-dimensional settings. Moreover, we present in Section 2.4 an efficient active set algorithm that scales well to high dimensions. Finally, we illustrate in Section 2.6 the behavior of our norms with synthetic examples on specific geometric settings, such as lines and two-dimensional grids.

**Notation.** For any finite set $A$ with cardinality $|A|$, we also define the $|A|$-tuple $(\mathbf{y}^a)_{a\in A} \in \mathbb{R}^{p\times|A|}$ as the collection of $p$-dimensional vectors $\mathbf{y}^a$ indexed by the elements of $A$. Furthermore, for two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^p$, we denote by $\mathbf{x}\circ\mathbf{y} = (\mathbf{x}_1\mathbf{y}_1, \ldots, \mathbf{x}_p\mathbf{y}_p)^\top \in \mathbb{R}^p$ the elementwise product of $\mathbf{x}$ and $\mathbf{y}$.

## 2.2 Regularized Risk Minimization

We consider the problem of predicting a random variable $Y \in \mathcal{Y}$ from a (potentially non random) vector $\mathbf{x} \in \mathbb{R}^p$, where $\mathcal{Y}$ is the set of responses, typically a subset of $\mathbb{R}$. We assume that we are given $n$ observations $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$, for $i \in [\![1; n]\!]$. We define the *empirical risk* of a loading vector $\mathbf{w} \in \mathbb{R}^p$ as $L(w) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \mathbf{w}^\top \mathbf{x}_i)$, where $\ell : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}_+$ is a *loss function*. We assume that $\ell$ is *convex and continuously differentiable* with respect to the second parameter. Typical examples of loss functions are the square loss for least squares regression, i.e., $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ with $y \in \mathbb{R}$, and the logistic loss $\ell(y, \hat{y}) = \log(1 + e^{-y\hat{y}})$ for logistic regression, with $y \in \{-1, 1\}$.

We focus on a general family of sparsity-inducing norms that allow the penalization of subsets of variables grouped together. Let us denote by $\mathcal{G}$ a subset of the power set of $[\![1; p]\!]$ such that $\bigcup_{g \in \mathcal{G}} g = [\![1; p]\!]$, i.e., a spanning set of subsets of $[\![1; p]\!]$. Note that $\mathcal{G}$ does not necessarily define a partition of $[\![1; p]\!]$, and therefore, *it is possible for elements of* $\mathcal{G}$ *to overlap*. We consider the norm $\Omega$ defined by

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \left( \sum_{j \in g} (\omega_j^g)^2 |\mathbf{w}_j|^2 \right)^{\frac{1}{2}} = \sum_{g \in \mathcal{G}} \|\omega^g \circ w\|_2, \tag{2.1}$$

where $(\omega^g)_{g \in \mathcal{G}}$ is a $|\mathcal{G}|$-tuple of $p$-dimensional vectors such that $\omega_j^g > 0$ if $j \in g$ and $\omega_j^g = 0$ otherwise. A same variable $\mathbf{w}_j$ belonging to two different groups $g_1, g_2 \in \mathcal{G}$ is allowed to be weighted differently in $g_1$ and $g_2$ (by respectively $\omega_j^{g_1}$ and $\omega_j^{g_2}$). We do not study the more general setting where each $\omega^g$ would be a (non-diagonal) positive-definite matrix, which we defer to future work. Note that a larger family of penalties with similar properties may be obtained by replacing the $\ell_2$-norm in Eq. (2.1) by other $\ell_q$-norm, $q > 1$ (Zhao et al., 2009). Moreover, non-convex alternatives to Eq. (2.1) with quasi-norms in place of norms may also be interesting, in order to yield sparsity more aggressively (see, e.g., Jenatton et al., 2010b).

This general formulation has several important sub-cases that we present below, the goal of this chapter being to go beyond these, and to consider norms capable to incorporate richer prior knowledge.

- **$\ell_2$-norm**: $\mathcal{G}$ is composed of one element, the full set $[\![1; p]\!]$.
- **$\ell_1$-norm**: $\mathcal{G}$ is the set of all singletons, leading to the Lasso (Tibshirani, 1996) for the square loss.
- **$\ell_2$-norm and $\ell_1$-norm**: $\mathcal{G}$ is the set of all singletons and the full set $[\![1; p]\!]$, leading (up to the squaring of the $\ell_2$-norm) to the Elastic net (Zou and Hastie, 2005) for the square loss.
- **Group $\ell_1$-norm**: $\mathcal{G}$ is any partition of $[\![1; p]\!]$, leading to the group-Lasso for the square loss (Yuan and Lin, 2006).
- **Hierarchical norms**: when the set $[\![1; p]\!]$ is embedded into a tree (Zhao et al., 2009) or more generally into a directed acyclic graph (Bach, 2008a), then a set of $p$ groups, each of them composed of descendants of a given variable, is considered.

We study the following regularized problem:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) + \mu\Omega(\mathbf{w}), \tag{2.2}$$

where $\mu \geq 0$ is a regularization parameter. Note that a non-regularized constant term could be included in this formulation, but it is left out for simplicity. We denote by $\hat{\mathbf{w}}$ any solution of Eq. 2.2. Regularizing by linear combinations of (non-squared) $\ell_2$-norms is known to induce sparsity in $\hat{\mathbf{w}}$ (Zhao et al., 2009); our grouping leads to specific patterns that we describe in the next section.

## 2.3 Groups and Sparsity Patterns

We now study the relationship between the norm $\Omega$ defined in 2.1 and the nonzero patterns the estimated vector $\hat{\mathbf{w}}$ is allowed to have. We first characterize the set of nonzero patterns, then we provide forward and backward procedures to go back and forth from groups to patterns.

### 2.3.1 Stable Patterns Generated by $\mathcal{G}$

The regularization term $\Omega(\mathbf{w}) = \sum_{g\in\mathcal{G}} \|\omega^g \circ \mathbf{w}\|_2$ is a mixed $(\ell_1, \ell_2)$-norm (Zhao et al., 2009). At the group level, it behaves like an $\ell_1$-norm and therefore, $\Omega$ induces group sparsity. In other words, each $\omega^g \circ \mathbf{w}$, and equivalently each $\mathbf{w}_g$ (since the support of $\omega^g$ is exactly $g$), is encouraged to go to zero. On the other hand, within the groups $g \in \mathcal{G}$, the $\ell_2$-norm does not promote sparsity. Intuitively, for a certain subset of groups $\mathcal{G}' \subseteq \mathcal{G}$, the vectors $\mathbf{w}_g$ associated with the groups $g \in \mathcal{G}'$ will be exactly equal to zero, leading to a set of zeros which is the union of these groups, $\bigcup_{g\in\mathcal{G}'} g$. Thus, the set of allowed zero patterns should be the *union-closure* of $\mathcal{G}$, i.e. (see Figure 2.1 for an example):

$$\mathcal{Z} = \Big\{ \bigcup_{g\in\mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \Big\}. \tag{2.3}$$

The situation is however slightly more subtle as some zeros can be created by chance (just as regularizing by the $\ell_2$-norm may lead, though it is unlikely, to some zeros). Nevertheless, Theorem 1 shows that, under mild conditions, the previous intuition about the set of zero patterns is correct. Note that instead of considering the set of zero patterns $\mathcal{Z}$, it is also convenient to manipulate nonzero patterns, and we define

$$\mathcal{P} = \Big\{ \bigcap_{g\in\mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \Big\} = \{z^c;\ z \in \mathcal{Z}\}. \tag{2.4}$$

We can equivalently use $\mathcal{P}$ or $\mathcal{Z}$ by taking the complement of each element of these sets.

The following two results characterize the solutions of the problem (2.2). We first gives sufficient conditions under which this problem has a unique solution. We then

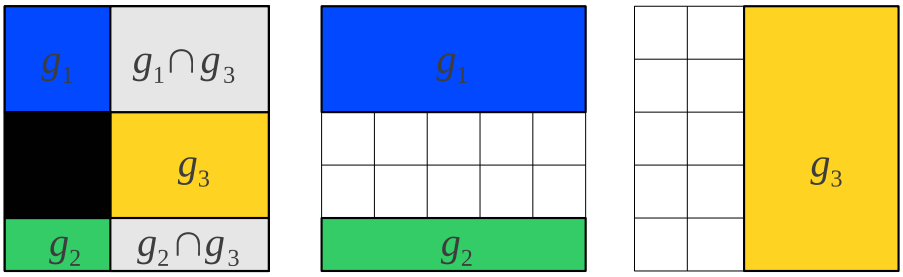



Figure 2.1: Groups and induced nonzero pattern: three sparsity-inducing groups (middle and right, denoted by $\{g_1, g_2, g_3\}$) with the associated nonzero pattern which is the complement of the union of groups, i.e., $(g_1 \cup g_2 \cup g_3)^c$ (left, in black).

formally prove the aforementioned intuition about the zero patterns of the solutions of (2.2), namely they should belong to $\mathcal{Z}$. In the following two results (see proofs in Appendix A.1.1 and Appendix A.1.2), we assume that $\ell : (y, y') \mapsto \ell(y, y')$ is nonnegative, twice continuously differentiable with positive second derivative with respect to the second variable and non-vanishing mixed derivative, i.e., for any $y, y'$ in $\mathbb{R}$, $\frac{\partial^2 \ell}{\partial y'^2}(y, y') > 0$ and $\frac{\partial^2 \ell}{\partial y \partial y'}(y, y') \neq 0$.

**Proposition 5** (Uniqueness)
*Let* $\mathbf{Q}$ *denote the Gram matrix* $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. *We consider the optimization problem in (2.2) with* $\mu > 0$. *If* $\mathbf{Q}$ *is invertible or if the group* $[\![1; p]\!]$ *belongs to* $\mathcal{G}$, *then the problem in (2.2) admits a unique solution.*

Note that the invertibility of the matrix $\mathbf{Q}$ requires $p \leq n$. For high-dimensional settings, the uniqueness of the solution will hold when $\{1, \dots, p\}$ belongs to $\mathcal{G}$, or as further discussed at the end of the proof, as soon as for any $j, k \in \{1, \dots, p\}$, there exists a group $g \in \mathcal{G}$ which contains both $j$ and $k$. Adding the group $\{1, \dots, p\}$ to $\mathcal{G}$ will in general not modify $\mathcal{P}$ (and $\mathcal{Z}$), but it will cause $\mathcal{G}$ to lose its minimality (in a sense introduced in the next subsection). Furthermore, adding the full group $\{1, \dots, p\}$ has to be put in parallel with the equivalent (up to the squaring) $\ell_2$-norm term in the elastic-net penalty (Zou and Hastie, 2005), whose effect is to notably ensure strong convexity. For more sophisticated uniqueness conditions that we have not explored here, we refer the readers to Osborne et al. (2000a, Theorem 1, 4 and 5), Rosset et al. (2004, Theorem 5) or Dossal (2007, Theorem 3) in the Lasso case, and Roth and Fischer (2008) for the group Lasso setting. We now turn to the result about the zero patterns of the solution of the problem in (2.2):

**Theorem 1** (Stable patterns)
*Assume that* $Y = (y_1, \dots, y_n)^\top$ *is a realization of an absolutely continuous probability distribution. Let* $k$ *be the maximal number such that any* $k$ *rows of the matrix* $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$ *are linearly independent. For* $\mu > 0$, *any solution of the problem in (2.2) with at most* $k-1$ *nonzero coefficients has a zero pattern in* $\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$

*almost surely.*

In other words, when $Y = (y_1, \dots, y_n)^\top$ is a realization of an absolutely continuous probability distribution, the sparse solutions have a zero pattern in $\mathcal{Z} = \left\{\bigcup_{g\in\mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G}\right\}$ almost surely. As a corollary of our two results, if the Gram matrix $\mathbf{Q}$ is invertible, the problem in (2.2) has a unique solution, whose zero pattern belongs to $\mathcal{Z}$ almost surely. Note that with the assumption made on $Y$, Theorem 1 is not directly applicable to the classification setting. Based on these previous results, we can look at the following usual special cases from Section 2.2 (we give more examples in Section 2.3.5):

- **$\ell_2$-norm**: the set of allowed nonzero patterns is composed of the empty set and the full set $[\![1; p]\!]$.
- **$\ell_1$-norm**: $\mathcal{P}$ is the set of all possible subsets.
- **$\ell_2$-norm and $\ell_1$-norm**: $\mathcal{P}$ is also the set of all possible subsets.
- **Group $\ell_1$-norm**: $\mathcal{P} = \mathcal{Z}$ is the set of all possible unions of the elements of the partition defining $\mathcal{G}$.
- **Hierarchical norms**: the set of patterns $\mathcal{P}$ is then all sets $J$ for which all ancestors of elements in $J$ are included in $J$ (Bach, 2008a).

Two natural questions now arise: (1) starting from the groups $\mathcal{G}$, is there an efficient way to generate the set of nonzero patterns $\mathcal{P}$; (2) conversely, and more importantly, given $\mathcal{P}$, how can the groups $\mathcal{G}$—and hence the norm $\Omega(w)$—be designed?

### 2.3.2 General Properties of $\mathcal{G}$, $\mathcal{Z}$ and $\mathcal{P}$

We now study the different properties of the set of groups $\mathcal{G}$ and its corresponding sets of patterns $\mathcal{Z}$ and $\mathcal{P}$.

**Closedness.** The set of zero patterns $\mathcal{Z}$ (respectively, the set of nonzero patterns $\mathcal{P}$) is closed under union (respectively, intersection), that is, for all $K \in \mathbb{N}$ and all $z_1, \dots, z_K \in \mathcal{Z}$, $\bigcup_{k=1}^K z_k \in \mathcal{Z}$ (respectively, $p_1, \dots, p_K \in \mathcal{P}$, $\bigcap_{k=1}^K p_k \in \mathcal{P}$). This implies that when "reverse-engineering" the set of nonzero patterns, we have to assume it is closed under intersection. Otherwise, the best we can do is to deal with its intersection-closure. For instance, if we consider a sequence (see Figure 2.4), we cannot take $\mathcal{P}$ to be the set of contiguous patterns with length two, since the intersection of such two patterns may result in a singleton (that does not belong to $\mathcal{P}$).

**Minimality.** If a group in $\mathcal{G}$ is the union of other groups, it may be removed from $\mathcal{G}$ without changing the sets $\mathcal{Z}$ or $\mathcal{P}$. This is the main argument behind the pruning backward algorithm in Section 2.3.3. Moreover, this leads to the notion of a *minimal* set $\mathcal{G}$ of groups, which is such that for all $\mathcal{G}' \subseteq \mathcal{Z}$ whose union-closure spans $\mathcal{Z}$, we have $\mathcal{G} \subseteq \mathcal{G}'$. The existence and uniqueness of a minimal set is a consequence of classical results in set theory (Doignon and Falmagne, 1998). The elements of this minimal set are usually referred to as the *atoms* of $\mathcal{Z}$.

Minimal sets of groups are attractive in our setting because they lead to a smaller number of groups and lower computational complexity—for example, for 2 dimensional-

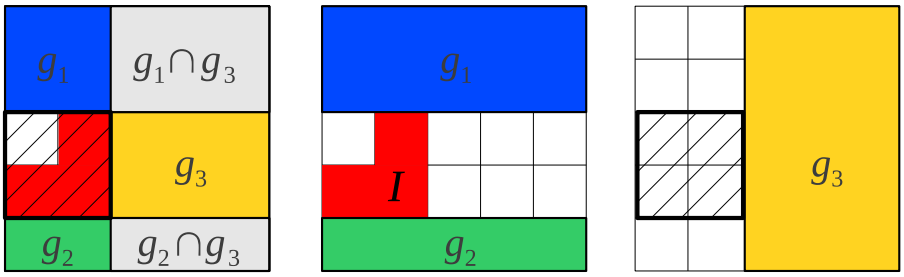



Figure 2.2: $\mathcal{G}$-adapted hull: the pattern of variables $I$ (left and middle, in red) and its hull (left and right, hatched square) that is defined by the complement of the union of groups that do not intersect $I$, i.e., $(g_1 \cup g_2 \cup g_3)^c$.

grids with rectangular patterns, we have a quadratic possible number of rectangles, i.e., $|\mathcal{Z}| = O(p^2)$, that can be generated by a minimal set $\mathcal{G}$ whose size is $|\mathcal{G}| = O(\sqrt{p})$.

**Hull.** Given a set of groups $\mathcal{G}$, we can define for any subset $I \subseteq [\![1; p]\!]$ the *$\mathcal{G}$-adapted hull*, or simply *hull*, as:

$$\text{Hull}(I) = \left\{ \bigcup_{g \in \mathcal{G},\ g \cap I = \varnothing} g \right\}^c,$$

which is the smallest set in $\mathcal{P}$ containing $I$ (see Figure 2.2); we always have $I \subseteq \text{Hull}(I)$ with equality if and only if $I \in \mathcal{P}$. The hull has a clear geometrical interpretation for specific sets $\mathcal{G}$ of groups. For instance, if the set $\mathcal{G}$ is formed by all vertical and horizontal half-spaces when the variables are organized in a 2 dimensional-grid (see Figure 2.5), the hull of a subset $I \subset \{1, \ldots, p\}$ is simply the axis-aligned bounding box of $I$. Similarly, when $\mathcal{G}$ is the set of all half-spaces with all possible orientations (e.g., orientations $\pm\pi/4$ are shown in Figure 2.6), the hull becomes the regular convex hull [1]. Note that those interpretations of the hull are possible and valid only when we have geometrical information at hand about the set of variables.

**Graphs of patterns.** We consider the directed acyclic graph (DAG) stemming from the *Hasse diagram* (Cameron, 1994) of the partially ordered set (poset) $(\mathcal{G}, \supset)$. By definition, the nodes of this graph are the elements $g$ of $\mathcal{G}$ and there is a directed edge from $g_1$ to $g_2$ if and only if $G_1 \supset G_2$ and there exists no $g \in \mathcal{G}$ such that $g_1 \supset g \supset g_2$ (Cameron, 1994). We can also build the corresponding DAG for the set of zero patterns $\mathcal{Z} \supset \mathcal{G}$, which is a super-DAG of the DAG of groups (see Figure 2.3 for examples). Note that we obtain also the isomorphic DAG for the nonzero patterns $\mathcal{P}$, although it corresponds to the poset $(\mathcal{P}, \subset)$: this DAG will be used in the active set algorithm presented in Section 2.4.

1. We use the term *convex* informally here. It can however be made precise with the notion of convex subgraphs (Chung, 1997).

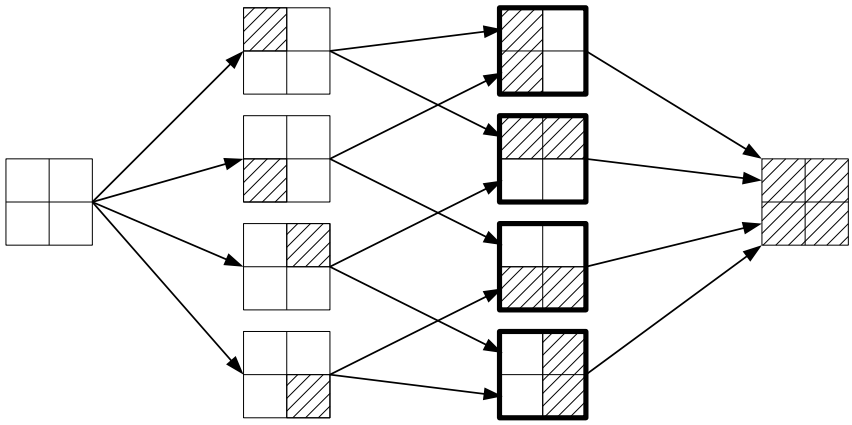


Figure 2.3: The DAG for the set $\mathcal{Z}$ associated with the $2\times 2$-grid. The members of $\mathcal{Z}$ are the complement of the areas hatched in black. The elements of $\mathcal{G}$ (i.e., the atoms of $\mathcal{Z}$) are highlighted by bold edges.

Prior works with nested groups (Zhao et al., 2009; Bach, 2008a; Kim and Xing, 2010; Jenatton et al., 2010a; Schmidt and Murphy, 2010) have also used a similar DAG structure, with the slight difference that in these works, the corresponding hierarchy of variables is built from the prior knowledge about the problem at hand (e.g., the tree of wavelets in Zhao et al. (2009), the decomposition of kernels in Bach (2008a) or the hierarchy of genes in Kim and Xing (2010)). The DAG we introduce here on the set of groups naturally and always comes up, with no assumption on the variables themselves (for which no DAG is defined in general).

### 2.3.3 From Patterns to Groups

We now assume that we want to impose a priori knowledge on the sparsity structure of a solution $\hat{\mathbf{w}}$ of our regularized problem in Eq. 2.2. This information can be exploited by restricting the patterns allowed by the norm $\Omega$. Namely, from an intersection-closed set of zero patterns $\mathcal{Z}$, we can build back a minimal set of groups $\mathcal{G}$ by iteratively pruning away in the DAG corresponding to $\mathcal{Z}$, all sets which are unions of their parents. See Algorithm 1. This algorithm can be found under a different form in Doignon and Falmagne (1998)—we present it through a pruning algorithm on the DAG, which is natural in our context (the proof of the minimality of the procedure can be found in Appendix A.1.3). The complexity of Algorithm 1 is $O(p|\mathcal{Z}|^2)$. The pruning may reduce significantly the number of groups necessary to generate the whole set of zero patterns, sometimes from exponential in $p$ to polynomial in $p$ (e.g., the $\ell_1$-norm). In Section 2.3.5, we give other examples of interest where $|\mathcal{G}|$ (and $|\mathcal{P}|$) is also polynomial in $p$.

**Algorithm 1** Backward procedure

**Input:** Intersection-closed family of nonzero patterns $\mathcal{P}$.
**Output:** Set of groups $\mathcal{G}$.
**Initialization:** Compute $\mathcal{Z} = \{P^c;\ P \in \mathcal{P}\}$ and set $\mathcal{G} = \mathcal{Z}$.
Build the Hasse diagram for the poset $(\mathcal{Z}, \supset)$.
**for** $t = \min_{g \in \mathcal{Z}} |g|$ **to** $\max_{g \in \mathcal{Z}} |g|$ **do**
  **for** each node $g \in \mathcal{Z}$ such that $|g| = t$ **do**
    **if** $\left(\bigcup_{C \in \text{Children}(g)} C = g\right)$ **then**
      **if** $(\text{Parents}(g) \neq \varnothing)$ **then**
        Connect children of $g$ to parents of $g$.
      **end if**
      Remove $g$ from $\mathcal{G}$.
    **end if**
  **end for**
**end for**

**Algorithm 2** Forward procedure

**Input:** Set of groups $\mathcal{G} = \{g_1, \ldots, g_M\}$.
**Output:** Collection of zero patterns $\mathcal{Z}$ and nonzero patterns $\mathcal{P}$.
**Initialization:** $\mathcal{Z} = \{\varnothing\}$.
**for** $m = 1$ **to** $M$ **do**
  $C = \{\varnothing\}$
  **for each** $Z \in \mathcal{Z}$ **do**
    **if** $(g_m \not\subseteq Z)$ **and** $(\forall g \in \{g_1, \ldots, g_{m-1}\},\ g \subseteq Z \cup g_m \Rightarrow g \subseteq Z)$ **then**
      $C \leftarrow C \cup \{Z \cup g_m\}$ .
    **end if**
  **end for**
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup C$.
**end for**
$\mathcal{P} = \{Z^c;\ Z \in \mathcal{Z}\}$.

### 2.3.4 From Groups to Patterns

The *forward* procedure presented in Algorithm 2, taken from Doignon and Falmagne (1998), allows the construction of $\mathcal{Z}$ from $\mathcal{G}$. It iteratively builds the collection of patterns by taking unions, and has complexity $O(p|\mathcal{Z}||\mathcal{G}|^2)$. The general scheme is straightforward. Namely, by considering increasingly larger sub-families of $\mathcal{G}$ and the collection of patterns already obtained, all possible unions are formed. However, some attention needs to be paid while checking we are not generating a pattern already encountered. Such a verification is performed by the *if* condition within the inner loop of the algorithm. Indeed, we do not have to scan the whole collection of patterns already obtained (whose size can be exponential in $|\mathcal{G}|$), but we rather use the fact that $\mathcal{G}$ generates $\mathcal{Z}$.

Note that in general, it is not possible to upper bound the size of $|\mathcal{Z}|$ by a polynomial term in $p$, even when $\mathcal{G}$ is very small (indeed, $|\mathcal{Z}| = 2^p$ and $|\mathcal{G}| = p$ for the $\ell_1$-norm).

### 2.3.5 Examples

We now present several examples of sets of groups $\mathcal{G}$, especially suited to encode geometric and temporal prior information.

**Sequences.** Given $p$ variables organized in a sequence, if we want only contiguous nonzero patterns, the backward algorithm will lead to the set of groups which are intervals $[1, k]_{k\in\{1,\ldots,p-1\}}$ and $[k, p]_{k\in\{2,\ldots,p\}}$, with both $|\mathcal{Z}| = O(p^2)$ and $|\mathcal{G}| = O(p)$ (see Figure 2.4). Imposing the contiguity of the nonzero patterns is for instance relevant for the diagnosis of tumors, based on the profiles of arrayCGH (Rapaport et al., 2008).

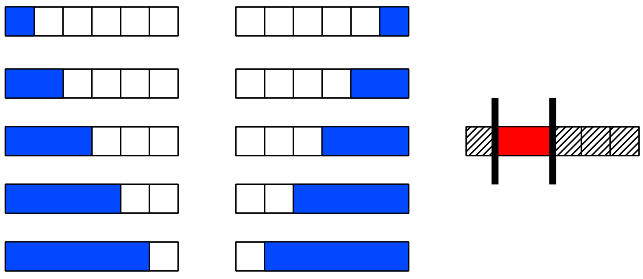


Figure 2.4: (Left) The set of blue groups to penalize in order to select contiguous patterns in a sequence. (Right) In red, an example of such a nonzero pattern with its corresponding zero pattern (hatched area).

**Two-dimensional grids.** In Section 2.6, we notably consider for $\mathcal{P}$ the set of all rectangles in two dimensions, leading by the previous algorithm to the set of axis-aligned half-spaces for $\mathcal{G}$ (see Figure 2.5), with $|\mathcal{Z}| = O(p^2)$ and $|\mathcal{G}| = O(\sqrt{p})$. This type of structure is encountered in object or scene recognition, where the selected rectangle would correspond to a certain box inside an image, that concentrates the predictive power for a given class of object/scene (Harzallah et al., 2009).

Larger set of convex patterns can be obtained by adding in $\mathcal{G}$ half-planes with other orientations than vertical and horizontal. For instance, if we use planes with angles that are multiples of $\pi/4$, the nonzero patterns of $\mathcal{P}$ can have polygonal shapes with up to 8 faces. In this sense, if we keep on adding half-planes with finer orientations, the nonzero patterns of $\mathcal{P}$ can be described by polygonal shapes with an increasingly larger number of faces. The standard notion of convexity defined in $\mathbb{R}^2$ would correspond to the situation where an infinite number of orientations is considered (Soille, 2003). See Figure 2.6. The number of groups is linear in $\sqrt{p}$ with constant growing linearly with the number of angles, while $|\mathcal{Z}|$ grows more rapidly (typically non-polynomially in the number of angles). Imposing such convex-like regions turns out to be useful in computer vision. For

instance, in face recognition, it enables the design of localized features that improve upon the robustness to occlusions (Jenatton et al., 2010b). In the same vein, regularizations with similar two-dimensional sets of groups have led to good performances in background subtraction tasks (Mairal et al., 2010b), where the pixel spatial information is crucial to avoid scattered results. Another application worth being mentioned is the design of topographic dictionaries in the context of image processing (Kavukcuoglu et al., 2009; Mairal et al., 2011). In this case, dictionaries self-organize and adapt to the underlying geometrical structure encoded by the two-dimensional set of groups.

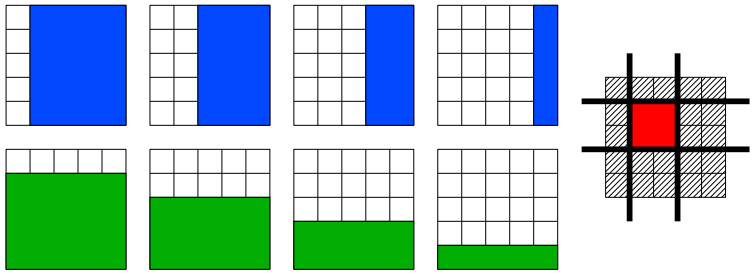


Figure 2.5: Vertical and horizontal groups: (Left) the set of blue and green groups with their (not displayed) complements to penalize in order to select rectangles. (Right) In red, an example of nonzero pattern recovered in this setting, with its corresponding zero pattern (hatched area).

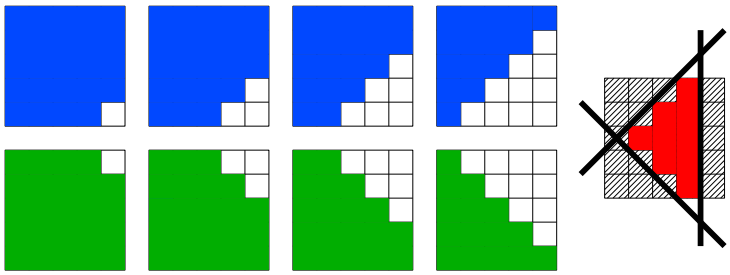


Figure 2.6: Groups with $\pm\pi/4$ orientations: (Left) the set of blue and green groups with their (not displayed) complements to penalize in order to select diamond-shaped patterns. (Right) In red, an example of nonzero pattern recovered in this setting, with its corresponding zero pattern (hatched area).

**Extensions.** The sets of groups presented above can be straightforwardly extended to more complicated topologies, such as three-dimensional spaces discretized in cubes or spherical volumes discretized in slices. Similar properties hold for such settings. For instance, if all the axis-aligned half-spaces are considered for $\mathcal{G}$ in a three-dimensional space, then $\mathcal{P}$ is the set of all possible rectangular boxes with $|\mathcal{P}| = O(p^2)$ and $|\mathcal{G}| =$

$O(p^{1/3})$. Such three-dimensional structures are interesting to retrieve discriminative and local sets of voxels from fMRI/MEEG responses. In particular, they have recently proven useful for modeling brain resting-state activity (Varoquaux et al., 2010c). Moreover, while the two-dimensional rectangular patterns described previously are adapted to find bounding boxes in static images (Harzallah et al., 2009), scene recognition in videos requires to deal with a third temporal dimension (Dalal et al., 2006). This may be achieved by designing appropriate sets of groups, embedded in the three-dimensional space obtained by tracking the frames over time. Finally, in the context of matrix-based optimization problems, e.g., multi-task learning and dictionary learning, sets of groups $\mathcal{G}$ can also be designed to encode *structural constraints* the solutions must respect. This notably encompasses banded structures (Levina et al., 2008) and *simultaneous* row/column sparsity for CUR matrix factorization (Mairal et al., 2011).

**Representation and computation of $\mathcal{G}$.** The sets of groups described so far can actually be represented in a same form, that lends itself well to the analysis of the next section. When dealing with a discrete sequence of length $l$ (see Figure 2.4), we have

$$\begin{aligned}\mathcal{G} &= \{g_-^k;\ k \in \{1,\dots,l-1\}\} \cup \{g_+^k;\ k \in \{2,\dots,l\}\},\\ &= \mathcal{G}_{\text{left}} \cup \mathcal{G}_{\text{right}},\end{aligned}$$

with $g_-^k = \{i;\ 1 \leq i \leq k\}$ and $g_+^k = \{i;\ l \geq i \geq k\}$. In other words, the set of groups $\mathcal{G}$ can be rewritten as a partition [2] in two sets of nested groups, $\mathcal{G}_{\text{left}}$ and $\mathcal{G}_{\text{right}}$.

The same goes for a two-dimensional grid, with dimensions $h \times l$ (see Figure 2.5 and Figure 2.6). In this case, the nested groups we consider are defined based on the following groups of variables

$$g^{k,\theta} = \{(i,j) \in \{1,\dots,l\} \times \{1,\dots,h\};\ \cos(\theta)i + \sin(\theta)j \leq k\},$$

where $k \in \mathbb{Z}$ is taken in an appropriate range. The nested groups we obtain in this way are therefore parameterized by an angle [3] $\theta$, $\theta \in (-\pi;\pi]$. We refer to this angle as an *orientation*, since it defines the normal vector $\binom{\cos(\theta)}{\sin(\theta)}$ to the line $\{(i,j) \in \mathbb{R}^2; \cos(\theta)i + \sin(\theta)j = k\}$. In the example of the rectangular groups (see Figure 2.5), we have four orientations, with $\theta \in \{0, \pi/2, -\pi/2, \pi\}$. More generally, if we denote by $\Theta$ the set of the orientations, we have

$$\mathcal{G} = \bigcup_{\theta \in \Theta} \mathcal{G}_\theta,$$

where $\theta \in \Theta$ indexes the partition of $\mathcal{G}$ in sets $\mathcal{G}_\theta$ of nested groups of variables. Although we have not detailed the case of $\mathbb{R}^3$, we likewise end up with a similar partition of $\mathcal{G}$.

2. Note the subtlety: the sets $\mathcal{G}_\theta$ are disjoint, that is $\mathcal{G}_\theta \cap \mathcal{G}_{\theta'} = \varnothing$ for $\theta \neq \theta'$, but groups in $\mathcal{G}_\theta$ and $\mathcal{G}_{\theta'}$ can overlap.

3. Due to the discrete nature of the underlying geometric structure of $\mathcal{G}$, angles $\theta$ that are not multiple of $\pi/4$ (i.e., such that $\tan(\theta) \notin \mathbb{Z}$) are dealt with by rounding operations.

## 2.4 Optimization and Active Set Algorithm

For moderate values of $p$, one may obtain a solution for Eq. (2.2) using generic toolboxes for second-order cone programming (SOCP) whose time complexity is equal to $O(p^{3.5} + |\mathcal{G}|^{3.5})$ (Boyd and Vandenberghe, 2004), which is not appropriate when $p$ or $|\mathcal{G}|$ are large. This time complexity corresponds to the computation of Eq. (2.2) for a single value of the regularization parameter $\mu$.

We present in this section an *active set algorithm* (Algorithm 3) that finds a solution for Eq. (2.2) by considering increasingly larger active sets and checking global optimality at each step. When the rectangular groups are used, the total complexity of this method is in $O(s \max\{p^{1.75}, s^{3.5}\})$, where $s$ is the size of the active set at the end of the optimization. Here, the sparsity prior is exploited for computational advantages. Our active set algorithm needs an underlying *black-box* SOCP solver; in this chapter, we consider both a first order approach (see Appendix A.1.8) and a SOCP method [4] — in our experiments, we use `SDPT3` (Toh et al., 1999; Tütüncü et al., 2003). Our active set algorithm extends to general overlapping groups the work of Bach (2008a), by further assuming that it is computationally possible to have a time complexity polynomial in the number of variables $p$.

We primarily focus here on finding an efficient active set algorithm; we defer to future work the design of specific SOCP solvers, e.g., based on proximal techniques (see, e.g., Nesterov, 2007; Beck and Teboulle, 2009; Combettes and Pesquet, 2010, and numerous references therein), adapted to such non-smooth sparsity-inducing penalties.

### 2.4.1 Optimality Conditions: from Reduced Problems to Full Problems

It is simpler to derive the algorithm for the following regularized optimization problem [5] which has the same solution set as the regularized problem of Eq. (2.2) when $\mu$ and $\lambda$ are allowed to vary (Borwein and Lewis, 2006, see Section 3.2):

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) + \frac{\lambda}{2}\left[\Omega(\mathbf{w})\right]^2 . \tag{2.5}$$

In active set methods, the set of nonzero variables, denoted by $J$, is built incrementally, and the problem is solved only for this reduced set of variables, adding the constraint $\mathbf{w}_{J^c} = 0$ to Eq. (2.5). In the subsequent analysis, we will use arguments based on duality to monitor the optimality of our active set algorithm. We denote by $L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i)$ the empirical risk (which is by assumption convex and

4. The C++/Matlab code used in the experiments may be downloaded from the authors website.

5. It is also possible to derive the active set algorithm for the constrained formulation $\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, \mathbf{w}^\top \mathbf{x}_i\big)$ such that $\Omega(\mathbf{w}) \leq \lambda$. We however observed that the penalized formulation tends to be empirically easier to tune (e.g., via cross-validation), as the performance is usually quite robust to small changes in $\mu$, while it is not robust to small changes in $\lambda$.

continuously differentiable) and by $L^*$ its *Fenchel-conjugate*, defined as (Boyd and Vandenberghe, 2004; Borwein and Lewis, 2006):

$$L^*(u) = \sup_{\mathbf{w}\in\mathbb{R}^p} \{\mathbf{w}^\top \mathbf{u} - L(\mathbf{w})\}.$$

The restriction of $L$ to $\mathbb{R}^{|J|}$ is denoted $L_J(\mathbf{w}_J) = L(\tilde{\mathbf{w}})$ for $\tilde{\mathbf{w}}_J = \mathbf{w}_J$ and $\tilde{\mathbf{w}}_{J^c} = 0$, with Fenchel-conjugate $L_J^*$. Note that, as opposed to $L$, we do not have in general $L_J^*(\boldsymbol{\kappa}_J) = L^*(\tilde{\boldsymbol{\kappa}})$ for $\tilde{\boldsymbol{\kappa}}_J = \boldsymbol{\kappa}_J$ and $\tilde{\boldsymbol{\kappa}}_{J^c} = 0$.

For a potential active set $J \subset [\![1;p]\!]$ which belongs to the set of allowed nonzero patterns $\mathcal{P}$, we denote by $\mathcal{G}_J$ the set of active groups, i.e., the set of groups $G \in \mathcal{G}$ such that $G \cap J \neq \varnothing$. We consider the reduced norm $\Omega_J$ defined on $\mathbb{R}^{|J|}$ as

$$\Omega_J(\mathbf{w}_J) = \sum_{g\in\mathcal{G}} \|\omega_J^g \circ \mathbf{w}_J\|_2 = \sum_{g\in\mathcal{G}_J} \|\omega_J^g \circ \mathbf{w}_J\|_2,$$

and its *dual norm* $\Omega_J^*(\boldsymbol{\kappa}_J) = \max_{\Omega_J(\mathbf{w}_J)\leq 1} \mathbf{w}_J^\top \boldsymbol{\kappa}_J$, also defined on $\mathbb{R}^{|J|}$. The next proposition (see proof in Appendix A.1.4) gives the optimization problem dual to the reduced problem (Eq. 2.6 below):

**Proposition 6** (Dual Problems)
*Let $J \subseteq [\![1;p]\!]$. The following two problems*

$$\min_{\mathbf{w}_J\in\mathbb{R}^{|J|}} L_J(\mathbf{w}_J) + \frac{\lambda}{2}\left[\Omega_J(\mathbf{w}_J)\right]^2, \tag{2.6}$$

$$\max_{\boldsymbol{\kappa}_J\in\mathbb{R}^{|J|}} -L_J^*(-\boldsymbol{\kappa}_J) - \frac{1}{2\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2, \tag{2.7}$$

*are dual to each other and strong duality holds. The pair of primal-dual variables $\{\mathbf{w}_J, \boldsymbol{\kappa}_J\}$ is optimal if and only if we have*

$$\begin{cases} \boldsymbol{\kappa}_J & = -\nabla L_J(\mathbf{w}_J), \\ \mathbf{w}_J^\top \boldsymbol{\kappa}_J & = \frac{1}{\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 = \lambda\left[\Omega_J(\mathbf{w}_J)\right]^2. \end{cases}$$

As a brief reminder, the duality gap of a minimization problem is defined as the difference between the primal and dual objective functions, evaluated for a feasible pair of primal/dual variables (Boyd and Vandenberghe, 2004, see Section 5.5). This gap serves as a certificate of (sub)optimality: if it is equal to zero, then the optimum is reached, and provided that strong duality holds, the converse is true as well (Boyd and Vandenberghe, 2004, see Section 5.5).

The previous proposition enables us to derive the duality gap for the optimization problem Eq. 2.6, that is reduced to the active set of variables $J$. In practice, this duality gap will always vanish (up to the precision of the underlying SOCP solver), since we will sequentially solve (2.6) for increasingly larger active sets $J$. We now study how, starting from the optimality of the problem in (2.6), we can control the optimality, or

equivalently the duality gap, for the full problem Eq. (2.5). More precisely, the duality gap of the optimization problem Eq. 2.6 is

$$\begin{aligned} & L_J(\mathbf{w}_J) + L_J^*(-\boldsymbol{\kappa}_J) + \frac{\lambda}{2}\left[\Omega_J(\mathbf{w}_J)\right]^2 + \frac{1}{2\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 \\ = & \left\{L_J(\mathbf{w}_J) + L_J^*(-\boldsymbol{\kappa}_J) + \mathbf{w}_J^\top \boldsymbol{\kappa}_J\right\} + \left\{\frac{\lambda}{2}\left[\Omega_J(\mathbf{w}_J)\right]^2 + \frac{1}{2\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 - \mathbf{w}_J^\top \boldsymbol{\kappa}_J\right\}, \end{aligned}$$

which is a sum of two nonnegative terms, the nonnegativity coming from the Fenchel-Young inequality (Borwein and Lewis, 2006; Boyd and Vandenberghe, 2004, Proposition 3.3.4 and Section 3.3.2 respectively). We can think of this duality gap as the sum of two duality gaps, respectively relative to $L_J$ and $\Omega_J$. Thus, if we have a primal candidate $\mathbf{w}_J$ and we choose $\boldsymbol{\kappa}_J = -\nabla L_J(\mathbf{w}_J)$, the duality gap relative to $L_J$ vanishes and the total duality gap then reduces to

$$\frac{\lambda}{2}\left[\Omega_J(\mathbf{w}_J)\right]^2 + \frac{1}{2\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 - \mathbf{w}_J^\top \boldsymbol{\kappa}_J.$$

In order to check that the reduced solution $\mathbf{w}_J$ is optimal for the full problem in Eq. (2.5), we pad $\mathbf{w}_J$ with zeros on $J^c$ to define $\mathbf{w}$ and compute $\boldsymbol{\kappa} = -\nabla L(\mathbf{w})$, which is such that $\boldsymbol{\kappa}_J = -\nabla L_J(\mathbf{w}_J)$. For our given candidate pair of primal/dual variables $\{\mathbf{w}, \boldsymbol{\kappa}\}$, we then get a duality gap for the full problem in Eq. (2.5) equal to

$$\begin{aligned} & \frac{\lambda}{2}\left[\Omega(\mathbf{w})\right]^2 + \frac{1}{2\lambda}\left[\Omega^*(\boldsymbol{\kappa})\right]^2 - \mathbf{w}^\top \boldsymbol{\kappa} \\ = & \frac{\lambda}{2}\left[\Omega(\mathbf{w})\right]^2 + \frac{1}{2\lambda}\left[\Omega^*(\boldsymbol{\kappa})\right]^2 - \mathbf{w}_J^\top \boldsymbol{\kappa}_J \\ = & \frac{\lambda}{2}\left[\Omega(\mathbf{w})\right]^2 + \frac{1}{2\lambda}\left[\Omega^*(\boldsymbol{\kappa})\right]^2 - \frac{\lambda}{2}\left[\Omega_J(\mathbf{w}_J)\right]^2 - \frac{1}{2\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 \\ = & \frac{1}{2\lambda}\left(\left[\Omega^*(\boldsymbol{\kappa})\right]^2 - \left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2\right) \\ = & \frac{1}{2\lambda}\left(\left[\Omega^*(\boldsymbol{\kappa})\right]^2 - \lambda \mathbf{w}_J^\top \boldsymbol{\kappa}_J\right). \end{aligned}$$

Computing this gap requires computing the dual norm which itself is as hard as the original problem, prompting the need for upper and lower bounds on $\Omega^*$ (see Propositions 7 and 8 for more details).

### 2.4.2 Active set algorithm

We can interpret the active set algorithm as a walk through the DAG of nonzero patterns allowed by the norm $\Omega$. The parents $\Pi_{\mathcal{P}}(J)$ of $J$ in this DAG are exactly the patterns containing the variables that may enter the active set at the next iteration of Algorithm 3. The groups that are exactly at the boundaries of the active set (referred to as the *fringe groups*) are $\mathcal{F}_J = \{g \in (\mathcal{G}_J)^c \,;\, \nexists g' \in (\mathcal{G}_J)^c,\ g \subseteq g'\}$, i.e., the groups that are not contained by any other inactive groups.

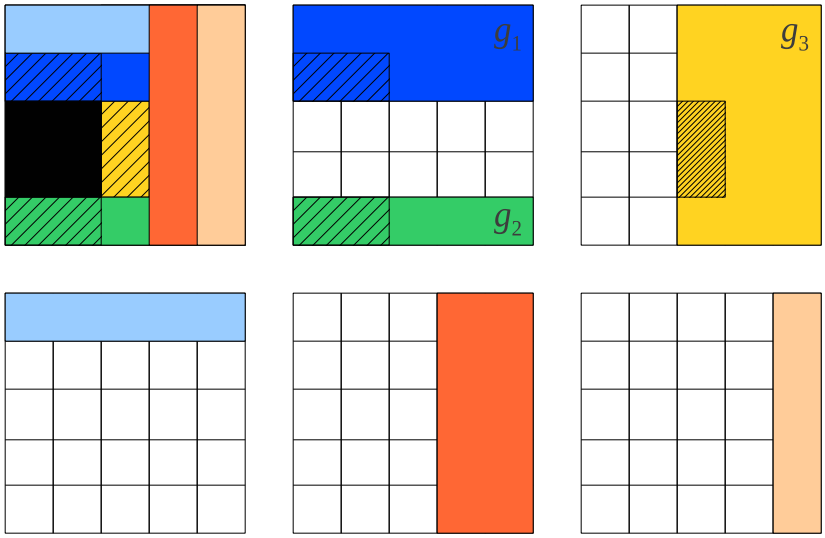



Figure 2.7: The active set (black) and the candidate patterns of variables, i.e. the variables in $K \backslash J$ (hatched in black) that can become active. The fringe groups are exactly the groups that have the hatched areas (i.e., here we have $\mathcal{F}_J = \bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J = \{g_1, g_2, g_3\}$).

In simple settings, e.g., when $\mathcal{G}$ is the set of rectangular groups, the correspondence between groups and variables is straightforward since we have $\mathcal{F}_J = \bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$ (see Figure 2.7). However, in general, we just have the inclusion $(\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J) \subseteq \mathcal{F}_J$ and some elements of $\mathcal{F}_J$ might not correspond to any patterns of variables in $\Pi_{\mathcal{P}}(J)$ (see Figure 2.8).

We now present the optimality conditions (see proofs in Appendix A.1.5) that monitor the progress of Algorithm 3:

**Proposition 7** (Necessary condition)
*If* $\mathbf{w}$ *is optimal for the full problem in Eq. (2.5), then*

$$\max_{K \in \Pi_{\mathcal{P}}(J)} \frac{\|\nabla L(\mathbf{w})_{K \backslash J}\|_2}{\sum_{h \in \mathcal{G}_K \backslash \mathcal{G}_J} \|\omega^h_{K \backslash J}\|_\infty} \leq \{-\lambda \mathbf{w}^\top \nabla L(\mathbf{w})\}^{\frac{1}{2}}. \tag{N}$$

**Proposition 8** (Sufficient condition)
*If*

$$\max_{g \in \mathcal{F}_J} \left\{ \sum_{k \in g} \left\{ \frac{\nabla L(\mathbf{w})_k}{\sum_{h \ni k,\ h \in (\mathcal{G}_J)^c} \omega^h_k} \right\}^2 \right\}^{\frac{1}{2}} \leq \{\lambda(2\varepsilon - w^\top \nabla L(\mathbf{w}))\}^{\frac{1}{2}}, \tag{$S_\varepsilon$}$$

*then* $\mathbf{w}$ *is an approximate solution for Eq. (2.5) whose duality gap is less than* $\varepsilon \geq 0$.

Note that for the Lasso, the conditions $(N)$ and $(S_0)$ (i.e., the sufficient condition taken with $\varepsilon = 0$) are both equivalent (up to the squaring of $\Omega$) to the condition

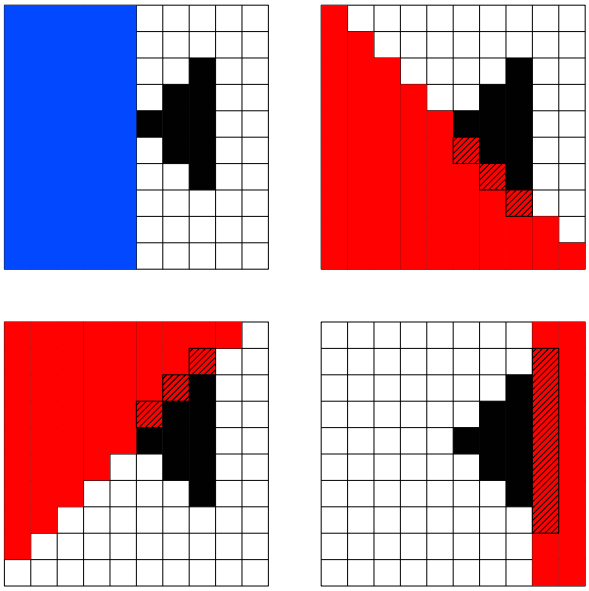


Figure 2.8: The active set (black) and the candidate patterns of variables, i.e. the variables in $K \backslash J$ (hatched in black) that can become active. The groups in red are those in $\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$, while the blue group is in $\mathcal{F}_J \backslash (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J)$. The blue group does not intersect with any patterns in $\Pi_{\mathcal{P}}(J)$.

$\|\nabla L(\mathbf{w})_{J^c}\|_\infty \leq -\mathbf{w}^\top \nabla L(\mathbf{w})$, which is the usual optimality condition (Fuchs, 2005; Tibshirani, 1996; Wainwright, 2009). Moreover, when they are not satisfied, our two conditions provide good heuristics for choosing which $K \in \Pi_{\mathcal{P}}(J)$ should enter the active set.

More precisely, since the necessary condition ($N$) directly deals with the *variables* (as opposed to groups) that can become active at the next step of Algorithm 3, it suffices to choose the pattern $K \in \Pi_{\mathcal{P}}(J)$ that violates most the condition.

The heuristics for the sufficient condition ($S_\varepsilon$) implies that, to go from groups to variables, we simply consider the group $g \in \mathcal{F}_J$ violating the sufficient condition the most and then take all the patterns of variables $K \in \Pi_{\mathcal{P}}(J)$ such that $K \cap g \neq \varnothing$ to enter the active set. If $g \cap (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} K) = \varnothing$, we look at all the groups $h \in \mathcal{F}_J$ such that $h \cap g \neq \varnothing$ and apply the scheme described before (see Algorithm 4).

A direct consequence of this heuristics is that it is possible for the algorithm to *jump over* the right active set and to consider instead a (slightly) larger active set as optimal. However if the active set is larger than the optimal set, then (it can be proved that) the sufficient condition ($S_0$) is satisfied, and the reduced problem, which we solve exactly, will still output the correct nonzero pattern.

Moreover, it is worthwhile to notice that in Algorithm 3, the active set may sometimes be increased only to make sure that the current solution is optimal (we only check a sufficient condition of optimality).

**Algorithm 3** Active set algorithm

**Input:** Data $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$, regularization parameter $\lambda$,
Duality gap precision $\varepsilon$, maximum number of variables $s$.
**Output:** Active set $J$, loading vector $\hat{w}$.
**Initialization:** $J = \{\varnothing\}$, $\hat{w} = 0$.
**while** ( $(N)$ is not satisfied ) **and** ( $|J| \leq s$ ) **do**
Replace $J$ by violating $K \in \Pi_{\mathcal{P}}(J)$ in $(N)$.
Solve the reduced problem $\min_{\mathbf{w}_J \in \mathbb{R}^{|J|}} L_J(\mathbf{w}_J) + \frac{\lambda}{2} \left[\Omega_J(\mathbf{w}_J)\right]^2$ to get $\hat{w}$.
**end while**
**while** ( $(S_\varepsilon)$ is not satisfied ) **and** ( $|J| \leq s$ ) **do**
Update $J$ according to Algorithm 4.
Solve the reduced problem $\min_{\mathbf{w}_J \in \mathbb{R}^{|J|}} L_J(\mathbf{w}_J) + \frac{\lambda}{2} \left[\Omega_J(\mathbf{w}_J)\right]^2$ to get $\hat{w}$.
**end while**

**Convergence of the active set algorithm.** The procedure described in Algorithm 3 can terminate in two different states. If the procedure stops because of the limit on the number of active variables $s$, the solution might be suboptimal. Note that, in any case, we have at our disposal a upperbound on the duality gap.

Otherwise, the procedure always converges to an optimal solution, either (1) by validating both the necessary and sufficient conditions (see Propositions 7 and 8), ending up with fewer than $p$ active variables and a precision of (at least) $\varepsilon$, or (2) by running until the $p$ variables become active, the precision of the solution being given by the underlying solver.

**Algorithm 4** Heuristics for the sufficient condition $(S_\varepsilon)$

Let $g \in \mathcal{F}_J$ be the group that violates $(S_\varepsilon)$ most.
**if** $(g \cap (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} K) \neq \varnothing)$ **then**
**for** $K \in \Pi_{\mathcal{P}}(J)$ such that $K \cap g \neq \varnothing$ **do**
$J \leftarrow J \cup K$.
**end for**
**else**
**for** $H \in \mathcal{F}_J$ such that $H \cap g \neq \varnothing$ **do**
**for** $K \in \Pi_{\mathcal{P}}(J)$ such that $K \cap H \neq \varnothing$ **do**
$J \leftarrow J \cup K$.
**end for**
**end for**
**end if**

**Algorithmic complexity.** We analyze in detail the time complexity of the active set algorithm when we consider sets of groups $\mathcal{G}$ such as those presented in the examples of Section 2.3.5. We recall that we denote by $\Theta$ the set of orientations in $\mathcal{G}$ (for more details, see Section 2.3.5). For such choices of $\mathcal{G}$, the fringe groups $\mathcal{F}_J$ reduces to the

largest groups of each orientation and therefore $|\mathcal{F}_J| \leq |\Theta|$. We further assume that the groups in $\mathcal{G}_\theta$ are sorted by cardinality, so that computing $\mathcal{F}_J$ costs $O(|\Theta|)$.

Given an active set $J$, both the necessary and sufficient conditions require to have access to the direct parents $\Pi_\mathcal{P}(J)$ of $J$ in the DAG of nonzero patterns. In simple settings, e.g., when $\mathcal{G}$ is the set of rectangular groups, this operation can be performed in $O(1)$ (it just corresponds to scan the (up to) four patterns at the edges of the current rectangular hull).

However, for more general orientations, computing $\Pi_\mathcal{P}(J)$ requires to find the smallest nonzero patterns that we can generate from the groups in $\mathcal{F}_J$, reduced to the stripe of variables around the current hull. This stripe of variables can be computed as $\left[\bigcup_{g \in (\mathcal{G}_J)^c \backslash \mathcal{F}_J} g\right]^c \backslash J$, so that getting $\Pi_\mathcal{P}(J)$ costs $O(s2^{|\Theta|} + p|\mathcal{G}|)$ in total.

Thus, if the number of active variables is upper bounded by $s \ll p$ (which is true if our target is actually sparse), the time complexity of Algorithm 3 is the sum of:

- the computation of the gradient, $O(snp)$ for the square loss.
- if the underlying solver called upon by the active set algorithm is a standard SOCP solver, $O(s \max_{J \in \mathcal{P}, |J| \leq s} |\mathcal{G}_J|^{3.5} + s^{4.5})$ (note that the term $s^{4.5}$ could be improved upon by using warm-restart strategies for the sequence of reduced problems).
- $t_1$ times the computation of $(N)$, that is $O(t_1(s2^{|\Theta|} + p|\mathcal{G}| + sn_\theta^2) + p|\mathcal{G}|) = O(t_1 p|\mathcal{G}|)$. During the initialization (i.e., $J = \varnothing$), we have $|\Pi_\mathcal{P}(\varnothing)| = O(p)$ (since we can start with any singletons), and $|\mathcal{G}_K \backslash \mathcal{G}_J| = |\mathcal{G}_K| = |\mathcal{G}|$, which leads to a complexity of $O(p|\mathcal{G}|)$ for the sum $\sum_{G \in \mathcal{G}_K \backslash \mathcal{G}_J} = \sum_{G \in \mathcal{G}_K}$. Note however that this sum does not depend on $J$ and can therefore be cached if we need to make several runs with the same set of groups $\mathcal{G}$.
- $t_2$ times the computation of $(S_\varepsilon)$, that is $O(t_2(s2^{|\Theta|} + p|\mathcal{G}| + |\Theta|^2 + |\Theta|p + p|\mathcal{G}|)) = O(t_2 p|\mathcal{G}|)$, with $t_1 + t_2 \leq s$.

We finally get complexity with a leading term in $O(sp|\mathcal{G}| + s \max_{J \in \mathcal{P}, |J| \leq s} |\mathcal{G}_J|^{3.5} + s^{4.5})$, which is much better than $O(p^{3.5} + |\mathcal{G}|^{3.5})$, without an active set method. In the example of the two-dimensional grid (see Section 2.3.5), we have $|\mathcal{G}| = O(\sqrt{p})$ and $O(s \max\{p^{1.75}, s^{3.5}\})$ as total complexity. The simulations of Section 2.6 confirm that the active set strategy is indeed useful when $s$ is much smaller than $p$. Moreover, the two extreme cases where $s \approx p$ or $p \ll 1$ are also shown not to be advantageous for the active set strategy, since either it is cheaper to use the SOCP solver directly on the $p$ variables, or we uselessly pay the additional fixed-cost of the active set machinery (such as computing the optimality conditions). Note that we have derived here the *theoretical* complexity of the active set algorithm when we use an interior point method as underlying solver. With the first order method presented in Appendix A.1.8, we would instead get a total complexity in $O(sp^{1.5})$.

### 2.4.3 Intersecting Nonzero Patterns

We have seen so far how overlapping groups can encore prior information about a desired set of (non)zero patterns. In practice, controlling these overlaps may be delicate and hinges on the choice of the weights $(\omega^g)_{G \in \mathcal{G}}$ (see the experiments in Section 2.6).

In particular, the weights have to take into account that some variables belonging to several overlapping groups are penalized multiple times.

However, it is possible to keep the benefit of overlapping groups whilst limiting their side effects, by taking up the idea of support intersection (Bach, 2008c; Meinshausen and Bühlmann, 2010). First introduced to stabilize the set of variables recovered by the Lasso, we reuse this technique in a different context, based on the fact that $\mathcal{Z}$ is closed under union.

If we deal with the same sets of groups as those considered in Section 2.3.5, it is natural to rewrite $\mathcal{G}$ as $\bigcup_{\theta\in\Theta}\mathcal{G}_\theta$, where $\Theta$ is the set of the orientations of the groups in $\mathcal{G}$ (for more details, see Section 2.3.5). Let us denote by $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}^\theta$ the solutions of Eq. (2.5), where the regularization term $\Omega$ is respectively defined by the groups in $\mathcal{G}$ and by the groups [6] in $\mathcal{G}_\theta$.

The main point is that, since $\mathcal{P}$ is closed under intersection, the two procedures described below actually lead to the same set of allowed nonzero patterns:

a) Simply considering the nonzero pattern of $\hat{\mathbf{w}}$.

b) Taking the *intersection* of the nonzero patterns obtained for each $\hat{\mathbf{w}}^\theta$, $\theta$ in $\Theta$.

With the latter procedure, although the learning of several models $\hat{\mathbf{w}}^\theta$ is required (a number of times equals to the number of orientations considered, e.g., 2 for the sequence, 4 for the rectangular groups and more generally $|\Theta|$ times), each of those learning tasks involves a smaller number of groups (that is, just the ones belonging to $\mathcal{G}_\theta$). In addition, this procedure is a *variable selection* technique that therefore needs a second step for estimating the loadings (restricted to the selected nonzero pattern). In the experiments, we follow Bach (2008c) and we use an ordinary least squares (OLS). The simulations of Section 2.6 will show the benefits of this variable selection approach.

## 2.5 Pattern Consistency

In this section, we analyze the model consistency of the solution of Eq. (2.2) for the square loss. Considering the set of nonzero patterns $\mathcal{P}$ derived in Section 2.3, we can only hope to estimate the correct hull of the generating sparsity pattern, since Theorem 1 states that other patterns occur with zero probability. We derive necessary and sufficient conditions for model consistency in a low-dimensional setting, and then consider a high-dimensional result.

We consider the square loss and a fixed-design analysis (i.e., $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed). The extension of the following consistency results to other loss functions is beyond the scope of the chapter (see for instance Bach, 2009). We assume that for all $i \in [\![1; n]\!]$, $y_i = \mathbf{x}_i^\top \mathbf{w}^\star + \boldsymbol{\varepsilon}_i$ where the vector $\boldsymbol{\varepsilon}$ is an i.i.d. vector with Gaussian distributions with mean zero and variance $\sigma^2 > 0$, and $\mathbf{w}^\star \in \mathbb{R}^p$ is the population sparse vector; we denote by $\mathrm{J}^\star$ the $\mathcal{G}$-adapted hull of its nonzero pattern. Note that estimating the $\mathcal{G}$-adapted hull of $\mathbf{w}^\star$ is equivalent to estimating the nonzero pattern of $\mathbf{w}^\star$ if and only if this nonzero pattern belongs to $\mathcal{P}$. This happens when our prior information has led us to consider

6. To be more precise, in order to regularize every variable, we add the full group $[\![1; p]\!]$ to $\mathcal{G}_\theta$, which does not modify $\mathcal{P}$.

an appropriate set of groups $\mathcal{G}$. Conversely, if $\mathcal{G}$ is misspecified, recovering the hull of the nonzero pattern of $\mathbf{w}$ may be irrelevant, which is for instance the case if $\mathbf{w}^\star = \binom{\mathbf{w}_1^\star}{0} \in \mathbb{R}^2$ and $\mathcal{G} = \{\{1\}, \{1,2\}\}$. Finding the appropriate structure of $\mathcal{G}$ *directly from the data* would therefore be interesting future work.

### 2.5.1 Consistency Condition

We begin with the low-dimensional setting where $n$ is tending to infinity with $p$ *fixed*. In addition, we also assume that the design is *fixed* and that the Gram matrix $\mathbf{Q} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ is invertible with positive-definite (i.e., invertible) limit: $\lim_{n\to\infty} \mathbf{Q} = \mathbf{Q}^\star \succ \mathbf{0}$. In this setting, the noise is the only source of randomness. We denote by $\mathbf{r}^\star_{\mathrm{J}^\star}$ the vector defined as

$$\forall j \in \mathrm{J}^\star,\ \mathbf{r}_j^\star = \mathbf{w}_j^\star \Bigg( \sum_{g \in \mathcal{G}_{\mathrm{J}^\star}, g \ni j} (\omega_j^g)^2 \|\omega^g \circ \mathbf{w}^\star\|_2^{-1} \Bigg).$$

In the Lasso and group Lasso setting, the vector $\mathbf{r}^\star_{\mathrm{J}^\star}$ is respectively the sign vector $\mathrm{sign}(\mathbf{w}^\star_{\mathrm{J}})$ and the vector defined by the blocks $(\frac{\mathbf{w}_g}{\|\mathbf{w}_g\|_2})_{g\in\mathcal{G}^\star_{\mathrm{J}}}$.

We define $\Omega^c_{\mathrm{J}^\star}(\mathbf{w}_{[\mathrm{J}^\star]^c}) = \sum_{g\in(\mathcal{G}^\star_{\mathrm{J}})^c} \|\omega^g_{[\mathrm{J}^\star]^c} \circ \mathbf{w}_{[\mathrm{J}^\star]^c}\|_2$ (which is the norm composed of inactive groups) with its dual norm $(\Omega^c_{\mathrm{J}^\star})^*$; note the difference with the norm reduced to $[\mathrm{J}^\star]^c$, defined as $\Omega_{[\mathrm{J}^\star]^c}(\mathbf{w}_{[\mathrm{J}^\star]^c}) = \sum_{g\in\mathcal{G}} \|\omega^g_{[\mathrm{J}^\star]^c} \circ \mathbf{w}_{[\mathrm{J}^\star]^c}\|_2$.

The following Theorem gives the sufficient and necessary conditions under which the hull of the generating pattern is consistently estimated. Those conditions naturally extend the results of Zhao and Yu (2006) and Bach (2008b) for the Lasso and the group Lasso respectively (see proof in Appendix A.1.6).

**Theorem 2** (Consistency condition)
*Assume $\mu \to 0$, $\mu\sqrt{n} \to \infty$ in Eq. (2.2). If the hull is consistently estimated, then $(\Omega^c_{\mathrm{J}^\star})^*\Big[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}[\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}]^{-1}\mathbf{r}^\star_{\mathrm{J}^\star}\Big] \leq 1$. Conversely, if $(\Omega^c_{\mathrm{J}^\star})^*\Big[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}[\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}]^{-1}\mathbf{r}^\star_{\mathrm{J}^\star}\Big] < 1$, then the hull is consistently estimated, i.e.,*

$$\Pr\Big(\{j \in [\![1;p]\!], \hat{\mathbf{w}}_j \neq 0\} = \mathrm{J}^\star\Big) \underset{n\to+\infty}{\longrightarrow} 1.$$

The two previous propositions bring into play the dual norm $(\Omega^c_{\mathrm{J}^\star})^*$ that we cannot compute in closed form, but requires to solve an optimization problem as complex as the initial problem in Eq. (2.5). However, we can prove bounds similar to those obtained in Propositions 7 and 8 for the necessary and sufficient conditions.

**Comparison with the Lasso and group Lasso.** For the $\ell_1$-norm, our two bounds lead to the usual consistency conditions for the Lasso, i.e., the quantity

$$\|\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}[\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}]^{-1}\mathrm{sign}(\mathbf{w}^\star_{\mathrm{J}})\|_\infty$$

must be less or strictly less than one. Similarly, when $\mathcal{G}$ defines a partition of $[\![1;p]\!]$ and if all the weights equal one, our two bounds lead in turn to the consistency conditions

for the group Lasso, i.e., the quantity

$$\max_{g \in (\mathcal{G}_J^\star)^c} \left\| \mathbf{Q}^\star_{g\,\mathrm{Hull}(J^\star)} [\mathbf{Q}^\star_{\mathrm{Hull}(J^\star)\mathrm{Hull}(J^\star)}]^{-1} \left[ \frac{\mathbf{w}^\star_g}{\|\mathbf{w}^\star_g\|_2} \right]_{g \in \mathcal{G}_J^\star} \right\|_2$$

must be less or strictly less than one.

### 2.5.2 High-Dimensional Analysis

We prove a high-dimensional variable consistency result (see proof in Appendix A.1.7) that extends the corresponding result for the Lasso (Zhao and Yu, 2006; Wainwright, 2009), by assuming that the consistency condition in Theorem 2 is satisfied.

**Theorem 3**
*Assume that* $\mathbf{Q}$ *has unit diagonal,* $\kappa = \lambda_{\min}(\mathbf{Q}_{J^\star J^\star}) > 0$ *and* $(\Omega^c_{J^\star})^* [\mathbf{Q}_{[J^\star]^c J^\star} \mathbf{Q}^{-1}_{J^\star J^\star} \mathbf{r}^\star_{J^\star}] < 1 - \tau$*, for some* $\tau > 0$*. If* $\tau \mu \sqrt{n} \geq \sigma C_3(\mathcal{G}, J^\star)$*, and* $\mu |J^\star|^{1/2} \leq C_4(\mathcal{G}, J^\star)$*, then the probability of incorrect hull selection is upper bounded by:*

$$\exp\left( -\frac{n\mu^2\tau^2 C_1(\mathcal{G}, J^\star)}{2\sigma^2} \right) + 2|J| \exp\left( -\frac{n C_2(\mathcal{G}, J^\star)}{2|J^\star|\sigma^2} \right),$$

*where* $C_1(\mathcal{G}, J^\star)$*,* $C_2(\mathcal{G}, J^\star)$*,* $C_3(\mathcal{G}, J^\star)$ *and* $C_4(\mathcal{G}, J^\star)$ *are constants defined in Appendix A.1.7, which essentially depend on the groups, the smallest nonzero coefficient of* $\mathbf{w}^\star$ *and how close the support* $\{j \in J^\star : \mathbf{w}^\star_j \neq 0\}$ *of* $\mathbf{w}^\star$ *is to its hull* $J^\star$*, that is the relevance of the prior information encoded by* $\mathcal{G}$*.*

In the Lasso case, we have $C_1(\mathcal{G}, J^\star) = O(1)$, $C_2(\mathcal{G}, J^\star) = O(|J^\star|^{-2})$, $C_3(\mathcal{G}, J^\star) = O((\log p)^{1/2})$ and $C_4(\mathcal{G}, J^\star) = O(|J^\star|^{-1})$, leading to the usual scaling $n \approx \log p$ and $\mu \approx \sigma(\log p/n)^{1/2}$.

We can also give the scaling of these constants in simple settings where groups overlap. For instance, let us consider that the variables are organized in a sequence (see Figure 2.4). Let us further assume that the weights $(\omega^g)_{g \in \mathcal{G}}$ satisfy the following two properties:

a) The weights take into account the overlaps, that is,

$$\omega^g_j = \beta(|\{h \in \mathcal{G}\,;\, h \ni j,\ h \subset g \text{ and } h \neq g\}|),$$

with $t \mapsto \beta(t) \in (0, 1]$ a non-increasing function such that $\beta(0) = 1$,

b) The term

$$\max_{j \in [\![1;p]\!]} \sum_{g \ni j, g \in \mathcal{G}} \omega^g_j$$

is upper bounded by a constant $\mathcal{K}$ independent of $p$.

Note that we consider such weights in the experiments (see Section 2.6). Based on these assumptions, some algebra directly leads to

$$\|\mathbf{u}\|_1 \leq \Omega(\mathbf{u}) \leq 2\mathcal{K}\|\mathbf{u}\|_1, \quad \text{for all } \mathbf{u} \in \mathbb{R}^p.$$

We thus obtain a scaling similar to the Lasso (with, *in addition*, a control of the allowed nonzero patterns). With stronger assumptions on the possible positions of $J^\star$, we may have better scalings, but these are problem-dependent (a careful analysis of the group-dependent constants would still be needed in all cases).

## 2.6 Experiments

In this section, we carry out several experiments to illustrate the behavior of the sparsity-inducing norm $\Omega$. We denote by *Structured-lasso*, or simply *Slasso*, the models regularized by the norm $\Omega$. In addition, the procedure (introduced in Section 2.4.3) that consists in intersecting the nonzero patterns obtained for different models of Slasso will be referred to as *Intersected Structured-lasso*, or simply *ISlasso*.

Throughout the experiments, we consider noisy linear models $\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \varepsilon$, where $\mathbf{w} \in \mathbb{R}^p$ is the generating loading vector and $\boldsymbol{\varepsilon}$ is a standard Gaussian noise vector with its variance set to satisfy $\|\mathbf{X}\mathbf{w}\|_2 = 3\|\boldsymbol{\varepsilon}\|_2$. This consequently leads to a fixed signal-to-noise ratio. We assume that the vector $\mathbf{w}$ is sparse, i.e., it has only a few nonzero components, that is, $|J^\star| \ll p$. We further assume that these nonzero components are either organized on a sequence or on a two-dimensional grid (see Figure 2.9). Moreover, we consider sets of groups $\mathcal{G}$ such as those presented in Section 2.3.5. We also consider different choices for the weights $(\omega^g)_{g\in\mathcal{G}}$ that we denote by **(W1)**, **(W2)** and **(W3)** (we will keep this notation throughout the following experiments):

**(W1)**: Uniform weights, $\omega_j^g = 1$,

**(W2)**: Weights depending on the size of the groups, $\omega_j^g = |g|^{-2}$,

**(W3)**: Weights for overlapping groups, $\omega_j^g = \rho^{|\{h\in\mathcal{G}\,;\,h\ni j,\ h\subset g \text{ and } h\neq g\}|}$, for some $\rho \in (0,1)$.

For each orientation in $\mathcal{G}$, the third type of weights **(W3)** aims at reducing the unbalance caused by the overlapping groups. Specifically, given a group $g \in \mathcal{G}$ and a variable $j \in g$, the corresponding weight $\omega_j^g$ is all the more small as the variable $j$ already belongs to other groups with the same orientation. Unless otherwise specified, we use the third type of weights **(W3)** with $\rho = 0.5$. In the following experiments, the loadings $\mathbf{w}_J^\star$, as well as the design matrices, are generated from a standard Gaussian distribution with identity covariance matrix. The positions of $J^\star$ are also random and are uniformly drawn.

### 2.6.1 Consistent hull estimation

We first illustrate Theorem 2 that establishes necessary and sufficient conditions for consistent hull estimation. To this end, we compute the probability of correct hull estimation when we consider diamond-shaped generating patterns of $|J^\star| = 24$ variables on a $20\times20$-dimensional grid (see Figure 2.9h). Specifically, we generate 500 covariance matrices $\mathbf{Q}^\star$ distributed according to a Wishart distribution with $\delta$ degrees of freedom,

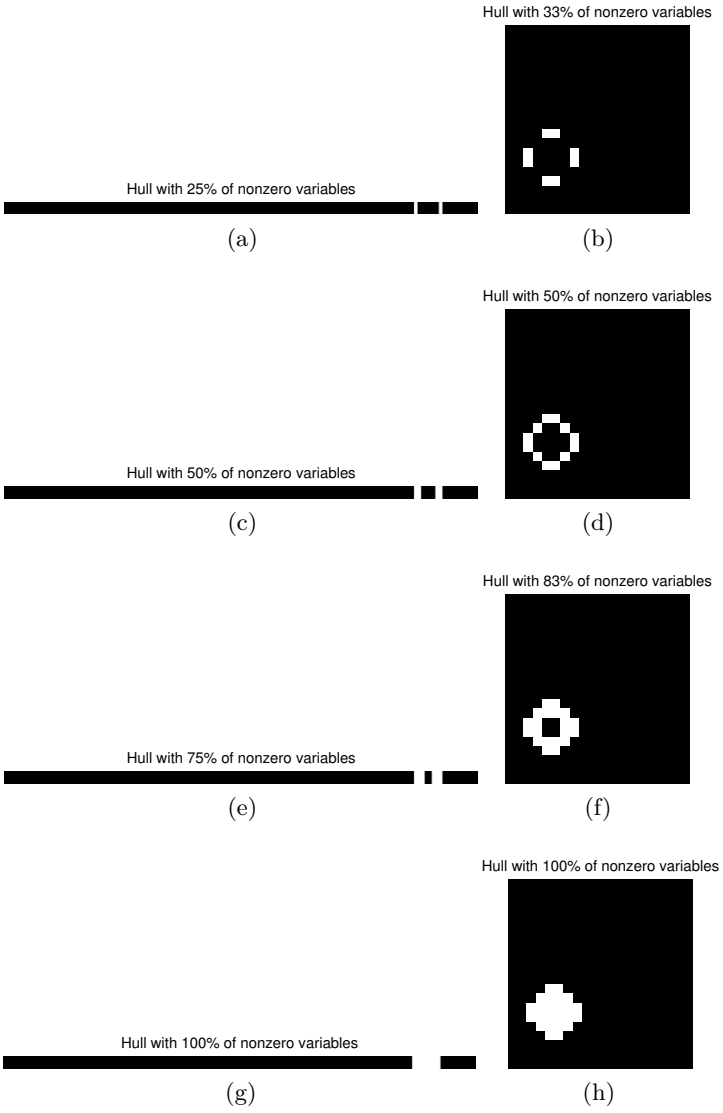



Figure 2.9: Examples of generating patterns (the zero variables are represented in black, while the nonzero ones are in white): (Left column, in white) generating patterns that are used for the experiments on 400-dimensional sequences; those patterns all form the same hull of 24 variables, i.e., the contiguous sequence in (g). (Right column, in white) generating patterns that we use for the 20×20-dimensional grid experiments; again, those patterns all form the same hull of 24 variables, i.e., the diamond-shaped convex in (h). The positions of these generating patterns are randomly selected during the experiments. For the grid setting, the hull is defined based on the set of groups that are half-planes, with orientations that are multiple of $\pi/4$ (see Section 2.3.5).

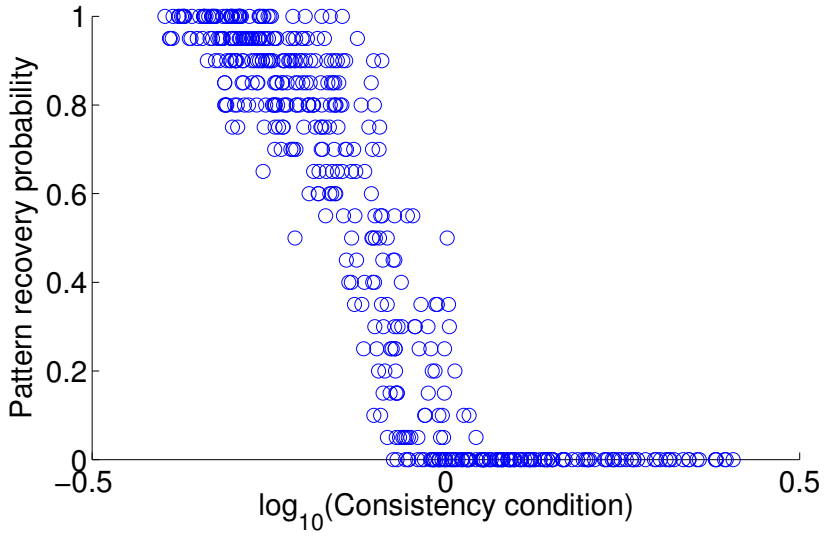



Figure 2.10: Consitent hull estimation: probability of correct hull estimation versus the consistency condition $(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}[\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}]^{-1}\mathbf{r}^\star_{\mathrm{J}^\star}]$. The transition appears at zero, in good agreement with Theorem 2.

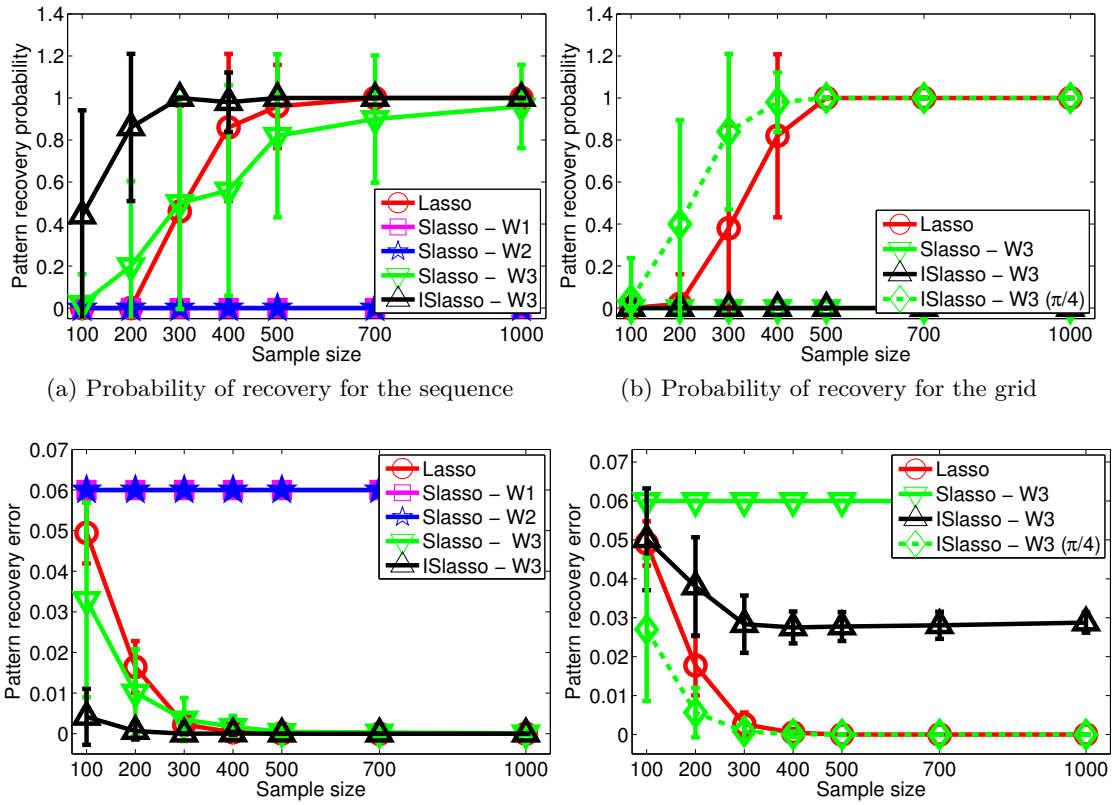
where $\delta$ is uniformly drawn in $\{1, 2, \ldots, 10p\}$.[7] The diagonal terms of $\mathbf{Q}$ are then re-normalized to one. For each of these covariance matrices, we compute an entire regularization path based on one realization of $\{\mathrm{J}^\star, \mathbf{w}^\star, \mathbf{X}, \boldsymbol{\varepsilon}\}$, with $n = 3000$ samples. The quantities $\{\mathrm{J}^\star, \mathbf{w}, \varepsilon\}$ are generated as described previously, while the $n$ rows of $\mathbf{X}$ are gaussian with covariance $\mathbf{Q}^\star$. After repeating 20 times this computation for each $\mathbf{Q}^\star$, we eventually report in Figure 2.10 the probabilites of correct hull estimation versus the consistency condition $(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}[\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}]^{-1}\mathbf{r}^\star_{\mathrm{J}^\star}]$. In good agreement with Theorem 2, comparing $(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}[\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}]^{-1}\mathbf{r}^\star_{\mathrm{J}^\star}]$ to 1 determines whether the hull is consistently estimated.

### 2.6.2 Structured variable selection

We show in this experiment that the prior information we put through the norm $\Omega$ improves upon the ability of the model to recover spatially structured nonzero patterns. We are looking at two situations where we can express such a prior through $\Omega$, namely (1) the selection of a contiguous pattern on a sequence (see Figure 2.9g) and (2) the selection of a convex pattern on a grid (see Figure 2.9h).

In what follows, we consider $p = 400$ variables with generating patterns $\mathbf{w}^\star$ whose hulls are composed of $|\mathrm{J}^\star| = 24$ variables. For different sample sizes $n$ ranges in the set $\{100, 200, 300, 400, 500, 700, 1000\}$, we consider the probabilites of correct recovery and

7. We have empirically observed that this choice of degrees of freedom enables to cover well the consistency transition regime around zero in Figure 2.10.




(a) Probability of recovery for the sequence

(b) Probability of recovery for the grid

(c) Distance to the true pattern for the sequence

(d) Distance to the true pattern for the grid

Figure 2.11: For different sample sizes, the probabilites of correct recovery and the (normalized) Hamming distance to the true nonzero patterns are displayed. In the grid case, two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). The points and the error bars on the curves respectively represent the mean and the standard deviation, based on 50 random settings $\{\mathrm{J}^\star, \mathbf{w}^\star, \mathbf{X}, \varepsilon\}$.

the (normalized) Hamming distance to the true nonzero patterns. For the realization of a random setting $\{\mathrm{J}^\star, \mathbf{w}^\star, \mathbf{X}, \varepsilon\}$, we compute an entire regularization path over which we collect the closest Hamming distance to the true nonzero pattern and whether it has been exactly recovered for some $\mu$. After repeating 50 times this computation for each sample size $n$, we report the results in Figure 2.11.

First and foremost, the simulations highlight how important the weights $(\omega^g)_{g\in\mathcal{G}}$ are. In particular, the uniform **(W1)** and size-dependent weights **(W2)** perform poorly since they do not take into account the overlapping groups. The models learned with such weights do not manage to recover the correct nonzero patterns (in that case, the best model found on the path corresponds to the empty solution, with a normalized Hamming distance of $|\mathrm{J}^\star|/p = 0.06$—see Figure 2.11c).

Although groups that moderately overlap do help (e.g., see Slasso with the weights

**(W3)** on Figure 2.11c), it remains delicate to handle groups with many overlaps, even with an appropriate choice of $(\omega^g)_{g\in\mathcal{G}}$ (e.g., see Slasso on Figure 2.11d). The ISlasso procedure does not suffer from this issue since it reduces the number of overlaps whilst keeping the desirable effects of overlapping groups. Another way to yield a better level of sparsity, even with many overlaps, would be to consider non-convex alternatives to $\Omega$ (see, e.g., Jenatton et al., 2010b). Moreover, adding the $\pm\pi/4$-groups to the rectangular groups enables to recover a nonzero pattern closer to the generating pattern. This is illustrated on Figure 2.11d where the error of ISlasso with only rectangular groups (in black) corresponds to the selection of the smallest rectangular box around the generating pattern.

### 2.6.3 Prediction error and relevance of the structured prior

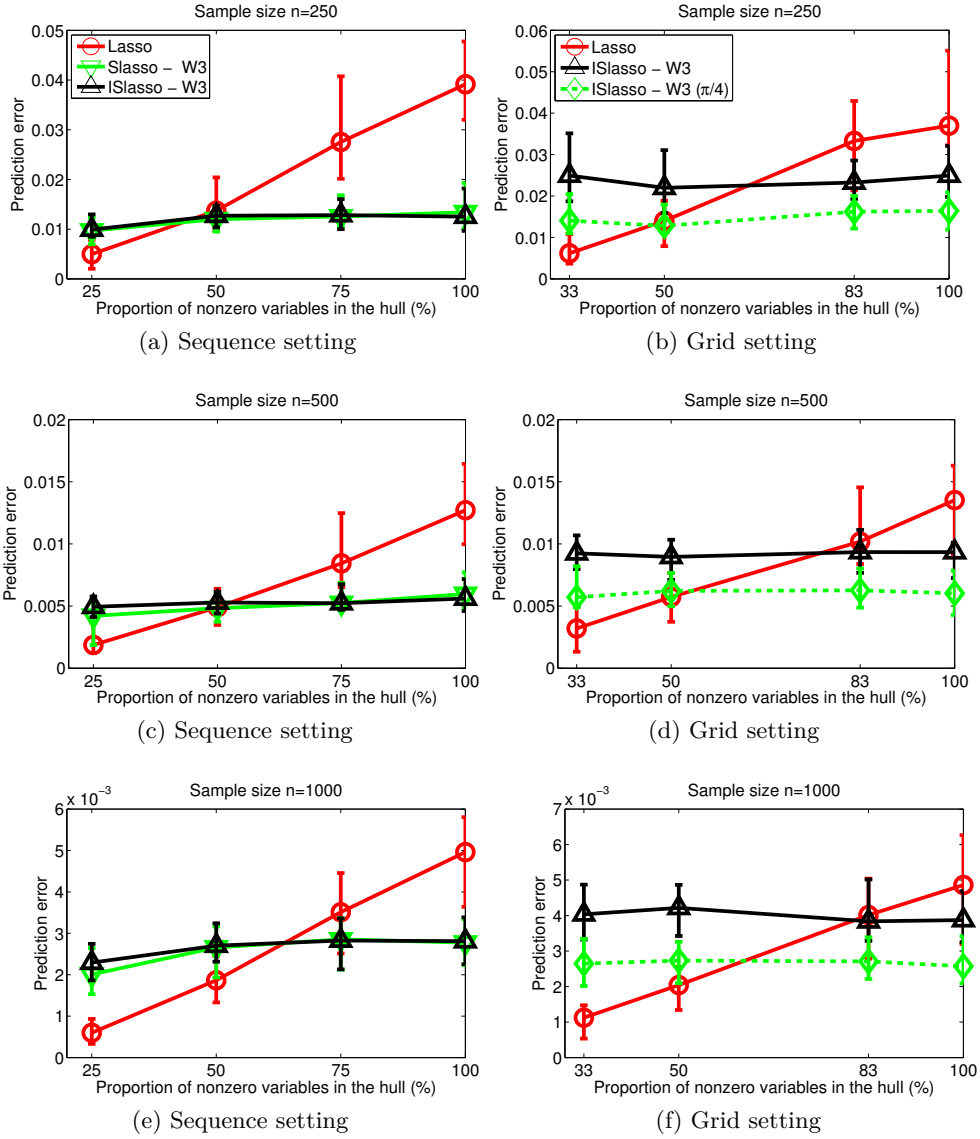
In the next simulation, we start from the same setting as Section 2.6.2 where we additionally evaluate the relevance of the contiguous (or convex) prior by varying the number of zero variables that are contained in the hull (see Figure 2.9). We then compute the prediction error for different sample sizes $n \in \{250, 500, 1000\}$. The prediction error is understood here as $\|\mathbf{X}^{\text{test}}(\mathbf{w}^\star - \hat{\mathbf{w}})\|_2^2 / \|\mathbf{X}^{\text{test}}\mathbf{w}^\star\|_2^2$, where $\hat{\mathbf{w}}$ denotes the OLS estimate, performed on the nonzero pattern found by the model considered (i.e., either Lasso, Slasso or ISlasso). The regularization parameter is chosen by 5-fold cross-validation and the test set consists of 500 samples. For each value of $n$, we display on Figure 2.12 the median errors over 50 random settings $\{\mathrm{J}^\star, \mathbf{w}^\star, \mathbf{X}, \boldsymbol{\varepsilon}\}$, for respectively the sequence and the grid. Note that we have dropped for clarity the models that performed already poorly in Section 2.6.2.

The experiments show that if the prior about the generating pattern is relevant, then our structured approach performs better that the standard Lasso. Indeed, as soon as the hull of the generating pattern does not contain too many zero variables, Slasso/ISlasso outperform Lasso. In fact, the sample complexity of the Lasso depends on the number of nonzero variables in $\mathbf{w}^\star$ (Wainwright, 2009) as opposed to the size of the hull for Slasso/ISlasso. This also explains why the error for Slasso/ISlasso is almost constant with respect to the number of nonzero variables (since the hull has a constant size).

### 2.6.4 Active set algorithm

We finally focus on the active set algorithm (see Section 2.4) and compare its time complexity to the SOCP solver when we are looking for a sparse structured target. More precisely, for a fixed level of sparsity $|\mathrm{J}^\star| = 24$ and a fixed number of observations $n = 3500$, we analyze the complexity with respect to the number of variables $p$ that varies in $\{100, 225, 400, 900, 1600, 2500\}$. We consider the same experimental protocol as above except that we display the median CPU time based only [8] on 5 random settings $\{\mathrm{J}^\star, \mathbf{w}, X, \varepsilon\}$. We assume that we have a rough idea of the level of sparsity of the

8. Note that it already corresponds to several hundreds of runs for both the SOCP and the active set algorithms since we compute a 5-fold cross-validation for each regularization parameter of the (approximate) regularization path.




(a) Sequence setting (b) Grid setting

(c) Sequence setting (d) Grid setting

(e) Sequence setting (f) Grid setting

Figure 2.12: For the sample size $n \in \{250, 500, 1000\}$, we plot the prediction error versus the proportion of nonzero variables in the hull of the generating pattern. In the grid case, two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). The points, the lower and upper error bars on the curves respectively represent the median, the first and third quartile, based on 50 random settings $\{\mathrm{J}^\star, \mathbf{w}^\star, \mathbf{X}, \varepsilon\}$.

true vector and we set the stopping criterion $s = 4|\mathrm{J}^\star|$ (see Algorithm 3), which is a rather conservative choice. We show on Figure 2.13 that we considerably lower the

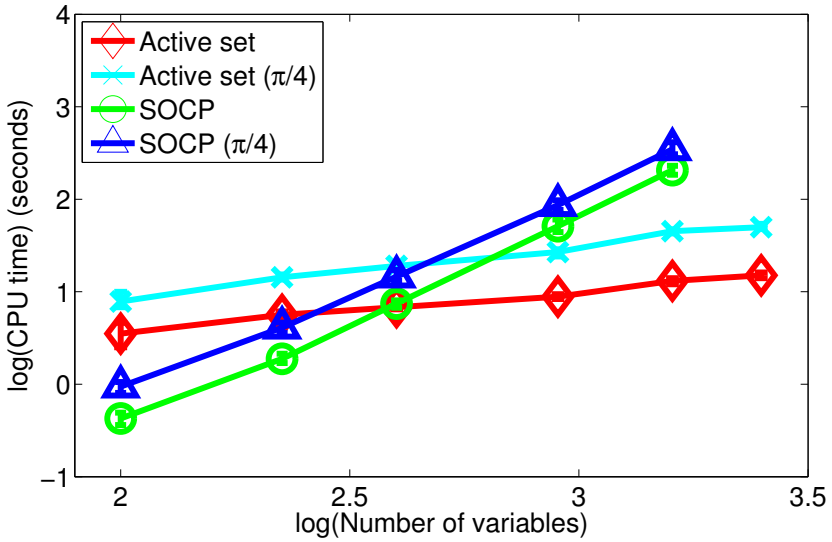



Figure 2.13: Computational benefit of the active set algorithm: CPU time (in seconds) versus the number of variables $p$, displayed in log-log scale. The points, the lower and upper error bars on the curves respectively represent the median, the first and third quartile. Two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). Due to the computational burden, we could not obtain the SOCP's results for $p = 2500$.

computational cost for the same level of performance [9]. As predicted by the complexity analysis of the active set algorithm (see the end of Section 2.4), considering the set of rectangular groups with or without the $\pm\pi/4$-groups results in the same complexity (up to constant terms). We empirically obtain an average complexity of $\approx O(p^{2.13})$ for the SOCP solver and of $\approx O(p^{0.45})$ for the active set algorithm.

Not surprisingly, for small values of $p$, the SOCP solver is faster than the active set algorithm, since the latter has to check its optimality by computing necessary and sufficient conditions (see Algorithm 3 and the discussion in the algorithmic complexity paragraph of Section 2.4).

## 2.7 Conclusion

We have shown how to incorporate prior knowledge on the form of nonzero patterns for linear supervised learning. Our solution relies on a regularizing term which linearly combines $\ell_2$-norms of possibly overlapping groups of variables. Our framework brings into play intersection-closed families of nonzero patterns, such as all rectangles on a two-

9. We have not displayed this second figure that is just the superposition of the error curves for the SOCP and the active set algorithms.

dimensional grid. We have studied the design of these groups, efficient algorithms and theoretical guarantees of the structured sparsity-inducing method. Our experiments have shown to which extent our model leads to better prediction, depending on the relevance of the prior information.

A natural extension to this work is to consider bootstrapping since this may improve theoretical guarantees and result in better variable selection (Bach, 2008c; Meinshausen and Bühlmann, 2010). In order to deal with broader families of (non)zero patterns, it would be interesting to combine our approach with recent work on structured sparsity: for instance, Baraniuk et al. (2010); Jacob et al. (2009) consider union-closed collections of nonzero patterns, He and Carin (2009) exploit structure through a Bayesian prior while Huang et al. (2009) handle non-convex penalties based on information-theoretic criteria.

More generally, our regularization scheme could also be used for various learning tasks, as soon as prior knowledge on the structure of the sparse representation is available, e.g., for multiple kernel learning (Micchelli and Pontil, 2006), multi-task learning (Argyriou et al., 2008; Obozinski et al., 2009; Kim and Xing, 2010) and sparse matrix factorization problems (Mairal et al., 2010a; Jenatton et al., 2010b, 2011c).

Finally, although we have mostly explored in this chapter the algorithmic and theoretical issues related to these norms, this type of prior knowledge is of clear interest for the spatially and temporally structured data typical in bioinformatics (Kim and Xing, 2010), computer vision (Jenatton et al., 2010b; Mairal et al., 2010b) and neuroscience applications (see, e.g., Varoquaux et al., 2010c).

# 3

# Structured Sparse Principal Component Analysis

**Chapter abstract:** We present an extension of sparse principal component analysis (PCA), or sparse dictionary learning, where the sparsity patterns of all dictionary elements are structured and constrained to belong to a prespecified set of shapes. This *structured sparse PCA* is based on a structured regularization recently introduced by Jenatton et al. (2011a). While classical sparse priors only deal with *cardinality*, the regularization we use encodes higher-order information about the data. We propose an efficient and simple optimization procedure to solve this problem. Experiments with two practical tasks, the denoising of sparse structured signals and face recognition, demonstrate the benefits of the proposed structured approach over unstructured approaches.

The material of this chapter is based on the following publication:

R. Jenatton, G. Obozinski, F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2010

## 3.1 Introduction

Principal component analysis (PCA) is an essential tool for data analysis and unsupervised dimensionality reduction. Its goal is to find, among linear combinations of the data variables, a sequence of orthogonal factors that most efficiently explain the variance of the observations.

One of PCA's main shortcomings is that, even if it finds a small number of important factors, the factor themselves typically involve all original variables. In the last decade, several alternatives to PCA which find sparse and potentially interpretable factors have been proposed, notably non-negative matrix factorization (NMF) (Lee and Seung, 1999) and sparse PCA (SPCA) (Jolliffe et al., 2003; Zou et al., 2006; Zass and Shashua, 2007; Witten et al., 2009).

However, in many applications, only constraining the size of the factors does not seem appropriate because the considered factors are not only expected to be sparse but also to have a certain structure. In fact, the popularity of NMF for face image analysis owes essentially to the fact that the method happens to retrieve sets of variables that are partly localized on the face and capture some features or parts of the face which

seem intuitively meaningful given our a priori. We might therefore gain in the quality of the factors induced by enforcing directly this a priori in the matrix factorization constraints. More generally, it would be desirable to encode higher-order information about the supports that reflects the *structure* of the data. For example, in computer vision, features associated to the pixels of an image are naturally organized on a grid and the supports of factors explaining the variability of images could be expected to be localized, connected or have some other regularity with respect to that grid. Similarly, in genomics, factors explaining the gene expression patterns observed on a microarray could be expected to involve groups of genes corresponding to biological pathways or set of genes that are neighbors in a protein-protein interaction network.

Recent research on structured sparsity has highlighted the benefit of exploiting such structure in the context of regression and classification (Jenatton et al., 2011a; Jacob et al., 2009; Huang et al., 2009), compressed sensing (Baraniuk et al., 2010), as well as within Bayesian frameworks (He and Carin, 2009). In particular, Jenatton et al. (2011a) show that, given any intersection-closed family of patterns $\mathcal{P}$ of variables, such as all the rectangles on a 2-dimensional grid of variables, it is possible to build an ad hoc regularization norm $\Omega$ that enforces that the support of the solution of a least-squares regression regularized by $\Omega$ belongs to the family $\mathcal{P}$.

Capitalizing on these results, we aim in this chapter to go beyond sparse PCA and propose *structured sparse PCA* (SSPCA), which explains the variance of the data by factors that are not only sparse but also respect some a priori structural constraints deemed relevant to model the data at hand. We show how slight variants of the regularization term from Jenatton et al. (2011a) can be used successfully to yield a structured and sparse formulation of principal component analysis for which we propose a simple and efficient optimization scheme.

The rest of the chapter is organized as follows: Section 3.2 casts the SSPCA problem in the dictionary learning framework, summarizes the regularization considered by Jenatton et al. (2011a) and its essential properties, and presents some simple variants which are more effective in the context of PCA. Section 3.3 is dedicated to our optimization scheme for solving SSPCA. Our experiments in Section 3.4 illustrate the benefits of our approach through the denoising of sparse structured synthetic signals and an application to face recognition.

**Notations.** For any matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$, we write $\mathbf{y}^j \in \mathbb{R}^n$ for the $j$-th column of $\mathbf{Y}$, while we write $\mathbf{Y}_i \in \mathbb{R}^p$ for its $i$-th row. We refer to the set $\{j \in [\![1; p]\!] \, ; \, \mathbf{w}_j \neq 0\}$ as the *support*, or *nonzero pattern* of the vector $\mathbf{w} \in \mathbb{R}^p$. For any finite set $A$ with cardinality $|A|$, we also define the $|A|$-tuple $(\mathbf{y}^a)_{a \in A} \in \mathbb{R}^{p \times |A|}$ as the collection of $p$-dimensional vectors $\mathbf{y}^a$ indexed by the elements of $A$. Furthermore, for two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^p$, we denote by $\mathbf{x} \circ \mathbf{y} = (\mathbf{x}_1 \mathbf{y}_1, \ldots, \mathbf{x}_p \mathbf{y}_p)^\top \in \mathbb{R}^p$ the elementwise product of $\mathbf{x}$ and $y$. Finally, we extend $b \mapsto \frac{a}{b}$ by continuity in zero with $\frac{a}{0} = \infty$ if $a \neq 0$ and 0 otherwise.

## 3.2 Problem Statement

It is useful to distinguish two conceptually different interpretations of PCA. In terms of *analysis*, PCA sequentially projects the data on subspaces that explain the largest fraction of the variance of the data. In terms of *synthesis*, PCA finds a basis, or orthogonal dictionary, such that all signals observed admit decompositions with low reconstruction error. These two interpretations recover the same basis of principal components for PCA but lead to different formulations for *sparse* PCA. The *analysis* interpretation leads to sequential formulations (d'Aspremont et al., 2008; Moghaddam et al., 2006; Jolliffe et al., 2003) that consider components one at a time and perform a *deflation* of the covariance matrix at each step (see Mackey, 2009). The *synthesis* interpretation leads to non-convex global formulations (Zou et al., 2006; Mairal et al., 2010a; Moghaddam et al., 2006; Lee et al., 2007) which estimate simultaneously all principal components, often drop the orthogonality constraints, and are referred to as matrix factorization problems (Singh and Gordon, 2008) in machine learning, and dictionary learning in signal processing.

The approach we propose fits more naturally in the framework of dictionnary learning, whose terminology we now introduce.

### 3.2.1 Matrix Factorization and Dictionary Learning

Given a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ of $n$ columns corresponding to $n$ observations in $\mathbb{R}^m$, the dictionary learning problem is to find a matrix $\mathbf{D} \in \mathbb{R}^{m \times p}$, called the *dictionary*, such that each observation can be well approximated by a linear combination of the $p$ columns $(\mathbf{d}^k)_{k \in [\![1;p]\!]}$ of $\mathbf{D}$ called the *dictionary elements*. If $\mathbf{A} \in \mathbb{R}^{p \times n}$ is the matrix of the linear combination coefficients or *decomposition coefficients*, the matrix product $\mathbf{DA}$ is called a decomposition of $\mathbf{X}$.

Learning simultaneously the dictionary $\mathbf{D}$ and the decomposition $\mathbf{A}$ corresponds to a matrix factorization problem (see Witten et al., 2009, and reference therein). As formulated by Bach et al. (2008) or Witten et al. (2009), it is natural, when learning a decomposition, to penalize or constrain some norms or quasi-norms of $\mathbf{A}$ and $\mathbf{D}$, say $\Omega_{\mathbf{A}}$ and $\Omega_{\mathbf{D}}$ respectively, to encode prior information — typically sparsity — about the decomposition of $\mathbf{X}$. This can be written generally as

$$\min_{\substack{\mathbf{A} \in \mathbb{R}^{p \times n}, \\ \mathbf{D} \in \mathbb{R}^{m \times p}}} \frac{1}{2nm} \left\| \mathbf{X} - \mathbf{DA} \right\|_{\mathrm{F}}^2 + \lambda \sum_{k=1}^{p} \Omega_{\mathbf{D}}(\mathbf{d}^k), \text{ such that } \Omega_{\mathbf{A}}(\mathbf{A}_k) \leq 1, \ \forall k \in [\![1;p]\!], \quad (3.1)$$

where the regularization parameter $\lambda \geq 0$ controls to which extent the dictionary is regularized [1]. If we assume that both regularizations $\Omega_{\mathbf{A}}$ and $\Omega_{\mathbf{D}}$ are convex, problem (3.1) is convex with respect to $\mathbf{A}$ for fixed $\mathbf{D}$ and vice versa. It is however not *jointly* convex in the pair $(\mathbf{A}, \mathbf{D})$.

The formulation of sparse PCA considered by Lee et al. (2007) corresponds to a particular instance of this problem, where the dictionary elements are required to be sparse

1. From Bach et al. (2008), we know that our formulation is also equivalent to two other ones, penalized respectively by $\frac{\lambda}{2} \sum_{k=1}^{p} [\Omega_{\mathbf{D}}(\mathbf{d}^k)]^2 + [\Omega_{\mathbf{A}}(\mathbf{A}_k)]^2$ and $\lambda \sum_{k=1}^{p} \Omega_{\mathbf{D}}(\mathbf{d}^k) \Omega_{\mathbf{A}}(\mathbf{A}_k)$.

(without the orthogonality constraint $\mathbf{D}^\top\mathbf{D} = \mathbf{I}$). This can be achieved by penalizing the columns of $\mathbf{D}$ by a sparsity-inducing norm, such as the $\ell_1$-norm: $\Omega_{\mathbf{D}}(\mathbf{d}^k) = \|\mathbf{d}^k\|_1$. In the next section we consider a regularization $\Omega_{\mathbf{D}}$ which controls not only the sparsity but also the structure of the supports of dictionary elements.

### 3.2.2 Structured Sparsity-Inducing Norms

The work of Jenatton et al. (2011a) considered a norm which induces structured sparsity in the following sense: the solutions to a learning problem regularized by this norm have a sparse support which moreover belongs to a certain set of groups of variables. Interesting sets of possible supports include sets of variables forming rectangles when arranged on a grid and more generally convex subsets [2].

The framework of Jenatton et al. (2011a) can be summarized as follows: if we denote by $\mathcal{G}$ a subset of the power set of $[\![1;p]\!]$, such that $\bigcup_{g\in\mathcal{G}} g = [\![1;p]\!]$, we define the mixed $\ell_1/\ell_2$ norm $\Omega$ on a vector $\mathbf{y} \in \mathbb{R}^p$ as

$$\Omega(\mathbf{y}) = \sum_{g\in\mathcal{G}} \left\{ \sum_{j\in g} (\omega_j^g)^2 |\mathbf{y}_j|^2 \right\}^{\frac{1}{2}} = \sum_{g\in\mathcal{G}} \|\omega^g \circ \mathbf{y}\|_2,$$

where $(\omega^g)_{g\in\mathcal{G}} \in \mathbb{R}p \times |\mathcal{G}|$ is a $|\mathcal{G}|$-tuple of $p$-dimensional vectors such that $\omega_j^g > 0$ if $j \in g$ and $\omega_j^g = 0$ otherwise. This norm $\Omega$ linearly combines the $\ell_2$ norms of possibly overlapping groups of variables, with variables in each group being weighted by $(\omega^g)_{g\in\mathcal{G}}$. Note that a same variable $\mathbf{y}_j$ belonging to two different groups $g_1, g_2 \in \mathcal{G}$ is allowed to be weighted differently in $g_1$ and $g_2$ (by respectively $\omega_j^{g_1}$ and $\omega_j^{g_2}$).

For specific choices of $\mathcal{G}$, $\Omega$ leads to standard sparsity-inducing norms. For example, when $\mathcal{G}$ is the set of all singletons, $\Omega$ is the usual $\ell_1$ norm (assuming that all the weights are equal to one).

We focus on the case of a 2-dimensional grid where the set of groups $\mathcal{G}$ is the set of all horizontal and vertical half-spaces (see Figure 3.1 taken from Jenatton et al., 2011a). As proved by Jenatton et al. (2011a, Theorem 3.1), the $\ell_1/\ell_2$ norm $\Omega$ sets to zero some groups of variables $\|\omega^g \circ \mathbf{y}\|_2$, i.e., some entire horizontal and vertical half-spaces of the grid, and therefore induces rectangular nonzero patterns. Note that a larger set of convex patterns can be obtained by adding in $\mathcal{G}$ half-planes with other orientations. In practice, we use planes with angles that are multiples of $\frac{\pi}{4}$, which enables the nonzero patterns to have polygonal shapes with up to 8 faces.

Among sparsity inducing regularizations, the $\ell_1$ norm is often privileged since it is convex. However, so-called concave penalizations, such as penalization by an $\ell_\alpha$ quasi-norm, which are closer to the $\ell_0$ quasi-norm and penalize more aggressively small coefficients can be preferred, especially in a context where the unregularized problem, here dictionary learning is itself non convex. In light of recent work showing the advantages of addressing sparse problems through concave penalization (e.g., see Zou and Li, 2008),

2. Although we use the term *convex* informally here, it can however be made precise with the notion of convex subgraphs (Chung, 1997).

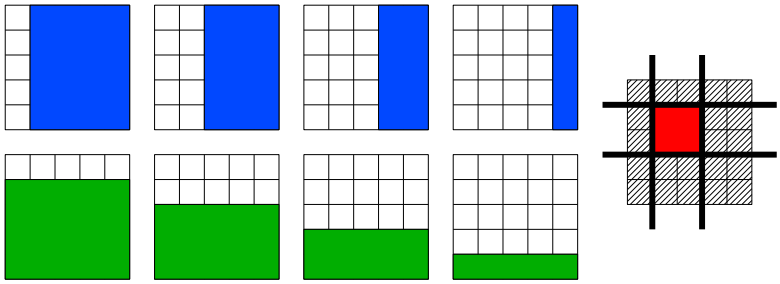


Figure 3.1: (Left) The set of blue and green groups with their (not displayed) complements to penalize to select rectangles. (Right) In red, an example of recovered pattern in this setting.

we therefore generalize $\Omega$ to a family of non-convex regularizers as follows: for $\alpha \in (0,1)$, we define the quasi-norm $\Omega^\alpha$ for all vectors $\mathbf{y} \in \mathbb{R}^p$ as

$$\Omega^\alpha(\mathbf{y}) = \left\{ \sum_{g \in \mathcal{G}} \|\omega^g \circ \mathbf{y}\|_2^\alpha \right\}^{\frac{1}{\alpha}} = \| \left( \|\omega^g \circ \mathbf{y}\|_2 \right)_{g \in \mathcal{G}} \|_\alpha,$$

where we denote by $\left(\|\omega^g \circ \mathbf{y}\|_2\right)_{g \in \mathcal{G}} \in \mathbb{R}^{1 \times |\mathcal{G}|}$ the $|\mathcal{G}|$-tuple composed of the different blocks $\|\omega^g \circ \mathbf{y}\|_2$. We thus replace the (convex) $\ell_1/\ell_2$ norm $\Omega$ by the (neither convex, nor concave) $\ell_\alpha/\ell_2$ quasi-norm $\Omega^\alpha$. While leading to the same set of (non)zero patterns, the $\ell_\alpha$ quasi-norm yields sparsity at the group level more aggressively.

## 3.3 Optimization

We consider the optimization of Eq. (3.1) where we use $\Omega_{\mathbf{D}} = \Omega^\alpha$ to regularize the dictionary $\mathbf{D}$. We discuss in Section 3.3.3 which norms $\Omega_{\mathbf{A}}$ we can handle in this optimization framework.

### 3.3.1 Formulation as a Sequence of Convex Problems

We now consider Eq. (3.1) where we take $\Omega_{\mathbf{D}}$ to be $\Omega^\alpha$, $\alpha \in (0,1)$, that is,

$$\min_{\substack{\mathbf{A} \in \mathbb{R}p \times n, \\ \mathbf{D} \in \mathbb{R}^{m \times p}}} \frac{1}{2nm} \left\| \mathbf{X} - \mathbf{D}\mathbf{A} \right\|_{\mathrm{F}}^2 + \lambda \sum_{k=1}^{p} \Omega^\alpha(\mathbf{d}^k), \text{ such that } \Omega_{\mathbf{A}}(\mathbf{A}_k) \leq 1, \ \forall k \in [\![1; p]\!], \quad (3.2)$$

Although the minimization problem in Eq. (3.2) is still convex in $\mathbf{A}$ for $\mathbf{D}$ fixed, the converse is not true anymore because of $\Omega^\alpha$. Indeed, the formulation in $\mathbf{D}$ is non-

differentiable and non-convex. To address this problem, we use the variational equality based on the following lemma that is related [3] to ideas from Micchelli and Pontil (2006):

**Lemma 1** (Variational formulation)
*Let $\alpha \in (0,2)$ and $\beta = \frac{\alpha}{2-\alpha}$. For any vector $\mathbf{y} \in \mathbb{R}^p$, we have the following equality*

$$\|\mathbf{y}\|_\alpha = \min_{\mathbf{z}\in\mathbb{R}^p_+} \Big[\frac{1}{2}\sum_{j=1}^p \frac{\mathbf{y}_j^2}{\mathbf{z}_j} + \frac{1}{2}\|\mathbf{z}\|_\beta\Big],$$

*and the minimum is uniquely attained for $\mathbf{z}_j = |\mathbf{y}_j|^{2-\alpha}\|\mathbf{y}\|_\alpha^{\alpha-1}$, $\forall j \in [\![1;p]\!]$.*

*Proof.* Let $\psi : \mathbf{z} \mapsto \sum_{j=1}^p \mathbf{y}_j^2 \mathbf{z}_j^{-1} + \|\mathbf{z}\|_\beta$ be the continuously differentiable function defined on $\mathbb{R}^p_+$. We have $\lim_{\|\mathbf{z}\|_\beta \to \infty} \psi(\mathbf{z}) = +\infty$ and $\lim_{\mathbf{z}_j \to 0} \psi(\mathbf{z}) = +\infty$ if $\mathbf{y}_j \neq 0$ (for $\mathbf{y}_j = 0$, note that $\min_{\mathbf{z}\in\mathbb{R}^p_+} \psi(\mathbf{z}) = \min_{\mathbf{z}\in\mathbb{R}^p_+, \mathbf{z}_j = 0} \psi(\mathbf{z})$). Thus, the infimum exists and it is attained. Taking the derivative with respect to $\mathbf{z}_j$ (for $\mathbf{z}_j > 0$) leads to the expression of the unique minimum, expression that is still correct for $\mathbf{z}_j = 0$. □

To reformulate problem (3.2), let us consider the $|\mathcal{G}|$-tuple $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}} \in \mathbb{R}^{p\times|\mathcal{G}|}$ of $r$-dimensional vectors $\boldsymbol{\eta}^g$ that satisfy for all $k \in [\![1;p]\!]$ and $g \in \mathcal{G}$, $\boldsymbol{\eta}^g_k \geq 0$. It follows from Lemma 1 that $2\sum_{k=1}^p \Omega^\alpha(\mathbf{d}^k)$ is equal to

$$\min_{(\boldsymbol{\eta}^g)_{G\in\mathcal{G}} \in \mathbb{R}^{p\times|\mathcal{G}|}_+} \sum_{k=1}^p \Big[\|(\boldsymbol{\eta}^g_k)_{g\in\mathcal{G}}\|_\beta + \sum_{g\in\mathcal{G}} \|\mathbf{d}^k \circ \omega^g\|_2^2 (\boldsymbol{\eta}^g_k)^{-1}\Big],$$

that can be rewritten in turn as

$$\min_{(\boldsymbol{\eta}^g)_{g\in\mathcal{G}} \in \mathbb{R}^{p\times|\mathcal{G}|}_+} \sum_{k=1}^p (\mathbf{d}^k)^\top \mathrm{Diag}(\boldsymbol{\zeta}^k)^{-1} \mathbf{d}^k + \|(\boldsymbol{\eta}^g_k)_{g\in\mathcal{G}}\|_\beta,$$

where we have introduced the matrix $\boldsymbol{\zeta} \in \mathbb{R}^{m\times p}$ defined by [4] $\boldsymbol{\zeta}_{jk} \triangleq \Big\{\sum_{\substack{g\in\mathcal{G},\\ g\ni j}} (\omega^g_j)^2 (\boldsymbol{\eta}^g_k)^{-1}\Big\}^{-1}$. This leads to the following formulation

$$\min_{\substack{\mathbf{A},\ \mathbf{D},\ \Omega_\mathbf{A}(\mathbf{A}_k)\leq 1\\ (\boldsymbol{\eta}^g)_{g\in\mathcal{G}} \in \mathbb{R}p\times|\mathcal{G}|_+}} \frac{1}{2nm}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_\mathrm{F}^2 + \frac{\lambda}{2}\sum_{k=1}^p \Big[(\mathbf{d}^k)^\top \mathrm{Diag}\Big(\boldsymbol{\zeta}^k\Big)^{-1} \mathbf{d}^k + \|(\boldsymbol{\eta}^g_k)_{g\in\mathcal{G}}\|_\beta\Big], \quad (3.3)$$

which is equivalent to Eq. (3.2) and which is now *quadratic* with respect to $\mathbf{D}$.

3. Note that we depart from Micchelli and Pontil (2006) who consider a quadratic upperbound on the *squared* norm. We prefer to remain in the standard dictionary learning framework where the penalization is not squared.

4. For the sake of clarity, we do not specify the dependence of $\boldsymbol{\zeta}$ on $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$.

### 3.3.2 Sharing Structure among Dictionary Elements

So far, the regularization quasi-norm $\Omega^\alpha$ has been used to induce a structure *inside* each dictionary element taken separately. Nonetheless, some applications may also benefit from a control of the structure *across* dictionary elements. For instance it can be desirable to impose the constraint that several dictionary elements share the exact same nonzero patterns. In the context of face recognition, this could be relevant to model the variability of faces as the combined variability of several parts, with each part having a small support (such as eyes), and having its variance itself explained by *several* dictionary elements (corresponding for example to the color and the shape of the eyes).

To this end, we consider $\mathcal{M}$, a partition of $[\![1;p]\!]$. Imposing that two dictionary elements $\mathbf{d}^k$ and $\mathbf{d}^{k'}$ share the same sparsity pattern is equivalent to imposing that $\mathbf{d}_i^k$ and $\mathbf{d}_i^{k'}$ are simultaneously zero or non-zero. Following the approach used for joint feature selection (Obozinski et al., 2009) where the $\ell_1$ norm is composed with an $\ell_2$ norm, we compose the norm $\Omega^\alpha$ with the $\ell_2$ norm $\mathbf{d}_i^M = \|(\mathbf{d}_i^k)_{k\in M}\|_2$, of all $i^{\text{th}}$ entries of each dictionary element of a class $M$ of the partition $\mathcal{M}$, leading to the regularization:

$$\sum_{M\in\mathcal{M}} \Omega^\alpha(\mathbf{d}_i^M) = \sum_{M\in\mathcal{M}} \Big[\sum_{g\in\mathcal{G}} \|(\omega_i^g \mathbf{d}_i^k)_{i\in g,\, k\in M}\|_2^\alpha\Big]^{\frac{1}{\alpha}}. \tag{3.4}$$

In fact, not surprisingly given that similar results hold for the group Lasso (Bach, 2008b), it can be shown that the above extension is equivalent to the variational formulation

$$\min_{\substack{\mathbf{A},\,\mathbf{D},\,\Omega_{\mathbf{A}}(\mathbf{A}_k)\le 1\\ (\boldsymbol{\eta}^g)_{g\in\mathcal{G}}\in\mathbb{R}_+^{|\mathcal{M}|\times|\mathcal{G}|}}} \frac{1}{2nm}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\sum_{M\in\mathcal{M}}\Big[\sum_{k\in M}(\mathbf{d}^k)^\top \mathrm{Diag}\Big(\boldsymbol{\zeta}^M\Big)^{-1}\mathbf{d}^k + \|(\boldsymbol{\eta}_M^g)_{g\in\mathcal{G}}\|_\beta\Big],$$

with class specific variables $\{\boldsymbol{\eta}_M\}_{M\in\mathcal{M}}$, $\{\boldsymbol{\zeta}^M\}_{M\in\mathcal{M}}$ defined in a similar way to $\{\boldsymbol{\eta}_k\}_{k\in[\![1;p]\!]}$ and $\{\boldsymbol{\zeta}^k\}_{k\in[\![1;p]\!]}$.

### 3.3.3 Algorithm

The main optimization procedure described in Algorithm 5 is based on a cyclic optimization over the three variables involved, namely $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$, $\mathbf{A}$ and $\mathbf{D}$. We use Lemma 1 to solve (3.2) through a sequence of problems that are convex in $\mathbf{A}$ for fixed $\mathbf{D}$ (and conversely, convex in $\mathbf{D}$ for fixed $\mathbf{A}$). For this sequence of problems, we then present efficient optimization procedures based on block coordinate descent (BCD) (Bertsekas, 1999, Section 2.7). We describe these in detail in Algorithm 5. Note that we depart from the approach of Jenatton et al. (2011a) who use an active set algorithm. Their approach does not indeed allow warm restarts, which is crucial in our alternating optimization scheme.

**Update of $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$.** The update of $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$ is straightforward (even if the underlying minimization problem is non-convex), since the minimizer $(\boldsymbol{\eta}^g)^*$ in Lemma 1 is given in closed-form. In practice, following Micchelli and Pontil (2006), we avoid numerical instability near zero with the smoothed update $\boldsymbol{\eta}_k^g \leftarrow \max\{(\boldsymbol{\eta}_k^g)^*, \varepsilon\}$, with $\varepsilon \ll 1$.

**Update of A.** The update of $\mathbf{A}$ follows the technique suggested by Mairal et al. (2010a). Each row $\mathbf{A}_k$ of $\mathbf{A}$ is constrained separately through $\Omega_{\mathbf{A}}(\mathbf{A}_k)$. Furthermore, if we assume that $\mathbf{D}$ and $\{\mathbf{A}_j\}_{j\neq k}$ are fixed, some basic algebra leads to

$$\begin{aligned} & \arg\min_{\Omega_{\mathbf{A}}(\mathbf{A}_k)\leq 1} \quad \frac{1}{2nm}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 \\ = \; & \arg\min_{\Omega_{\mathbf{A}}(\mathbf{A}_k)\leq 1} \quad \|\mathbf{A}_k - \|\mathbf{d}^k\|_2^{-2}[\mathbf{d}^k]^\top(\mathbf{X}-\sum_{j\neq k}\mathbf{d}^j\mathbf{A}_j)\|_2^2 \\ = \; & \arg\min_{\Omega_{\mathbf{A}}(\mathbf{A}_k)\leq 1} \quad \|\mathbf{A}_k - \mathbf{w}\|_2^2, \end{aligned} \tag{3.5}$$

which is simply the Euclidean projection $\Pi_{\Omega_{\mathbf{A}}}(\mathbf{w})$ of the row vector $\mathbf{w}$ onto the unit ball of $\Omega_{\mathbf{A}}$. Consequently, the cost of the BCD update of $\mathbf{A}$ depends on how fast we can perform this projection; the $\ell_1$ and $\ell_2$ norms are typical cases where the projection can be computed efficiently (Brucker, 1984; Duchi et al., 2008). In the experiments, we take $\Omega_{\mathbf{A}}$ to be the $\ell_2$ norm.

In addition, since the function $\mathbf{A}_k \mapsto \frac{1}{2nm}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2$ is continuously differentiable on the (closed convex) unit ball of $\Omega_{\mathbf{A}}$, the convergence of the BCD procedure is guaranteed since the minimum in Eq. (3.5) is unique (Bertsekas, 1999, Proposition 2.7.1). The complete update of $\mathbf{A}$ is given in Algorithm 5.

**Update of D.** A fairly natural way to update $\mathbf{D}$ would be to compute the closed form solutions available for each row of $\mathbf{D}$. Indeed, both the loss $\frac{1}{2nm}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2$ and the penalization on $\mathbf{D}$ are separable in the rows of $\mathbf{D}$, leading to $m$ independent ridge-regression problems, implying in turn $m$ matrix inversions.

However, in light of the update of $\mathbf{A}$, we consider again a BCD scheme on the columns of $\mathbf{D}$ that turns out to be much more efficient, without requiring any non-diagonal matrix inversion. The detailed procedure is given in Algorithm 5. The convergence follows along the same arguments as those used for $\mathbf{A}$.

**Algorithm 5** Main procedure for solving Eq. (3.3).

**Input:** Dictionary size $p$, data matrix $\mathbf{X}$.
**Initialization:** Initialization of $\mathbf{A}, \mathbf{D}$ (possibly random).
**while** ( *stopping criterion* not reached )
**Update** $(\boldsymbol{\eta}^g)_{G\in\mathcal{G}}$: closed-form solution.
**Update A** by BCD:
**for** $t = 1$ **to** $T_u$, **for** $k = 1$ **to** $p$**:**
$\mathbf{A}_k \leftarrow \Pi_{\Omega_{\mathbf{A}}}(\mathbf{A}_k + \|\mathbf{d}^k\|_2^{-2}[\mathbf{d}^k]^\top(\mathbf{X}-\mathbf{D}\mathbf{A}))$.
**Update D** by BCD:
**for** $t = 1$ **to** $T_v$, **for** $k = 1$ **to** $p$**:**
$\mathbf{d}^k \leftarrow \mathrm{Diag}\left(\boldsymbol{\zeta}^k\right)\mathrm{Diag}\left(\|\mathbf{A}_k\|_2^2\boldsymbol{\zeta}^k + nm\lambda\mathbf{1}\right)^{-1}(\mathbf{X}^\top\mathbf{A}_k^\top - \mathbf{D}\mathbf{A}\mathbf{A}_k^\top + \|\mathbf{A}_k\|_2^2\mathbf{d}^k)$.
**Output:** Decomposition $\mathbf{A}, \mathbf{D}$.

Our problem is not *jointly* convex in $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$, $\mathbf{A}$ and $\mathbf{D}$, which raises the question of the sensitivity of the optimization to its initialization. This point will be discussed in

Section 3.4. In practice, the stopping criterion relies on the relative decrease (typically $10^{-3}$) in the cost function in Eq. (3.2).

**Algorithmic complexity.** The complexity of Algorithm 5 can be decomposed into three terms, corresponding to the update procedures of $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$, $\mathbf{A}$ and $\mathbf{D}$. We denote by $T_u$ (respectively $T_v$) the number of updates of $\mathbf{A}$ (respectively $\mathbf{D}$) in Algorithm 5. First, computing $(\boldsymbol{\eta}^g)_{g\in\mathcal{G}}$ and $\boldsymbol{\zeta}$ costs $O(p|\mathcal{G}|+(|\mathcal{G}|+p)\sum_{g\in\mathcal{G}}|g|) = O(mp|\mathcal{G}|+m|\mathcal{G}|^2)$. The update of $U$ requires $O((m+T_u n)p^2+(nm+C_\Pi T_u)p)$ operations, where $C_\Pi$ is the cost of projecting onto the unit ball of $\Omega_\mathbf{A}$. Similarly, we get for the update of $\mathbf{D}$ a complexity of $O((n+T_v m)p^2+nmp)$. In practice, we notice that the BCD updates for both $\mathbf{A}$ and $\mathbf{D}$ require only few steps, so that we choose $T_u = T_v = 5$. In our experiments, the algorithmic complexity simplifies to $O(m^2+p^2\max\{n,m\}+pm\max\{m^{1/2},n\})$ times the number of iterations in Algorithm 5. Note that the complexity is linear in $n$ and is quadratic in $p$, which is empirically the computational bottleneck.

**Extension to NMF.** Our formalism does not cover the positivity constraints of non-negative matrix factorization, but it is straightforward to extend it at the cost of an additional cheap threshold operation (to project onto the positive orthant) in the BCD updates of $\mathbf{A}$ and $\mathbf{D}$.

## 3.4 Experiments

We first consider the denoising of synthetic signals to illustrate the effect of our regularization. We then focus on the application of SSPCA to a face recognition problem and we show that, by adding a sparse structured prior instead of a simple sparse prior, we gain in robustness to occlusions. In preliminary experiments, we considered the exact regularization from Jenatton et al. (2011a), i.e., with $\alpha = 1$, but found that the obtained patterns were not sufficiently sparse and salient. We therefore turned to the setting where the parameter $\alpha$ is in $(0,1)$. We chose $\alpha = 0.5$, since much smaller or larger values yield either not sparse enough solutions or numerical instability.

By definition, dictionary learning belongs to unsupervised learning; in that sense, our method may appear first as a tool for exploratory data analysis, which leads us naturally to *qualitatively* analyze the results of our decompositions (e.g., by visualizing the learned dictionaries). This is obviously a difficult and subjective exercise, beyond the assessment of the consistency of the method in artificial examples where the "true" dictionnary is known. For that reason, we endeavor in the experiments to compare our method objectively and *quantitatively* with other techniques. Specifically, we apply our method within either a denoising or a classification setting, and assess its performance respectively by the obtained increase in explained variance or classification accuracy.

A Matlab toolbox implementing our method can be downloaded from `http://www.di.ens.fr/~jenatton/`.

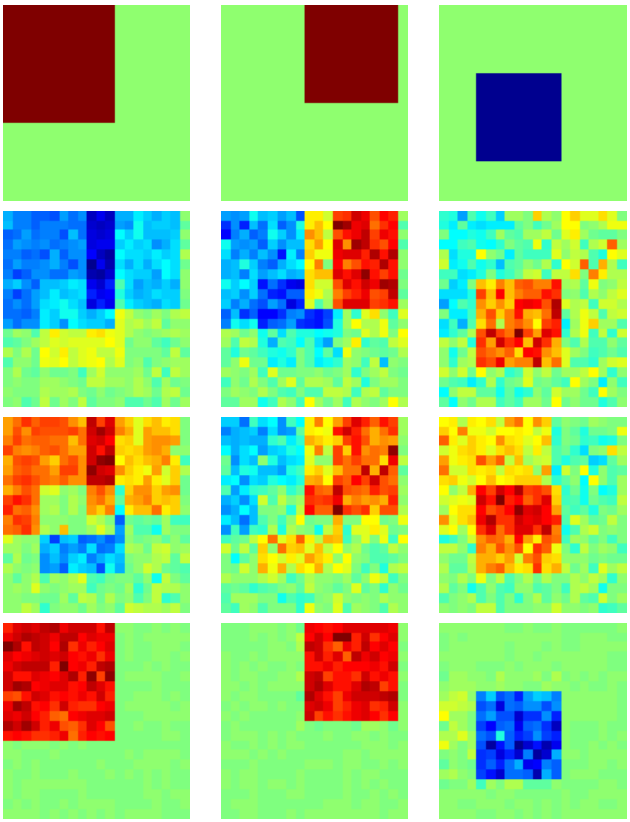


Figure 3.2: Top row: dictionary $\mathbf{D} = [\mathbf{d}^1, \mathbf{d}^2, \mathbf{d}^3] \in \mathbb{R}^{400\times 3}$ used to generate the signals Eq. (3.6). From the second to the bottom row: dictionary elements recovered from 250 signals by PCA, SPCA and SSPCA (best seen in color).

### 3.4.1 Denoising of Synthetic Signals

In this first experiment, we consider signals generated by the following noisy linear model

$$\boldsymbol{\alpha}_1\mathbf{d}^1 + \boldsymbol{\alpha}_2\mathbf{d}^2 + \boldsymbol{\alpha}_3\mathbf{d}^3 + \boldsymbol{\varepsilon} \in \mathbb{R}^{400}, \tag{3.6}$$

where $\mathbf{D} = [\mathbf{d}^1, \mathbf{d}^2, \mathbf{d}^3] \in \mathbb{R}^{400\times 3}$ are sparse and structured dictionary elements organized on a $20 \times 20$-dimensional grid ($\mathbf{D}$ is represented on the top row of Figure 3.2). The components of the noise vector $\varepsilon$ are independent and identically distributed according to a centered Gaussian distribution with its variance set to obtain a signal-to-noise ratio (SNR) of 0.5. The coefficients $[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3]$ that linearly combine the dictionary elements of $\mathbf{D}$ are generated according to a centered Gaussian distribution, with the following covariance matrix

$$\begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

From 250 of such signals, we learn a decomposition $\hat{\mathbf{D}}\hat{\mathbf{A}}$ with $p = 3$ dictionary elements, which seems a reasonable choice of $p$ in an attempt to recover the underlying (in this case, known) structure of $\mathbf{D}$. For SPCA and SSPCA, the regularization parameter $\lambda$ is selected by 5-fold cross-validation on the reconstruction error. Based on the learned dictionary $\hat{\mathbf{D}}$, we denoise 1000 new signals generated in the same way. We report in Table 3.1 the results of the denoising, for PCA, SPCA and SSPCA.

The difficulty of this task is essentially twofold and lies in (1) the high level of noise and in (2) the small number of signals (i.e., 250 signals against 400 variables) available to learn the decomposition.

As displayed on Figure 3.2, PCA and SPCA learn very scattered and uninterpretable dictionary elements. On the other hand, the sparse structured prior we put through $\Omega^\alpha$ helps to recover the initial structure of $\mathbf{D}$, which, in turn, improves upon the denoising performance of SSPCA (see Table 3.1). Note that in order to assess the statistical significance of the differences between the average denoising performances of Table 3.1, one has to consider the sample standard deviation *divided* by $\sqrt{1000}$ (Lehmann and Romano, 2005), i.e., roughly $\approx 0.007$.

The setting we consider here raises the interesting question of *model identifiability*, i.e., whether we can recover the true dictionary elements that generated the signals, which we defer to future work.

| | PCA | SPCA | SSPCA |
|---|---|---|---|
| Estimation error: | $0.41 \pm 0.22$ | $0.40 \pm 0.22$ | $0.34 \pm 0.21$ |

Table 3.1: Average and standard deviation of the normalized estimation error, computed over 1000 signals for PCA, SPCA and SSPCA.

### 3.4.2 Face Recognition

We apply SSPCA on the cropped AR Face Database (Martinez and Kak, 2001) that consists of 2600 face images, corresponding to 100 individuals (50 women and 50 men). For each subject, there are 14 non-occluded poses and 12 occluded ones (the occlusions are due to sunglasses and scarfs). We reduce the resolution of the images from $165 \times 120$ pixels to $38 \times 27$ pixels for computational reasons.

Figure 3.3 shows examples of learned dictionaries (for $p = 36$ elements), for NMF, SSPCA and SSPCA with shared structure (see Section 3.3.2). While NMF finds sparse but spatially unconstrained patterns, SSPCA select sparse convex areas that correspond to a more natural segment of faces. For instance, meaningful parts such as the mouth and the eyes are recovered by the dictionary.

We now quantitatively compare SSPCA, SPCA, PCA and NMF on a face recognition problem. We first split the data into 2 parts, the occluded faces and non-occluded ones. For different sizes of the dictionary, we apply each of the aforementioned dimensionality reduction techniques to the non-occluded faces. Keeping the learned dictionary $\mathbf{D}$, we

decompose both non-occluded and occluded faces on $\mathbf{D}$. We then classify the occluded faces with a k-nearest-neighbors classifier (k-NN), based on the obtained low-dimensional representations $\mathbf{A}$. Given the size of the dictionary, we choose the number of neighbor(s) and the amount of regularization $\lambda$ by cross-validation [5] on the non-occluded faces.

The formulations of NMF, SPCA and SSPCA are non-convex and as a consequence, the local minima reached by those methods might a priori be sensitive to the initialization. To evaluate this sensitivity, we repeat the protocol described above 10 times and display in Figure 3.4 the median, first and third quartile of the classification scores obtained in this way. In practice we found the performance on the test set to be pretty stable as a function of the initialization. We denote by shared-SSPCA (resp. shared-SPCA) the models where we impose, on top of the structure of $\Omega^\alpha$, to have only 10 different nonzero patterns among the learned dictionaries (see Section 3.3.2). We performed a Wilcoxon signed-rank test (Lehmann and Romano, 2005) between the classification scores of NMF and SSPCA, and for dictionary sizes greater than 100 (up to 150), our approach performs better than NMF at the 5% significance level. For smaller dictionaries, NMF and SSPCA perform similarly. The other methods, including PCA and SPCA, obtained overall lower scores than NMF and can also be shown to perform significantly worse than SSPCA.

As a baseline, we also plot the classification score that we obtain when we directly apply k-NN on the raw data, without preprocessing. Because of its local dictionary, SSPCA proves to be more robust to occlusions and therefore outperforms the other methods on this classification task. On the other hand, SPCA, that yields sparsity without a structured prior, performs poorly. Sharing structure across the dictionary elements (see Section 3.3.2) seems to help SPCA for which no structure information is otherwise available. The goal of our chapter is not to compete with state-of-the-art techniques of face recognition, but to demonstrate the improvement obtained between the $\ell_1$ norm and more structured norms. We could still improve upon our results using non-linear classification (e.g., with a SVM) or by refining our features (e.g., with a Laplacian filter).

5. We perform 5-fold cross-validation and the number of nearest neighbor(s) is searched in $\{1, 3, 5\}$ while $\log_{10}(\lambda)$ is in $\{-11, -10.5, \ldots, -7\}$. For the dictionary, we consider the sizes $p \in \{10, 20, \ldots, 150\}$.

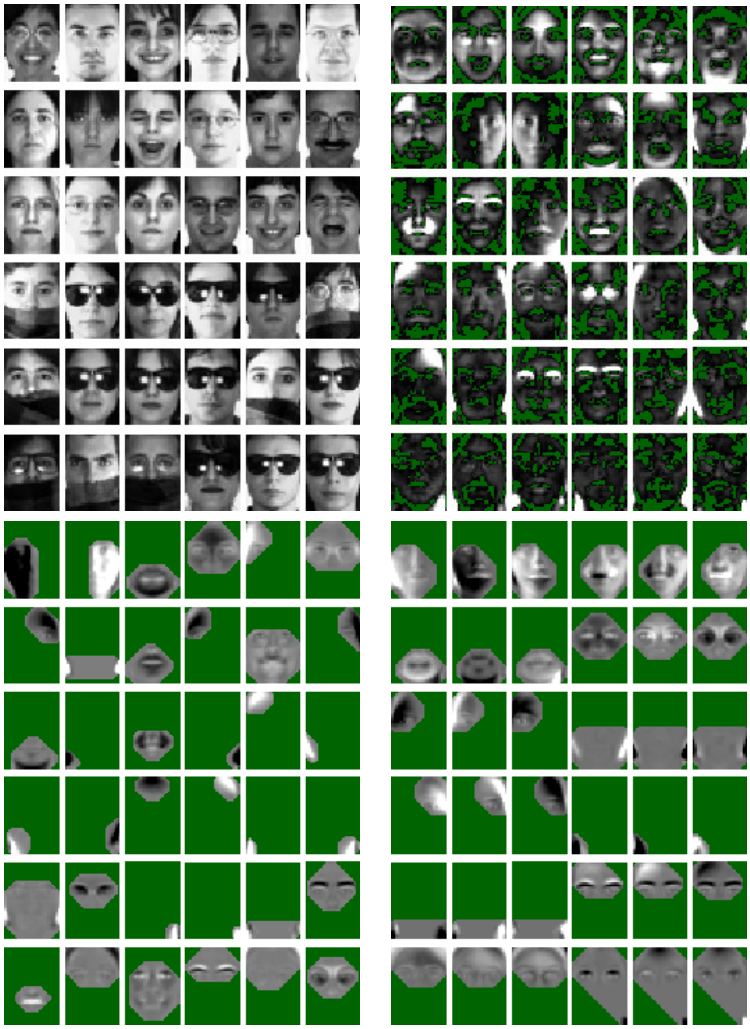


Figure 3.3: Top left, examples of faces in the datasets. The three remaining images represent learned dictionaries of faces with $p = 36$: NMF (top right), SSPCA (bottom left) and shared-SSPCA (bottom roght) (i.e., SSPCA with $|\mathcal{M}| = 12$ different patterns of size 3). The dictionary elements are sorted in decreasing order of explained variance. While NMF gives sparse spatially unconstrained patterns, SSPCA finds convex areas that correspond to more natural face segments. SSPCA captures the left/right illuminations and retrieves pairs of symmetric patterns. Some displayed patterns do not seem to be convex, e.g., nonzero patterns located at two opposite corners of the grid. However, a closer look at these dictionary elements shows that convex shapes are indeed selected, and that small numerical values (just as regularizing by $\ell_2$ norm may lead to) give the visual impression of having zeroes in convex nonzero patterns. This also shows that if a nonconvex pattern has to be selected, it will be, by considering its convex hull.

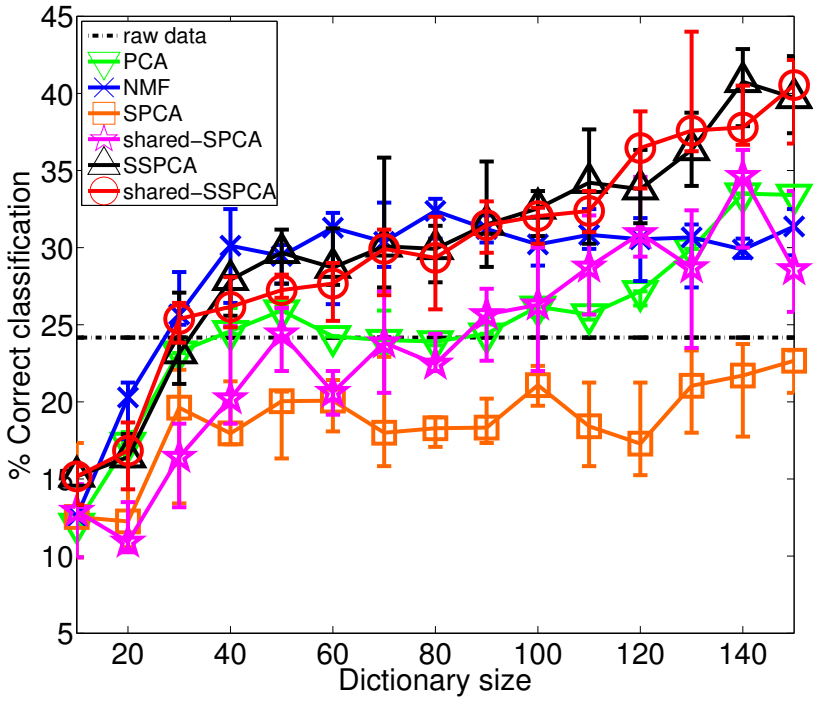



Figure 3.4: Classification accuracy versus dictionary size: each dimensionality reduction technique is used with k-NN to classify occluded faces. SSPCA shows better robustness to occlusions. The points, lower and upper error bars on the curves respectively represent the median, first and third quartile, based on 10 runs.

## 3.5 Conclusions

We proposed to apply a non-convex variant of the regularization introduced by Jenatton et al. (2011a) to the problem of structured sparse dictionary learning. We present an efficient block-coordinate descent algorithm with closed-form updates. In a denoising task of sparse structured signals, our approach led to better performance and to a more interpretable decomposition of the data. For face recognition, the dictionaries learned have increased robustness to occlusions compared to NMF.

In future work, we would like to investigate Bayesian frameworks that would define similar structured priors and allow the principled choice of the regularization parameter and the number of dictionary elements (Zhou et al., 2009). Moreover, although we focus in this work on controlling the structure of the dictionary $\mathbf{D}$, we could instead impose structure on the decompostion coefficients $\mathbf{A}$ and study the induced effect on the dictionary $\mathbf{D}$ (Kavukcuoglu et al., 2009). This could be straightforward to do with the same formulation, by transposing the data matrix $\mathbf{X}$. Finally, we intend to apply this structured sparsity-inducing regularization for multi-task learning, in order to take advantage of the structure between tasks.

# 4

# Proximal Methods for Structured Sparsity-Inducing Norms

**Abstract of the chapter:** Sparse coding consists of representing signals as sparse linear combinations of atoms selected from a dictionary. We consider an extension of this framework where the atoms are further assumed to be embedded in a tree. This is achieved using a recently introduced tree-structured sparse regularization norm, which has proven useful in several applications. This norm leads to regularized problems that are difficult to optimize, and in this chapter, we propose efficient algorithms for solving them. More precisely, we show that the proximal operator associated with this norm is computable exactly via a dual approach that can be viewed as the composition of elementary proximal operators. Our procedure has a complexity linear, or close to linear, in the number of atoms, and allows the use of accelerated gradient techniques to solve the tree-structured sparse approximation problem at the same computational cost as traditional ones using the $\ell_1$-norm. Our method is efficient and scales gracefully to millions of variables, which we illustrate in two types of applications: first, we consider *fixed* hierarchical dictionaries of wavelets to denoise natural images. Then, we apply our optimization tools in the context of *dictionary learning*, where learned dictionary elements naturally organize in a prespecified arborescent structure, leading to better performance in reconstruction of natural image patches. When applied to text documents, our method learns hierarchies of topics, thus providing a competitive alternative to probabilistic topic models.

The work presented in this chapter was achieved with the collaboration of Julien Mairal, Guillaume Obozinski and Francis Bach, with equal contribution between Julien Mairal and myself. The material of this chapter is based on the following work:

R. Jenatton*, J. Mairal*, G. Obozinski, F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. 2010

R. Jenatton*, J. Mairal*, G. Obozinski, F. Bach. Proximal Methods for Hierarchical Sparse Coding. In *Journal of Machine Learning Research*, 12, 2297-2334. 2011 (long version of the previous article)

(*equal contributions)

## 4.1 Introduction

Modeling signals as sparse linear combinations of atoms selected from a dictionary has become a popular paradigm in many fields, including signal processing, statistics, and machine learning. This line of research, also known as *sparse coding*, has witnessed the development of several well-founded theoretical frameworks (Tibshirani, 1996; Chen et al., 1998; Mallat, 1999; Tropp, 2004, 2006; Wainwright, 2009; Bickel et al., 2009) and the emergence of many efficient algorithmic tools (Efron et al., 2004; Nesterov, 2007; Beck and Teboulle, 2009; Wright et al., 2009; Needell and Tropp, 2009; Yuan et al., 2010).

In many applied settings, the structure of the problem at hand, such as, e.g., the spatial arrangement of the pixels in an image, or the presence of variables corresponding to several levels of a given factor, induces relationships between dictionary elements. It is appealing to use this a priori knowledge about the problem *directly* to constrain the possible sparsity patterns. For instance, when the dictionary elements are partitioned into predefined groups corresponding to different types of features, one can enforce a similar block structure in the sparsity pattern—that is, allow only that either all elements of a group are part of the signal decomposition or that all are dismissed simultaneously (see Yuan and Lin, 2006; Stojnic et al., 2009).

This example can be viewed as a particular instance of *structured sparsity*, which has been lately the focus of a large amount of research (Baraniuk et al., 2010; Zhao et al., 2009; Huang et al., 2009; Jacob et al., 2009; Jenatton et al., 2011a; Micchelli et al., 2010). In this chapter, we concentrate on a specific form of structured sparsity, which we call *hierarchical sparse coding*: the dictionary elements are assumed to be embedded in a directed tree $\mathcal{T}$, and the sparsity patterns are constrained to form a *connected and rooted subtree* of $\mathcal{T}$ (Donoho, 1997; Baraniuk, 1999; Baraniuk et al., 2002, 2010; Zhao et al., 2009; Huang et al., 2009). This setting extends more generally to a forest of directed trees. [1]

In fact, such a hierarchical structure arises in many applications. Wavelet decompositions lend themselves well to this tree organization because of their multiscale structure, and benefit from it for image compression and denoising (Shapiro, 1993; Crouse et al., 1998; Baraniuk, 1999; Baraniuk et al., 2002, 2010; He and Carin, 2009; Zhao et al., 2009; Huang et al., 2009). In the same vein, edge filters of natural image patches can be represented in an arborescent fashion (Zoran and Weiss, 2009). Imposing these sparsity patterns has further proven useful in the context of hierarchical variable selection, e.g., when applied to kernel methods (Bach, 2008a), to log-linear models for the selection of potential orders (Schmidt and Murphy, 2010), and to bioinformatics, to exploit the tree structure of gene networks for multi-task regression (Kim and Xing, 2010). Hierarchies of latent variables, typically used in neural networks and deep learning architectures (see Bengio, 2009, and references therein) have also emerged as a natural structure in several applications, notably to model text documents. In particular, in the context of *topic models* (Blei et al., 2003), a hierarchical model of latent variables based on Bayesian

1. A tree is defined as a connected graph that contains no cycle (see Ahuja et al., 1993).

non-parametric methods has been proposed by Blei et al. (2010) to model hierarchies of topics.

To perform hierarchical sparse coding, our work builds upon the approach of Zhao et al. (2009) who first introduced a sparsity-inducing norm $\Omega$ leading to this type of tree-structured sparsity pattern. We tackle the resulting nonsmooth convex optimization problem with proximal methods (e.g., Nesterov, 2007; Beck and Teboulle, 2009; Wright et al., 2009; Combettes and Pesquet, 2010) and we show in this chapter that its key step, the computation of the *proximal operator*, can be solved exactly with a complexity linear, or close to linear, in the number of dictionary elements—that is, with the same complexity as for classical $\ell_1$-sparse decomposition problems (Tibshirani, 1996; Chen et al., 1998). Concretely, given an $m$-dimensional signal $\mathbf{x}$ along with a dictionary $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p] \in \mathbb{R}^{m \times p}$ composed of $p$ atoms, the optimization problem at the core of our chapter can be written as

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\Omega(\boldsymbol{\alpha}), \text{ with } \lambda \geq 0.$$

In this formulation, the sparsity-inducing norm $\Omega$ encodes a hierarchical structure among the atoms of $\mathbf{D}$, where this structure is assumed to be known beforehand. The precise meaning of *hierarchical structure* and the definition of $\Omega$ will be made more formal in the next sections. A particular instance of this problem—known as the *proximal problem*—is central to our analysis and concentrates on the case where the dictionary $\mathbf{D}$ is orthogonal.

In addition to a speed benchmark that evaluates the performance of our proposed approach compared to other convex optimization techniques, two types of applications and experiments are carried out. First, we consider settings where the dictionary is fixed and given a priori, corresponding for instance to a basis of wavelets for the denoising of natural images. Second, we show how one can take advantage of this hierarchical sparse coding in the context of dictionary learning (Olshausen and Field, 1997; Aharon et al., 2006; Mairal et al., 2010a), where the dictionary is learned to adapt to the predefined tree structure. This extension of dictionary learning is notably shown to share interesting connections with hierarchical probabilistic topic models.

To summarize, the contributions of this chapter are threefold:

- We show that the proximal operator for a tree-structured sparse regularization can be computed exactly in a finite number of operations using a dual approach. Our approach is equivalent to computing a particular sequence of elementary proximal operators, and has a complexity linear, or close to linear, in the number of variables. Accelerated gradient methods (e.g., Nesterov, 2007; Beck and Teboulle, 2009; Combettes and Pesquet, 2010) can then be applied to solve large-scale tree-structured sparse decomposition problems at the same computational cost as traditional ones using the $\ell_1$-norm.
- We propose to use this regularization scheme to learn dictionaries embedded in a tree, which, to the best of our knowledge, has not been done before in the context of structured sparsity.
- Our method establishes a bridge between hierarchical dictionary learning and hierarchical topic models (Blei et al., 2010), which builds upon the interpretation

of topic models as multinomial PCA (Buntine, 2002), and can learn similar hierarchies of topics. This point is discussed in Sections 4.5.5 and 4.6.

The rest of this chapter is organized as follows: Section 4.2 presents related work and the problem we consider. Section 4.3 is devoted to the algorithm we propose, and Section 4.4 introduces the dictionary learning framework and shows how it can be used with tree-structured norms. Section 4.5 presents several experiments demonstrating the effectiveness of our approach and Section 4.6 concludes the chapter.

## 4.2 Problem Statement and Related Work

Let us consider an input signal of dimension $m$, typically an image described by its $m$ pixels, which we represent by a vector $\mathbf{x}$ in $\mathbb{R}^m$. In traditional sparse coding, we seek to approximate this signal by a sparse linear combination of atoms, or dictionary elements, represented here by the columns of a matrix $\mathbf{D} \triangleq [\mathbf{d}^1, \ldots, \mathbf{d}^p]$ in $\mathbb{R}^{m \times p}$. This can equivalently be expressed as $\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$ for some sparse vector $\boldsymbol{\alpha}$ in $\mathbb{R}^p$, i.e, such that the number of nonzero coefficients $\|\boldsymbol{\alpha}\|_0$ is small compared to $p$. The vector $\boldsymbol{\alpha}$ is referred to as the code, or decomposition, of the signal $\mathbf{x}$.

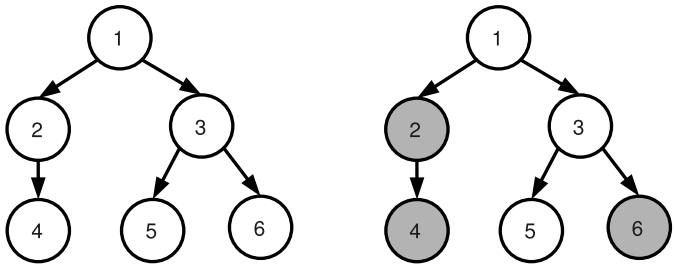



Figure 4.1: Example of a tree $\mathcal{T}$ when $p = 6$. With the rule we consider for the nonzero patterns, if we have $\boldsymbol{\alpha}_5 \neq 0$, we must also have $\boldsymbol{\alpha}_k \neq 0$ for $k$ in ancestors$(5) = \{1, 3, 5\}$.

In the rest of the chapter, we focus on specific sets of nonzero coefficients—or simply, nonzero patterns—for the decomposition vector $\boldsymbol{\alpha}$. In particular, we assume that we are given a tree[2] $\mathcal{T}$ whose $p$ nodes are indexed by $j$ in $\{1, \ldots, p\}$. We want the nonzero patterns of $\boldsymbol{\alpha}$ to form a *connected and rooted subtree* of $\mathcal{T}$; in other words, if ancestors$(j) \subseteq \{1, \ldots, p\}$ denotes the set of indices corresponding to the ancestors[3] of the node $j$ in $\mathcal{T}$ (see Figure 4.1), the vector $\boldsymbol{\alpha}$ obeys the following rule

$$\boldsymbol{\alpha}_j \neq 0 \Rightarrow [\, \boldsymbol{\alpha}_k \neq 0 \text{ for all } k \text{ in ancestors}(j) \,]. \tag{4.1}$$

Informally, we want to exploit the structure of $\mathcal{T}$ in the following sense: the decomposition of any signal $\mathbf{x}$ can involve a dictionary element $\mathbf{d}^j$ *only if the ancestors of* $\mathbf{d}^j$ *in the tree* $\mathcal{T}$ *are themselves part of the decomposition.*

2. Our analysis straightforwardly extends to the case of a forest of trees; for simplicity, we consider a single tree $\mathcal{T}$.

3. We consider that the set of ancestors of a node also contains the node itself.

We now review previous work that has considered the sparse approximation problem with tree-structured constraints (4.1). Similarly to traditional sparse coding, there are basically two lines of research, that either (A) deal with nonconvex and combinatorial formulations that are in general computationally intractable and addressed with greedy algorithms, or (B) concentrate on convex relaxations solved with convex programming methods.

### 4.2.1 Nonconvex Approaches

For a given sparsity level $s \geq 0$ (number of nonzero coefficients), the following nonconvex problem

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^p \\ \|\boldsymbol{\alpha}\|_0 \leq s}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{such that condition (4.1) is respected,} \tag{4.2}$$

has been addressed by Baraniuk (1999); Baraniuk et al. (2002) in the context of wavelet approximations with a greedy procedure. A penalized version of problem (4.2) (that adds $\lambda\|\boldsymbol{\alpha}\|_0$ to the objective function in place of the constraint $\|\boldsymbol{\alpha}\|_0 \leq s$) has been considered by Donoho (1997), while studying the more general problem of best approximation from dyadic partitions (see Section 6 in Donoho, 1997). Interestingly, the algorithm we introduce in Section 4.3 shares conceptual links with the dynamic-programming approach of Donoho (1997), which was also used by Baraniuk et al. (2010), in the sense that the same order of traversal of the tree is used in both procedures. We investigate more thoroughly the relations between our algorithm and this approach in Appendix A.2.1.

Problem (4.2) has been further studied for structured compressive sensing (Baraniuk et al., 2010), with a greedy algorithm that builds upon Needell and Tropp (2009). Finally, Huang et al. (2009) have proposed a formulation related to (4.2), with a nonconvex penalty based on an information-theoretic criterion.

### 4.2.2 Convex Approach

We now turn to a convex reformulation of the constraint (4.1), which is the starting point for the convex optimization tools we develop in Section 4.3.

#### Hierarchical Sparsity-Inducing Norms

Condition (4.1) can be equivalently expressed by taking its contrapositive, thus leading to an intuitive way of penalizing the vector $\boldsymbol{\alpha}$ to obtain tree-structured nonzero patterns. More precisely, defining descendants$(j) \subseteq \{1, \ldots, p\}$ analogously to ancestors$(j)$ for $j$ in $\{1, \ldots, p\}$, condition (4.1) amounts to saying that *if a dictionary element is not used in the decomposition, its descendants in the tree should not be used either*. Formally, this writes down

$$\boldsymbol{\alpha}_j = 0 \Rightarrow [\,\boldsymbol{\alpha}_k = 0 \text{ for all } k \text{ in descendants}(j)\,]. \tag{4.3}$$

From now on, we denote by $\mathcal{G}$ the set defined by $\mathcal{G} \triangleq \{\text{descendants}(j); j \in \{1,\dots,p\}\}$, and refer to each member $g$ of $\mathcal{G}$ as a *group* (Figure 4.2). To obtain a decomposition with the desired property (4.3), one can naturally penalize the number of groups $g$ in $\mathcal{G}$ that are "involved" in the decomposition of $\mathbf{x}$, i.e., that record at least one nonzero coefficient of $\boldsymbol{\alpha}$:

$$\sum_{g\in\mathcal{G}} \delta^g, \text{ with } \delta^g \triangleq \begin{cases} 1 & \text{if there exists } j \in g \text{ such that } \boldsymbol{\alpha}_j \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

While this intuitive penalization is nonconvex (and not even continuous), a convex proxy has been introduced by Zhao et al. (2009). It was further considered by Bach (2008a); Kim and Xing (2010); Schmidt and Murphy (2010) in several different contexts. For any vector $\boldsymbol{\alpha} \in \mathbb{R}^p$, let us define

$$\Omega(\boldsymbol{\alpha}) \triangleq \sum_{g\in\mathcal{G}} \omega_g \|\boldsymbol{\alpha}_{|g}\|,$$

where $\boldsymbol{\alpha}_{|g}$ is the vector of size $p$ whose coordinates are equal to those of $\boldsymbol{\alpha}$ for indices in the set $g$, and 0 otherwise [4]. The notation $\|.\|$ stands in practice either for the $\ell_2$- or $\ell_\infty$-norm, and $(\omega_g)_{g\in\mathcal{G}}$ denotes some positive weights [5]. As analyzed by Zhao et al. (2009) and Jenatton et al. (2011a), when penalizing by $\Omega$, some of the vectors $\boldsymbol{\alpha}_{|g}$ are set to zero for some $g \in \mathcal{G}$. [6] Therefore, the components of $\boldsymbol{\alpha}$ corresponding to some complete subtrees of $\mathcal{T}$ are set to zero, which exactly matches condition (4.3), as illustrated in Figure 4.2.

Note that although we have presented for simplicity this hierarchical norm in the context of a single tree with a single element at each node, it can easily be extended to the case of forests of trees, and/or trees containing arbitrary numbers of dictionary elements at each node (with nodes possibly containing no dictionary element). More broadly, this formulation can be extended with the notion of *tree-structured* groups, which we now present:

**Definition 1** (Tree-structured set of groups.)
*A set of groups $\mathcal{G} \triangleq \{g\}_{g\in\mathcal{G}}$ is said to be tree-structured in $\{1,\dots,p\}$, if $\bigcup_{g\in\mathcal{G}} g = \{1,\dots,p\}$ and if for all $g,h \in \mathcal{G}$, $(g \cap h \neq \emptyset) \Rightarrow (g \subseteq h \text{ or } h \subseteq g)$. For such a set of groups, there exists a (non-unique) total order relation $\preceq$ such that:*

$$g \preceq h \;\Rightarrow\; \{g \subseteq h \quad \text{or} \quad g \cap h = \emptyset\}.$$

Given such a tree-structured set of groups $\mathcal{G}$ and its associated norm $\Omega$, we are interested throughout the chapter in the following hierarchical sparse coding problem,

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda\Omega(\boldsymbol{\alpha}), \tag{4.5}$$

4. Note the difference with the notation $\boldsymbol{\alpha}_g$, which is often used in the literature on structured sparsity, where $\boldsymbol{\alpha}_g$ is a vector of size $|g|$.

5. For a complete definition of $\Omega$ for any $\ell_q$-norm, a discussion of the choice of $q$, and a strategy for choosing the weights $\omega_g$ (see Zhao et al., 2009; Kim and Xing, 2010).

6. It has been further shown by Bach (2010a) that the convex envelope of the nonconvex function of Eq. (4.4) is in fact $\Omega$ with $\|.\|$ being the $\ell_\infty$-norm.

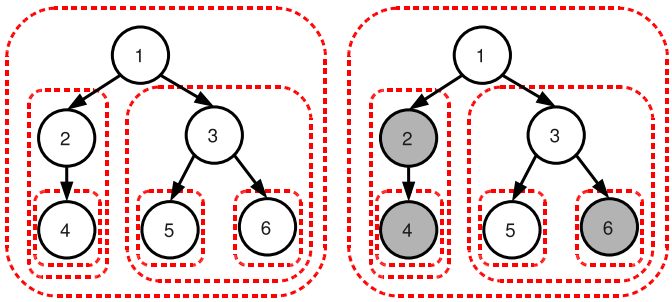



Figure 4.2: Left: example of a tree-structured set of groups $\mathcal{G}$ (dashed contours in red), corresponding to a tree $\mathcal{T}$ with $p = 6$ nodes represented by black circles. Right: example of a sparsity pattern induced by the tree-structured norm corresponding to $\mathcal{G}$: the groups $\{2, 4\}, \{4\}$ and $\{6\}$ are set to zero, so that the corresponding nodes (in gray) that form subtrees of $\mathcal{T}$ are removed. The remaining nonzero variables $\{1, 3, 5\}$ form a rooted and connected subtree of $\mathcal{T}$. This sparsity pattern obeys the following equivalent rules: (i) if a node is selected, the same goes for all its ancestors. (ii) if a node is not selected, then its descendant are not selected.

where $\Omega$ is the tree-structured norm we have previously introduced, the non-negative scalar $\lambda$ is a regularization parameter controlling the sparsity of the solutions of (4.5), and $f$ a smooth convex loss function (see Section 4.3 for more details about the smoothness assumptions on $f$). In the rest of the chapter, we will mostly use the square loss $f(\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2$, with a dictionary $\mathbf{D}$ in $\mathbb{R}^{m \times p}$, but the formulation of Eq. (4.5) extends beyond this context. In particular one can choose $f$ to be the logistic loss, which is commonly used for classification problems (e.g., see Hastie et al., 2009).

Before turning to optimization methods for the hierarchical sparse coding problem, we consider a particular instance. The *sparse group Lasso* was recently considered by Sprechmann et al. (2010b) and Friedman et al. (2010) as an extension of the group Lasso of Yuan and Lin (2006). To induce sparsity both groupwise and within groups, Sprechmann et al. (2010b) and Friedman et al. (2010) add an $\ell_1$ term to the regularization of the group Lasso, which given a partition $\mathcal{P}$ of $\{1, \ldots, p\}$ in disjoint groups yields a regularized problem of the form

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{g \in \mathcal{P}} \|\boldsymbol{\alpha}_{|g}\|_2 + \lambda' \|\boldsymbol{\alpha}\|_1.$$

Since $\mathcal{P}$ is a partition, the set of groups in $\mathcal{P}$ and the singletons form together a tree-structured set of groups according to definition 1 and the algorithm we will develop will therefore be applicable to this problem.

### Optimization for Hierarchical Sparsity-Inducing Norms

While generic approaches like interior-point methods (Boyd and Vandenberghe, 2004) and subgradient descent schemes (Bertsekas, 1999) might be used to deal with the non-

smooth norm $\Omega$, several dedicated procedures have been proposed.

In Zhao et al. (2009), a boosting-like technique is used, with a path-following strategy in the specific case where $\|.\|$ is the $\ell_\infty$-norm. Based on the variational equality

$$\|\mathbf{u}\|_1 = \min_{\mathbf{z}\in\mathbb{R}_+^p} \frac{1}{2}\Big[\sum_{j=1}^{p} \frac{\mathbf{u}_j^2}{\mathbf{z}_j} + \mathbf{z}_j\Big], \tag{4.6}$$

Kim and Xing (2010) follow a reweighted least-square scheme that is well adapted to the square loss function. To the best of our knowledge, a formulation of this type is however not available when $\|.\|$ is the $\ell_\infty$-norm. In addition it requires an appropriate smoothing to become provably convergent. The same approach is considered by Bach (2008a), but built upon an active-set strategy. Other proposed methods consist of a projected gradient descent with approximate projections onto the ball $\{\mathbf{u} \in \mathbb{R}^p;\ \Omega(\mathbf{u}) \leq \lambda\}$ (Schmidt and Murphy, 2010), and an augmented-Lagrangian based technique (Sprechmann et al., 2010b) for solving a particular case with two-level hierarchies.

While the previously listed first-order approaches are (1) loss-function dependent, and/or (2) not guaranteed to achieve optimal convergence rates, and/or (3) not able to yield sparse solutions without a somewhat arbitrary post-processing step, we propose to resort to proximal methods [7] that do not suffer from any of these drawbacks.

## 4.3 Optimization

We begin with a brief introduction to proximal methods, necessary to present our contributions. From now on, we assume that $f$ is convex and continuously differentiable with Lipschitz-continuous gradient. It is worth mentioning that there exist various proximal schemes in the literature that differ in their settings (e.g., batch versus stochastic) and/or the assumptions made on $f$. For instance, the material we develop in this chapter could also be applied to online/stochastic frameworks (Duchi and Singer, 2009; Hu et al., 2009; Xiao, 2010) and to possibly nonsmooth functions $f$ (e.g., Duchi and Singer, 2009; Xiao, 2010; Combettes and Pesquet, 2010, and references therein). Finally, most of the technical proofs of this section are presented in Appendix A.2.2 for readability.

### 4.3.1 Proximal Operator for the Norm $\Omega$

Proximal methods have drawn increasing attention in the signal processing (e.g., Becker et al., 2009; Wright et al., 2009; Combettes and Pesquet, 2010, and numerous references therein) and the machine learning communities (e.g., Bach et al., 2011, and references therein), especially because of their convergence rates (optimal for the class of first-order techniques) and their ability to deal with large nonsmooth convex problems (e.g., Nesterov, 2007; Beck and Teboulle, 2009). In a nutshell, these methods can be

7. Note that the authors of Chen et al. (2010) have considered proximal methods for general group structure $\mathcal{G}$ when $\|.\|$ is the $\ell_2$-norm; due to a smoothing of the regularization term, the convergence rate they obtained is suboptimal.

seen as a natural extension of gradient-based techniques when the objective function to minimize has a nonsmooth part. Proximal methods are iterative procedures. The simplest version of this class of methods linearizes at each iteration the function $f$ around the current estimate $\hat{\boldsymbol{\alpha}}$, and this estimate is updated as the (unique by strong convexity) solution of the *proximal problem*, defined as follows:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \; f(\hat{\boldsymbol{\alpha}}) + (\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})^\top \nabla f(\hat{\boldsymbol{\alpha}}) + \lambda\Omega(\boldsymbol{\alpha}) + \frac{L}{2}\|\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\|_2^2.$$

The quadratic term keeps the update in a neighborhood where $f$ is close to its linear approximation, and $L > 0$ is a parameter which is an upper bound on the Lipschitz constant of $\nabla f$. This problem can be equivalently rewritten as:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \; \frac{1}{2}\Big\|\boldsymbol{\alpha} - (\hat{\boldsymbol{\alpha}} - \frac{1}{L}\nabla f(\hat{\boldsymbol{\alpha}}))\Big\|_2^2 + \frac{\lambda}{L}\Omega(\boldsymbol{\alpha}).$$

Solving *efficiently* and *exactly* this problem is crucial to enjoy the fast convergence rates of proximal methods. In addition, when the nonsmooth term $\Omega$ is not present, the previous proximal problem exactly leads to the standard gradient update rule. More generally, we define the *proximal operator*:

**Definition 2** (Proximal Operator)
*The proximal operator associated with our regularization term $\lambda\Omega$, which we denote by $Prox_{\lambda\Omega}$, is the function that maps a vector $\mathbf{u} \in \mathbb{R}^p$ to the unique solution of*

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\Omega(\mathbf{v}). \tag{4.7}$$

This operator was initially introduced by Moreau (1962) to generalize the projection operator onto a convex set. What makes proximal methods appealing for solving sparse decomposition problems is that this operator can be often computed in closed-form. For instance,

- When $\Omega$ is the $\ell_1$-norm—that is, $\Omega(\mathbf{u}) = \|\mathbf{u}\|_1$, the proximal operator is the well-known elementwise soft-thresholding operator,

$$\forall j \in [\![1;p]\!], \quad \mathbf{u}_j \mapsto \operatorname{sign}(\mathbf{u}_j)(|\mathbf{u}_j|-\lambda)_+ \quad = \begin{cases} 0 & \text{if } \; |\mathbf{u}_j| \leq \lambda \\ \operatorname{sign}(\mathbf{u}_j)(|\mathbf{u}_j|-\lambda) & \text{otherwise.} \end{cases}$$

- When $\Omega$ is a group-Lasso penalty with $\ell_2$-norms—that is, $\Omega(\mathbf{u}) = \sum_{g\in\mathcal{G}} \|\mathbf{u}_{|g}\|_2$, with $\mathcal{G}$ being a partition of $[\![1;p]\!]$, the proximal problem is *separable* in every group, and the solution is a generalization of the soft-thresholding operator to groups of variables:

$$\forall g \in \mathcal{G} \quad , \mathbf{u}_{|g} \mapsto \mathbf{u}_{|g} - \Pi_{\|.\|_2\leq\lambda}[\mathbf{u}_{|g}] = \begin{cases} 0 & \text{if } \; \|\mathbf{u}_{|g}\|_2 \leq \lambda \\ \frac{\|\mathbf{u}_{|g}\|_2-\lambda}{\|\mathbf{u}_{|g}\|_2}\mathbf{u}_{|g} & \text{otherwise,} \end{cases}$$

  where $\Pi_{\|.\|_2\leq\lambda}$ denotes the orthogonal projection onto the ball of the $\ell_2$-norm of radius $\lambda$.

- When $\Omega$ is a group-Lasso penalty with $\ell_\infty$-norms—that is, $\Omega(\mathbf{u}) = \sum_{g\in\mathcal{G}} \|\mathbf{u}_{|g}\|_\infty$, the solution is also a group-thresholding operator:

$$\forall g \in \mathcal{G}, \quad \mathbf{u}_{|g} \mapsto \mathbf{u}_{|g} - \Pi_{\|.\|_1\leq\lambda}[\mathbf{u}_{|g}],$$

  where $\Pi_{\|.\|_1\leq\lambda}$ denotes the orthogonal projection onto the $\ell_1$-ball of radius $\lambda$, which can be solved in $O(p)$ operations (Brucker, 1984; Maculan and Galdino de Paula, 1989). Note that when $\|\mathbf{u}_{|g}\|_1 \leq \lambda$, we have a group-thresholding effect, with $\mathbf{u}_{|g} - \Pi_{\|.\|_1\leq\lambda}[\mathbf{u}_{|g}] = 0$.

More generally, a classical result (see, e.g., Combettes and Pesquet, 2010; Wright et al., 2009) says that the proximal operator for a norm $\|.\|$ can be computed as the residual of the projection of a vector onto a ball of the dual-norm denoted by $\|.\|_*$, and defined for any vector $\boldsymbol{\kappa}$ in $\mathbb{R}^p$ by $\|\boldsymbol{\kappa}\|_* \triangleq \max_{\|\mathbf{z}\|\leq 1} \mathbf{z}^\top\boldsymbol{\kappa}$.[8] This is a classical duality result for proximal operators leading to the different closed forms we have just presented. We have indeed that $\text{Prox}_{\lambda\|.\|_2} = \text{Id} - \Pi_{\|.\|_2\leq\lambda}$ and $\text{Prox}_{\lambda\|.\|_\infty} = \text{Id} - \Pi_{\|.\|_1\leq\lambda}$, where Id stands for the identity operator. Obtaining such closed forms is, however, not possible anymore as soon as some groups in $\mathcal{G}$ overlap, which is always the case in our hierarchical setting with tree-structured groups.

### 4.3.2 A Dual Formulation of the Proximal Problem

We now show that Eq. (4.7) can be solved using a dual approach, as described in the following lemma. The result relies on conic duality (Boyd and Vandenberghe, 2004), and does not make any assumption on the choice of the norm $\|.\|$:

**Lemma 2** (Dual of the proximal problem)
*Let $\mathbf{u} \in \mathbb{R}^p$ and let us consider the problem*

$$\begin{aligned} &\max_{\boldsymbol{\xi}\in\mathbb{R}^{p\times|\mathcal{G}|}} -\frac{1}{2}\Big[\Big\|\mathbf{u} - \sum_{g\in\mathcal{G}} \boldsymbol{\xi}^g\Big\|_2^2 - \|\mathbf{u}\|_2^2\Big] \\ &\text{s.t. } \forall g \in \mathcal{G}, \ \|\boldsymbol{\xi}^g\|_* \leq \lambda\omega_g \text{ and } \boldsymbol{\xi}^g_j = 0 \text{ if } j \notin g, \end{aligned} \tag{4.8}$$

*where $\boldsymbol{\xi} = (\boldsymbol{\xi}^g)_{g\in\mathcal{G}}$ and $\boldsymbol{\xi}^g_j$ denotes the $j$-th coordinate of the vector $\boldsymbol{\xi}^g$ in $\mathbb{R}^p$. Then, problems (4.7) and (4.8) are dual to each other and strong duality holds. In addition, the pair of primal-dual variables $\{\mathbf{v}, \boldsymbol{\xi}\}$ is optimal if and only if $\boldsymbol{\xi}$ is a feasible point of the optimization problem (4.8), and*

$$\mathbf{v} = \mathbf{u} - \textstyle\sum_{g\in\mathcal{G}} \boldsymbol{\xi}^g \quad \text{and} \quad \forall g \in \mathcal{G}, \ \boldsymbol{\xi}^g = \Pi_{\|.\|_*\leq\lambda\omega_g}(\mathbf{v}_{|g} + \boldsymbol{\xi}^g), \tag{4.9}$$

*where we denote by $\Pi_{\|.\|_*\leq\lambda\omega_g}$ the Euclidean projection onto the ball $\{\boldsymbol{\kappa} \in \mathbb{R}^p; \ \|\boldsymbol{\kappa}\|_* \leq \lambda\omega_g\}$.*

8. It is easy to show that the dual norm of the $\ell_2$-norm is the $\ell_2$-norm itself. The dual norm of the $\ell_\infty$ is the $\ell_1$-norm.

Note that we focus here on specific tree-structured groups, but the previous lemma is valid regardless of the nature of $\mathcal{G}$. The rationale of introducing such a dual formulation is to consider an equivalent problem to (4.7) that removes the issue of overlapping groups at the cost of a larger number of variables. In Eq. (4.7), one is indeed looking for a vector $\mathbf{v}$ of size $p$, whereas one is considering a matrix $\boldsymbol{\xi}$ in $\mathbb{R}^{p\times|\mathcal{G}|}$ in Eq. (4.8) with $\sum_{g\in\mathcal{G}}|g|$ nonzero entries, but with separable (convex) constraints for each of its columns.

This specific structure makes it possible to use block coordinate ascent (Bertsekas, 1999). Such a procedure is presented in Algorithm 6. It optimizes sequentially Eq. (4.8) with respect to the variable $\boldsymbol{\xi}^g$, while keeping fixed the other variables $\boldsymbol{\xi}^h$, for $h \neq g$. It is easy to see from Eq. (4.8) that such an update of a column $\boldsymbol{\xi}^g$, for a group $g$ in $\mathcal{G}$, amounts to computing the orthogonal projection of the vector $\mathbf{u}_{|g} - \sum_{h\neq g}\boldsymbol{\xi}^h_{|g}$ onto the ball of radius $\lambda\omega_g$ of the dual norm $\|.\|_*$.

**Algorithm 6** Block coordinate ascent in the dual

Inputs: $\mathbf{u} \in \mathbb{R}^p$ and set of groups $\mathcal{G}$.
Outputs: $(\mathbf{v}, \boldsymbol{\xi})$ (primal-dual solutions).
Initialization: $\boldsymbol{\xi} = \mathbf{0}$.
**while** ( *maximum number of iterations not reached* ) **do**
  **for** $g \in \mathcal{G}$ **do**
    $\boldsymbol{\xi}^g \leftarrow \Pi_{\|.\|_*\leq\lambda\omega_g}([\mathbf{u} - \sum_{h\neq g}\boldsymbol{\xi}^h]_{|g})$.
  **end for**
**end while**
$\mathbf{v} \leftarrow \mathbf{u} - \sum_{g\in\mathcal{G}}\boldsymbol{\xi}^g$.

### 4.3.3 Convergence in One Pass

In general, Algorithm 6 is not guaranteed to solve exactly Eq. (4.7) in a finite number of iterations. However, when $\|.\|$ is the $\ell_2$- or $\ell_\infty$-norm, and provided that the groups in $\mathcal{G}$ are appropriately ordered, we now prove that only *one pass* of Algorithm 6, i.e., only one iteration over all groups, is sufficient to obtain the exact solution of Eq. (4.7). This result constitutes the main technical contribution of the chapter and is the key for the efficiency of our procedure.

Before stating this result, we need to introduce a lemma showing that, given two nested groups $g, h$ such that $g \subseteq h \subseteq \{1,\ldots,p\}$, if $\boldsymbol{\xi}^g$ is updated before $\boldsymbol{\xi}^h$ in Algorithm 6, then the optimality condition for $\boldsymbol{\xi}^g$ is not perturbed by the update of $\boldsymbol{\xi}^h$.

**Lemma 3** (Projections with nested groups)
*Let $\|.\|$ denote either the $\ell_2$- or $\ell_\infty$-norm, and $g$ and $h$ be two nested groups—that is, $g \subseteq h \subseteq \{1,\ldots,p\}$. Let $\mathbf{u}$ be a vector in $\mathbb{R}^p$, and let us consider the successive projections*

$$\boldsymbol{\xi}^g \triangleq \Pi_{\|.\|_*\leq t_g}(\mathbf{u}_{|g}) \quad \textit{and} \quad \boldsymbol{\xi}^h \triangleq \Pi_{\|.\|_*\leq t_h}(\mathbf{u}_{|h} - \boldsymbol{\xi}^g),$$

*with $t_g, t_h > 0$. Let us introduce $\mathbf{v} = \mathbf{u} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h$. The following relationships hold*

$$\boldsymbol{\xi}^g = \Pi_{\|.\|_*\leq t_g}(\mathbf{v}_{|g} + \boldsymbol{\xi}^g) \quad \textit{and} \quad \boldsymbol{\xi}^h = \Pi_{\|.\|_*\leq t_h}(\mathbf{v}_{|h} + \boldsymbol{\xi}^h).$$

The previous lemma establishes the convergence in one pass of Algorithm 6 in the case where $\mathcal{G}$ only contains two nested groups $g \subseteq h$, provided that $\boldsymbol{\xi}^g$ is computed before $\boldsymbol{\xi}^h$. Let us illustrate this fact more concretely. After initializing $\boldsymbol{\xi}^g$ and $\boldsymbol{\xi}^h$ to zero, Algorithm 6 first updates $\boldsymbol{\xi}^g$ with the formula $\boldsymbol{\xi}^g \leftarrow \Pi_{\|.\|_* \leq \lambda\omega_g}(\mathbf{u}_{|g})$, and then performs the following update: $\boldsymbol{\xi}^h \leftarrow \Pi_{\|.\|_* \leq \lambda\omega_h}(\mathbf{u}_{|h} - \boldsymbol{\xi}^g)$ (where we have used that $\boldsymbol{\xi}^g = \boldsymbol{\xi}^g_{|h}$ since $g \subseteq h$). We are now in position to apply Lemma 3 which states that the primal/dual variables $\{\mathbf{v}, \boldsymbol{\xi}^g, \boldsymbol{\xi}^h\}$ satisfy the optimality conditions (4.9), as described in Lemma 2. In only one pass over the groups $\{g, h\}$, we have in fact reached a solution of the dual formulation presented in Eq. (4.8), and in particular, the solution of the proximal problem (4.7).

In the following proposition, this lemma is extended to general tree-structured sets of groups $\mathcal{G}$:

**Proposition 9** (Convergence in one pass)
*Suppose that the groups in $\mathcal{G}$ are ordered according to the total order relation $\preceq$ of Definition 1 and that the norm $\|.\|$ is either the $\ell_2$- or $\ell_\infty$-norm. Then, after initializing $\boldsymbol{\xi}$ to $\mathbf{0}$, a single pass of Algorithm 6 over $\mathcal{G}$ with the order $\preceq$ yields the solution of the proximal problem (4.7).*

*Proof.* The proof largely relies on Lemma 3 and proceeds by induction. By definition of Algorithm 6, the feasibility of $\boldsymbol{\xi}$ is always guaranteed. We consider the following induction hypothesis

$$\mathcal{H}(h) \triangleq \{\forall g \preceq h, \text{ it holds that } \boldsymbol{\xi}^g = \Pi_{\|.\|_* \leq \lambda\omega_g}([\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'}]_{|g} + \boldsymbol{\xi}^g)\}.$$

Since the dual variables $\boldsymbol{\xi}$ are initially equal to zero, the summation over $g' \preceq h$, $g' \neq g$ is equivalent to a summation over $g' \neq g$. We initialize the induction with the first group in $\mathcal{G}$, that, by definition of $\preceq$, does not contain any other group. The first step of Algorithm 6 easily shows that the induction hypothesis $\mathcal{H}$ is satisfied for this first group.

We now assume that $\mathcal{H}(h)$ is true and consider the next group $h'$, $h \preceq h'$, in order to prove that $\mathcal{H}(h')$ is also satisfied. We have for each group $g \subseteq h$,

$$\boldsymbol{\xi}^g = \Pi_{\|.\|_* \leq \lambda\omega_g}([\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'}]_{|g} + \boldsymbol{\xi}^g) = \Pi_{\|.\|_* \leq \lambda\omega_g}([\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'} + \boldsymbol{\xi}^g]_{|g}).$$

Since $\boldsymbol{\xi}^g_{|h'} = \boldsymbol{\xi}^g$ for $g \subseteq h'$, we have

$$[\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'}]_{|h'} = [\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'}]_{|h'} + \boldsymbol{\xi}^g - \boldsymbol{\xi}^g = [\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'} + \boldsymbol{\xi}^g]_{|h'} - \boldsymbol{\xi}^g,$$

and following the update rule for the group $h'$,

$$\boldsymbol{\xi}^{h'} = \Pi_{\|.\|_* \leq \lambda\omega_{h'}}([\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'}]_{|h'}) = \Pi_{\|.\|_* \leq \lambda\omega_{h'}}([\mathbf{u} - \textstyle\sum_{g' \preceq h} \boldsymbol{\xi}^{g'} + \boldsymbol{\xi}^g]_{|h'} - \boldsymbol{\xi}^g).$$

At this point, we can apply Lemma 3 for each group $g \subseteq h$, which proves that the induction hypothesis $\mathcal{H}(h')$ is true. Let us introduce $\mathbf{v} \triangleq \mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g$. We have shown that for all $g$ in $\mathcal{G}$, $\boldsymbol{\xi}^g = \Pi_{\|.\|_* \leq \lambda\omega_g}(\mathbf{v}_{|g} + \boldsymbol{\xi}^g)$. As a result, the pair $\{\mathbf{v}, \boldsymbol{\xi}\}$ satisfies the optimality conditions (4.9) of problem (4.8). Therefore, after one complete pass over $g \in \mathcal{G}$, the primal/dual pair $\{\mathbf{v}, \boldsymbol{\xi}\}$ is optimal, and in particular, $\mathbf{v}$ is the solution of problem (4.7). □

We recall that the total order relation $\preceq$ introduced in Definition 1 is defined so that when a group $h$ is included in a group $g$, then $h$ should be processed before $g$. We illustrate in Figure 4.3 the practical implications of Proposition 9. More precisely, we consider Algorithm 6 with both the "right" order for $\mathcal{G}$ (as advocated by Proposition 9), and random orders. We then monitor the cost function of the primal proximal problem and its dual counterpart, respectively given in (4.7) and (4.8). Using conic duality, we

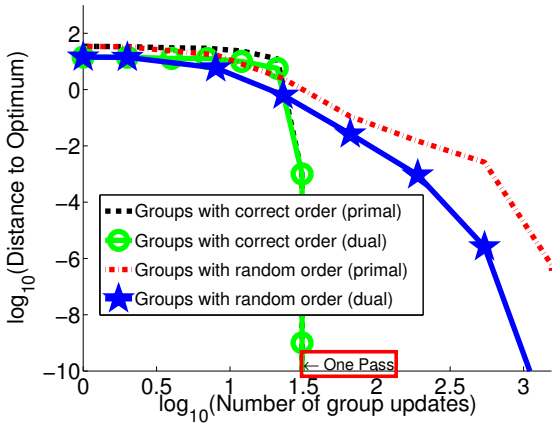



Figure 4.3: One pass convergence: the cost function of the primal proximal problem and its dual counterpart are monitored with respect to the number of group updates in Algorithm 6. In this setting, $\mathcal{G}$ corresponds to a complete binary tree of depth 4, with a total of $p = |\mathcal{G}| = 31$ nodes. With the correct order, one pass is sufficient to reach the exact solution (note that $\log_{10}(31) \approx 1.49$). For the random orders of $\mathcal{G}$, we display the average of the cost functions based on 20 different orders.

have derived a dual formulation of the proximal operator, leading to Algorithm 6 which is generic and works for any norm $\|.\|$, as long as one is able to perform projections onto balls of the dual norm $\|.\|_*$. We have further shown that when $\|.\|$ is the $\ell_2$- or the $\ell_\infty$-norm, a single pass provides the exact solution when the groups $\mathcal{G}$ are correctly ordered. We show however in Appendix A.2.3, that, perhaps surprisingly, the conclusions of Proposition 9 do not hold for general $\ell_q$-norms, if $q \notin \{1, 2, \infty\}$. Next, we give another interpretation of this result.

### 4.3.4 Interpretation in Terms of Composition of Proximal Operators

In Algorithm 6, since all the vectors $\boldsymbol{\xi}^g$ are initialized to $\mathbf{0}$, when the group $g$ is considered, we have by induction $\mathbf{u} - \sum_{h \neq g} \boldsymbol{\xi}^h = \mathbf{u} - \sum_{h \preceq g} \boldsymbol{\xi}^h$. Thus, to maintain at each iteration of the inner loop $\mathbf{v} = \mathbf{u} - \sum_{h \neq g} \boldsymbol{\xi}^h$ one can instead update $\mathbf{v}$ after updating $\boldsymbol{\xi}^g$ according to $\mathbf{v} \leftarrow \mathbf{v} - \boldsymbol{\xi}^g$. Moreover, since $\boldsymbol{\xi}^g$ is no longer needed in the algorithm, and since only the entries of $\mathbf{v}$ indexed by $g$ are updated, we can combine the two updates into $\mathbf{v}_{|g} \leftarrow \mathbf{v}_{|g} - \Pi_{\|.\|_* \leq \lambda \omega_g}(\mathbf{v}_{|g})$, leading to a simplified Algorithm 7 equivalent to Algorithm 6.

**Algorithm 7** Practical Computation of the Proximal Operator for $\ell_2$- or $\ell_\infty$-norms.
Inputs: $\mathbf{u} \in \mathbb{R}^p$ and an ordered tree-structured set of groups $\mathcal{G}$.
Outputs: $\mathbf{v}$ (primal solution).
Initialization: $\mathbf{v} = \mathbf{u}$.
**for** $g \in \mathcal{G}$, following the order $\preceq$, **do**
  $\mathbf{v}_{|g} \leftarrow \mathbf{v}_{|g} - \Pi_{\|.\|_* \leq \lambda\omega_g}(\mathbf{v}_{|g})$.
**end for**

Actually, in light of the classical relationship between proximal operator and projection (as discussed in Section 4.3.1), it is easy to show that each update $\mathbf{v}_{|g} \leftarrow \mathbf{v}_{|g} - \Pi_{\|.\|_* \leq \lambda\omega_g}(\mathbf{v}_{|g})$ is equivalent to $\mathbf{v}_{|g} \leftarrow \mathrm{Prox}_{\lambda\omega_g\|.\|}[\mathbf{v}_{|g}]$. To simplify the notations, we define the proximal operator for a group $g$ in $\mathcal{G}$ as $\mathrm{Prox}^g(\mathbf{u}) \triangleq \mathrm{Prox}_{\lambda\omega_g\|.\|}(\mathbf{u}_{|g})$ for every vector $\mathbf{u}$ in $\mathbb{R}^p$.

Thus, Algorithm 7 in fact performs a sequence of $|\mathcal{G}|$ proximal operators, and we have shown the following corollary of Proposition 9:

**Corollary 4** (Composition of Proximal Operators)
*Let $g_1 \preccurlyeq \ldots \preccurlyeq g_m$ such that $\mathcal{G} = \{g_1, \ldots, g_m\}$. The proximal operator $Prox_{\lambda\Omega}$ associated with the norm $\Omega$ can be written as the composition of elementary operators:*

$$Prox_{\lambda\Omega} = Prox^{g_m} \circ \ldots \circ Prox^{g_1}.$$

### 4.3.5 Efficient Implementation and Complexity

Since Algorithm 7 involves $|\mathcal{G}|$ projections on the dual balls (respectively the $\ell_2$- and the $\ell_1$-balls for the $\ell_2$- and $\ell_\infty$-norms) of vectors in $\mathbb{R}^p$, in a first approximation, its complexity is at most $O(p^2)$, because each of these projections can be computed in $O(p)$ operations (Brucker, 1984; Maculan and Galdino de Paula, 1989). But in fact, the algorithm performs one projection for each group $g$ involving $|g|$ variables, and the total complexity is therefore $O\Big(\sum_{g\in\mathcal{G}} |g|\Big)$. By noticing that if $g$ and $h$ are two groups with the same depth in the tree, then $g \cap h = \emptyset$, it is easy to show that the number of variables involved in all the projections is less than or equal to $dp$, where $d$ is the depth of the tree:

**Lemma 4** (Complexity of Algorithm 7)
*Algorithm 7 gives the solution of the primal problem Eq. (4.7) in $O(pd)$ operations, where $d$ is the depth of the tree.*

Lemma 4 should not suggest that the complexity is linear in $p$, since $d$ could depend of $p$ as well, and in the worst case the hierarchy is a chain, yielding $d = p - 1$. However, in a balanced tree, $d = O(\log(p))$. In practice, the structures we have considered experimentally are relatively flat, with a depth not exceeding $d = 5$, and the complexity is therefore almost linear.

Moreover, in the case of the $\ell_2$-norm, it is actually possible to propose an algorithm with complexity $O(p)$. Indeed, in that case each of the proximal operators $\mathrm{Prox}^g$ is a

**Algorithm 8** Fast computation of the Proximal operator for $\ell_2$-norm case.

**Require:** $\mathbf{u} \in \mathbb{R}^p$ (input vector), set of groups $\mathcal{G}$, $(\omega_g)_{g\in\mathcal{G}}$ (positive weights), and $g_0$ (root of the tree).

1: Variables: $\boldsymbol{\rho} = (\rho_g)_{g\in\mathcal{G}}$ in $\mathbb{R}^{|\mathcal{G}|}$ (scaling factors); $\mathbf{v}$ in $\mathbb{R}^p$ (output, primal variable).
2: `computeSqNorm`($g_0$).
3: `recursiveScaling`($g_0$,1).
4: **return** $\mathbf{v}$ (primal solution).

**Procedure** `computeSqNorm`($g$)

1: Compute the squared norm of the group: $\eta_g \leftarrow \|\mathbf{u}_{\text{root}(g)}\|_2^2 + \sum_{h\in\text{children}(g)}$ `computeSqNorm`($h$).
2: Compute the scaling factor of the group: $\rho_g \leftarrow \left(1 - \lambda\omega_g/\sqrt{\eta_g}\right)_+$.
3: **return** $\eta_g\rho_g^2$.

**Procedure** `recursiveScaling`($g$,$t$)

1: $\rho_g \leftarrow t\rho_g$.
2: $\mathbf{v}_{\text{root}(g)} \leftarrow \rho_g\mathbf{u}_{\text{root}(g)}$.
3: **for** $h \in$ children($g$) **do**
4: `recursiveScaling`($h$,$\rho_g$).
5: **end for**

scaling operation: $\mathbf{v}_{|g} \leftarrow \left(1 - \lambda\omega_g/\|\mathbf{v}_{|g}\|_2\right)_+\mathbf{v}_{|g}$. The composition of these operators in Algorithm 6 thus corresponds to performing sequences of scaling operations. The idea behind Algorithm 8 is that the corresponding scaling factors depend only on the norms of the successive residuals of the projections and that these norms can be computed recursively in one pass through all nodes in $O(p)$ operations; finally, computing and applying all scalings to each entry takes then again $O(p)$ operations.

To formulate the algorithm, two new notations are used: for a group $g$ in $\mathcal{G}$, we denote by root($g$) the indices of the variables that are at the root of the subtree corresponding to $g$, [9] and by children($g$) the set of groups that are the children of root($g$) in the tree. For example, in the tree presented in Figure 4.2, root($\{3,5,6\}$)$=\{3\}$, root($\{1,2,3,4,5,6\}$)$=\{1\}$, children($\{3,5,6\}$) $=\{\{5\},\{6\}\}$, and children($\{1,2,3,4,5,6\}$) $=\{\{2,4\},\{3,5,6\}\}$. Note that all the groups of children($g$) are necessarily included in $g$. The next lemma is proved in Appendix A.2.2.

**Lemma 5** (Correctness and complexity of Algorithm 8)
*When $\|.\|$ is chosen to be the $\ell_2$-norm, Algorithm 8 gives the solution of the primal problem Eq. (4.7) in $O(p)$ operations.*

So far the dictionary $\mathbf{D}$ was fixed to be for example a wavelet basis. In the next section, we apply the tools we developed for solving efficiently problem (4.5) to learn a dictionary $\mathbf{D}$ adapted to our hierarchical sparse coding formulation.

9. As a reminder, root($g$) is not a singleton when several dictionary elements are considered per node.

## 4.4 Application to Dictionary Learning

We start by briefly describing dictionary learning.

### 4.4.1 The Dictionary Learning Framework

Let us consider a set $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$ in $\mathbb{R}^{m \times n}$ of $n$ signals of dimension $m$. Dictionary learning is a matrix factorization problem which aims at representing these signals as linear combinations of the dictionary elements, that are the columns of a matrix $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p]$ in $\mathbb{R}^{m \times p}$. More precisely, the dictionary $\mathbf{D}$ is *learned* along with a matrix of decomposition coefficients $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n]$ in $\mathbb{R}^{p \times n}$, so that $\mathbf{x}^i \approx \mathbf{D}\boldsymbol{\alpha}^i$ for every signal $\mathbf{x}^i$.

While learning simultaneously $\mathbf{D}$ and $\mathbf{A}$, one may want to encode specific prior knowledge about the problem at hand, such as, for example, the positivity of the decomposition (Lee and Seung, 1999), or the sparsity of $\mathbf{A}$ (Olshausen and Field, 1997; Aharon et al., 2006; Lee et al., 2007; Mairal et al., 2010a). This leads to penalizing or constraining $(\mathbf{D}, \mathbf{A})$ and results in the following formulation:

$$\min_{\mathbf{D} \in \mathcal{D}, \mathbf{A} \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Big[ \frac{1}{2} \|\mathbf{x}^i - \mathbf{D}\boldsymbol{\alpha}^i\|_2^2 + \lambda \Psi(\boldsymbol{\alpha}^i) \Big], \tag{4.10}$$

where $\mathcal{A}$ and $\mathcal{D}$ denote two convex sets and $\Psi$ is a regularization term, usually a norm or a squared norm, whose effect is controlled by the regularization parameter $\lambda > 0$. Note that $\mathcal{D}$ is assumed to be bounded to avoid any degenerate solutions of Problem (4.10). For instance, the standard sparse coding formulation takes $\Psi$ to be the $\ell_1$-norm, $\mathcal{D}$ to be the set of matrices in $\mathbb{R}^{m \times p}$ whose columns have unit $\ell_2$-norm, with $\mathcal{A} = \mathbb{R}^{p \times n}$ (Olshausen and Field, 1997; Lee et al., 2007; Mairal et al., 2010a).

However, this classical setting treats each dictionary element independently from the others, and does not exploit possible relationships between them. To embed the dictionary in a tree structure, we therefore replace the $\ell_1$-norm by our hierarchical norm and set $\Psi = \Omega$ in Eq. (4.10).

A question of interest is whether hierarchical priors are more appropriate in supervised settings or in the matrix-factorization context in which we use it. It is not so common in the supervised setting to have strong prior information that allows us to organize the features in a hierarchy. On the contrary, in the case of dictionary learning, since the atoms are learned, one can argue that the dictionary elements learned will *have to* match well the hierarchical prior that is imposed by the regularization. In other words, combining structured regularization with dictionary learning has precisely the advantage that the dictionary elements will *self-organize* to match the prior.

### 4.4.2 Learning the Dictionary

Optimization for dictionary learning has already been intensively studied. We choose in this chapter a typical alternating scheme, which optimizes in turn $\mathbf{D}$ and $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n]$ while keeping the other variable fixed (Aharon et al., 2006; Lee et al.,

2007; Mairal et al., 2010a).[10] Of course, the convex optimization tools we develop in this chapter do not change the intrinsic non-convex nature of the dictionary learning problem. However, they solve the underlying convex subproblems efficiently, which is crucial to yield good results in practice. In the next section, we report good performance on some applied problems, and we show empirically that our algorithm is stable and does not seem to get trapped in bad local minima. The main difficulty of our problem lies in the optimization of the vectors $\boldsymbol{\alpha}^i$, $i$ in $\{1,\dots,n\}$, for the dictionary $\mathbf{D}$ kept fixed. Because of $\Omega$, the corresponding convex subproblem is nonsmooth and has to be solved for each of the $n$ signals considered. The optimization of the dictionary $\mathbf{D}$ (for $\mathbf{A}$ fixed), which we discuss first, is in general easier.

**Updating the dictionary D.** We follow the matrix-inversion free procedure of Mairal et al. (2010a) to update the dictionary. This method consists in iterating block-coordinate descent over the columns of $\mathbf{D}$. Specifically, we assume that the domain set $\mathcal{D}$ has the form

$$\mathcal{D}_\mu \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p},\ \mu\|\mathbf{d}^j\|_1 + (1-\mu)\|\mathbf{d}^j\|_2^2 \leq 1,\ \text{for all } j \in \{1,\dots,p\}\}, \tag{4.11}$$

or $\mathcal{D}_\mu^+ \triangleq \mathcal{D}_\mu \cap \mathbb{R}_+^{m\times p}$, with $\mu \in [0,1]$. The choice for these particular domain sets is motivated by the experiments of Section 4.5. For natural image patches, the dictionary elements are usually constrained to be in the unit $\ell_2$-norm ball (i.e., $\mathcal{D} = \mathcal{D}_0$), while for topic modeling, the dictionary elements are distributions of words and therefore belong to the simplex (i.e., $\mathcal{D} = \mathcal{D}_1^+$). The update of each dictionary element amounts to performing a Euclidean projection, which can be computed efficiently (Mairal et al., 2010a). Concerning the stopping criterion, we follow the strategy from the same authors and go over the columns of $\mathbf{D}$ only a few times, typically 5 times in our experiments. Although we have not explored locality constraints on the dictionary elements, these have been shown to be particularly relevant to some applications such as patch-based image classification (Yu et al., 2009). Combining tree structure and locality constraints is an interesting future research.

**Updating the vectors $\boldsymbol{\alpha}^i$.** The procedure for updating the columns of $\mathbf{A}$ is based on the results derived in Section 4.3.3. Furthermore, positivity constraints can be added on the domain of $\mathbf{A}$, by noticing that for our norm $\Omega$ and any vector $\mathbf{u}$ in $\mathbb{R}^p$, adding these constraints when computing the proximal operator is equivalent to solving $\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|[\mathbf{u}]_+ - \mathbf{v}\|_2^2 + \lambda\Omega(\mathbf{v})$. This equivalence is proved in Appendix A.2.2. We will indeed use positive decompositions to model text corpora in Section 4.5. Note that by constraining the decompositions $\boldsymbol{\alpha}^i$ to be nonnegative, some entries $\boldsymbol{\alpha}_j^i$ may be set to zero in addition to those already zeroed out by the norm $\Omega$. As a result, the sparsity patterns obtained in this way might not satisfy the tree-structured condition (4.1) anymore.

10. Note that although we use this classical scheme for simplicity, it would also be possible to use the stochastic approach proposed by Mairal et al. (2010a).

## 4.5 Experiments

We next turn to the experimental validation of our hierarchical sparse coding.

### 4.5.1 Implementation Details

In Section 4.3.3, we have shown that the proximal operator associated to $\Omega$ can be computed exactly and efficiently. The problem is therefore amenable to fast proximal algorithms that are well suited to nonsmooth convex optimization. Specifically, we tried the accelerated scheme from both Nesterov (2007) and Beck and Teboulle (2009), and finally opted for the latter since, for a comparable level of precision, fewer calls of the proximal operator are required. The basic proximal scheme presented in Section 4.3.1 is formalized by Beck and Teboulle (2009) as an algorithm called ISTA; the same authors propose moreover an accelerated variant, FISTA, which is a similar procedure, except that the operator is not directly applied on the current estimate, but on an auxiliary sequence of points that are linear combinations of past estimates. This latter algorithm has an optimal convergence rate in the class of first-order techniques, and also allows for warm restarts, which is crucial in the alternating scheme of dictionary learning. [11]

Finally, we monitor the convergence of the algorithm by checking the relative decrease in the cost function. [12] Unless otherwise specified, all the algorithms used in the following experiments are implemented in `C/C++`, with a `Matlab` interface. Our implementation is freely available at `http://www.di.ens.fr/willow/SPAMS/`.

### 4.5.2 Speed Benchmark

To begin with, we conduct speed comparisons between our approach and other convex programming methods, in the setting where $\Omega$ is chosen to be a linear combination of $\ell_2$-norms. The algorithms that take part in the following benchmark are:

- Proximal methods, with ISTA and the accelerated FISTA methods (Beck and Teboulle, 2009).
- A reweighted-least-square scheme (Re-$\ell_2$), as described by Jenatton et al. (2011a); Kim and Xing (2010). This approach is adapted to the square loss, since closed-form updates can be used. [13]
- Subgradient descent, whose step size is taken to be equal either to $a/(k+b)$ or $a/(\sqrt{k}+b)$ (respectively referred to as SG and $\text{SG}_{\text{sqrt}}$), where $k$ is the iteration number, and $(a, b)$ are the best [14] parameters selected on the logarithmic grid $(a, b) \in \{10^{-4}, \ldots, 10^{3}\} \times \{10^{-2}, \ldots, 10^{5}\}$.

11. Unless otherwise specified, the initial stepsize in ISTA/FISTA is chosen as the maximum eigenvalue of the sampling covariance matrix divided by 100, while the growth factor in the line search is set to 1.5.

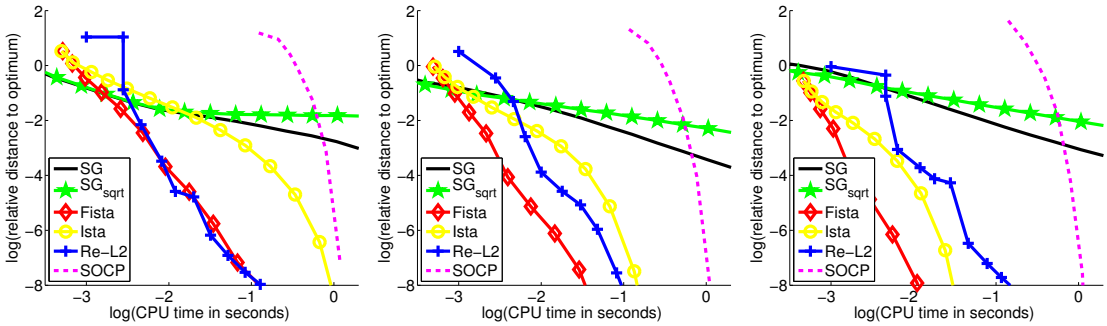
12. We are currently investigating algorithms for computing duality gaps based on network flow optimization tools (Mairal et al., 2010b).

13. The computation of the updates related to the variational formulation (4.6) also benefits from the hierarchical structure of $\mathcal{G}$, and can be performed in $O(p)$ operations.

14. "The best step size" is understood as being the step size leading to the smallest cost function after 500 iterations.

- A commercial software (`Mosek`, available at `http://www.mosek.com/`) for second-order cone programming (SOCP).

Moreover, the experiments we carry out cover various settings, with notably different sparsity regimes, i.e., low, medium and high, respectively corresponding to about $50\%, 10\%$ and $1\%$ of the total number of dictionary elements. Eventually, all reported results are obtained on a single core of a 3.07Ghz CPU with 8GB of memory.



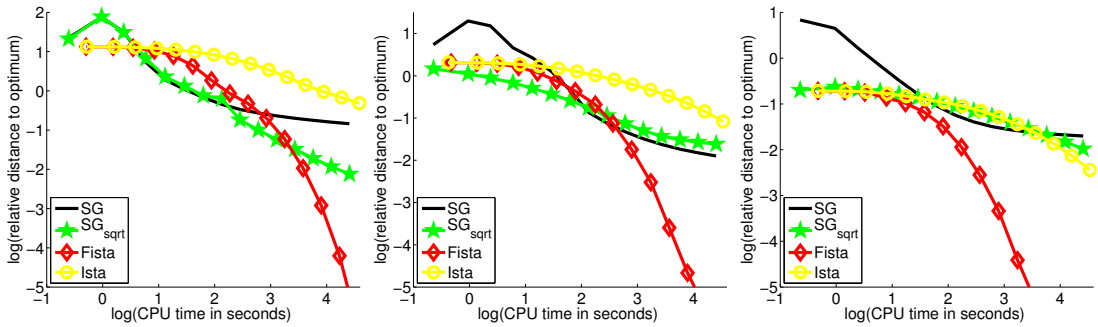

(a) scale: small, regul.: low (b) scale: small, regul.: medium (c) scale: small, regul.: high

Figure 4.4: Benchmark for solving a least-squares regression problem regularized by the hierarchical norm $\Omega$. The experiment is small scale, $m = 256, p = 151$, and shows the performances of six optimization methods (see main text for details) for three levels of regularization. The curves represent the relative value of the objective to the optimal value as a function of the computational time in second on a $\log_{10} / \log_{10}$ scale. All reported results are obtained by averaging 5 runs.

### Hierarchical dictionary of natural image patches

In this first benchmark, we consider a least-squares regression problem regularized by $\Omega$ that arises in the context of denoising of natural image patches, as further exposed in Section 4.5.4. In particular, based on a hierarchical dictionary, we seek to reconstruct noisy $16\times16$-patches. The dictionary we use is represented on Figure 4.9. Although the problem involves a small number of variables, i.e., $p = 151$ dictionary elements, it has to be solved repeatedly for tens of thousands of patches, at moderate precision. It is therefore crucial to be able to solve this problem quickly and efficiently.

We can draw several conclusions from the results of the simulations reported in Figure 4.4. First, we observe that in most cases, the accelerated proximal scheme performs better than the other approaches. In addition, unlike FISTA, ISTA seems to suffer in non-sparse scenarios. In the least sparse setting, the reweighted-$\ell_2$ scheme is the only method that competes with FISTA. It is however not able to yield truly sparse solutions, and would therefore need a subsequent (somewhat arbitrary) thresholding operation. As expected, the generic techniques such as SG and SOCP do not compete with dedicated algorithms.




(a) scale: large, regul.: low (b) scale: large, regul.: medium (c) scale: large, regul.: high

Figure 4.5: Benchmark for solving a large-scale multi-class classification problem for four optimization methods (see details about the datasets and the methods in the main text). Three levels of regularization are considered. The curves represent the relative value of the objective to the optimal value as a function of the computational time in second on a $\log_{10} / \log_{10}$ scale. In the highly regularized setting, tuning the step-size for the subgradient turned out to be difficult, which explains the behavior of SG in the first iterations.

### Multi-class classification of cancer diagnosis

The second benchmark explores a different supervised learning setting, where $f$ is no longer the square loss function. The goal is to demonstrate that our optimization tools apply in various scenarios, beyond traditional sparse approximation problems. To this end, we consider a gene expression dataset [15] in the context of cancer diagnosis. More precisely, we focus on a multi-class classification problem where the number $m$ of samples to be classified is small compared to the number $p$ of gene expressions that characterize these samples. Each atom thus corresponds to a gene expression across the $m$ samples, whose class labels are recorded in the vector $\mathbf{x}$ in $\mathbb{R}^m$.

The dataset contains $m = 308$ samples, $p = 30\,017$ variables and 26 classes. In addition, the data exhibit highly-correlated dictionary elements. Inspired by Kim and Xing (2010), we build the tree-structured set of groups $\mathcal{G}$ using Ward's hierarchical clustering (Johnson, 1967) on the gene expressions. The norm $\Omega$ built in this way aims at capturing the hierarchical structure of gene expression networks (Kim and Xing, 2010).

Instead of the square loss function, we consider the multinomial logistic loss function that is better suited to deal with multi-class classification problems (see, e.g., Hastie et al., 2009). As a direct consequence, algorithms whose applicability crucially depends on the choice of the loss function $f$ are removed from the benchmark. This is the case with reweighted-$\ell_2$ schemes that do not have closed-form updates anymore. Importantly, the choice of the multinomial logistic loss function leads to an optimization problem over a matrix with dimensions $p$ times the number of classes (i.e., a total of $30\,017 \times 26 \approx$

15. The dataset we use is *14_Tumors*, which is freely available at `http://www.gems-system.org/`.

780 000 variables). Also, due to scalability issues, generic interior point solvers could not be considered here.

The results in Figure 4.5 highlight that the accelerated proximal scheme performs overall better that the two other methods. Again, it is important to note that both proximal algorithms yield sparse solutions, which is not the case for SG.

### 4.5.3 Denoising with Tree-Structured Wavelets

We demonstrate in this section how a tree-structured sparse regularization can improve classical wavelet representation, and how our method can be used to efficiently solve the corresponding large-scale optimization problems. We consider two wavelet orthonormal bases, Haar and Daubechies3 (see Mallat, 1999), and choose a classical quad-tree structure on the coefficients, which has notably proven to be useful for image compression problems (Baraniuk, 1999). This experiment follows the approach of Zhao et al. (2009) who used the same tree-structured regularization in the case of small one-dimensional signals, and the approach of Baraniuk et al. (2010) and Huang et al. (2009) images where images were reconstructed from compressed sensing measurements with a hierarchical nonconvex penalty.

We compare the performance for image denoising of both nonconvex and convex approaches. Specifically, we consider the following formulation

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\psi(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{D}^\top\mathbf{x} - \boldsymbol{\alpha}\|_2^2 + \lambda\psi(\boldsymbol{\alpha}),$$

where $\mathbf{D}$ is one of the orthonormal wavelet basis mentioned above, $\mathbf{x}$ is the input noisy image, $\mathbf{D}\boldsymbol{\alpha}$ is the estimate of the denoised image, and $\psi$ is a sparsity-inducing regularization. Note that in this case, $m = p$. We first consider classical settings where $\psi$ is either the $\ell_1$-norm— this leads to the wavelet soft-thresholding method of Donoho and Johnstone (1995)— or the $\ell_0$-pseudo-norm, whose solution can be obtained by hard-thresholding (see Mallat, 1999). Then, we consider the convex tree-structured regularization $\Omega$ defined as a sum of $\ell_2$-norms ($\ell_\infty$-norms), which we denote by $\Omega_{\ell_2}$ (respectively $\Omega_{\ell_\infty}$). Since the basis is here orthonormal, solving the corresponding decomposition problems amounts to computing a single instance of the proximal operator. As a result, when $\psi$ is $\Omega_{\ell_2}$, we use Algorithm 8 and for $\Omega_{\ell_\infty}$, Algorithm 7 is applied. Finally, we consider the nonconvex tree-structured regularization used by Baraniuk et al. (2010) denoted here by $\ell_0^{\text{tree}}$, which we have presented in Eq. (4.4); the implementation details for $\ell_0^{\text{tree}}$ can be found in Appendix A.2.1. Compared to Zhao et al. (2009), the novelty of our approach is essentially to be able to solve efficiently and exactly large-scale instances of this problem. We use 12 classical standard test images, [16] and generate noisy versions of them corrupted by a white Gaussian noise of variance $\sigma$. For each image, we test several values of $\lambda = 2^{\frac{i}{4}}\sigma\sqrt{\log m}$, with $i$ taken in a specific range. [17] We then keep

16. These images are used in classical image denoising benchmarks. See Mairal et al. (2009b).

17. For the convex formulations, $i$ ranges in $\{-15, -14, \ldots, 15\}$, while in the nonconvex case $i$ ranges in $\{-24, \ldots, 48\}$.

| | | Haar | | | | |
|---|---|---|---|---|---|---|
| | $\sigma$ | $\ell_0$ [0.0012] | $\ell_0^{\text{tree}}$ [0.0098] | $\ell_1$ [0.0016] | $\Omega_{\ell_2}$ [0.0125] | $\Omega_{\ell_\infty}$ [0.0221] |
| PSNR | 5 | 34.48 | 34.78 | 35.52 | **35.89** | 35.79 |
| | 10 | 29.63 | 30.24 | 30.74 | **31.40** | 31.23 |
| | 25 | 24.44 | 25.27 | 25.30 | **26.41** | 26.14 |
| | 50 | 21.53 | 22.37 | 20.42 | **23.41** | 23.05 |
| | 100 | 19.27 | 20.09 | 19.43 | **20.97** | 20.58 |
| IPSNR | 5 | - | $.30 \pm .23$ | $1.04 \pm .31$ | $\mathbf{1.41 \pm .45}$ | $1.31 \pm .41$ |
| | 10 | - | $.60 \pm .24$ | $1.10 \pm .22$ | $\mathbf{1.76 \pm .26}$ | $1.59 \pm .22$ |
| | 25 | - | $.83 \pm .13$ | $.86 \pm .35$ | $\mathbf{1.96 \pm .22}$ | $1.69 \pm .21$ |
| | 50 | - | $.84 \pm .18$ | $.46 \pm .28$ | $\mathbf{1.87 \pm .20}$ | $1.51 \pm .20$ |
| | 100 | - | $.82 \pm .14$ | $.15 \pm .23$ | $\mathbf{1.69 \pm .19}$ | $1.30 \pm .19$ |

| | | Daub3 | | | | |
|---|---|---|---|---|---|---|
| | $\sigma$ | $\ell_0$ [0.0013] | $\ell_0^{\text{tree}}$ [0.0099] | $\ell_1$ [0.0017] | $\Omega_{\ell_2}$ [0.0129] | $\Omega_{\ell_\infty}$ [0.0204] |
| PSNR | 5 | 34.64 | 34.95 | 35.74 | **36.14** | 36.00 |
| | 10 | 30.03 | 30.63 | 31.10 | **31.79** | 31.56 |
| | 25 | 25.04 | 25.84 | 25.76 | **26.90** | 26.54 |
| | 50 | 22.09 | 22.90 | 22.42 | **23.90** | 23.41 |
| | 100 | 19.56 | 20.45 | 19.67 | **21.40** | 20.87 |
| IPSNR | 5 | - | $.31 \pm .21$ | $1.10 \pm .23$ | $\mathbf{1.49 \pm .34}$ | $1.36 \pm .31$ |
| | 10 | - | $.60 \pm .16$ | $1.06 \pm .25$ | $\mathbf{1.76 \pm .19}$ | $1.53 \pm .17$ |
| | 25 | - | $.80 \pm .10$ | $.71 \pm .28$ | $\mathbf{1.85 \pm .17}$ | $1.50 \pm .18$ |
| | 50 | - | $.81 \pm .15$ | $.33 \pm .24$ | $\mathbf{1.80 \pm .11}$ | $1.33 \pm .12$ |
| | 100 | - | $.89 \pm .13$ | $0.11 \pm .24$ | $\mathbf{1.82 \pm .24}$ | $1.30 \pm .17$ |

Table 4.1: Top part of the tables: Average PSNR measured for the denoising of 12 standard images, when the wavelets are Haar or Daubechies3 wavelets (see Mallat, 1999), for two nonconvex approaches ($\ell_0$ and $\ell_0^{\text{tree}}$) and three different convex regularizations—that is, the $\ell_1$-norm, the tree-structured sum of $\ell_2$-norms ($\Omega_{\ell_2}$), and the tree-structured sum of $\ell_\infty$-norms ($\Omega_{\ell_\infty}$). Best results for each level of noise and each wavelet type are in bold. Bottom part of the tables: Average improvement in PSNR with respect to the $\ell_0$ nonconvex method (the standard deviations are computed over the 12 images). CPU times (in second) averaged over all images and noise realizations are reported in brackets next to the names of the methods they correspond to.

the parameter $\lambda$ giving the best reconstruction error. The factor $\sigma\sqrt{\log m}$ is a classical heuristic for choosing a reasonable regularization parameter (see Mallat, 1999). We provide reconstruction results in terms of PSNR in Table 4.1.[18] We report in this table

18. Denoting by MSE the mean-squared-error for images whose intensities are between 0 and 255, the PSNR is defined as $\text{PSNR} = 10\log_{10}(255^2/\text{MSE})$ and is measured in dB. A gain of 1dB reduces the MSE by approximately 20%.

the results when $\Omega$ is chosen to be a sum of $\ell_2$-norms or $\ell_\infty$-norms with weights $\omega_g$ all equal to one. Each experiment was run 5 times with different noise realizations. In every setting, we observe that the tree-structured norm significantly outperforms the $\ell_1$-norm and the nonconvex approaches. We also present a visual comparison on two images on Figure 4.6, showing that the tree-structured norm reduces visual artefacts (these artefacts are better seen by zooming on a computer screen). The wavelet transforms in our experiments are computed with the matlabPyrTools software. [19]

(a) Lena, $\sigma = 25$, $\ell_1$ (b) Lena, $\sigma = 25$, $\Omega_{\ell_2}$ (c) Barb., $\sigma = 50$, $\ell_1$ (d) Barb., $\sigma = 50$, $\Omega_{\ell_2}$

Figure 4.6: Visual comparison between the wavelet shrinkage model with the $\ell_1$-norm and the tree-structured model, on cropped versions of the images Lena and Barb.. Haar wavelets are used.

This experiment does of course not provide state-of-the-art results for image denoising (see Mairal et al., 2009b, and references therein), but shows that the tree-structured regularization significantly improves the reconstruction quality for wavelets. In this experiment the convex setting $\Omega_{\ell_2}$ and $\Omega_{\ell_\infty}$ also outperforms the nonconvex one $\ell_0^{\text{tree}}$. [20] We also note that the speed of our approach makes it scalable to real-time applications. Solving the proximal problem for an image with $m = 512 \times 512 = 262\,144$ pixels takes approximately 0.013 seconds on a single core of a 3.07GHz CPU if $\Omega$ is a sum of $\ell_2$-norms, and 0.02 seconds when it is a sum of $\ell_\infty$-norms. By contrast, unstructured approaches have a speed-up factor of about 7-8 with respect to the tree-structured methods.

### 4.5.4 Dictionaries of Natural Image Patches

This experiment studies whether a hierarchical structure can help dictionaries for denoising natural image patches, and in which noise regime the potential gain is significant. We aim at reconstructing *corrupted* patches from a test set, after having learned dictionaries on a training set of *non-corrupted* patches. Though not typical in machine

19. `http://www.cns.nyu.edu/~eero/steerpyr/`.

20. It is worth mentioning that comparing convex and nonconvex approaches for sparse regularization is a bit difficult. This conclusion holds for the classical formulation we have used, but might not hold in other settings such as Coifman and Donoho (1995).

learning, this setting is reasonable in the context of images, where lots of non-corrupted patches are easily available. [21]

| noise | 50 % | 60 % | 70 % | 80 % | 90 % |
|---|---|---|---|---|---|
| flat | $19.3 \pm 0.1$ | $26.8 \pm 0.1$ | $36.7 \pm 0.1$ | $50.6 \pm 0.0$ | $72.1 \pm 0.0$ |
| tree | $18.6 \pm 0.1$ | $25.7 \pm 0.1$ | $35.0 \pm 0.1$ | $48.0 \pm 0.0$ | $65.9 \pm 0.3$ |

Table 4.2: Quantitative results of the reconstruction task on natural image patches. First row: percentage of missing pixels. Second and third rows: mean square error multiplied by 100, respectively for classical sparse coding, and tree-structured sparse coding.

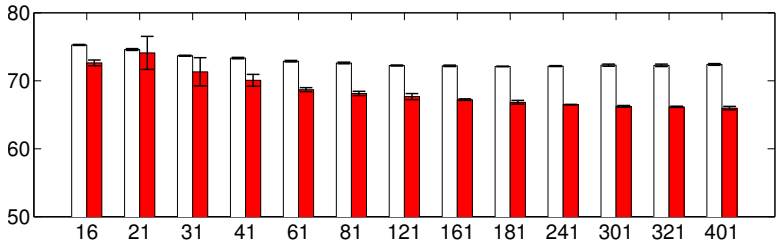



Figure 4.7: Mean square error multiplied by 100 obtained with 13 structures with error bars, sorted by number of dictionary elements from 16 to 401. Red plain bars represents the tree-structured dictionaries. White bars correspond to the flat dictionary model containing the same number of dictionary as the tree-structured one. For readability purpose, the $y$-axis of the graph starts at 50.

We extracted 100 000 patches of size $m = 8 \times 8$ pixels from the Berkeley segmentation database of natural images (Martin et al., 2001), which contains a high variability of scenes. We then split this dataset into a training set $\mathbf{X}_{tr}$, a validation set $\mathbf{X}_{val}$, and a test set $\mathbf{X}_{te}$, respectively of size 50 000, 25 000, and 25 000 patches. All the patches are centered and normalized to have unit $\ell_2$-norm.

For the first experiment, the dictionary $\mathbf{D}$ is learned on $\mathbf{X}_{tr}$ using the formulation of Eq. (4.10), with $\mu = 0$ for $\mathcal{D}_\mu$ as defined in Eq. (4.11). The validation and test sets are corrupted by removing a certain percentage of pixels, the task being to reconstruct the missing pixels from the known pixels. We thus introduce for each element $\mathbf{x}$ of the validation/test set, a vector $\tilde{\mathbf{x}}$, equal to $\mathbf{x}$ for the known pixel values and 0 otherwise. Similarly, we define $\tilde{\mathbf{D}}$ as the matrix equal to $\mathbf{D}$, except for the rows corresponding to missing pixel values, which are set to 0. By decomposing $\tilde{\mathbf{x}}$ on $\tilde{\mathbf{D}}$, we obtain a sparse code $\boldsymbol{\alpha}$, and the estimate of the reconstructed patch is defined as $\mathbf{D}\boldsymbol{\alpha}$. Note that this procedure assumes that we know which pixel is missing and which is not for every element $\mathbf{x}$.

21. Note that we study the ability of the model to reconstruct independent patches, and additional work is required to apply our framework to a full image processing task, where patches usually overlap (Elad and Aharon, 2006; Mairal et al., 2009b).

The parameters of the experiment are the regularization parameter $\lambda_{tr}$ used during the training step, the regularization parameter $\lambda_{te}$ used during the validation/test step, and the structure of the tree. For every reported result, these parameters were selected by taking the ones offering the best performance on the *validation* set, before reporting any result from the *test* set. The values for the regularization parameters $\lambda_{tr}, \lambda_{te}$ were selected on a logarithmic scale $\{2^{-10}, 2^{-9}, \ldots, 2^{2}\}$, and then further refined on a finer logarithmic scale with multiplicative increments of $2^{-1/4}$. For simplicity, we chose arbitrarily to use the $\ell_\infty$-norm in the structured norm $\Omega$, with all the weights equal to one. We tested 21 balanced tree structures of depth 3 and 4, with different *branching factors* $p_1, p_2, \ldots, p_{d-1}$, where $d$ is the depth of the tree and $p_k$, $k \in \{1, \ldots, d-1\}$ is the number of children for the nodes at depth $k$. The branching factors tested for the trees of depth 3 where $p_1 \in \{5, 10, 20, 40, 60, 80, 100\}$, $p_2 \in \{2, 3\}$, and for trees of depth 4, $p_1 \in \{5, 10, 20, 40\}$, $p_2 \in \{2, 3\}$ and $p_3 = 2$, giving 21 possible structures associated with dictionaries with at most 401 elements. For each tree structure, we evaluated the performance obtained with the tree-structured dictionary along with a non-structured dictionary containing the same number of elements. These experiments were carried out four times, each time with a different initialization, and with a different noise realization.

Quantitative results are reported in Table 4.2. For all fractions of missing pixels considered, the tree-structured dictionary outperforms the "unstructured one", and the most significant improvement is obtained in the noisiest setting. Note that having more dictionary elements is worthwhile when using the tree structure. To study the influence of the chosen structure, we report in Figure 4.7 the results obtained with the 13 tested structures of depth 3, along with those obtained with unstructured dictionaries containing the same number of elements, when 90% of the pixels are missing. For each dictionary size, the tree-structured dictionary significantly outperforms the unstructured one. An example of a learned tree-structured dictionary is presented on Figure 4.9. Dictionary elements naturally organize in groups of patches, often with low frequencies near the root of the tree, and high frequencies near the leaves.

### 4.5.5 Text Documents

This last experimental section shows that our approach can also be applied to model text corpora. The goal of probabilistic topic models is to find a low-dimensional representation of a collection of documents, where the representation should provide a semantic description of the collection. Approaching the problem in a parametric Bayesian framework, latent Dirichlet allocation (LDA) Blei et al. (2003) model documents, represented as vectors of word counts, as a mixture of a predefined number of *latent topics* that are distributions over a fixed vocabulary. LDA is fundamentally a matrix factorization problem: Buntine (2002) shows that LDA can be interpreted as a Dirichlet-multinomial counterpart of factor analysis. The number of topics is usually small compared to the size of the vocabulary (e.g., 100 against 10 000), so that the topic proportions of each document provide a compact representation of the corpus. For instance, these new features can be used to feed a classifier in a subsequent classification task. We similarly use our dictionary learning approach to find low-dimensional representations of text corpora.

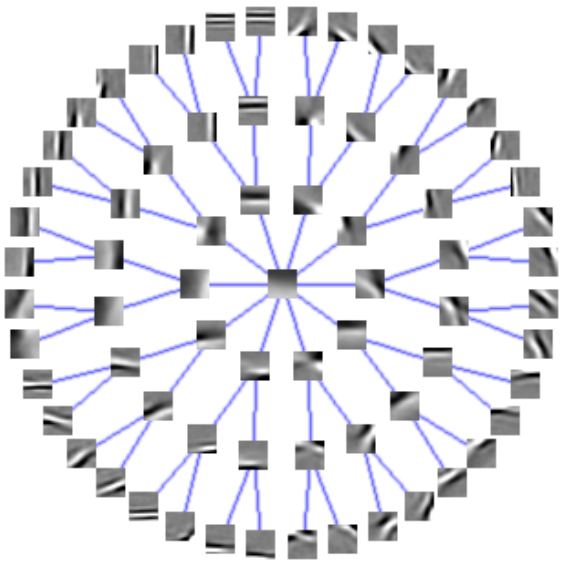


Figure 4.8: Learned dictionary with tree structure of depth 4. The root of the tree is in the middle of the figure. The branching factors are $p_1 = 10$, $p_2 = 2$, $p_3 = 2$. The dictionary is learned on $50,000$ patches of size $16 \times 16$ pixels.

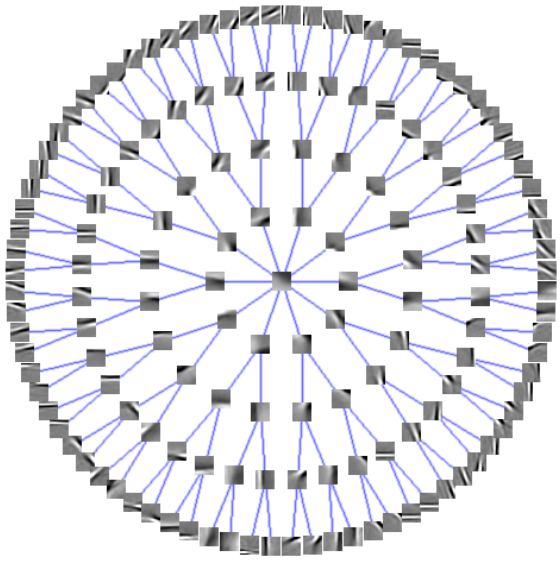


Figure 4.9: Learned dictionary with a tree structure of depth 5. The root of the tree is in the middle of the figure. The branching factors are $p_1 = 10$, $p_2 = 2$, $p_3 = 2$, $p_4 = 2$. The dictionary is learned on $50,000$ patches of size $16 \times 16$ pixels.

Suppose that the signals $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$ in $\mathbb{R}^{m \times n}$ are each the *bag-of-word* representation of each of $n$ documents over a vocabulary of $m$ words, the $k$-th component of $\mathbf{x}^i$ standing for the frequency of the $k$-th word in the document $i$. If we further assume that the entries of $\mathbf{D}$ and $\mathbf{A}$ are nonnegative, and that the dictionary elements $\mathbf{d}^j$ have unit $\ell_1$-norm, the decomposition $(\mathbf{D}, \mathbf{A})$ can be interpreted as the parameters of a topic-mixture model. The regularization $\Omega$ induces the organization of these topics on a tree, so that, if a document involves a certain topic, then all ancestral topics in the tree are also present in the topic decomposition. Since the hierarchy is shared by all documents, the topics at the top of the tree participate in every decomposition, and should therefore gather the lexicon which is common to all documents. Conversely, the deeper the topics in the tree, the more specific they should be. An extension of LDA to model topic hierarchies was proposed by Blei et al. (2010), who introduced a non-parametric Bayesian prior over trees of topics and modelled documents as convex combinations of topics selected along a path in the hierarchy. We plan to compare our approach with this model in future work.

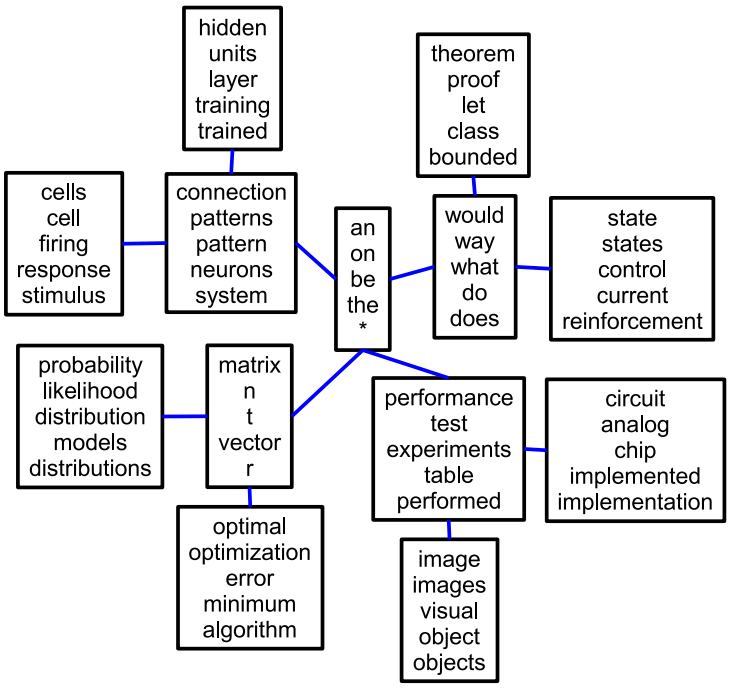



Figure 4.10: Example of a topic hierarchy estimated from 1714 NIPS proceedings papers (from 1988 through 1999). Each node corresponds to a topic whose 5 most important words are displayed. Single characters such as $n, t, r$ are part of the vocabulary and often appear in NIPS papers, and their place in the hierarchy is semantically relevant to children topics.

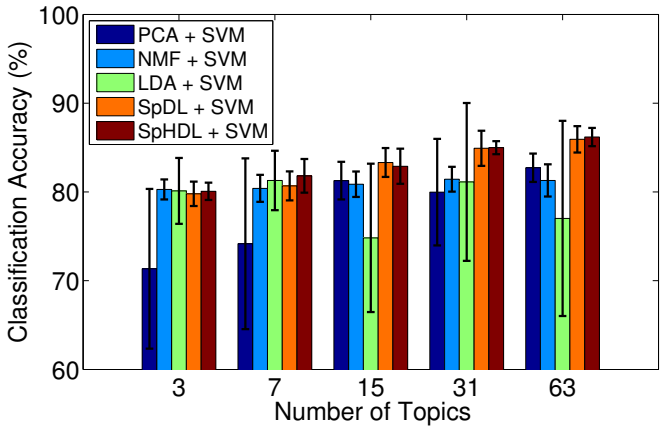



Figure 4.11: Binary classification of two newsgroups: classification accuracy for different dimensionality reduction techniques coupled with a linear SVM classifier. The bars and the errors are respectively the mean and the standard deviation, based on 10 random splits of the dataset. Best seen in color.

**Visualization of NIPS proceedings** We qualitatively illustrate our approach on the NIPS proceedings from 1988 through 1999 (Griffiths and Steyvers, 2004). After removing words appearing fewer than 10 times, the dataset is composed of 1714 articles, with a vocabulary of 8274 words. As explained above, we consider $\mathcal{D}_1^+$ and take $\mathcal{A}$ to be $\mathbb{R}_+^{p\times n}$. Figure 4.10 displays an example of a learned dictionary with 13 topics, obtained by using the $\ell_\infty$-norm in $\Omega$ and selecting manually $\lambda = 2^{-15}$. As expected and similarly to Blei et al. (2010), we capture the stopwords at the root of the tree, and topics reflecting the different subdomains of the conference such as neurosciences, optimization or learning theory.

**Posting classification** We now consider a binary classification task of $n$ postings from the 20 Newsgroups data set. [22] We learn to discriminate between the postings from the two newsgroups *alt.atheism* and *talk.religion.misc*, following the setting of Lacoste-Julien et al. (2008) and Zhu et al. (2009). After removing words appearing fewer than 10 times and standard stopwords, these postings form a data set of 1425 documents over a vocabulary of 13312 words. We compare different dimensionality reduction techniques that we use to feed a linear SVM classifier, i.e., we consider (i) LDA, with the code from Blei et al. (2003), (ii) principal component analysis (PCA), (iii) nonnegative matrix factorization (NMF), (iv) standard sparse dictionary learning (denoted by SpDL) and (v) our sparse hierarchical approach (denoted by SpHDL). Both SpDL and SpHDL are optimized over $\mathcal{D}_1^+$ and $\mathcal{A} = \mathbb{R}_+^{p\times n}$, with the weights $\omega_g$ equal to 1. We proceed as follows: given a random split into a training/test set of 1 000/425 postings, and given a number of topics $p$ (also the number of components for PCA, NMF, SpDL and

22. Available at `http://people.csail.mit.edu/jrennie/20Newsgroups/`.

SpHDL), we train an SVM classifier based on the low-dimensional representation of the postings. This is performed on a training set of 1 000 postings, where the parameters, $\lambda \in \{2^{-26}, \ldots, 2^{-5}\}$ and/or $C_{\text{svm}} \in \{4^{-3}, \ldots, 4^{1}\}$ are selected by 5-fold cross-validation. We report in Figure 4.11 the average classification scores on the test set of 425 postings, based on 10 random splits, for different number of topics. Unlike the experiment on image patches, we consider only complete binary trees with depths in $\{1, \ldots, 5\}$. The results from Figure 4.11 show that SpDL and SpHDL perform better than the other dimensionality reduction techniques on this task. As a baseline, the SVM classifier applied directly to the raw data (the 13312 words) obtains a score of $90.9 \pm 1.1$, which is better than all the tested methods, but without dimensionality reduction (as already reported by Blei et al., 2003). Moreover, the error bars indicate that, though nonconvex, SpDL and SpHDL do not seem to suffer much from instability issues. Even if SpDL and SpHDL perform similarly, SpHDL has the advantage to provide a more interpretable topic mixture in terms of hierarchy, which standard unstructured sparse coding does not.

## 4.6 Discussion

We have applied this approach in various settings, with fixed/learned dictionaries, and based on different types of data, namely, natural images and text documents. A line of research to pursue is to develop other optimization tools for structured norms with general overlapping groups. For instance, Mairal et al. (2010b) have used network flow optimization techniques for that purpose, and Bach (2010a) submodular function optimization. This framework can also be used in the context of hierarchical kernel learning (Bach, 2008a), where we believe that our method can be more efficient than existing ones.

This work establishes a connection between dictionary learning and probabilistic topic models, which should prove fruitful as the two lines of work have focused on different aspects of the same unsupervised learning problem: Our approach is based on convex optimization tools, and provides experimentally more stable data representations. Moreover, it can be easily extended with the same tools to other types of structures corresponding to other norms (Jenatton et al., 2011a; Jacob et al., 2009). It should be noted, however, that, unlike some Bayesian methods, dictionary learning by itself does not provide mechanisms for the automatic selection of model hyper-parameters (such as the dictionary size or the topology of the tree). An interesting common line of research to pursue could be the supervised design of dictionaries, which has been proved useful in the two frameworks (Mairal et al., 2009a; Bradley and Bagnell, 2009a; Blei and McAuliffe, 2008).

## 4.7 Extension: General Overlapping Groups and $\ell_1/\ell_\infty$-norms

The work presented in this extension section was achieved with the collaboration of Julien Mairal, Guillaume Obozinski and Francis Bach, with equal contribution between Julien Mairal and myself. The material we present below is based on the following work:

J. Mairal*, R. Jenatton*, G. Obozinski, F. Bach. Network Flow Algorithms for Structured Sparsity. Advances in Neural Information Processing Systems. 2010

J. Mairal*, R. Jenatton*, G. Obozinski, F. Bach. Convex and Network Flow Optimization for Structured Sparsity. In *Journal of Machine Learning Research*, 12, 2681-2720. 2011 (long version of the previous article)

(*equal contributions)

In this section, we consider an extension of the setting studied so far in the chapter. In particular, we now take $\Omega$ to be a linear combination of $\ell_\infty$-norms, while we do not assume anymore that $\mathcal{G}$ is tree-structured; $\mathcal{G}$ thus corresponds to a *general set of overlapping groups*. As it will be discussed at length later, the $\ell_\infty$-norm is piecewise linear, a property we will fully take advantage of. From now on, we focus on the computation of the following proximal operator:

$$\min_{\mathbf{w}\in\mathbb{R}^p}\Big[\frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2+\lambda\Omega(\mathbf{w})=\frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2+\lambda\sum_{g\in\mathcal{G}}\eta_g\|\mathbf{w}_g\|_\infty\Big]. \tag{4.12}$$

For all the detailed proofs of the results presented subsequently, we refer the interested readers to the long version Mairal et al. (2011).

We start by specifying the dual formulation from Lemma 2 to the setting of Eq. (4.12); we shall see that this dual problem can be reformulated as a *quadratic min-cost flow problem* for which we present an efficient algorithm.

**Lemma 6** (Dual of the proximal problem (4.12))
*Given* $\mathbf{u}$ *in* $\mathbb{R}^p$*, consider the problem*

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^{p\times|\mathcal{G}|}}\frac{1}{2}\|\mathbf{u}-\sum_{g\in\mathcal{G}}\boldsymbol{\xi}^g\|_2^2 \quad s.t. \quad \forall g\in\mathcal{G},\ \|\boldsymbol{\xi}^g\|_1\leq\lambda\eta_g \quad and \quad \boldsymbol{\xi}^g_j=0 \ \text{ if } j\notin g, \tag{4.13}$$

*where* $\boldsymbol{\xi}=(\boldsymbol{\xi}^g)_{g\in\mathcal{G}}$ *is in* $\mathbb{R}^{p\times|\mathcal{G}|}$*, and* $\boldsymbol{\xi}^g_j$ *denotes the* $j$*-th coordinate of the vector* $\boldsymbol{\xi}^g$*. Then, every solution* $\boldsymbol{\xi}^\star=(\boldsymbol{\xi}^{\star g})_{g\in\mathcal{G}}$ *of Eq. (4.13) satisfies* $\mathbf{w}^\star=\mathbf{u}-\sum_{g\in\mathcal{G}}\boldsymbol{\xi}^{\star g}$*, where* $\mathbf{w}^\star$ *is the solution of Eq. (4.12).*

Without loss of generality,[23] we assume from now on that the scalars $\mathbf{u}_j$ are all non-negative, and we constrain the entries of $\boldsymbol{\xi}$ to be non-negative. Such a formulation introduces $p|\mathcal{G}|$ dual variables which can be much greater than $p$, the number of primal variables, but it removes the issue of overlapping regularization. We now associate a graph with problem (4.13), on which the variables $\boldsymbol{\xi}_j^g$, for $g$ in $\mathcal{G}$ and $j$ in $g$, can be interpreted as measuring the components of a flow.

### 4.7.1 Graph Model

Let $G$ be a directed graph $G = (V, E, s, t)$, where $V$ is a set of vertices, $E \subseteq V \times V$ a set of arcs, $s$ a source, and $t$ a sink. Let $c : E \to \mathbb{R}_+$ and $c' : E \to \mathbb{R}_+ \cup \{+\infty\}$ be two functions on the arcs, where $c$ is a *cost function* and $c'$ is a non-negative *capacity function*. As done classically in the network flow literature (Ahuja et al., 1993; Bertsekas, 1991), we define a *flow* as a non-negative function on arcs that satisfies capacity constraints on all arcs (the value of the flow on an arc is less than or equal to the arc capacity) and conservation constraints on all vertices (the sum of incoming flows at a vertex is equal to the sum of outgoing flows) except for the source and the sink. We now introduce the *canonical* graph $G$ associated with our optimization problem:

**Definition 3** (Canonical Graph)

*Let $\mathcal{G} \subseteq 2^{[\![1;p]\!]}$ be a set of groups, and $(\eta_g)_{g\in\mathcal{G}}$ be positive weights. The canonical graph $G = (V, E, s, t)$ is the unique graph defined as follows:*

1. *$V = V_u \cup V_{gr}$, where $V_u$ is a vertex set of size $p$, one vertex being associated to each index $j$ in $[\![1;p]\!]$, and $V_{gr}$ is a vertex set of size $|\mathcal{G}|$, one vertex per group $g$ in $\mathcal{G}$. We thus have $|V| = |\mathcal{G}| + p$. For simplicity, we identify groups $g$ in $\mathcal{G}$ and indices $j$ in $[\![1;p]\!]$ with vertices of the graph, such that one can from now on refer to "vertex $j$" or "vertex $g$".*
2. *For every group $g$ in $\mathcal{G}$, $E$ contains an arc $(s, g)$. These arcs have capacity $\lambda\eta_g$ and zero cost.*
3. *For every group $g$ in $\mathcal{G}$, and every index $j$ in $g$, $E$ contains an arc $(g, j)$ with zero cost and infinite capacity. We denote by $\boldsymbol{\xi}_j^g$ the flow on this arc.*
4. *For every index $j$ in $[\![1;p]\!]$, $E$ contains an arc $(j, t)$ with infinite capacity and a cost $\frac{1}{2}(\mathbf{u}_j - \overline{\boldsymbol{\xi}}_j)^2$, where $\overline{\boldsymbol{\xi}}_j$ is the flow on $(j, t)$.*

Examples of canonical graphs are given in Figures 4.12a-(c) for three simple group structures. The flows $\boldsymbol{\xi}_j^g$ associated with $G$ can now be identified with the variables of problem (4.13). Since we have assumed the entries of $\mathbf{u}$ to be non-negative, we can now

23. Let $\boldsymbol{\xi}^\star$ denote a solution of Eq. (4.13). Optimality conditions of Eq. (4.13) derived in Jenatton et al. (2010a, 2011c) show that for all $j$ in $[\![1;p]\!]$, the signs of the non-zero coefficients $\boldsymbol{\xi}_j^{\star g}$ for $g$ in $\mathcal{G}$ are the same as the signs of the entries $\mathbf{u}_j$. To solve Eq. (4.13), one can therefore flip the signs of the negative variables $\mathbf{u}_j$, then solve the modified dual formulation (with non-negative variables), which gives the magnitude of the entries $\boldsymbol{\xi}_j^{\star g}$ (the signs of these being known).

reformulate Eq. (4.13) as

$$\min_{\boldsymbol{\xi}\in\mathbb{R}_+^{p\times|\mathcal{G}|},\overline{\boldsymbol{\xi}}\in\mathbb{R}^p} \sum_{j=1}^{p} \frac{1}{2}(\mathbf{u}_j - \overline{\boldsymbol{\xi}}_j)^2 \text{ s.t. } \overline{\boldsymbol{\xi}} = \sum_{g\in\mathcal{G}} \boldsymbol{\xi}^g \text{ and } \forall g \in \mathcal{G}, \begin{cases} \sum_{j\in g} \boldsymbol{\xi}_j^g \leq \lambda\eta_g, \\ \boldsymbol{\xi}_j^g = 0 \text{ for } j \notin g. \end{cases} \tag{4.14}$$

Indeed,

- the only arcs with a cost are those leading to the sink, which have the form $(j, t)$, where $j$ is the index of a variable in $[\![1; p]\!]$. The sum of these costs is $\sum_{j=1}^{p} \frac{1}{2}(\mathbf{u}_j - \overline{\boldsymbol{\xi}}_j)^2$, which is the objective function minimized in Eq. (4.14);
- by flow conservation, we necessarily have $\overline{\boldsymbol{\xi}}_j = \sum_{g\in\mathcal{G}} \boldsymbol{\xi}_j^g$ in the canonical graph;
- the only arcs with a capacity constraints are those coming out of the source, which have the form $(s, g)$, where $g$ is a group in $\mathcal{G}$. By flow conservation, the flow on an arc $(s, g)$ is $\sum_{j\in g} \boldsymbol{\xi}_j^g$ which should be less than $\lambda\eta_g$ by capacity constraints;
- all other arcs have the form $(g, j)$, where $g$ is in $\mathcal{G}$ and $j$ is in $g$. Thus, $\boldsymbol{\xi}_j^g = 0$ for $j \notin g$.

Therefore we have shown that finding a flow *minimizing the sum of the costs* on such a graph is equivalent to solving problem (4.13). When some groups are included in others, the canonical graph can be simplified to yield a graph with a smaller number of edges. Specifically, if $h$ and $g$ are groups with $h \subset g$, the edges $(g, j)$ for $j \in h$ carrying a flow $\boldsymbol{\xi}_j^g$ can be removed and replaced by a single edge $(g, h)$ of infinite capacity and zero cost, carrying the flow $\sum_{j\in h} \boldsymbol{\xi}_j^g$. This simplification is illustrated in Figure 4.12d, with a graph equivalent to the one of Figure 4.12c. This does not change the optimal value of $\overline{\boldsymbol{\xi}}^\star$, which is the quantity of interest for computing the optimal primal variable $\mathbf{w}^\star$. These simplifications are useful in practice, since they reduce the number of edges in the graph and improve the speed of our algorithms.

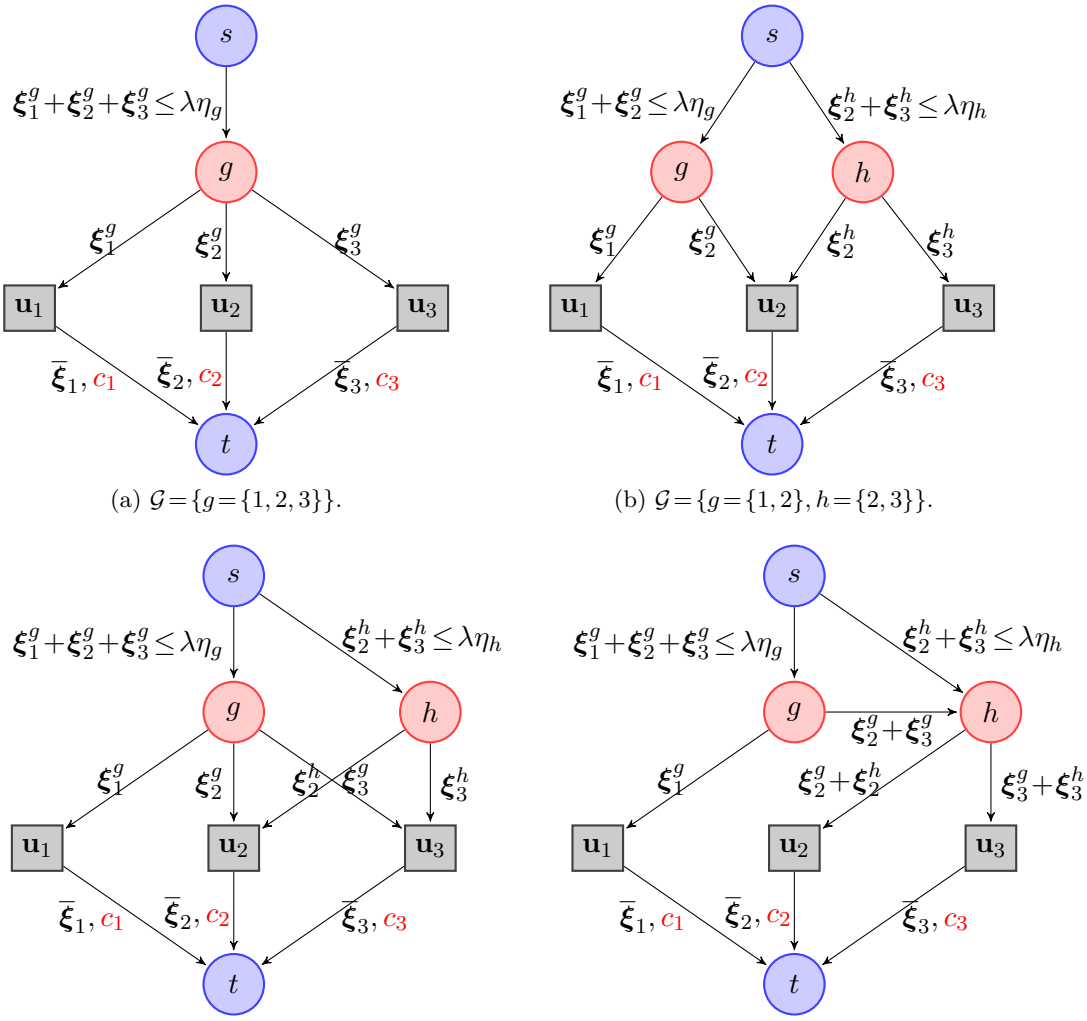
### 4.7.2 Computation of the Proximal Operator

Quadratic min-cost flow problems have been well studied in the operations research literature (Hochbaum and Hong, 1995). One of the simplest cases, where $\mathcal{G}$ contains a single group as in Figure 4.12a, is solved by an orthogonal projection on the $\ell_1$-ball of radius $\lambda\eta_g$. It has been shown, both in machine learning (Duchi et al., 2008) and operations research (Hochbaum and Hong, 1995; Brucker, 1984), that such a projection can be computed in $O(p)$ operations. When the group structure is a tree as in Figure 4.12d, strategies developed in the two communities are also similar (Jenatton et al., 2010a; Hochbaum and Hong, 1995), [24] and solve the problem in $O(pd)$ operations, where $d$ is the depth of the tree.

The general case of overlapping groups is more difficult. Hochbaum and Hong (1995) have shown that *quadratic min-cost flow problems* can be reduced to a specific *parametric max-flow* problem, for which an efficient algorithm exists (Gallo et al., 1989). [25] While

24. Note however that, while Hochbaum and Hong (1995) only consider a tree-structured sum of $\ell_\infty$-norms, the results from Jenatton et al. (2010a) also apply for a sum of $\ell_2$-norms.

25. By definition, a parametric max-flow problem consists in solving, for every value of a parameter, a max-flow problem on a graph whose arc capacities depend on this parameter.




(a) $\mathcal{G}=\{g=\{1,2,3\}\}$. (b) $\mathcal{G}=\{g=\{1,2\},h=\{2,3\}\}$.

(c) $\mathcal{G}=\{g=\{1,2,3\},h=\{2,3\}\}$. (d) $\mathcal{G}=\{g=\{1\}\cup h,h=\{2,3\}\}$.

Figure 4.12: Graph representation of simple proximal problems with different group structures $\mathcal{G}$. The three indices $1,2,3$ are represented as grey squares, and the groups $g,h$ in $\mathcal{G}$ as red discs. The source is linked to every group $g,h$ with respective maximum capacity $\lambda\eta_g,\lambda\eta_h$ and zero cost. Each variable $\mathbf{u}_j$ is linked to the sink $t$, with an infinite capacity, and with a cost $c_j \triangleq \frac{1}{2}(\mathbf{u}_j-\overline{\xi}_j)^2$. All other arcs in the graph have zero cost and infinite capacity. They represent inclusion relations in-between groups, and between groups and variables. The graphs (c) and (d) correspond to a special case of tree-structured hierarchy in the sense of Jenatton et al. (2010a). Their min-cost flow problems are equivalent.

this generic approach could be used to solve Eq. (4.13), we propose to use Algorithm 9 that also exploits the fact that our graphs have non-zero costs only on edges leading to the sink. As further shown in Mairal et al. (2011) through speed benchmarks, it it has a significantly better performance in practice. This algorithm clearly shares some similarities with existing approaches in network flow optimization such as the simplified version of Gallo et al. (1989) presented by Babenko and Goldberg (2006) that uses a divide and conquer strategy. Moreover, an equivalent algorithm exists for minimizing convex functions over polymatroid sets (Groenevelt, 1991). This equivalence, a priori non trivial, is uncovered through a representation of structured sparsity-inducing norms via submodular functions, which was recently proposed by Bach (2010a).

**Algorithm 9** Computation of the proximal operator for overlapping groups.

1: **Input:** $\mathbf{u} \in \mathbb{R}^p$, a set of groups $\mathcal{G}$, positive weights $(\eta_g)_{g\in\mathcal{G}}$, and $\lambda$ (regularization parameter).
2: Build the initial graph $G_0 = (V_0, E_0, s, t)$ as explained in Section 4.7.2.
3: Compute the optimal flow: $\overline{\boldsymbol{\xi}} \leftarrow \texttt{computeFlow}(V_0, E_0)$.
4: **Return:** $\mathbf{w} = \mathbf{u} - \overline{\boldsymbol{\xi}}$ (optimal solution of the proximal problem).

**Function** `computeFlow`$(V = V_u \cup V_{gr}, E)$

1: Projection step: $\boldsymbol{\gamma} \leftarrow \arg\min_{\gamma} \sum_{j\in V_u} \frac{1}{2}(\mathbf{u}_j - \boldsymbol{\gamma}_j)^2$ s.t. $\sum_{j\in V_u} \boldsymbol{\gamma}_j \leq \lambda \sum_{g\in V_{gr}} \eta_g$.
2: For all nodes $j$ in $V_u$, set $\boldsymbol{\gamma}_j$ to be the capacity of the arc $(j,t)$.
3: Max-flow step: Update $(\overline{\boldsymbol{\xi}}_j)_{j\in V_u}$ by computing a max-flow on the graph $(V, E, s, t)$.
4: **if** $\exists\, j \in V_u$ s.t. $\overline{\boldsymbol{\xi}}_j \neq \boldsymbol{\gamma}_j$ **then**
5: Denote by $(s, V^+)$ and $(V^-, t)$ the two disjoint subsets of $(V, s, t)$ separated by the minimum $(s,t)$-cut of the graph, and remove the arcs between $V^+$ and $V^-$. Call $E^+$ and $E^-$ the two remaining disjoint subsets of $E$ corresponding to $V^+$ and $V^-$.
6: $(\overline{\boldsymbol{\xi}}_j)_{j\in V_u^+} \leftarrow \texttt{computeFlow}(V^+, E^+)$.
7: $(\overline{\boldsymbol{\xi}}_j)_{j\in V_u^-} \leftarrow \texttt{computeFlow}(V^-, E^-)$.
8: **end if**
9: **Return:** $(\overline{\boldsymbol{\xi}}_j)_{j\in V_u}$.

The intuition behind our algorithm, `computeFlow` (see Algorithm 9), is the following: since $\overline{\boldsymbol{\xi}} = \sum_{g\in\mathcal{G}} \boldsymbol{\xi}^g$ is the only value of interest to compute the solution of the proximal operator $\mathbf{w} = \mathbf{u} - \overline{\boldsymbol{\xi}}$, the first step looks for a candidate value $\boldsymbol{\gamma}$ for $\overline{\boldsymbol{\xi}}$ by solving the following relaxed version of problem (4.14):

$$\arg\min_{\gamma\in\mathbb{R}^p} \sum_{j\in V_u} \frac{1}{2}(\mathbf{u}_j - \boldsymbol{\gamma}_j)^2 \text{ s.t. } \sum_{j\in V_u} \boldsymbol{\gamma}_j \leq \lambda \sum_{g\in V_{gr}} \eta_g. \tag{4.15}$$

The cost function here is the same as in problem (4.14), but the constraints are weaker: Any feasible point of problem (4.14) is also feasible for problem (4.15). This problem can be solved in linear time (Brucker, 1984). Its solution, which we denote $\boldsymbol{\gamma}$ for simplicity, provides the lower bound $\|\mathbf{u} - \boldsymbol{\gamma}\|_2^2/2$ for the optimal cost of problem (4.14).

The second step tries to construct a feasible flow $(\boldsymbol{\xi}, \overline{\boldsymbol{\xi}})$, satisfying additional capacity constraints equal to $\gamma_j$ on arc $(j,t)$, and whose cost matches this lower bound; this latter problem can be cast as a max-flow problem (Goldberg and Tarjan, 1988). If such a flow exists, the algorithm returns $\overline{\boldsymbol{\xi}} = \boldsymbol{\gamma}$, the cost of the flow reaches the lower bound, and is therefore optimal. If such a flow does not exist, we have $\overline{\boldsymbol{\xi}} \neq \boldsymbol{\gamma}$, the lower bound is not achievable, and we build a minimum $(s,t)$-cut of the graph (Ford and Fulkerson, 1987) defining two disjoints sets of nodes $V^+$ and $V^-$; $V^+$ is the part of the graph that could potentially have received more flow from the source (the arcs between $s$ and $V^+$ are not saturated), whereas $V^-$ could not (all arcs linking $s$ to $V^-$ are saturated). At this point, it is possible to show that the value of the optimal min-cost flow on all arcs between $V^+$ and $V^-$ is necessary zero. Thus, removing them yields an equivalent optimization problem, which can be decomposed into two independent problems of smaller sizes and solved recursively by the calls to `computeFlow`$(V^+, E^+)$ and `computeFlow`$(V^-, E^-)$. A formal proof of correctness of Algorithm 9 and further details can be found in Mairal et al. (2011).

### 4.7.3 Experiments

In this section, we present two experiments demonstrating the applicability and the benefits of our methods for solving large-scale sparse and structured regularized problems (additional applications can be found in Mairal et al., 2011).

#### CUR-like Matrix Factorization

In this experiment, we show how our tools can be used to perform the so-called CUR matrix decomposition (Mahoney and Drineas, 2009). It consists of a low-rank approximation of a data matrix $\mathbf{X}$ in $\mathbb{R}^{n\times p}$ in the form of a product of three matrices—that is, $\mathbf{X} \approx \mathbf{CUR}$. The particularity of the CUR decomposition lies in the fact that the matrices $\mathbf{C} \in \mathbb{R}^{n\times c}$ and $\mathbf{R} \in \mathbb{R}^{r\times p}$ are constrained to be respectively a subset of $c$ columns and $r$ rows of the original matrix $\mathbf{X}$. The third matrix $\mathbf{U} \in \mathbb{R}^{c\times r}$ is then given by $\mathbf{C}^+\mathbf{X}\mathbf{R}^+$, where $\mathbf{A}^+$ denotes a Moore-Penrose generalized inverse of the matrix $\mathbf{A}$ (Horn and Johnson, 1990). Such a matrix factorization is particularly appealing when the interpretability of the results matters (Mahoney and Drineas, 2009). For instance, when studying gene-expression datasets, it is easier to gain insight from the selection of actual patients and genes, rather than from linear combinations of them.

In Mahoney and Drineas (2009), CUR decompositions are computed by a sampling procedure based on the singular value decomposition of $\mathbf{X}$. In a recent work, Bien et al. (2010) have shown that *partial* CUR decompositions, i.e., the selection of either rows or columns of $\mathbf{X}$, can be obtained by solving a convex program with a group-Lasso penalty. We propose to extend this approach to the simultaneous selection of both rows and columns of $\mathbf{X}$, with the following convex problem:

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda_{\mathrm{row}} \sum_{i=1}^{n} \|\mathbf{W}^i\|_\infty + \lambda_{\mathrm{col}} \sum_{j=1}^{p} \|\mathbf{W}_j\|_\infty. \tag{4.16}$$

In this formulation, the two sparsity-inducing penalties controlled by the parameters $\lambda_{\text{row}}$ and $\lambda_{\text{col}}$ set to zero some entire rows and columns of the solutions of problem (4.16). Now, let us denote by $\mathbf{W}_{\mathrm{I}\,\mathrm{J}}$ in $\mathbb{R}^{|\mathrm{I}|\times|\mathrm{J}|}$ the submatrix of $\mathbf{W}$ reduced to its nonzero rows and columns, respectively indexed by $\mathrm{I} \subseteq \{1,\ldots,p\}$ and $\mathrm{J} \subseteq \{1,\ldots,n\}$. We can then readily identify the three components of the CUR decomposition of $\mathbf{X}$, namely

$$\mathbf{X}\mathbf{W}\mathbf{X} = \mathbf{C}\mathbf{W}_{\mathrm{I}\,\mathrm{J}}\mathbf{R} \approx \mathbf{X}.$$

Problem (4.16) has a smooth convex data-fitting term and brings into play a sparsity-inducing norm with overlapping groups of variables (the rows and the columns of $\mathbf{W}$). As a result, it can be handled with the optimization tools introduced in this section. We now compare the performance of the sampling procedure from Mahoney and Drineas (2009) with our proposed sparsity-based approach. To this end, we consider the four gene-expression datasets `9_Tumors`, `Brain_Tumors1`, `Leukemia1` and `SRBCT`, with respective dimensions $(n,p) \in \{(60,5727),(90,5921),(72,5328),(83,2309)\}$.[26] In the sequel, the matrix $\mathbf{X}$ is normalized to have unit Frobenius-norm while each of its columns is centered. To begin with, we run our approach[27] over a grid of values for $\lambda_{\text{row}}$ and $\lambda_{\text{col}}$ in order to obtain solutions with different sparsity levels, i.e., ranging from $|\mathrm{I}| = p$ and $|\mathrm{J}| = n$ down to $|\mathrm{I}| = |\mathrm{J}| = 0$. For each pair of values $[|\mathrm{I}|,|\mathrm{J}|]$, we then apply the sampling procedure from Mahoney and Drineas (2009). Finally, the variance explained by the CUR decompositions is reported in Figure 4.13 for both methods. Since the sampling approach involves some randomness, we show the average and standard deviation of the results based on five initializations. The conclusions we can draw from the experiments match the ones already reported in Bien et al. (2010) for the partial CUR decomposition. We can indeed see that both schemes perform similarly. However, our approach has the advantage not to be randomized, which can be less disconcerting in the practical perspective of analyzing a single run of the algorithm. It is finally worth being mentioned that the convex approach we develop here is flexible and can be extended in different ways. For instance, we may imagine add further low-rank/sparsity constraints on $\mathbf{W}$ thanks to sparsity-promoting convex regularizations.

### 4.7.4 Background Subtraction

Following Cevher et al. (2008); Huang et al. (2009), we consider a background subtraction task. Given a sequence of frames from a fixed camera, we try to segment out foreground objects in a new image. If we denote by $\mathbf{y} \in \mathbb{R}^n$ this image composed of $n$ pixels, we model $\mathbf{y}$ as a sparse linear combination of $p$ other images $\mathbf{X} \in \mathbb{R}^{n\times p}$, plus an error term $\mathbf{e}$ in $\mathbb{R}^n$, i.e., $\mathbf{y} \approx \mathbf{X}\mathbf{w} + \mathbf{e}$ for some sparse vector $\mathbf{w}$ in $\mathbb{R}^p$. This approach is reminiscent of Wright et al. (2008) in the context of face recognition, where $\mathbf{e}$ is further made sparse to deal with small occlusions. The term $\mathbf{X}\mathbf{w}$ accounts for *background* parts

26. The datasets are freely available at `http://www.gems-system.org/`.

27. More precisely, since the penalties in problem (4.16) shrink the coefficients of $\mathbf{W}$, we follow a two-step procedure: We first run our approach to determine the sets of nonzero rows and columns, and then compute $\mathbf{W}_{\mathrm{I}\,\mathrm{J}} = \mathbf{C}^{+}\mathbf{X}\mathbf{R}^{+}$.

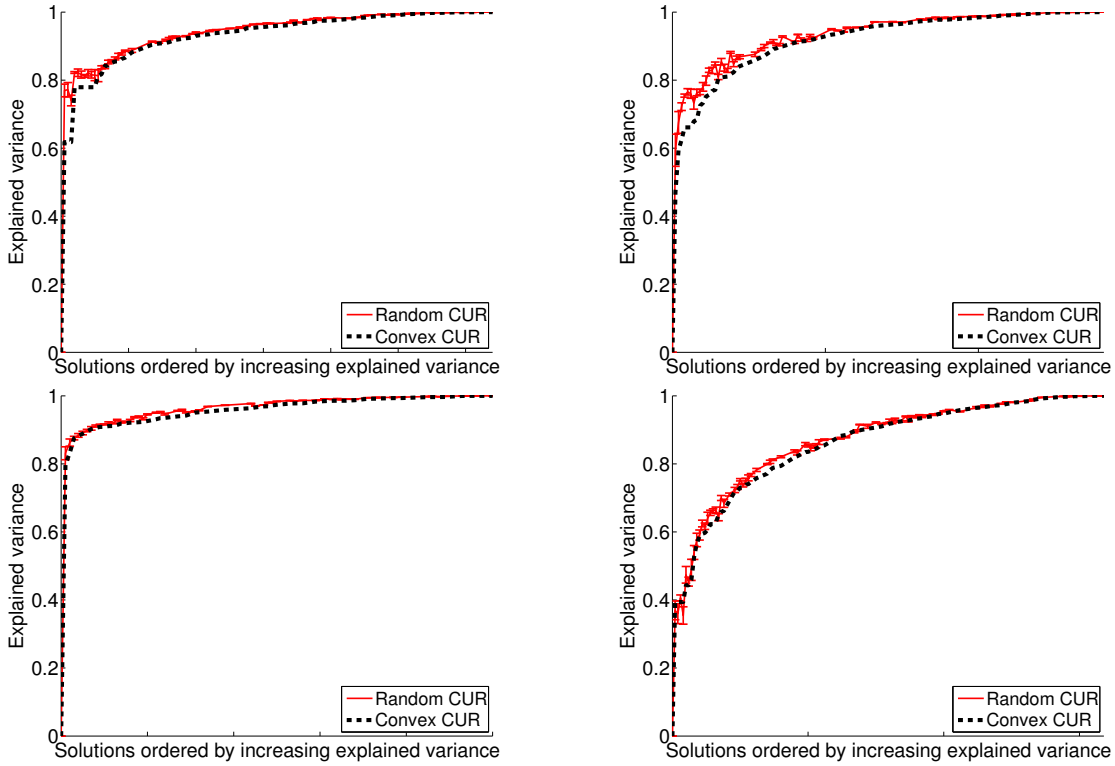



Figure 4.13: Explained variance of the CUR decompositions obtained for our sparsity-based approach and the sampling scheme from Mahoney and Drineas (2009). For the latter, we report the average and standard deviation of the results based on five initializations. From left to right and top to bottom, the curves correspond to the datasets `9_Tumors`, `Brain_Tumors1`, `Leukemia1` and `SRBCT`.

present in both $\mathbf{y}$ and $\mathbf{X}$, while $\mathbf{e}$ contains specific, or *foreground*, objects in $\mathbf{y}$. The resulting optimization problem is given by

$$\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{e}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}-\mathbf{e}\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\{\|\mathbf{e}\|_1 + \Omega(\mathbf{e})\}, \text{ with } \lambda_1,\lambda_2 \geq 0. \tag{4.17}$$

In this formulation, the only $\ell_1$-norm penalty does not take into account the fact that neighboring pixels in $\mathbf{y}$ are likely to share the same label (background or foreground), which may lead to scattered pieces of foreground and background regions (Figure 4.14). We therefore put an additional structured regularization term $\Omega$ on $\mathbf{e}$, where the groups in $\mathcal{G}$ are all the overlapping $3\times3$-squares on the image.

This optimization problem can be viewed as an instance of least-squares problem regularized by $\Omega$, with the particular design matrix $[\mathbf{X},\ \mathbf{I}]$ in $\mathbb{R}^{n\times(p+n)}$, defined as the columnwise concatenation of $\mathbf{X}$ and the identity matrix. As a result, we could directly apply the same procedure as the one used in the other experiments. Instead, we further exploit the specific structure of problem (4.17): Notice that for a fixed vector $\mathbf{e}$, the

optimization with respect to $\mathbf{w}$ is a standard Lasso problem (with the vector of observations $\mathbf{y} - \mathbf{e}$),[28] while for $\mathbf{w}$ fixed, we simply have a proximal problem associated to the sum of $\Omega$ and the $\ell_1$-norm. Alternating between these two simple and computationally inexpensive steps, i.e., optimizing with respect to one variable while keeping the other one fixed, is guaranteed to converge to a solution of (4.17).[29] In our simulations, this alternating scheme has led to a significant speed-up compared to the general procedure.

A dataset with hand-segmented images is used to illustrate the effect of $\Omega$.[30] For simplicity, we use a single regularization parameter, i.e., $\lambda_1 = \lambda_2$, chosen to maximize the number of pixels matching the ground truth. We consider $p = 200$ images with $n = 57600$ pixels (i.e., a resolution of $120 \times 160$, times 3 for the RGB channels). As shown in Figure 4.14, adding $\Omega$ improves the background subtraction results for the two tested images, by removing the scattered artifacts due to the lack of structural constraints of the $\ell_1$-norm, which encodes neither spatial nor color consistency.

28. Since successive frames might not change much, the columns of $\mathbf{X}$ exhibit strong correlations. As a result, we use the LARS algorithm (Efron et al., 2004) whose complexity is independent of the level of correlation in $\mathbf{X}$.

29. More precisely, the convergence is guaranteed since the non-smooth part in (4.17) is *separable* with respect to $\mathbf{w}$ and $\mathbf{e}$ (Tseng, 2001). The result from Bertsekas (1999) may also be applied here, after reformulating (4.17) as a smooth convex problem under separable conic constraints.

30. `http://research.microsoft.com/en-us/um/people/jckrumm/wallflower/testimages.htm`

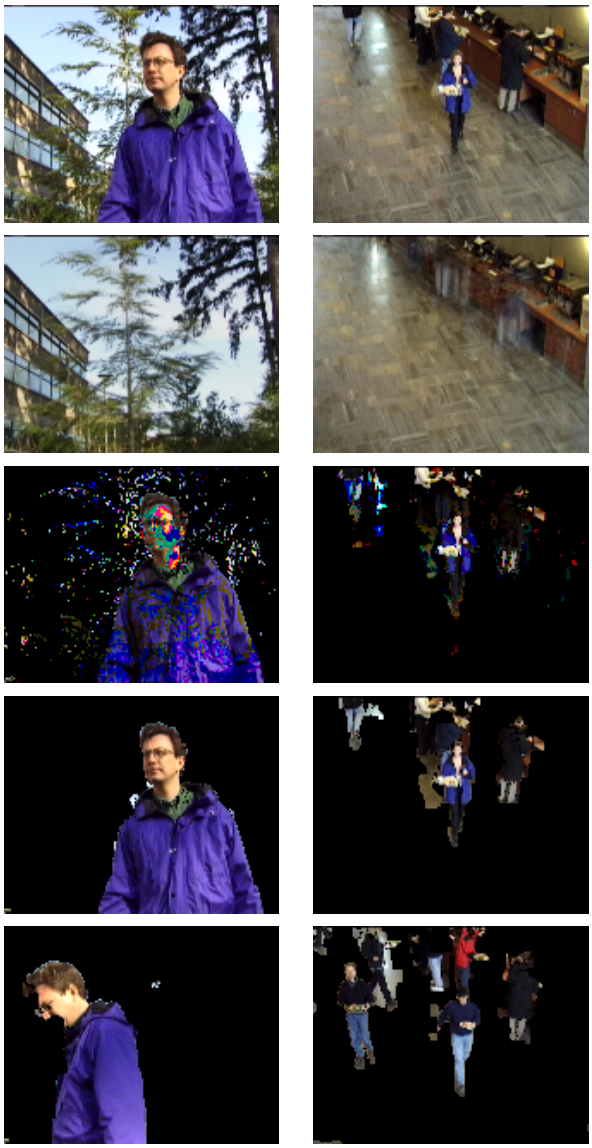


Figure 4.14: The original image $\mathbf{y}$ (line 1), the background (i.e., $\mathbf{Xw}$) reconstructed by our method (line 2), and the foreground (i.e., the sparsity pattern of $\mathbf{e}$ as a mask on the original image) detected with $\ell_1$ (line 3) and with $\ell_1 + \Omega$ (line 4). The bottom line is another foreground found with $\Omega$, on a different image, with the same values of $\lambda_1, \lambda_2$ as for the previous image. For the top left image, the percentage of pixels matching the ground truth is 98.8% with $\Omega$, 87.0% without. As for the top right image, the result is 93.8% with $\Omega$, 90.4% without (best seen in color).

# 5

# Application of Structured Sparsity to Neuroimaging

## 5.1 Multi-scale Mining of fMRI Data with Hierarchical Structured Sparsity

**Abstract of the first section of the chapter:** Inverse inference, or *"brain reading"*, is a recent paradigm for analyzing functional magnetic resonance imaging (fMRI) data, based on pattern recognition and statistical learning. By predicting some cognitive variables related to brain activation maps, this approach aims at decoding brain activity. Inverse inference takes into account the multivariate information between voxels and is currently the only way to assess how precisely some cognitive information is encoded by the activity of neural populations within the whole brain. However, it relies on a prediction function that is plagued by the curse of dimensionality, since there are far more features than samples, *i.e.*, more voxels than fMRI volumes. To address this problem, different methods have been proposed, such as, among others, univariate feature selection, feature agglomeration and regularization techniques. In this paper, we consider a sparse hierarchical structured regularization. Specifically, the penalization we use is constructed from a tree that is obtained by spatially-constrained agglomerative clustering. This approach encodes the spatial structure of the data at different scales into the regularization, which makes the overall prediction procedure more robust to inter-subject variability. The regularization used induces the selection of spatially coherent predictive brain regions simultaneously at different scales. We test our algorithm on real data acquired to study the mental representation of objects, and we show that the proposed algorithm non only delineates meaningful brain regions but yields as well better prediction accuracy than reference methods.

The material of this first section is based on the following papers:

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion. Multi-scale Mining of fMRI Data with Hierarchical Structured Sparsity. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. 2011

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multi-scale Mining of fMRI Data with Hierarchical Structured Sparsity. Preprint arXiv:1105.0363 *Submitted to SIAM Journal on Imaging Sciences*. 2011 (long version of the previous paper)

### 5.1.1 Introduction

Functional magnetic resonance imaging (or fMRI) is a widely used functional neuroimaging modality. Modeling and statistical analysis of fMRI data are commonly done through a linear model, called general linear model (GLM) in the community, that incorporates information about the different experimental conditions and the dynamics of the hemodynamic response in the design matrix. The experimental conditions are typically modelled by the type of stimulus presented, *e.g.*, visual and auditory stimulation, which are included as regressors in the design matrix. The resulting model parameters—one coefficient per voxel and regressor—are known as *activation maps*. They represent the local influence of the different experimental conditions on fMRI signals at the level of individual voxels. The most commonly used approach to analyze these activation maps is called classical inference. It relies on mass-univariate statistical tests (one for each voxel), and yields so-called statistical parametric maps (SPMs) (Friston et al., 1995). Such maps are useful for functional brain mapping, but classical inference has some limitations: it suffers from multiple comparisons issues and it is oblivious of the multivariate structure of fMRI data. Such data exhibit natural correlations between neighboring voxels forming clusters with different sizes and shapes, and also between distant but functionally connected brain regions.

To address these limitations, an approach called inverse inference (or "brain-reading") (Dehaene et al., 1998; Cox and Savoy, 2003) was recently proposed. Inverse inference relies on pattern recognition tools and statistical learning methods to explore fMRI data. Based on a set of activation maps, inverse inference estimates a function that can then be used to predict a target (typically, a variable representing a perceptual, cognitive or behavioral parameter) for a new set of images. The challenge is to capture the correlation structure present in the data in order to improve the performance of the mapping learnt, which is measured through the resulting prediction accuracy. Many standard statistical learning approaches have been used to construct prediction functions, among them kernel machines (SVM, RVM) (Schölkopf and Smola, 2002) or discriminant analysis (LDA, QDA) (Hastie et al., 2009). For the application considered in this chapter, earlier performance results (Cox and Savoy, 2003; LaConte et al., 2005) indicate that we can restrict ourselves to mappings that are linear functions of the data.

Throughout the chapter, we shall consider a training set composed of $n$ pairs $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathcal{Y}$, where $\mathbf{x}$ denotes a $p$-dimensional fMRI signal ($p$ voxels) and $y$ stands for the target we try to predict. In the experiments we carry out in Section 5.2.4, we will encounter both the regression and the multi-class classification settings, where $\mathcal{Y}$ denotes respectively the set of real numbers and a finite set of integers. In this chapter, we aim at learning a weight vector $\mathbf{w} \in \mathbb{R}^p$ and an intercept $b \in \mathbb{R}$ such that the prediction of $y$ can be based on the value of $\mathbf{w}^\top \mathbf{x} + b$. This is the case for the linear regression and logistic regression models that we use in Section 5.2.4. It is useful to rewrite these quantities in matrix form; more precisely, we denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix assembled from $n$ fMRI data points and by $\mathbf{y} \in \mathbb{R}^n$ the corresponding $n$ targets. In other words, each row of $\mathbf{X}$ is a $p$-dimensional sample, *i.e.*, an activation map of $p$ voxels related to one stimulus presentation.

Learning the parameters $(\mathbf{w}, b)$ remains challenging since the number of features ($10^4$ to $10^5$ voxels) exceeds by far the number of samples (a few hundreds of images). The prediction function is therefore prone to the phenomenon of overfitting in which the learning set is predicted precisely whereas the algorithm provides very inaccurate predictions on new samples (the test set). To address this issue, *dimensionality reduction* attempts to find a low dimensional subspace that concentrates as much of the predictive power of the original set as possible for the problem at hand.

Feature selection is a natural approach to perform dimensionality reduction in fMRI, since reducing the number of voxels potentially allows to identify a predictive region of the brain. This corresponds to discarding some columns of $\mathbf{X}$. This feature selection can be univariate, *e.g.*, analysis of variance (ANOVA) (Lehmann and Romano, 2005), or multivariate. While univariate methods ignore joint information between features, multivariate approaches are more adapted to inverse inference since they extract predictive patterns from the data as a whole. However, due to the huge number of possible patterns, these approaches suffer from combinatorial explosion, and some costly suboptimal heuristics (*e.g.*, recursive feature elimination (Guyon et al., 2002; Martino et al., 2008)) can be used. That is why ANOVA is usually preferred in fMRI. Alternatively, two more adapted solutions have been proposed: *regularization* and *feature agglomeration*.

Regularization is a way to encode a priori knowledge about the weight vector $\mathbf{w}$. Possible regularizers can promote for example spatial smoothness or sparsity which is a natural assumption for fMRI data. Indeed, only a few brain regions are assumed to be significantly activated during a cognitive task. Previous contributions on fMRI-based inverse inference include Carroll et al. (2009); Rissman et al. (2010); Ryali et al. (2010); Yamashita et al. (2008). They can be presented through the following minimization problem:

$$\min_{(\mathbf{w},b)\in\mathbb{R}^{p+1}} \mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}, b) + \lambda\Omega(\mathbf{w}) \quad \text{with} \quad \lambda \geq 0, \tag{5.1}$$

where $\lambda\Omega(\mathbf{w})$ is the regularization term, typically a non-Euclidean norm, and the fit to the data is measured through a convex loss function $(\mathbf{w}, b) \mapsto \mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}, b) \in \mathbb{R}_+$. The choice of the loss function will be made more specific and formal in the next sections. The coefficient of regularization $\lambda$ balances the loss and the penalization term. In this notation, a common regularization term in inverse inference is the so-called *Elastic net* (Zou and Hastie, 2005; Grosenick et al., 2009), which is a combined $\ell_1$ and $\ell_2$ penalization:

$$\lambda\Omega(\mathbf{w}) = \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2 = \sum_{j=1}^{p} \{\lambda_1|\mathbf{w}_j| + \lambda_2\mathbf{w}_j^2\}. \tag{5.2}$$

For the square loss, when setting $\lambda_1$ to 0, the model is called ridge regression, while when $\lambda_2 = 0$ it is known as Lasso (Tibshirani, 1996) or basis pursuit (Chen et al., 1998). The essential shortcoming of the Elastic net is that it does not take into account the spatial structure of the data, which is crucial in this context Michel et al. (2011). Indeed, due to the intrinsic smoothing of the complex metabolic pathway underlying the difference of blood oxygenation measured with fMRI (Ugurbil et al., 2003), statistical learning approaches should be informed by the 3D grid structure of the data.

In order to achieve dimensionality reduction, while taking into account the spatial structure of the data, one can resort to *feature agglomeration*. It constructs new features, called *parcels*, by averaging neighboring voxels exhibiting similar activations. The advantage of agglomeration is that no information is discarded a priori and that it is reasonable to hope that averaging might reduce noise. Although, this approach has been successfully used in previous work for brain mapping (Flandin et al., 2002; Thirion et al., 2006), it often does not consider the supervised information (*i.e.*, the target $\mathbf{y}$) while constructing the parcels. A recent approach has been proposed to address this issue, using a supervised greedy top-down exploration of a tree obtained by hierarchical clustering (Michel et al., 2010). This greedy approach has proven to be effective especially for inter-subject analyses, *i.e.*, when the training and the evaluation sets are related to different subjects. In this context, methods need to be robust to intrinsic spatial variations that exist across subjects: despite being co-registered into a common space, some variability remains between subjects, which implies that there is no perfect voxel-to-voxel correspondence between volumes. As a result, the performances of traditional voxel-based methods are strongly affected. Therefore, averaging in the form of parcels is a good way to cope with inter-subject variability. This greedy approach is nonetheless suboptimal, as it explores only a subpart of the whole tree.

Based on these considerations, we propose to integrate the multi-scale spatial structure of the data *within* the regularization term $\Omega$, while preserving convexity in the optimization. This notably guarantees global optimality and stability of the obtained solutions. To this end, we design a sparsity-inducing penalty that is directly built from the hierarchical structure of the spatial model obtained by Ward's algorithm (Ward, 1963). Such a penalty has already been successfully applied in several contexts, *e.g.*, in bioinformatics, to exploit the tree structure of gene networks for multi-task regression (Kim and Xing, 2010), and also for topic models and image inpainting (Jenatton et al., 2010a).

We summarize here the contributions of our chapter:

- We explain how the multi-scale spatial structure of fMRI data can be taken into account in the context of inverse inference through the combination of a spatially constrained hierarchical clustering procedure and a sparse hierarchical regularization.
- We provide a convex formulation of the problem and propose an efficient optimization procedure.
- We conduct a experimental comparison of several algorithms and formulations on fMRI data and illustrate the ability of the proposed method to localize in space and in scale some brain regions involved in the processing of visual stimuli.

The rest of the chapter is organized as follows: we first present the concept of structured sparsity-inducing regularization and then describe the different regression/classification formulations we are interested in. After exposing how we handle the resulting large-scale convex optimization problems thanks to proximal methods, we validate our approach on both a synthetic setting and a real dataset.

### 5.1.2 Combining agglomerative clustering with sparsity inducing regularizers

Hierarchical clustering allows to construct a tree-structured hierarchy of features on top of the original voxels features. Moreover, the underlying voxels corresponding to each of these features correspond to localized spatial patterns on the brain of the form we hope to retrieve (Chklovskii and Koulakov, 2004). Instead of selecting features in the tree greedily, we propose to cast the feature selection problem as supervised learning problem of the form (5.1). It is natural to require of the regularizer $\Omega$ that it should respect the tree structure of the hierarchy so as to induce the selection of localized patterns.

#### Constructing the sparsity-inducing norm

The structured sparsity-inducing term $\Omega$ is built from the result of the hierarchical clustering of the voxels. The latter yields a hierarchy of *clusters* represented as a tree $\mathcal{T}$ (or dendrogram) (Johnson, 1967). The root of the tree is the unique cluster that gathers all the voxels, while the leaves are the clusters with a single voxel. Among different *hierarchical agglomerative clustering* procedures, we use the variance-minimizing approach of Ward's algorithm (Ward, 1963), since it minimizes the loss of information at each step of clustering. In short, two *clusters* are merged if the resulting parcellation minimizes the sum of squared differences within all *clusters* (also known as *inertia criterion*).

In order to take into account the spatial information, we also add connectivity constraints in the hierarchical clustering algorithm, so that only neighboring clusters can be merged together. The resulting clusters are thus called *parcels*. Each node of the tree $\mathcal{T}$ either corresponds to a voxel if it is a leaf, or defines a *parcel*, as the union of its children's clusters of voxels (see Figure 5.1).

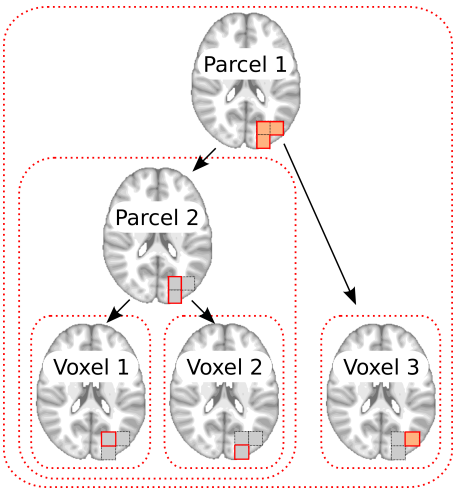



Figure 5.1: Example of a tree $\mathcal{T}$ when $p = 5$, with three voxels and two parcels. The parcel 2 is defined as the averaged intensity of the voxels $\{1, 2\}$, while the parcel 1 is obtained by averaging the parcel 2 and voxel 3. In red dashed lines are represented the five groups of variables that compose $\mathcal{G}$. For instance, if the group containing the parcel 2 is set to zero, the voxels $\{1, 2\}$ are also (and necessarily) zeroed out. Best seen in color.

We now consider the augmented space of variables (also known as features), formed by not only the voxels, but also by the parcels. This approximately doubles the number

of features of the fMRI signals. In other words, $p$ does not denote the number of voxels anymore, but instead, the total number of nodes of $\mathcal{T}$.[1] In the following, the level of activation of each parcel is (recursively) defined by the averaged intensity of the voxels it is composed of (*i.e.*, local averages) (Flandin et al., 2002; Thirion et al., 2006). This produces a multi-scale representation of the fMRI data that becomes increasingly invariant to spatial shifts of the encoding regions within the brain volume. More formally, if $j$ is a node of $\mathcal{T}$ and $P_j$ stands for the set of voxels of the corresponding parcel (i.e., the set of leaves of the subtree rooted at node $j$), we consider the mean of the parcel that we denote by $\langle \mathbf{x}_{P_j} \rangle$. In this notation, the linear model we use is of the form

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \tilde{\mathbf{x}} = \sum_{j \in \mathcal{T}} \mathbf{w}_j \langle \mathbf{x}_{P_j} \rangle = \sum_{i \in V} \Big[ \sum_{j \in A(i)} \frac{\mathbf{w}_j}{|P_j|} \Big] \mathbf{x}_i,$$

where $A(i)$ is the set of ancestors of a node $i$ in $\mathcal{T}$ (including itself), and $V$ corresponds to the leaves of the tree. To lighten notations, in the remainder of the chapter, we will denote by $\mathbf{X}$ instead of $\tilde{\mathbf{X}}$ the matrix of features from the augmented space.

In the perspective of inter-subject validation, the augmented space of variables can be exploited in the following way: since the information of single voxels may be unreliable, *the deeper the node in $\mathcal{T}$, the more variable the corresponding parcel's intensity is likely to be across subjects*. This property suggests that, while looking for sparse solutions of (5.1), we should preferentially select the variables near the root of $\mathcal{T}$, before trying to access smaller parcels located further down in $\mathcal{T}$.

Traditional sparsity-inducing penalties, *e.g.*, the $\ell_1$-norm $\Omega(\mathbf{w}) = \sum_{j=1}^{p} |\mathbf{w}_j|$, yield sparsity at the level of single variables $\mathbf{w}_j$, disregarding potential structures—for instance, spatial—existing between larger subsets of variables. We leverage here the concept of *structured sparsity* where $\Omega$ penalizes some predefined subsets, or *groups*, of variables that reflect prior information about the problem at hand (Baraniuk et al., 2010; Huang et al., 2009; Jenatton et al., 2011a; Jacob et al., 2009). In particular, we follow Zhao et al. (2009) that first introduced hierarchical sparsity-inducing penalties. Given a node $j$ of $\mathcal{T}$, we denote by $g_j \subseteq \{1, \ldots, p\}$ the set of indices that record all the descendants of $j$ in $\mathcal{T}$, including itself. In other words, $g_j$ contains the indices of the subtree rooted at $j$; see Figure 5.1. If we now denote by $\mathcal{G}$ the set of all $g_j$, $j \in \{1, \ldots, p\}$, that is, $\mathcal{G} \triangleq \{g_1, \ldots, g_p\}$, we can define our hierarchical penalty as

$$\Omega(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2 \triangleq \sum_{g \in \mathcal{G}} \Big[ \sum_{j \in g} \mathbf{w}_j^2 \Big]^{1/2}. \tag{5.3}$$

As shown in Jenatton et al. (2011a), $\Omega$ is a norm, and it promotes sparsity at the level of groups $g \in \mathcal{G}$, in the sense that it acts as a $\ell_1$-norm on the vector $(\|\mathbf{w}_g\|_2)_{g \in \mathcal{G}}$. Regularizing by $\Omega$ therefore causes some $\|\mathbf{w}_g\|_2$ (and equivalently $\mathbf{w}_g$) to be zeroed out for some $g \in \mathcal{G}$. Moreover, since the groups $g \in \mathcal{G}$ represent rooted subtrees of $\mathcal{T}$, this implies that if one node/parcel $j \in g$ is set to zero by $\Omega$, the same occurs for all its descendants (Zhao et al., 2009). To put it differently, *if one parcel is selected, then all*

1. We can then identify nodes (and parcels) of $\mathcal{T}$ with indices in $\{1, \ldots, p\}$.

*the ancestral parcels in $\mathcal{T}$ will also be selected.* This property is in accordance with our concern of robustness with respect to voxel misalignments between subjects, since large parcels are considered before smaller ones.

The family of norms with the previous property is actually slightly larger and we consider throughout the chapter norms $\Omega$ of the form (Zhao et al., 2009):

$$\Omega(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} \eta_g \|\mathbf{w}_g\|, \tag{5.4}$$

where $\|\mathbf{w}_g\|$ denotes either the $\ell_2$-norm $\|\mathbf{w}_g\|_2$ or the $\ell_\infty$-norm $\|\mathbf{w}_g\|_\infty \triangleq \max_{j \in g} |\mathbf{w}_j|$ and $(\eta_g)_{g \in \mathcal{G}}$ are (strictly) positive weights that can compensate for the fact that some features are overpenalized as a result of being included in a larger number of groups than others. In light of the results from Jenatton et al. (2010a), we will see in Section 5.1.4 that a large class of optimization problems regularized by $\Omega$—as defined in (5.4)— can be solved efficiently.

### 5.1.3 Supervised learning framework

In this section, we introduce the formulations we consider in our experiments. As further discussed in Section 5.2.4, the target $y$ we try to predict corresponds to (discrete) sizes of objects, *i.e.*, a one-dimensional *ordered* variable. It is therefore sensible to address this prediction task from both a regression and a classification viewpoint.

#### Regression

In this first setting, we naturally consider the square loss function, so that problem (5.1) can be reduced to

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \Omega(\mathbf{w}) \quad \text{with} \quad \lambda \geq 0.$$

Note that in this case, we have omitted the intercept $b$ since we can center the vector $\mathbf{y}$ and the columns of $\mathbf{X}$ instead.

#### Classification

We now look at our prediction task from a multi-class classification viewpoint. Specifically, we assume that $\mathcal{Y}$ is a finite set of integers $\{1, \ldots, c\}$, $c > 2$, and consider both multi-class and "one-versus-all" strategies (Rifkin and Klautau, 2004). We need to slightly extend the formulation (5.1): To this end, we introduce the weight matrix $\mathbf{W} \triangleq [\mathbf{w}^1, \ldots, \mathbf{w}^c] \in \mathbb{R}^{p \times c}$, composed of $c$ weight vectors, along with a vector of intercepts $\mathbf{b} \in \mathbb{R}^c$.

A standard way of addressing multi-class classification problems consists in using a multi-logit model, also known as multinomial logistic regression (see, *e.g.*, (Hastie et al.,

2009) and references therein). In this case, class-conditional probabilities are modeled for each class by a softmax function and leads to

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times c}\\ \mathbf{b}\in\mathbb{R}^c}} \frac{1}{n}\sum_{i=1}^{n}\log\Big[\sum_{k=1}^{c} e^{\mathbf{x}_i^\top(\mathbf{w}^k-\mathbf{w}^{\mathbf{y}_i})+\mathbf{b}_k-\mathbf{b}_{\mathbf{y}_i}}\Big] + \lambda\sum_{k=1}^{c}\Omega(\mathbf{w}^k) \ .$$

Whereas the regularization term is separable with respect to the different weight vectors $\mathbf{w}^k$, the loss function induces a coupling in the columns of $\mathbf{W}$. As a result, the optimization has to be carried out over the entire matrix $\mathbf{W}$.

In Section 5.2.4, we consider another multi-class classification scheme. The "one-versus-all" strategy (OVA) consists in training $c$ different (real-valued) binary classifiers, each one being trained to distinguish the examples in a single class from the observations in all remaining classes. In order to classify a new example, among the $c$ classifiers, the one which outputs the largest (most positive) value is chosen. In this framework, we consider binary classifiers built from both the square and the logistic loss functions. If we denote by $\bar{\mathbf{Y}} \in \{-1,1\}^{n\times c}$ the indicator response matrix defined as $\bar{\mathbf{Y}}_i^k \triangleq 1$ if $\mathbf{y}_i = k$ and $-1$ otherwise, we obtain

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times c}} \frac{1}{2n}\sum_{k=1}^{c}\|\bar{\mathbf{Y}}^k - \mathbf{X}\mathbf{w}^k\|_2^2 + \lambda\sum_{k=1}^{c}\Omega(\mathbf{w}^k),$$

and

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times c}\\ \mathbf{b}\in\mathbb{R}^c}} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}\log\Big[1+e^{-\bar{\mathbf{Y}}_i^k(\mathbf{x}_i^\top\mathbf{w}^k+\mathbf{b}_k)}\Big] + \lambda\sum_{k=1}^{c}\Omega(\mathbf{w}^k).$$

By invoking the same arguments as in Section 5.1.3, the vector of intercepts $\mathbf{b}$ is again omitted in the above problem with the square loss. The formulations we reviewed in this section can be solved efficiently within the same optimization framework we now introduce.

### 5.1.4 Optimization

The convex minimization problem (5.1) is challenging, since the penalty $\Omega$ as defined in (5.4) is non-smooth and the number of variables to deal with is large (about $p \approx 10^5$ voxels in the following experiments). To this end, we resort to *proximal methods* (see, e.g., Beck and Teboulle, 2009; Combettes and Pesquet, 2010; Nesterov, 2007; Wright et al., 2009)). In a nutshell, these methods can be seen as a natural extension of gradient-based techniques when the objective function to minimize has an amenable non-smooth part. They have increasingly drawn the attention of a broad research community because of their convergence rates (optimal within the class of first-order techniques) and their ability to deal with large non-smooth convex problems. We assume from now on that the convex loss function $\mathcal{L}(\mathbf{y},\mathbf{X},.)$ is differentiable with Lipschitz-continuous gradient, which notably covers the cases of the square and simple/multinomial logistic functions (introduced in Section 5.1.3).

The simplest version of this class of methods linearizes at each iteration the function $\mathcal{L}(\mathbf{y}, \mathbf{X}, .)$ around the current estimate $\mathbf{w}_0$, [2] and this estimate is then updated as the (unique by strong convexity) solution of the *proximal problem*:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}_0) + (\mathbf{w} - \mathbf{w}_0)^\top \nabla\mathcal{L}_{\mathbf{w}}(\mathbf{y}, \mathbf{X}, \mathbf{w}_0) + \lambda\Omega(\mathbf{w}) + \frac{L}{2}\|\mathbf{w} - \mathbf{w}_0\|_2^2.$$

The quadratic term keeps the update in a neighborhood where $\mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}_0)$ is close to its linear approximation, and $L > 0$ is a parameter which is a upper bound on the Lipschitz constant of the gradient of $\mathcal{L}$. This problem can be equivalently rewritten as:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{w}_0 - \frac{1}{L}\nabla\mathcal{L}_{\mathbf{w}}(\mathbf{y}, \mathbf{X}, \mathbf{w}_0) - \mathbf{w}\right\|_2^2 + \frac{\lambda}{L}\Omega(\mathbf{w}). \tag{5.5}$$

Solving efficiently and exactly this problem is crucial to enjoy the fastest convergence rates of proximal methods. In addition, when the non-smooth term $\Omega$ is not present, the previous proximal problem exactly leads to the standard gradient update rule. In simple settings, the solution of problem (5.5) is given in closed form: For instance, when the regularization $\Omega$ is chosen to be the $\ell_1$-norm, we get back the well-known soft-thresholding operator (Donoho and Johnstone, 1995).

The work of Jenatton et al. (2010a) recently showed that the proximal problem (5.5) could be solved efficiently and exactly with $\Omega$ as defined in (5.4). The underlying idea of this computation is to solve a *well-ordered* sequence of simple proximal problems associated with each of the terms $\|\mathbf{w}_g\|$ for $g \in \mathcal{G}$. We refer the interested readers to Jenatton et al. (2011c) for further details.

In our experiments, we will use the accelerated proximal gradient scheme (FISTA) taken from Beck and Teboulle (2009), which is a similar procedure as the one described above, except that the proximal problem (5.5) is not solved for the current estimate, but for an auxiliary sequence of points that are linear combinations of past estimates. [3] In terms of computational complexity, such proximal schemes are guaranteed to be $\varepsilon$ close to the optimal objective function in $O(\sqrt{L/\varepsilon})$ iterations (Beck and Teboulle, 2009; Nesterov, 2007). The cost of each iteration is dominated by the computation of the gradient (*e.g.*, $O(np)$ for the square loss) and the proximal operator, whose time complexity is linear, or close to linear, in $p$ for the tree-structured regularization (Jenatton et al., 2011c).

### 5.1.5 Experiments and results

We now present experimental results on simulated data and real fMRI data.

#### Simulations

In order to illustrate the proposed method, the hierarchical regularization with the $\ell_2$-norm and $\eta_g = 1$ for all $g$ was applied in a regression setting on a small two-dimensional

2. For simplicity and clarity of the presentation, we do not consider the optimization of the intercept that we let unregularized in all our experiments.

3. The `Matlab/C++` implementation we use is available at `http://www.di.ens.fr/willow/SPAMS/`.

simulated dataset consisting of 300 square images ($40 \times 40$ pixels i.e. $\mathbf{X} \in \mathbb{R}^{300\times1600}$). The weight vector $\mathbf{w}$ used in the simulation— itself an image of the same dimension— is presented in Fig. 5.2-a. It consists of three localized regions of two different sizes that are predictive of the output. The images $\mathbf{x}^{(i)}$ are sampled so as to obtain a correlation structure which mimics fMRI data. Precisely, each image $\mathbf{x}^{(i)}$ was obtained by smoothing a completely random image — where each pixel was drawn i.i.d from a normal distribution — with a Gaussian kernel, which introduces spatial correlations between neighboring pixels. Subsequently, correlations between the regions corresponding to the three patterns were introduced in order to simulate co-activations between different brain regions (0.3 correlation between the two bigger patterns, and $-0.2$ correlation between the smallest and lower-corner patterns).

The choice of the weights and of the correlation introduced in images aim at illustrating how the hierarchical regularization estimates weights at different resolutions in the image. The targets were simulated by forming $\mathbf{w}^\top x^{(i)}$ corrupted with an additive white noise (SNR=10dB). The loss used was the square loss as detailed in Section 5.1.3. The regularization parameter was estimated with two-fold cross-validation (150 images per fold) on a logarithmic grid of 30 values between $10^3$ and $10^{-3}$.

The weights estimated are presented in Fig. 5.2 at different scales, *i.e.,* different depths in the tree. It can be observed that all three patterns are present in the weight vector but at different depth in the tree. The small activation in the top-right hand corner shows up mainly in scale 3 while the bigger patterns appear higher in the tree in scales 5 and 6. This simulation clearly illustrates the ability of the method to capture informative spatial patterns at different scales. We now present results on real data.

### Description on the data

We apply the different methods to analyze the data of ten subjects from an fMRI study originally designed to investigate object coding in high-level visual cortex (see Eger et al., 2008, for details). During the experiment, twelve healthy volunteers viewed objects of two categories (each one of the two categories is used in half of the subjects) with four different exemplars in each category. Each exemplar was presented at three different sizes (yielding 12 different experimental conditions per subject). Each stimulus was presented four times in each of the six sessions. We averaged data from the four repetitions, resulting in a total of $n = 72$ images by subject (one image of each stimulus by session). Functional images were acquired on a 3-T MR system with eight-channel head coil (Siemens Trio, Erlangen, Germany) as T2*-weighted echo-planar image (EPI) volumes. Twenty transverse slices were obtained with a repetition time of 2s (echo time, 30ms; flip angle, 70°; $2 \times 2 \times 2$-mm voxels; 0.5-mm gap). Realignment, normalization to MNI space, and GLM fit were performed with the SPM5 software [4]. In the GLM, the time course of each of the 12 stimuli convolved with a standard hemodynamic response function was modeled separately, while accounting for serial auto-correlation with an AR(1) model and removing low-frequency drift terms with a high-pass filter with a cut-

4. `http://www.fil.ion.ucl.ac.uk/spm/software/spm5`.

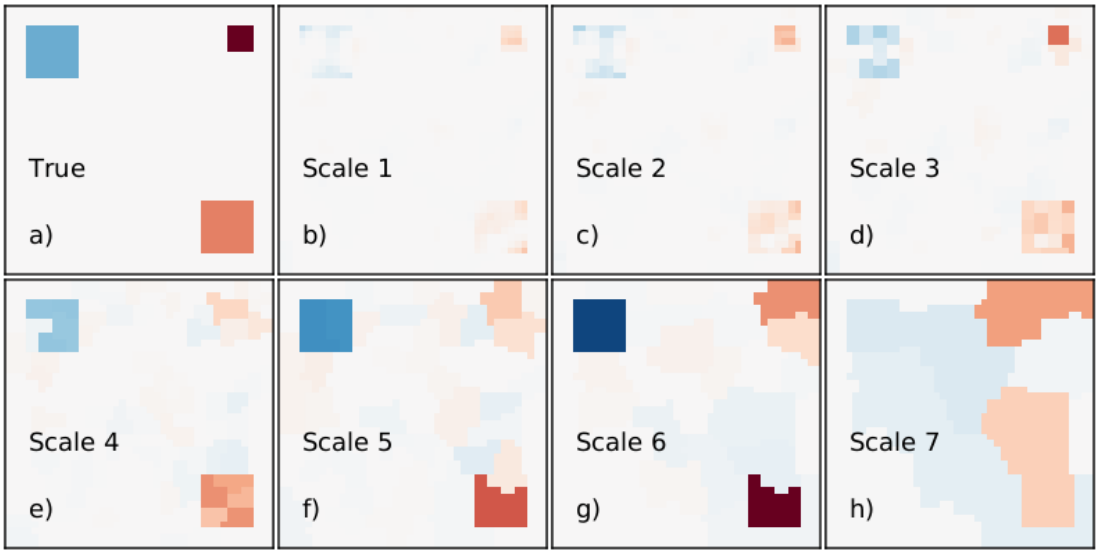



Figure 5.2: Weights estimated in the simulation study. The true coefficients are presented in a) and the estimated weights at different scales, *i.e.,* different depths in the tree, are presented in b)-h). The results are best seen in color.

off of 128s. In the present work we used the resulting session-wise parameter estimate images. All the analysis are performed on the whole brain volume.

The four different exemplars in each of the two categories were pooled, leading to images labeled according to the three possible sizes of the object. By doing so, we are interested in finding discriminative information to predict the size of the presented object.

This can be reduced to either a regression problem in which our goal is to predict a simple scalar factor (size or scale of the presented object), or a three-category classification problem, each size corresponding to a category. We perform an inter-subject analysis on the sizes both in regression and classification settings. This analysis relies on subject-specific fixed-effects activations, *i.e.*, for each condition, the six activation maps corresponding to the six sessions are averaged together. This yields a total of 12 images per subject, one for each experimental condition. The dimensions of the real data set are $p \approx 7 \times 10^4$ and $n = 120$ (divided into three different sizes). We evaluate the performance of the method by cross-validation with a natural data splitting, *leave-one-subject-out*. Each fold consists of 12 volumes. The parameter $\lambda$ of all methods is optimized over a grid of 30 values of the form $2^k$, with a nested leave-one-subject-out cross-validation on the training set. The exact scaling of the grid varies for each model to account for different $\Omega$.

### Methods involved in the comparisons

In addition to considering standard $\ell_1$- and squared $\ell_2$-regularizations in both our regression and multi-class classification tasks, we compare various methods that we now review.

First of all, when the regularization $\Omega$ as defined in (5.4) is employed, we consider three settings of values for $(\eta_g)_{g\in\mathcal{G}}$ which leverage the tree structure $\mathcal{T}$. More precisely, we set $\eta_g = \rho^{\text{depth}(g)}$ for $g$ in $\mathcal{G}$, with $\rho \in \{0.5, 1, 1.5\}$ and where depth$(g)$ denotes the depth of the root of the group $g$ in $\mathcal{T}$. In other words, the larger $\rho$, the more averse we are to selecting small (and variable) parcels located near the leaves of $\mathcal{T}$.

The greedy approach from Michel et al. (2010) is included in the comparisons, for both the regression and classification tasks. It relies on a top-down exploration of the tree $\mathcal{T}$. In short, starting from the root parcel that contains all the voxels, we choose at each step the split of the parcel that yields the highest prediction score. The exploration step is performed until a given number of parcels is reached, and yields a set of nested parcellations with increasing complexity. Similarly to a model selection step, we chose the best parcellation among those found in the exploration step. The selected parcellation is thus used on the test set. In the regression setting, this approach is combined with Bayesian ridge regression, while it is associated with a linear support vector machine for the classification task (whose value of $C$ is found by nested cross-validation in $\{0.01, 0.1, 1\}$).

**Regression setting.** In order to evaluate whether the level of sparsity is critical in our analysis, we implemented a reweighted $\ell_1$-scheme (Candes et al., 2008). In this case, sparsity is encouraged more aggressively as a multi-stage convex relaxation of a concave penalty. Specifically, it consists in using iteratively a weighted $\ell_1$-norm, whose weights are determined by the solution of previous iteration.

To better understand the added value of the hierarchical norm (5.4) over unstructured penalties, we consider another variant of weighted $\ell_1$-norm, this time defined in the augmented space of features. The weights are manually set and reflect the underlying tree structure $\mathcal{T}$. By analogy with the choice of $(\eta_g)_{g\in\mathcal{G}}$ made for the tree-structured regularization, we take exponential weights depending on the depth of the variable $j$, with $\rho = 1.5$.[5] We also tried weights $(\eta_g)_{g\in\mathcal{G}}$ that are linear with respect to the depths, but those led to worse results. We now turn to the models taking part in the classification task.

**Classification setting.** As discussed in Section 5.1.3, the optimization in the classification setting is carried out over a matrix of weights $\mathbf{W} \in \mathbb{R}^{p\times c}$. This makes it possible to consider other regularization schemes.

In particular, we apply ideas from *multi-task* learning (Obozinski et al., 2009) by viewing each class as a task. More precisely, we use a regularization norm defined by

5. Formally, the depth of the feature $j$ is equal to depth$(g_j)$, where $g_j$ is the smallest group in $\mathcal{G}$ that contains $j$ (*smallest* is understood here in the sense of the inclusion).

$\Omega_{\text{multi-task}}(\mathbf{W}) \triangleq \sum_{j=1}^{p} \|\mathbf{W}_j\|$, where $\|\mathbf{W}_j\|$ denotes either the $\ell_2$- or $\ell_\infty$-norm of the $j$-th row of $\mathbf{W}$. The rationale for the definition of $\Omega_{\text{multi-task}}$ is to assume that the set of relevant voxels is the same across the $c$ different classes, so that sparsity is induced simultaneously over the columns of $\mathbf{W}$. As a remark, in the "one-versus-all" setting, although the loss functions for the $c$ classes are decoupled, the use of $\Omega_{\text{multi-task}}$ induces a relationship that ties them together.

Note that the tree-structured regularization $\Omega$ we consider does not impose a joint pattern-selection across the $c$ different classes. Although a multi-task extension of $\Omega$ with $\ell_\infty$-norms has recently been proposed (Mairal et al., 2010b), the cost of the corresponding proximal operator is significantly higher, which is likely to raise some computational issues in our large-scale experiments.

### Results

We present result of the comparison of our approach based on the hierarchical sparsity-inducing norm (5.4) with the models presented in the previous section. For each method, we computed the cross-validated prediction accuracy and the percentage of non-zero coefficients, *i.e.*, the level of sparsity of the models.

**Regression results.** The results for the inter-subject regression analysis are given in Table 5.1. The lowest error in prediction accuracy is obtained by the proposed hierarchical structured sparsity approach (Tree $\ell_2$ with $\rho = 1$), that also yields one of the lowest (along with greedy) standard deviation indicating that the results are most stable. This can be explained by the fact that the use of local signal averages in the proposed algorithm is a good way to get some robustness to inter-subject variability. We also notice that the sparsity-inducing approaches (Lasso and reweighted $\ell_1$) have the highest error in prediction accuracy, probably because the obtained solutions are too sparse, and suffer from the absence of perfect voxel-to-voxel correspondences between subjects.

In terms of sparsity, we can see, as expected, that ridge regression does not yield any sparsity and that the Lasso solution is very sparse (in the feature space, with approximately $7 \times 10^4$ voxels). Our method yields a median value of 9.36% of non-zero coefficients (in the augmented space of features, with about $1.4 \times 10^5$ nodes in the tree). The maps of weights obtained with Lasso and the hierarchical regularization for one fold, are given in Fig. 5.3. The Lasso yields a scattered and overly sparse pattern of voxels, that is not easily readable, while our approach extracts a pattern of voxels with a compact structure, that clearly outlines brain regions expected to activate differentially for stimuli with different low-level visual properties, *e.g.*, sizes; the early visual cortex in the occipital lobe at the back of the brain. Interestingly, the patterns of voxels show some symmetry between left and right hemispheres, especially in the primary visual cortex which is located at the back and center of the brain. Such an observation matches very well with existing neurosciences knowledge of this brain region that processes the visual contents of both visual hemifields. The weights obtained at different depth level in the tree, corresponding to different scales, show that the largest coefficients are concentrated

| | Square | | |
|---|---|---|---|
| Loss function: | Square | | |
| | | | |
| | Error (mean,std) | P-value w.r.t. Tree $\ell_2$ ($\rho = 1$) | Median fraction of non-zeros (%) |
| Regularization: | | | |
| $\ell_2$ (Ridge) | (8.3, 4.6) | 0.096 | 100.00 |
| $\ell_1$ | (12.1, 6.6) | 0.013* | 0.11 |
| Reweighted $\ell_1$ | (11.3, 8.8) | 0.052 | 0.10 |
| $\ell_1$ (tree weights) | (8.3, 4.7) | 0.032* | 0.02 |
| Tree $\ell_2$ ($\rho = 0.5$) | (7.8, 4.4) | 0.137 | 99.99 |
| Tree $\ell_2$ ($\rho = 1$) | (**7.1**, **4.0**) | - | 9.36 |
| Tree $\ell_2$ ($\rho = 1.5$) | (8.1, 4.2) | 0.080 | 0.04 |
| Tree $\ell_\infty$ ($\rho = 0.5$) | (8.1, 4.7) | 0.080 | 99.99 |
| Tree $\ell_\infty$ ($\rho = 1$) | (7.7, 4.1) | 0.137 | 1.22 |
| Tree $\ell_\infty$ ($\rho = 1.5$) | (7.8,4.1) | 0.096 | 0.04 |
| | | | |
| | Error (mean,std) | P-value w.r.t. Tree $\ell_2$ ($\rho = 1$) | Median fraction of non-zeros (%) |
| Greedy | (7.2, 3.3) | 0.5 | 0.01 |

Table 5.1: Prediction results obtained on fMRI data (see text) for the regression setting. From the left, the first column contains the mean and standard deviation of the test error (unexplained variance), computed over leave-one-subject-out folds. The best performance is obtained with the hierarchical $\ell_2$ penalization ($\rho = 1$) constructed from the Ward tree. Statistical significance is assessed with a Wilcoxon two-sample paired signed rank test. The superscript * indicates a rejection at 5%.

at the higher scales (scale 6 in Fig. 5.3), showing that the object size cannot be well decoded at the voxel level but requires features formed by more macroscopic clusters of voxels.

**Classification results.** The results for the inter-subject classification analysis are given in Table 5.2. The best performance is obtained with a multinomial logistic loss function, also using the hierarchical $\ell_2$ penalization ($\rho = 1$).

For both $\ell_1$ and hierarchical regularizations, one of the three vectors of coefficients obtained for one fold are presented in Fig. 5.4. While for $\ell_1$, the active voxels are scattered all over the brain, the tree $\ell_2$ regularization yields clearly delineated sparsity patterns located in the visual areas of the brain. Like for the regression results, the highest coefficients are obtained at scale 6 showing how spatially extended is the brain region involved in the cognitive task. The symmetry of the pattern at this scale is also particularly striking in the primary visual areas. It also extends more anteriorly into the inferior temporal cortex, known for high-level visual processing.

| Loss function: | Square ("one-versus-all") | | |
|---|---|---|---|
| | | | |
| | Error (mean,std) | P-value w.r.t. Tree $\ell_2$ ($\rho = 1$)-ML | Median fraction of non-zeros (%) |
| Regularization: | | | |
| $\ell_2$ (Ridge) | (29.2, 5.9) | 0.004* | 100.00 |
| $\ell_1$ | (33.3, 6.8) | 0.004* | 0.10 |
| $\ell_1/\ell_2$ (Multi-task) | (31.7, 9.5) | 0.004* | 0.12 |
| $\ell_1/\ell_\infty$ (Multi-task) | (33.3,13.6) | 0.009* | 0.22 |
| Tree $\ell_2$ ($\rho = 0.5$) | (25.8, 9.2) | 0.004* | 99.93 |
| Tree $\ell_2$ ($\rho = 1$) | (25.0, 5.5) | 0.027* | 10.08 |
| Tree $\ell_2$ ($\rho = 1.5$) | (24.2, 9.9) | 0.130 | 0.05 |
| Tree $\ell_\infty$ ($\rho = 0.5$) | (30.8, 8.8) | 0.004* | 59.49 |
| Tree $\ell_\infty$ ($\rho = 1$) | (24.2, 7.3) | 0.058 | 1.21 |
| Tree $\ell_\infty$ ($\rho = 1.5$) | (25.8, 10.7) | 0.070 | 0.04 |
| | | | |
| Loss function: | Logistic ("one-versus-all") | | |
| | | | |
| | Error (mean,std) | P-value w.r.t. Tree $\ell_2$ ($\rho = 1$)-ML | Median fraction of non-zeros (%) |
| Regularization: | | | |
| $\ell_2$ (Ridge) | (25.0, 9.6) | 0.008* | 100.00 |
| $\ell_1$ | (34.2, 15.9) | 0.004* | 0.55 |
| $\ell_1/\ell_2$ (Multi-task) | (31.7, 8.6) | 0.002* | 47.35 |
| $\ell_1/\ell_\infty$ (Multi-task) | (33.3, 10.4) | 0.002* | 99.95 |
| Tree $\ell_2$ ($\rho = 0.5$) | (25.0, 9.6) | 0.007* | 99.93 |
| Tree $\ell_2$ ($\rho = 1$) | (20.0, 11.2) | 0.250 | 7.88 |
| Tree $\ell_2$ ($\rho = 1.5$) | (18.3, 6.6) | 0.500 | 0.06 |
| Tree $\ell_\infty$ ($\rho = 0.5$) | (30.8, 10.4) | 0.004* | 59.42 |
| Tree $\ell_\infty$ ($\rho = 1$) | (24.2, 6.1) | 0.035* | 0.60 |
| Tree $\ell_\infty$ ($\rho = 1.5$) | (21.7, 8.9) | 0.125 | 0.03 |
| | | | |
| Loss function: | Multinomial logistic (ML) | | |
| | | | |
| | Error (mean,std) | P-value w.r.t. Tree $\ell_2$ ($\rho = 1$)-ML | Median fraction of non-zeros (%) |
| Regularization: | | | |
| $\ell_2$ (Ridge) | (24.2, 9.2) | 0.035* | 100.00 |
| $\ell_1$ | (25.8, 12.0) | 0.004* | 97.95 |
| $\ell_1/\ell_2$ (Multi-task) | (26.7, 7.6) | 0.007* | 30.24 |
| $\ell_1/\ell_\infty$ (Multi-task) | (26.7, 11.6) | 0.002* | 99.98 |
| Tree $\ell_2$ ($\rho = 0.5$) | (22.5, 8.8) | 0.070 | 83.06 |
| Tree $\ell_2$ ($\rho = 1$) | (**16.7**, **10.4**) | - | 4.87 |
| Tree $\ell_2$ ($\rho = 1.5$) | (18.3, 10.9) | 0.445 | 0.02 |
| Tree $\ell_\infty$ ($\rho = 0.5$) | (26.7, 11.6) | 0.015* | 48.82 |
| Tree $\ell_\infty$ ($\rho = 1$) | (22.5, 13.0) | 0.156 | 0.34 |
| Tree $\ell_\infty$ ($\rho = 1.5$) | (21.7, 8.9) | 0.460 | 0.05 |
| | | | |
| | Error (mean,std) | P-value w.r.t. Tree $\ell_2$ ($\rho = 1$)-ML | Median fraction of non-zeros (%) |
| Greedy | (21.6, 14.5) | 0.001* | 0.01 |

Table 5.2: Prediction results obtained on fMRI data (see text) for the multi-class classification setting. From the left, the first column contains the mean and standard deviation of the test error (percentage of misclassification), computed over leave-one-subject-out folds. The best performance is obtained with the hierarchical $\ell_2$ penalization ($\rho = 1$) constructed from the Ward tree, coupled with the multinomial logistic loss function. Statistical significance is assessed with a Wilcoxon two-sample paired signed rank test. The superscript * indicates a rejection at 5%.

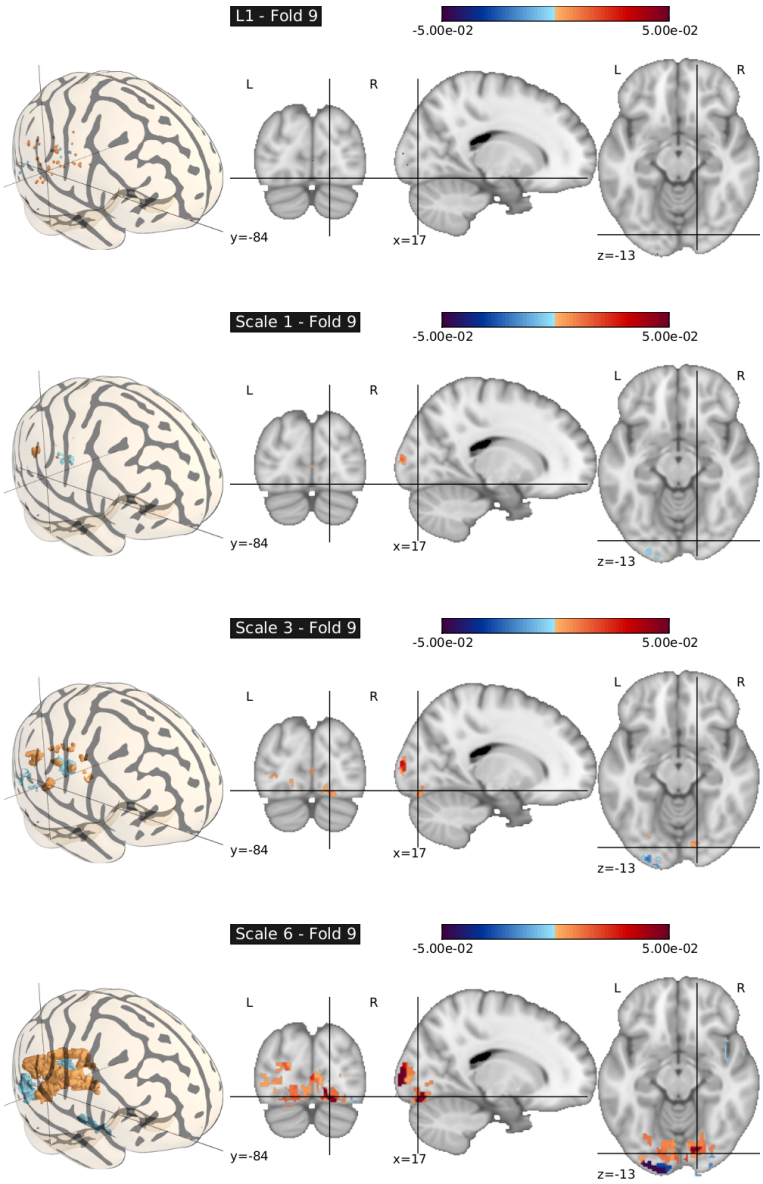



Figure 5.3: Maps of weights obtained using different regularizations in the regression setting. (a) $\ell_1$ regularization - We can notice that the predictive pattern obtained is excessively sparse, and is not easily readable despite being mainly located in the occipital cortex. (b-d) tree $\ell_2$ regularization ($\rho = 1$) at different scales - In this case, the regularization algorithm extracts a pattern of voxels with a compact structure, that clearly outlines early visual cortex which is expected to discriminate between stimuli of different sizes. 3D images were generated with Mayavi (Ramachandran and Varoquaux, 2011).

a)

b)

c)

d)

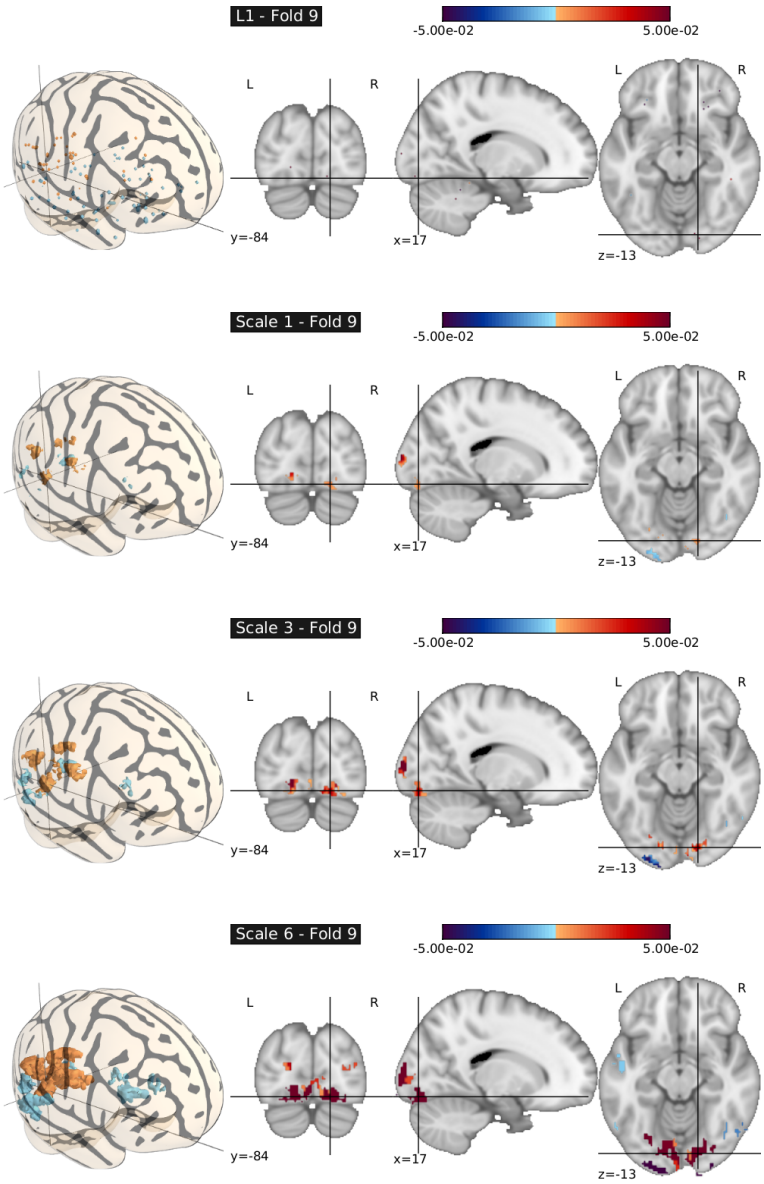



Figure 5.4: Maps of weights obtained using different regularizations in the classification setting. (a) $\ell_1$ regularization - We can notice that the predictive pattern obtained is excessively sparse, and is not easily readable with voxels scattered all over the brain. (b-d) tree regularization at different scales - In this case, the regularization algorithm extracts a pattern of voxels with a compact structure, that clearly outlines early visual cortex which is expected to discriminate between stimuli of different sizes.

### 5.1.6 Conclusion

In this chapter, we introduced a hierarchically structured regularization, which takes into account the spatial and multi-scale structure of fMRI data. This approach copes with inter-subject variability in a similar way as feature agglomeration, by averaging neighboring voxels. Although alternative agglomeration strategies do exist, we simply used the criterion which appears as the most natural, Ward's clustering, and which builds parcels with little variance.

Results on a real dataset show that the proposed algorithm is a promising tool for mining fMRI data. It yields higher prediction accuracy than reference methods, and the map of weights it obtains exhibit a cluster-like structure. It makes them easily readable compared to the overly sparse patterns found by classical sparsity-promoting approaches.

For the regression problem, both the greedy method from Michel et al. (2010) and the proposed algorithm yield better results than unstructured and non-hierarchical regularizations. However, in both regression and classification settings, the convex formulation introduced here leads to the best performance while enjoying the guarantees of convex optimization. In particular, while the greedy algorithm relies on a two-step approach that may be far from optimal, the hierarchical regularization induces simultaneously the selection of the optimal parcellation and the construction of the optimal predictive model, given the initial hierarchical clustering of the voxels. Moreover, convex methods yield predictors that are essentially stable with respect to perturbations of the design or the initial clustering, which is typically not the case of greedy methods.

Finally, it should be mentioned that the performance achieved by this approach in inter-subject problems suggests that it could potentially be used successfully in medical diagnosis problems, where brain images –not necessarily functional images– are used to classify individuals into diseased or control population. Indeed, for difficult problems of that sort, where the reliability of the diagnostic is essential, the stability of models obtained from convex formulations and the interpretability of sparse and localized solutions are useful properties to have in order to provide a credible diagnostic.

## 5.2 Sparse Structured Dictionary Learning for Brain Resting-State Activity Modeling

**Abstract of the second section of the chapter:** We introduce a generative model to study brain resting-state time series. These signals are represented as linear combinations of latent spatial maps, which we obtain via matrix factorization techniques developed for dictionary learning. In particular, we propose to learn the spatial components with specific structural constraints, e.g, small localized clusters of voxels, which can be achieved with sparsity-inducing regularization schemes recently used for dictionary learning. While brain resting-state time-series are generally the object of exploratory data analysis, our model provides a natural framework for model selection and quantitative evaluation. We show that our approach yields improved estimates

as assessed by the likelihood on unseen data, while exhibiting interpretable spatial components, that match known areas of interest in the brain.

The content of this second section is built upon the following work:

G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. Sparse Structured Dictionary Learning for Brain Resting-State Activity Modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*. 2010

### 5.2.1 Introduction

Functional magnetic resonance imaging (fMRI) yields in-vivo time-resolved measurements of brain activity. As the workhorse of neuroimaging, it has been the basis for much research on brain function, such as the investigation of neural coding corresponding to specific behavior. The study of intrinsic brain function is receiving increasing interest from the neuroscience community (Biswal et al., 2010). For this purpose, the spontaneous brain activity of subjects in the absence of a task is recorded in so-called *resting-state* experiments. In this setting, the statistical modeling of the signals is an unsupervised problem, and independent component analysis (ICA) has been widely used to study decomposition of the brain-activation time series in a set of interpretable spatial components (Beckmann and Smith, 2004; Beckmann et al., 2005; Varoquaux et al., 2010e). More recently, graphical models learned from the correlation matrices between distant activation time series have been shown to contain discriminative features of pathologies or cognitive processes (Cecchi et al., 2009; Varoquaux et al., 2010a; Richiardi et al., 2010). These two approaches are complementary: on the one hand, spatial ICA extracts spatial patterns of coactivation, but forgoes any modeling of the correlation structure of the time series, since it operates on whitened data; on the other hand, learning statistically sound correlation matrices with the small sample size of neuroimaging data must rely on the definition of regions of interest to extract the relevant signal (Varoquaux et al., 2010b).

Here we propose to *simultaneously* learn spatial patterns as well as the associated correlation structure by fitting a single model, leveraging recent work on sparse structured principal component analysis (SSPCA) (Jenatton et al., 2010b). Specifically, the time series are represented as linear combinations of factors, or *dictionary elements*, that are not simply constrained to be sparse (i.e., to use only a small number of voxels) but that are rather explicitly constrained to be supported by spatially-compact regions. The motivation for introducing this structure is the high degree of local spatial organization of the neuronal maps corresponding to different cognitive functions (Chklovskii and Koulakov, 2004). In addition, the spatial variability of function across different subjects limits the ability of brain imaging to resolve functional processes and dictates the use of brain regions learned from the data for inter-subject comparisons (Thirion et al., 2006). As detailed in the conclusion, an important consequence of selecting for structured com-

ponents, as with SSPCA, rather than for independent components, as with ICA, is that it gives a two level representation, segmenting localized brain regions regardless of the fluctuations of large scale cognitive networks that have been shown to encode cognitive or pathology-related brain states.

We introduce a new generative model for spontaneous brain activity that is suitable for model selection. This model provides a consistent framework for covariance estimation as well as various matrix decomposition methods, such as ICA, the reference method in neuroimaging (Beckmann and Smith, 2004). As illustrated in our experiments, our approach seeking structure in the data leads to improvements over other unstructured decompositions, with both (1) an increase in model likelihood on left-out time series corresponding to unseen subjects, as well as (2) decomposition in spatial components that matches known brain areas of interest and is thus interpretable. Our model is consistent with current neuroscientific understanding of spontaneous brain activity.

### 5.2.2 Model and Problem Statement

We start by introducing our generative model and by motivating the use of dictionary learning. We consider brain resting-state time series measured on $m$ voxels, and denote by $\mathbf{x} \in \mathbb{R}^m$ one of the time points, a three-dimensional image of brain activity. We assume that the signal $\mathbf{x}$ can be expressed as a linear combination of $p$ dictionary elements that are the columns of the *dictionary* matrix $\mathbf{D} \in \mathbb{R}^{m \times p}$. More precisely, we have

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{5.6}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ is a zero-mean Gaussian noise vector identically and independently distributed (i.i.d.) across the different data points acquired at different times. In the sequel, we shall assume that there is no spatial correlation between different voxels in the noise term, so that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is chosen to be diagonal. The decomposition of the signal $\mathbf{x}$ on the dictionary $\mathbf{D}$ is given by the vector $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ which follows a zero-mean multivariate Gaussian distribution, also i.i.d. across the observations. We further assume independence between $\boldsymbol{\varepsilon}$ and $\boldsymbol{\alpha}$, and $\mathbf{x}$ is thus Gaussian with covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{x}} \triangleq \mathbf{D}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\mathbf{D}^\top + \Sigma_{\boldsymbol{\varepsilon}}. \tag{5.7}$$

In the experiments, we will use this generative model to evaluate the quality of fit of our approach, as measured by the corresponding likelihood function.

Now, given a set of $n$ observations of the brain activity images (i.e., $n$ times points) represented by the rows of the data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, we are interested in expressing $\mathbf{X}$ as

$$\mathbf{X} \approx \mathbf{D}\mathbf{A}, \tag{5.8}$$

where the $n$ decomposition vectors are stacked in the matrix $\mathbf{A} \in \mathbb{R}^{p \times n}$. Finding such a pair of matrices $(\mathbf{A}, \mathbf{D})$ corresponds precisely to a matrix factorization problem (Singh and Gordon, 2008; Witten et al., 2009), also known as dictionary learning in signal processing (Mairal et al., 2010a). Various penalties and constraints can be imposed on $\mathbf{A}$

and/or $\mathbf{D}$ in order to specify prior knowledge about the target factorization, e.g., positivity (Lee and Seung, 1999) or sparsity (Lee et al., 2007; Mairal et al., 2010a). Independent Component Analysis (ICA), widely used in neuroimaging, can also be formulated as a dictionary-learning problem, with the constrain of minimum mutual information between the columns of $\mathbf{D}$.

The wide success that ICA has enjoyed in neuroimaging comes from the interpretability of the dictionary elements that it extracts. Indeed, unlike dictionary elements learned by PCA, ICA components display localized and contrasted features that match functional neuroanatomical knowledge. For instance the component shown on Figure 5.6 corresponds to the brain regions composing the well-known *default mode network*. More recently, it has been argued that the key to the success of ICA in fMRI is not the independence of the components, but their sparsity (Daubechies et al., 2009; Varoquaux et al., 2010d). In addition, systematic studies of all ICA components learn from resting-state brain activity show that they are composed of localized features, that correspond either to functional brain units or to structured artifacts in the signal generated by anatomical features (Varoquaux et al., 2010e). It thus appears that, for the interpretation of the decomposition on dictionary elements (i.e., the columns of $\mathbf{D}$), of the spatial compactness of the salient features that they display plays an important role: imaging neuroscientists think in terms of brain activation areas forming small localized clusters. Interestingly, this property can be encoded by using the structured sparsity-inducing regularization recently introduced in Jenatton et al. (2011a). We now briefly present this regularization scheme and explain how it has been further exploited for dictionary learning (Jenatton et al., 2010b).

### 5.2.3 Structured Sparse Dictionary Learning

The work of Jenatton et al. Jenatton et al. (2011a) considered a norm $\Omega$ which induces structured sparsity in the following sense: the solutions to a learning problem regularized by this norm have a sparse support which moreover belongs to a predefined set of possible nonzero patterns. Interesting examples of such sets of supports include sets of variables forming contiguous segments or rectangles when arranged respectively on a line or on a grid. More formally, the norm $\Omega$ can be defined on $\mathbb{R}^m$ by introducing a set $\mathcal{G}$ of subsets of $\{1, \ldots, m\}$ that we refer to as *groups*. The choice of the set of groups $\mathcal{G}$ defines the nature of the possible sparsity patterns associated with $\Omega$. For any vector $\mathbf{d} \in \mathbb{R}^m$,

$$\Omega(\mathbf{d}) \triangleq \sum_{g \in \mathcal{G}} \|\mathbf{d}_g\|_2 = \sum_{g \in \mathcal{G}} \left\{ \sum_{j \in g} \mathbf{d}_j^2 \right\}^{1/2}. \tag{5.9}$$

In other words, $\Omega$ consists of a sum of *non-squared* $\ell_2$-norms on subsets of variables specified by $g \in \mathcal{G}$. The definition (5.9) covers interesting subcases: for example, when $\mathcal{G}$ is the set of singletons, we get back the $\ell_1$-norm. As analyzed by (Jenatton et al., 2011a), $\Omega$ promotes sparsity at the level of groups, in the sense that it acts as a $\ell_1$-norm on the vector $(\|\mathbf{d}_g\|_2)_{g \in \mathcal{G}}$. Regularizing by $\Omega$ therefore causes some $\|\mathbf{d}_g\|_2$ (and equivalently $\mathbf{d}_g$) to be zeroed out for some $g$ in $\mathcal{G}$. In the experiments, we design a particular set of

overlapping groups $\mathcal{G}$ that is adapted to the 3-dimensional structure of the voxels. Our choice of $\mathcal{G}$ constrains the sparsity patterns to form rectangular boxes, in order to favor the selection of clustered voxels. This is achieved by considering for $\mathcal{G}$ all axis-aligned half-spaces of the discrete 3-dimensional space of voxels.

The norm $\Omega$ can naturally fit in the dictionary-learning framework (Jenatton et al., 2010b). Suppose that we look for a pair of matrices $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n] = [\mathbf{A}_1, \ldots, \mathbf{A}_p]$ and $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p]$ such that (5.8) holds. If in addition we require to obtain sparse structured dictionary elements, we can consider the formulation of Jenatton et al. (2010b), [6] that is,

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times n}, \mathbf{D}\in\mathbb{R}^{m\times p}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2+\lambda\sum_{k=1}^{p}\Omega(\mathbf{d}^k) \text{ such that } \|\mathbf{A}_k\|_2 \leq 1 \text{ for all } k \in [\![1;p]\!]. \quad (5.10)$$

The positive scalar $\lambda$ controls the strength of the regularization, while the constraints on the rows of the matrix $\mathbf{A}$ are required to avoid degenerate solutions. The specific choice of the $\ell_2$-norm follows Jenatton et al. (2010b). Problem (5.10) is not *jointly* convex in $\mathbf{A}$ and $\mathbf{D}$, but convex with respect to one matrix while the other one is kept fixed, and vice versa. This property calls for an alternate optimization scheme, more precisely a block coordinate descent scheme; we refer the reader to Jenatton et al. (2010b) and references therein for more details on the optimization procedure.

### 5.2.4 Applications and Experiments

We present in this section the experimental validation of our approach. To this end, we consider the set of brain resting-state time series used in Varoquaux et al. (2010e). Twelve healthy volunteers were scanned at rest, eyes closed, for a period of 20 minutes. Each individual dataset is made of $n = 820$ volumes (time points) with a 3 mm isotropic resolution, corresponding to approximately 50 000 voxels within the brain. Standard neuroimaging preprocessing was applied using the SPM5 software [7]: after slice-timing interpolation and motion correction, cerebral volumes were realigned to an inter-subject template and smoothed with a 6 mm isotropic Gaussian kernel.

The objective of the experiments is to compare our sparse structured approach (SSPCA) with a standard sparse dictionary learning formulation (Sparse PCA, SPCA)— i.e., using the $\ell_1$-norm in place of $\Omega$ in the formulation (5.10)— and with principal component analysis (PCA). This validation scheme is based solely on the decomposition $\mathbf{D}\mathbf{A}$ and thus makes no distinction between ICA and PCA. Indeed, the matrices $\mathbf{A}$, $\mathbf{D}$ found by ICA and PCA are identical up to a rotation, therefore leading to the same likelihood on unseen data [8]. For a set of dictionary sizes, $p \in \{10, 40, 70, \ldots, 220\}$, we perform a

6. In fact, Jenatton et al. (2010b) considers a weighted non-convex variant of $\Omega$ that leads to the same set of sparsity patterns, but more aggressively. In the experiments, we adopt the very same setting as Jenatton et al. (2010b); we do not detail this variant for simplicity. Moreover, though non-convex, the optimization with the variant of $\Omega$ follows along the same line thanks to a variational formulation.

7. Wellcome Department of Cognitive Neurology; www.fil.ion.ucl.ac.uk/spm

8. However, the spatial components obtained by these two methods are different.

6-fold cross-validation that respects splits of the data by subjects, i.e., each fold contains all the time points for two subjects. After learning a decomposition $(\hat{\mathbf{A}}, \hat{\mathbf{D}})$ on 5 out of the 6 folds, we compute an estimate $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ of $\mathbf{\Sigma}_{\mathbf{x}}$ as defined in (5.7), where $\hat{\mathbf{\Sigma}}_{\boldsymbol{\alpha}}$ is the sample covariance for $\hat{\mathbf{A}}$, and $\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}$ is obtained from the residuals. If we now denote by $\mathbf{\Sigma}_{\mathbf{x}}^{\text{held-out}}$ the sample covariance matrix of the time-series data points computed from the held-out fold, we evaluate the different approaches based on the likelihood $\mathcal{L}$ of the held out data in our Gaussian generative model, which is

$$\mathcal{L}(\hat{\mathbf{\Sigma}}_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}}^{\text{held-out}}) \triangleq -\log|\hat{\mathbf{\Sigma}}_{\mathbf{x}}| - \text{Tr}(\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\mathbf{\Sigma}_{\mathbf{x}}^{\text{held-out}}). \tag{5.11}$$

To lower the computational burden, the space of voxels is down-sampled to $m \approx 5\,000$, where the activation of a single voxel is replaced by the average activation of a $2\times2\times2$-cube of voxels. The cross-validation scores are displayed on Figure 5.5. Both SPCA and SSPCA are non-convex, and solutions therefore depend on the initialization of the algorithms. We experimentally observed that the cross-validation scores are quite stable under random initialization of $\mathbf{A}$ and $\mathbf{D}$. Moreover, we assess the statistical significance of the cross-validation scores in Table 5.3 and carry out paired t-tests.

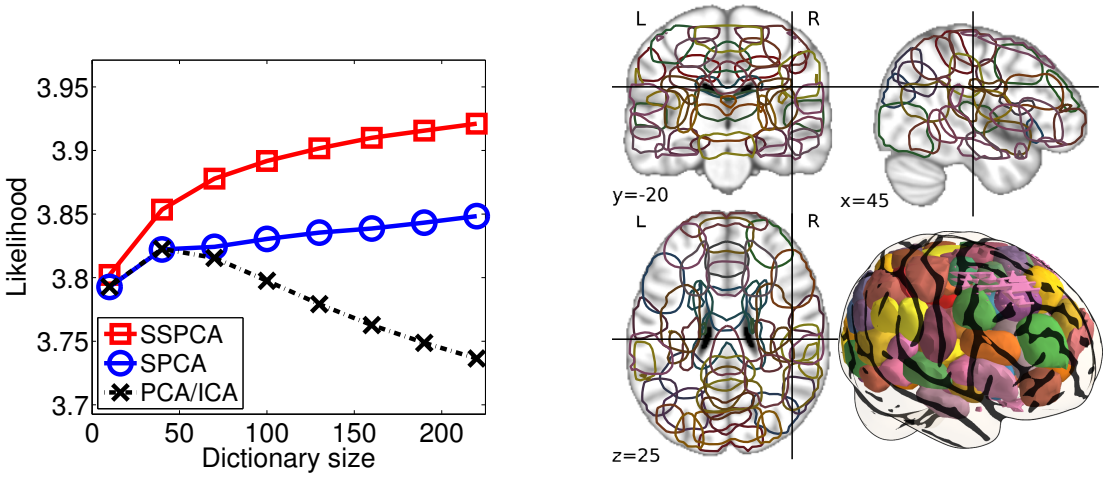



Figure 5.5: **(left)** Cross-validation scores: Average likelihood on unseen data obtained for SSPCA, SPCA and PCA/ICA over the 6 folds, with different dictionary sizes. The likelihood corresponds here to $\mathcal{L}$ normalized by $10^4$. No error bars are displayed on this figure, the statistical significance of the results being assessed in Table 5.3. **(right)** Outline of the regions defined by the different dictionary elements for $p = 100$ displayed on 2D cuts, as well as a 3D view with a representation of the cortical folds. The outlines drawn correspond to 25% of the maximum value of the dictionary elements. A poorly-structured pink region stands out outside the upper frontal regions: it is a well known movement-pattern often extracted in ICA analysis (map 9 on Figure 2 of Varoquaux et al. (2010e)).

On Figure 5.5, we can see that, as the number of elements in the dictionary learned increases, the various learning procedures extract a more detailed representation and the

| Dictionary sizes | | 10 | 40 | 70 | 100 | 130 | 170 | 200 | 220 |
|---|---|---|---|---|---|---|---|---|---|
| p-values | SSPCA vs. SPCA | $5.10^{-2}$ | $3.10^{-3}$ | $1.10^{-3}$ | $7.10^{-5}$ | $1.10^{-4}$ | $4.10^{-6}$ | $8.10^{-6}$ | $1.10^{-5}$ |
| | SSPCA vs. PCA/ICA | $4.10^{-2}$ | $3.10^{-3}$ | $5.10^{-3}$ | $2.10^{-3}$ | $1.10^{-3}$ | $8.10^{-4}$ | $6.10^{-4}$ | $6.10^{-4}$ |

Table 5.3: P-values from paired t-tests carried out on the series of likelihood results over the 6 cross-validation scores. For the models learned using SSPCA, the average likelihood on the held-out data is significantly higher than those obtained using SPCA and PCA at the 1% significance level for dictionary sizes greater than 40.

likelihood of unseen data increases. However for PCA and ICA, this generalization score reaches a plateau and decreases as the algorithms overfit and learn dictionary elements from signal that is not reproducible across folds, that is between different subjects. Interestingly, the optimal number of ICA components set by this model-selection procedure, 50, corresponds roughly to the number selected on the same dataset by a different procedure unrelated to the proposed generative model (Varoquaux et al., 2010e). For the penalized methods, SPCA and SSPCA, the generalization score keeps increasing with the number of elements in the dictionary: the penalization controls over-fitting. SSPCA generalizes best for large dictionary sizes and is thus best suited for obtaining a detailed description of resting-state brain covariance.

In our experiments, the generalization scores of models learned by SSPCA and SPCA do not reach a maximum. This could indicate that the best description of the resting brain involves a higher number of regions. Indeed, Tucholka et al. (2008) argues that, in a task-based fMRI study, the optimal number of regions lies around 500. As the computational cost of these methods scales in $p^2$, our experiments were limited by available computational resources and we plan to explore larger dictionary sizes in the future.

### 5.2.5 Discussion

The key idea behind the study of resting-state brain activity is that the correlations in the signal observed by fMRI reveal regions of the brain that function together. This is why modeling the covariance matrix of fMRI time series is receiving increasing attention (Varoquaux et al., 2010b; Cecchi et al., 2009). However, with the limited number of samples, learning a covariance between the time series of all the 50 000 brain voxels is an ill-posed problem. The generative model defined by equations (5.7) and (5.8) can be interpreted as decomposing the covariance of the observed signal in a structured part, given by the dictionary elements $\mathbf{D}$ and the covariance $\mathbf{\Sigma}_{\boldsymbol{\alpha}}$, and an unstructured part, given by the diagonal covariance $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$. As such, this model is well-suited for the study of resting-state functional connectivity.

The dictionary elements estimated by SSPCA are spatial maps displaying one or two localized features. As such, they define a set of regions that explain well the signal.

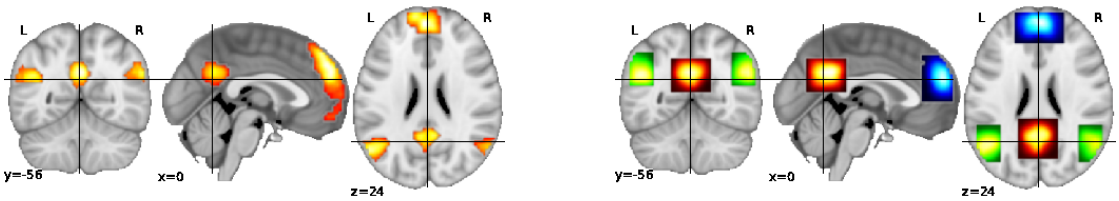



Figure 5.6: **(left)** Dictionary element estimated by ICA, corresponding to the brain network known as the *default mode network*. As it is usually done in neuroimaging, the spatial map learned by decomposition of the functional images is displayed, thresholded, on an anatomical brain image. **(right)** Three dictionary elements corresponding to the same brain regions. The different colormaps (red, green and blue) correspond to the different dictionary elements learned. The transparency reflects the structured sparsity, rather than an arbitrary threshold[9]. This is why the different elements displayed on the figure have a block-like outline that reflect the shape of the groups used in norm $\Omega$. Some displayed patterns do not seem to be rectangular boxes (e.g., the green patterns). However, a closer look at these dictionary elements shows that boxes are indeed selected, and that small numerical values (just as regularizing by $\ell_2$-norm may lead to) give the visual impression of having zeroes between the two green areas.

Figure 5.5 shows the outline of all the regions defined by such a decomposition. These regions either correspond to functional brain regions, located in the gray matter, or are characteristic of non-neural signals, such as movement, most-often considered as artifacts to be accounted for in fMRI studies. We find that the brain regions segmented by SSPCA are well positioned on functionally-relevant anatomical structures, such as cortical folds, and correspond well to the regions forming the well-known brain networks extracted most-often by ICA (see Figure 5.6).

Compared to standard ICA-based procedures for fMRI, our approach based on SSPCA offers several improvements that go beyond better cross-validation performance for large dictionaries. First of all, it naturally extracts regions: no thresholding is necessary to recover sparsity as in Varoquaux et al. (2010d) and spatially scattered dictionary elements are penalized. In addition, ICA has no intrinsic notion of noise: it forgoes explained variance and explains all the data at hand by estimating a rotation matrix, called the mixing matrix, that maps whitened signals to dictionary elements. As a result, ICA procedures used in fMRI rely on PCA to select a subspace maximizing the explained variance (Beckmann and Smith, 2004; Varoquaux et al., 2010e). On the contrary, SPCA and SSPCA perform subspace selection and signal decomposition in a single step, relying on the penalization to model noise. Thus, when using SSPCA, the proportion of the observed signal modelled as noise is set by the structured-sparsity prior.

9. The optimization procedure we rely on (Jenatton et al., 2010b) is not based on thresholding to set coefficients explicitly to zero, but on successive re-weighing of coefficients, some of which converge to zero. As a result the sparsity of the dictionary element is given by the error tolerance set in the optimization.

But the most important improvement of SSPCA over ICA is probably to differentiate the estimation of localized regions from the grouping distant regions in networks: $\mathbf{D}$ encodes only local information and all long-distance structure can be found in $\mathbf{\Sigma}_{\boldsymbol{\alpha}}$. Indeed, a large body of work has shown that the amount of correlation between distant brain regions is modulated by cognitive brain processes (Richiardi et al., 2010) or pathologies (Vanhaudenhuyse et al., 2010; Cecchi et al., 2009; Varoquaux et al., 2010a). When estimating dictionary elements on multiple subjects with different long-distance correlation structures, brain regions extracted by ICA will be affected while there is no reason to believe that the underlying localized functional units differ. SSPCA is more robust to such variations in the data. As such, it is more suited to extract regions that will be used in a second step to perform principled inter-subject inference on the correlation structure (Varoquaux et al., 2010a).

### 5.2.6 Conclusion

We introduced in this paper a generative model for brain resting-state activity. We use the dictionary-learning framework to learn this model from fMRI time-series. In these settings, we can compare several dictionary-learning techniques using cross-validation. We apply sparse-structured PCA, an approach with assumptions that match current understanding of brain spontaneous activity, and show that it generalizes better than alternative methods, including ICA, the reference method in neuroimaging. These preliminary results suggest that SSPCA is a good candidate for learning regions from the data, which is a challenging problem, critical for brain covariance modeling (Varoquaux et al., 2010b). Our proposed model could be extended in further work to estimating a subject-specific covariance matrix with common dictionary elements. This would enable principled inter-subject comparison on brain covariance using a consistent framework based on dictionary learning.

# 6

# Local Analysis of Sparse Coding in Presence of Noise

**Abstract of the chapter:** A popular approach within the signal processing and machine learning communities consists in modelling signals as sparse linear combinations of atoms selected from a *learned* dictionary. While this paradigm has led to numerous empirical successes in various fields ranging from image to audio processing, there have only been a few theoretical arguments supporting these evidences. In particular, sparse coding, or sparse dictionary learning, relies on a non-convex procedure whose local minima have not been fully analyzed yet. In this chapter, we consider a probabilistic model of sparse signals, and show that, with high probability, sparse coding admits a local minimum along some curves passing through the reference dictionary generating the signals.

Our study takes into account the case of over-complete dictionaries and noisy signals, thus extending previous work limited to noiseless settings and/or under-complete dictionaries. The analysis we conduct is non-asymptotic and makes it possible to understand how the key quantities of the problem, such as the coherence or the level of noise, are allowed to scale with respect to the dimension of the signals, the number of atoms, the sparsity and the number of observations.

This work in progress constitutes a first step towards the more involved proof of the existence of the local minimum over the entire manifold of normalized dictionaries.

The material of this chapter is based on some work in progress which has been achieved with the collaboration of Francis Bach and Rémi Gribonval.

## 6.1 Introduction

Modelling signals as sparse linear combinations of atoms selected from a dictionary has become a popular paradigm in many fields, including signal processing, statistics, and machine learning. This line of research has witnessed the development of several well-founded theoretical frameworks (e.g., see Tropp, 2006; Wainwright, 2009; Zhang, 2009) and efficient algorithmic tools (e.g., see Bach et al., 2011, and references therein).

However, the performance of such approaches hinges on the representation of the signals, which makes the question of designing "good" dictionaries prominent. A great deal of effort has been dedicated to come up with efficient *predefined* dictionaries, e.g., the various types of wavelets (Mallat, 1999). These representations have notably contributed to many successful image processing applications such as compression. More recently,

the idea of simultaneously *learning* the dictionary and the sparse decompositions of the signals —also known as *sparse dictionary learning*, or simply, *sparse coding*— has emerged as a powerful framework, with state-of-the-art performances in many tasks, including denoising, inpainting and image classification (e.g., see Mairal et al., 2010a, and references therein).

Although sparse dictionary learning can alternatively be formulated as convex (Bach et al., 2008; Bradley and Bagnell, 2009b), non-parametric Bayesian (Zhou et al., 2009) and submodular (Krause and Cevher, 2010) problems, the most popular and widely used definition of sparse coding brings into play a non-convex optimization problem. Despite its empirical and practical success, there has only been a little theoretical analysis of the properties of sparse dictionary learning. For instance, Maurer and Pontil (2010); Vainsencher et al. (2010) derive generalization bounds which quantify how much the *expected* signal-reconstruction error differs from the *empirical* one, computed from a random and finite-size sample of signals. The bounds obtained in Maurer and Pontil (2010); Vainsencher et al. (2010) are non-asymptotic and uniform with respect to the whole class of dictionaries considered (e.g., those with normalized atoms). As discussed later, the questions raised in this chapter explore a different and complementary direction.

Another theoretical aspect of interest consists of characterizing the local minima of sparse coding, in spite of the non-convexity of its formulation. This problem is closely related to the question of *identifiability*, that is, whether it is possible to recover a reference dictionary that is assumed to generate the observed signals. The authors of Gribonval and Schnass (2010) pioneered research in this direction by considering noiseless signals in the case where the reference dictionary forms a basis. Still in a noiseless setting, Geng et al. (2011) extended the analysis to *over-complete* dictionaries, i.e., these composed of more atoms than the dimension of the signals. To the best of our knowledge, comparable analysis have not been carried out yet for noisy signals. In particular, the structure of the proofs of Gribonval and Schnass (2010); Geng et al. (2011) hinges on the absence of noise and cannot be straightforwardly transposed to take into account some noise; this point will be made more formal in the subsequent sections.

In this chapter, we therefore analyze the local minima of sparse coding in presence of noise and make the following contributions:

- Within a probabilistic model of sparse signals, we derive a *non-asymptotic* characterization of the probability of finding a local minimum along some curves passing through the reference dictionary.
- The analysis we conduct makes it possible to better understand (a) how many signals are required to hope for identifiability, (b) what the impact of the degree of over-completeness is, (c) what parameters contribute to the curvature around minima, and (d) what level of noise appears as manageable.
- We show that under deterministic coherence-based assumptions, such a local minimum along a curve is guaranteed to exist with high probability.

This work is in progress and represents a first milestone towards the more involved

result guaranteeing the existence of a local minimum in a neighborhood of the reference dictionary. Importantly, the neighborhood has to be understood as being *within the entire manifold of normalized dictionaries*, as opposed to be just along single curves. Arguments based on (1) concentration of the measure uniformly over all such curves, and (2) the discretization of this set of curves through $\epsilon$-nets (Massart, 2003) seem to indicate that the current conclusions of this chapter shall still apply with small changes.

**Notation.** For all vectors $\mathbf{v} \in \mathbb{R}^p$, we denote by $\text{sign}(\mathbf{v}) \in \{-1, 0, 1\}^p$ the vector such that its $j$-th entry $[\text{sign}(\mathbf{v})]_j$ is equal to zero if $\mathbf{v}_j = 0$, and to one (respectively, minus one) if $\mathbf{v}_j > 0$ (respectively, $\mathbf{v}_j < 0$). We extensively manipulate matrix norms in the sequel. For any matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$, we define its Frobenius norm by $\|\mathbf{A}\|_{\text{F}} \triangleq [\sum_{i=1}^n \sum_{j=1}^p \mathbf{A}_{ij}^2]^{1/2}$; similarly, we denote the spectral norm of $\mathbf{A}$ by $\|\!|\mathbf{A}|\!\|_2 \triangleq \max_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{A}\mathbf{x}\|_2$, and refer to the operator $\ell_\infty$-norm as $\|\!|\mathbf{A}|\!\|_\infty \triangleq \max_{\|\mathbf{x}\|_\infty \leq 1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_{i \in [\![1;n]\!]} \sum_{j=1}^p |\mathbf{A}_{ij}|$.

For any square matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, we denote by $\text{diag}(\mathbf{B}) \in \mathbb{R}^n$ the vector formed by extracting the diagonal terms of $\mathbf{B}$, and conversely, for any $\mathbf{b} \in \mathbb{R}^n$, we use $\text{Diag}(\mathbf{b}) \in \mathbb{R}^{n \times n}$ to represent the (square) diagonal matrix whose diagonal elements are built from the vector $\mathbf{b}$.

Moreover, we introduce the ordering $\lesssim$ such that for any scalar $a, b \in \mathbb{R}$, the relationship $a \lesssim b$ holds if and only if there exists a non-negative universal constant $\omega$ satisfying $a \leq \omega b$.

## 6.2 Problem statement

We introduce in this section the material required to define our problem and state our results.

### 6.2.1 Background material on sparse coding

Let us consider a set of $n$ signals $\mathbf{X} \triangleq [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{m \times n}$ of dimension $m$, along with a dictionary $\mathbf{D} \triangleq [\mathbf{d}^1, \ldots, \mathbf{d}^p] \in \mathbb{R}^{m \times p}$ composed of $p$ atoms—also referred to as dictionary elements. Sparse coding simultaneously learns a dictionary $\mathbf{D}$ and a set of $n$ sparse vectors $\mathbf{A} \triangleq [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n] \in \mathbb{R}^{p \times n}$ such that each signal $\mathbf{x}^i$ can be well approximated by $\mathbf{x}^i \approx \mathbf{D}\boldsymbol{\alpha}^i$ for $i$ in $[\![1; n]\!]$. By sparse, we mean that the vector $\boldsymbol{\alpha}^i$ has $k \ll p$ non-zero coefficients, so that we aim at reconstructing $\mathbf{x}^i$ from only a few atoms. Before introducing the standard formulation of sparse coding (Olshausen and Field, 1997; Lee et al., 2007; Mairal et al., 2010a), we define the following elements:

**Definition 4**
*For any dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ and signal $\mathbf{x} \in \mathbb{R}^m$, we define the function $f_{\mathbf{x}}$ as*

$$f_{\mathbf{x}}(\mathbf{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{6.1}$$

*Similarly for any set of $n$ signals $\mathbf{X} \triangleq [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{m \times n}$, we introduce*

$$F_n(\mathbf{D}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{x}^i}(\mathbf{D}). \tag{6.2}$$

The standard approach to perform sparse coding (Olshausen and Field, 1997; Lee et al., 2007; Mairal et al., 2010a) consists in solving the minimization

$$\min_{\mathbf{D} \in \mathcal{D}} F_n(\mathbf{D}), \tag{6.3}$$

where the regularization parameter $\lambda$ in Eq. (6.1) controls the level of sparsity, while $\mathcal{D} \subseteq \mathbb{R}^{p \times n}$ denotes a compact (generally convex) set; in this chapter, $\mathcal{D}$ is chosen to be the set of dictionaries with unit $\ell_2$-norm atoms, which is a natural choice in image processing (Mairal et al., 2010a). Note however that other choices for the set $\mathcal{D}$ may also be relevant depending on the application at hand (e.g., see Jenatton et al., 2011c, where in the context of topic models, the atoms in $\mathcal{D}$ belong to the unit simplex).

The goal of the chapter is to characterize some local minima of the function $F_n$ under a generative model for the signals $\mathbf{x}^i$. Although the function $F_n$ is Lipschitz continuous (Mairal et al., 2010a), its minimization is challenging since it is non-convex and subject to the constraints of $\mathcal{D}$. Moreover, $F_n$ is defined through the minimization over the vectors $\mathbf{A}$, which, at first sight, does not lead to a simple and convenient expression. We next show that $F_n$ has a simple form in some specific favorable scenarios.

**Closed-form expression for $F_n$:** We leverage here a key property of the function $f_{\mathbf{x}}$. Let denote by $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ a solution of problem (6.1), that is, the minimization defining $f_{\mathbf{x}}$. If the sign of $\hat{\boldsymbol{\alpha}}$ is known in advance, say $\hat{\mathbf{s}} \in \{-1, 0, 1\}^p$ with *support* defined by $\mathrm{J} \triangleq \{j \in [\![1; p]\!]; \hat{\mathbf{s}}_j \neq 0\}$, then $\hat{\boldsymbol{\alpha}}$ has a simple closed-form expression (e.g., see Fuchs, 2005; Wainwright, 2009). In particular, if we use the notation $\mathbf{D}_{\mathrm{J}} \in \mathbb{R}^{p \times |\mathrm{J}|}$ to denote the dictionary restricted to the $|\mathrm{J}|$ atoms indexed by J, and assuming that $\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}$ is invertible, we have

$$\hat{\boldsymbol{\alpha}}_{\mathrm{J}} = \left[\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}\right]^{-1} \left[\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \hat{\mathbf{s}}_{\mathrm{J}}\right] \in \mathbb{R}^{|\mathrm{J}|} \quad \text{and} \quad \hat{\boldsymbol{\alpha}}_{\mathrm{J}^c} = \mathbf{0}.$$

This property is appealing in that it makes it possible to obtain a closed-form expression for $f_{\mathbf{x}}$ (and hence, $F_n$), provided that we can control the sign patterns of $\hat{\boldsymbol{\alpha}}$. In the light of this remark, it is natural to define the following function:

**Definition 5**
*Let $\mathbf{s} \in \{-1, 0, 1\}^p$ be a sign vector with support J. For any dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ such that $\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}$ is invertible, and for any signal $\mathbf{x} \in \mathbb{R}^m$, we define the function $\phi_{\mathbf{x}}$ as*

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) \triangleq \frac{1}{2}\left[\|\mathbf{x}\|_2^2 - (\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})^\top (\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}})^{-1} (\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})\right]. \tag{6.4}$$

*We define $\Phi_n$ analogously to $F_n$ for a sign matrix $\mathbf{S} \in \{-1, 0, 1\}^{p \times n}$.*

Based on the previous definition, note that we have the relationship

$$f_{\mathbf{x}}(\mathbf{D}) = \frac{1}{2}\|\mathbf{x} - \mathbf{D}\hat{\boldsymbol{\alpha}}\|_2^2 + \lambda\|\hat{\boldsymbol{\alpha}}\|_1 = \phi_{\mathbf{x}}(\mathbf{D}|\hat{\mathbf{s}}).$$

As it will be further discussed later, showing that the functions $F_n$ and $\Phi_n$ coincide will amount to studying the sign recovery property of $\ell_1$-regularized least-squares problems, which is a topic already well-understood (e.g., see Wainwright, 2009, and references therein). We now introduce the generative model of sparse signals.

### 6.2.2 Probabilistic model of sparse signals

Throughout the chapter, we assume the signals we observe are generated *independently* according to a specific probabilistic model. Let us consider a fixed reference dictionary $\mathbf{D}_0 \in \mathcal{D}$. Each signal $\mathbf{x} \in \mathbb{R}^m$ is built from the following *independent* steps:

(1) Draw without replacement $k$ atoms out of the $p$ available in $\mathbf{D}_0$. This procedure thus defines a support $\mathrm{J} \triangleq \{j \in [\![1;p]\!];\ \delta(j) = 1\}$ whose size is $|\mathrm{J}| = k$, and where $\delta(j)$ denotes the indicator function equal to one if the $j$-th atom is selected, zero otherwise. Note that $\mathbb{E}[\delta(j)] = \frac{k}{p}$, and for $i \neq j$, we further have $\mathbb{E}[\delta(j)\delta(i)] = \frac{k(k-1)}{p(p-1)} \leq \frac{k^2}{p^2}$. Our result also holds for any sampling scheme for which the conditions above on the expectation are satisfied.

(2) Define a sparse vector $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$ whose entries in J are generated i.i.d. according to a symmetric distribution with a compact support bounded away from zero. Formally, there exist some strictly positive scalars $(\underline{\alpha}, \overline{\alpha})$ such that for all $j \in J$, it holds that $|[\boldsymbol{\alpha}_0]_j| \in [\underline{\alpha}, \overline{\alpha}]$ almost surely. On the other hand, for $j$ not in J, $[\boldsymbol{\alpha}_0]_j$ is set to zero. As discussed later in Section 6.3.1, our analysis can probably be extended to the class of distributions with non-compact supports containing zero, with a sufficiently small mass at zero.

(3) Eventually generate the signal $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}$, where the entries of the additive noise $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ are assumed i.i.d. Gaussian with zero-mean and variance $\sigma^2$. The Gaussian assumption is made here for simplicity; our study can be similarly conducted for more general sub-Gaussian noises.

With this generative model in place, we can state more precisely our objective: we want to show that

$$\Pr(F_n \text{ has a local minimum in a "neighborhood" of } \mathbf{D}_0) \approx 1,$$

where the probability is with respect to the two sources of randomness, i.e., $\boldsymbol{\alpha}_0$ and $\boldsymbol{\varepsilon}$. Note that we loosely refer to a certain "neighborhood" since in our regularized formulation, a local minimum cannot appear exactly at $\mathbf{D}_0$. The proper meaning of this neighborhood is the subject of the next section.

Importantly, we have so far referred to $\mathbf{D}_0$ as *the* reference dictionary generating the signals. However, and as already discussed in Gribonval and Schnass (2010); Geng et al. (2011), we know that for any permutation matrix $\Sigma$ and any diagonal matrix $\mathrm{Diag}(\mathbf{s})$ with the vector $\mathbf{s}$ in $\{-1, 1\}^p$, we have $\mathbf{D}_0\boldsymbol{\alpha}_0 = (\mathbf{D}_0\Sigma^{-1}\mathrm{Diag}(\mathbf{s})^{-1})(\mathrm{Diag}(\mathbf{s})\Sigma\boldsymbol{\alpha}_0)$

with $\|\text{Diag}(\mathbf{s})\Sigma\boldsymbol{\alpha}_0\|_1 = \|\boldsymbol{\alpha}_0\|_1$. As a result, while solving (6.3), we cannot hope for the identifiability of the specific $\mathbf{D}_0$, but we focus instead on the identifiability of the whole *equivalence class* defined by the transformations described above. From now on, we simply refer to $\mathbf{D}_0$ to denote one element of this equivalence class.

### 6.2.3 Oblique manifold and tangent space

The minimization of $F_n$ is carried out over $\mathcal{D}$, which is the set of dictionaries with unit $\ell_2$-norm atoms. This set turns out to be a manifold, known as the *oblique manifold* (Absil et al., 2008). Since $\mathbf{D}_0$ is assumed to belong to $\mathcal{D}$, it is therefore natural to consider the behavior of $F_n$ according to the geometry of $\mathcal{D}$. More specifically, let us consider the set of matrices

$$\mathcal{W}_{\mathbf{D}_0} \triangleq \Big\{\mathbf{W} \in \mathbb{R}^{m\times p};\ \text{diag}(\mathbf{W}^\top\mathbf{D}_0) = \mathbf{0} \text{ and } \text{diag}(\mathbf{W}^\top\mathbf{W}) = \mathbf{1}\Big\}.$$

Now, for any matrix $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$, for any velocity vector $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 = 1$, and for all $t \in \mathbb{R}$, we introduce the parameterized dictionary:

$$\mathbf{D}(t) \triangleq \mathbf{D}_0\text{Diag}[\cos(\mathbf{v}t)] + \mathbf{W}\text{Diag}[\sin(\mathbf{v}t)], \tag{6.5}$$

where $\text{Diag}[\cos(\mathbf{v}t)]$ and $\text{Diag}[\sin(\mathbf{v}t)] \in \mathbb{R}^{p\times p}$ stand for the diagonal matrices with diagonal terms equal to $\{\cos(\mathbf{v}_j t)\}_{j\in[\![1;p]\!]}$ and $\{\sin(\mathbf{v}_j t)\}_{j\in[\![1;p]\!]}$ respectively. By construction, we have $\mathbf{D}(t) \in \mathcal{D}$ for all $t \in \mathbb{R}$ and $\mathbf{D}(0) = \mathbf{D}_0$. Moreover, in the notation of our parametrization, the set of matrices given by

$$\Big\{\mathbf{M} \in \mathbb{R}^{m\times p}; \mathbf{M} = \Big[\frac{\partial\mathbf{D}(t)}{\partial t}\Big]_{t=0} \triangleq \mathbf{W}\text{Diag}(\mathbf{v}),\ \text{with } \mathbf{W} \in \mathcal{W}_{\mathbf{D}_0} \text{ and } \|\mathbf{v}\|_2 = 1\Big\}$$

corresponds to the tangent space of $\mathcal{D}$ at $\mathbf{D}_0$ (Absil et al., 2008), intersected with the set of matrices in $\mathbb{R}^{m\times p}$ with unit Frobenius norm (indeed, we have $\|\mathbf{W}\text{Diag}(\mathbf{v})\|_{\text{F}} = 1$). Note that the parameter $\mathbf{v}$ is essential in that the set $\mathcal{W}_{\mathbf{D}_0}$ does not contain every direction from the tangent space of $\mathcal{D}$ at $\mathbf{D}_0$, as illustrated in the example below:

*Example: Why is the parameter* $\mathbf{v}$ *important?* Consider the dictionary $\mathbf{D}_0 = [\mathbf{d}_0^1, \mathbf{d}_0^2]$ with $m = p = 2$ and the following choices of atoms

$$\mathbf{d}_0^1 \triangleq \begin{bmatrix}0\\1\end{bmatrix} \quad \text{and} \quad \mathbf{d}_0^2 \triangleq \begin{bmatrix}1\\0\end{bmatrix}.$$

In this case, the set $\mathcal{W}_{\mathbf{D}_0}$ is given by

$$\mathcal{W}_{\mathbf{D}_0} = \Big\{\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2] \in \mathbb{R}^{2\times 2};\ \mathbf{w}^1 = \pm\begin{bmatrix}1\\0\end{bmatrix} \text{ and } \mathbf{w}^2 = \pm\begin{bmatrix}0\\1\end{bmatrix}\Big\},$$

while the entire tangent space at $\mathbf{D}_0$ is the following vector space

$$\Big\{\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2] \in \mathbb{R}^{2\times 2};\ (\mathbf{w}^1)^\top\mathbf{d}_0^1 = 0,\ \text{and } (\mathbf{w}^2)^\top\mathbf{d}_0^2 = 0\Big\}.$$

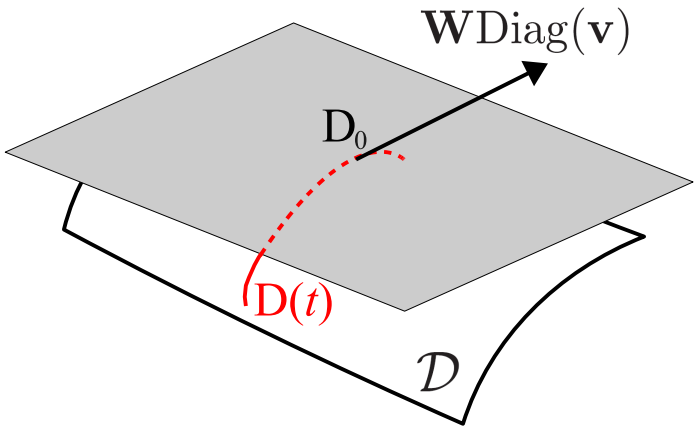



Figure 6.1: Illustration of the parametrization of the set $\mathcal{D}$ of dictionaries with unit $\ell_2$-norm atoms. The dictionary $\mathbf{D}(t)$ is a curve of $\mathcal{D}$ passing through the reference dictionary $\mathbf{D}_0$ (in red). The curve is parametrized by $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ and $\mathbf{v}$, so that $\mathbf{W}\mathrm{Diag}(\mathbf{v})$ defines a direction of the tangent space at $\mathbf{D}_0$ (in gray).

In particular, we can see that the matrix $\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ does belong to the tangent space whereas we obviously have $\mathbf{M} \notin \mathcal{W}_{\mathbf{D}_0}$. Moreover, the matrix $\mathbf{M}$ can be written as $\mathbf{M} = \mathbf{W}\mathrm{Diag}(\mathbf{v})$, with the vector $\mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\|\mathbf{v}\|_2 = 1$, and the matrix $\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in \mathcal{W}_{\mathbf{D}_0}$.

We can then describe the neighborhood of $\mathbf{D}_0$ through the triplet $(t, \mathbf{v}, \mathbf{W})$. In particular, instead of studying the local minima of $\mathbf{D} \mapsto F_n(\mathbf{D})$, we focus on the function $t \mapsto F_n(\mathbf{D}(t))$ for any pair $(\mathbf{v}, \mathbf{W})$. Loosely speaking, this corresponds to explore every curve of $\mathcal{D}$ passing through $\mathbf{D}_0$; see Figure 6.1. We next describe the last ingredient required to state our results.

### 6.2.4 Mutual coherence

Our analysis needs the reference dictionary $\mathbf{D}_0$ to respect some coherence-based properties. We thus define the mutual coherence of the dictionary $\mathbf{D}_0$ (e.g., see Donoho and Huo, 2001; Feuer and Nemirovski, 2003; Lounici, 2008) by

$$\mu_0 \triangleq \max_{i,j \in [\![1;p]\!], i \neq j} |[\mathbf{d}_0^i]^\top [\mathbf{d}_0^j]| \in [0, 1].$$

The term $\mu_0$ gives a measure of the level of correlation in $\mathbf{D}_0$, and it is for instance equal to zero in the case of an orthogonal dictionary. Similarly, we introduce $\mu(t)$ for the dictionary $\mathbf{D}(t)$. For any $t \in \mathbb{R}$, we notably have the simple inequality:

$$\mu(t) \triangleq \max_{i,j \in [\![1;p]\!], i \neq j} |[\mathbf{d}^i(t)]^\top [\mathbf{d}^j(t)]| \leq \mu_0 + 3|t|. \tag{6.6}$$

In particular, note that we have $\mu(0) = \mu_0$. We now see how the coherence is exploited in our study.

**Assumption on the reference dictionary $\mathbf{D}_0$:** For the theoretical analysis we conduct here, we consider a deterministic coherence-based assumption, such that the coherence $\mu_0$ and the level of sparsity $k$ should be inversely proportional. Such an assumption was considered for instance in the previous work by Geng et al. (2011).

The need for this constraint on $\mu_0$ appears twice in the analysis, first to control the supports of the solutions of (6.1) (see Section 6.3.3 for more details), and second, to be able to guarantee for the function $F_n$ some curvature in a certain neighborhood of $\mathbf{D}_0$. More specifically, we shall assume that

- **(Support recovery)** For some $\eta_0 \in (0,1)$, it holds that

$$\mu_0 \leq \frac{1-\eta_0}{2-\eta_0}\frac{1}{k}.$$

  Based on this assumption, we have for all $\mathrm{J} \subseteq [\![1;p]\!]$ with $|\mathrm{J}| \leq k$,

$$\left\| [\mathbf{D}_0]_{\mathrm{J}^c}^\top [\mathbf{D}_0]_{\mathrm{J}} \left[ [\mathbf{D}_0]_{\mathrm{J}}^\top [\mathbf{D}_0]_{\mathrm{J}} \right]^{-1} \right\|_\infty \overbrace{\leq}^{(*)} \frac{k\mu_0}{1-k\mu_0} \leq 1-\eta_0, \tag{6.7}$$

  where the first inequality $(*)$ is proved in Lemma 16, while the second inequality follows from the assumption. The condition displayed in Eq. (6.7) is well-known and is often referred to as the *irrepresentability condition* (e.g., see Fuchs, 2005; Zhao and Yu, 2006; Wainwright, 2009). As further discussed in Section 6.3.3, it is a central element to control the supports of the solutions of $\ell_1$-regularized least-squares problems. As a side comment, it worth noting that we impose the irrepresentability condition *via* a condition on the coherence, which is a stronger requirement (Van de Geer and Bühlmann, 2009).
- **(Curvature)** We also need the following condition to hold

$$1 - \frac{k}{p}\frac{1+p\mu_0}{1-k\mu_0} > 0, \text{ or equivalently, } \mu_0 < \frac{p-k}{2pk}. \tag{6.8}$$

  where the second inequality is equivalent to $\mu_0 \leq \frac{1-\eta_0}{2-\eta_0}\frac{1}{k}$. Note that the first inequality $(*)$ is proved in Lemma 16.
- **(Curvature)** We also need the following condition to hold

$$1 - \frac{k}{p}\frac{1+p\mu_0}{1-k\mu_0} > 0, \text{ or equivalently, } \mu_0 < \frac{p-k}{2pk}.$$

  This second assumption is required to show that the second derivative of $F_n$ is positive in some neighborhood of $\mathbf{D}_0$. Lemma 11 makes this statement more formal and precise.

In the main results from Section 6.3.1, we shall further simplify the assumptions above through numerical constants (for more details, see Assumption (1) in Theorem 5), so that $\frac{k}{p}\frac{1+p\mu_0}{1-k\mu_0}$ and terms depending only on $k\mu_0$ will be regarded as universal constants.

After having described the different components of our problem, we now present the core contribution of this chapter.

## 6.3 Main result and structure of the proof

This section contains the statement of the main result of this chapter (see Theorem 5 and Corollary 6) which shows that under appropriate scalings of $(n, p, k, m, \sigma, \mu_0)$, it is possible to prove that, with high probability, problem (6.3) admits a local minimum along some curves passing through $\mathbf{D}_0$. The proofs of the results of this section may be found in Appendix 6.6.

### 6.3.1 Main result

We begin with a complete statement of our result, before giving a simplified (but also easier to interpret) corollary:

**Theorem 5** (Local minimum of sparse coding along curves)
*Consider $n$ independent signals following the generative model from Section 6.2.2, with the reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$. Denoting by $N(n) \triangleq \log^2(n)/\sqrt{n}$ and $\sigma(t) \triangleq \max\{\sigma; \overline{\alpha}|t|\}$, let us introduce the following set of assumptions for some $t \in \mathbb{R}$:*

*(1)* ***[Coherence]:*** $k\mu_0 \leq \min\left\{\frac{1}{2}; \frac{p-2k}{3p}\right\}$

*(2)* ***[Dictionary perturbation]:*** $t_{\min} \lesssim |t| \lesssim \frac{1}{k}$ *with*

$$\begin{cases} t_{\min} & \gtrsim \frac{\lambda p}{\underline{\alpha}^2 k} \max\left\{\lambda; \overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0; \overline{\alpha}N(n)\right\} \\ t_{\min} & \gtrsim \left[\frac{p}{\underline{\alpha}^2 k}(\overline{\alpha}^2 k + \sigma^2 m)\frac{\lambda^2}{\sigma^2(t)}\exp(-\zeta\frac{\lambda^2}{\sigma^2(t)})\right]^{\frac{1}{2}} \text{for some universal constant } \zeta > 0 \end{cases}$$

*(3)* ***[Regularization]:*** $\sigma(t)\sqrt{\log(p)} \lesssim \lambda \lesssim \underline{\alpha}\min\left\{1; \frac{\underline{\alpha}}{\overline{\alpha}}\frac{k}{p}\right\}$

*(4)* ***[First/second/third-derivative conditions]:***

$$\lambda\max\left\{\lambda; \overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0; \overline{\alpha}N(n)\right\}\left(\lambda^2 m + \lambda\sqrt{m}\sqrt{\overline{\alpha}^2 k + \sigma^2 m} + \overline{\alpha}^2 k + \sigma^2 m\right) \lesssim \underline{\alpha}^4\frac{k^2}{p^2}$$

*(5)* ***[Taylor-expansion perturbation]:***

$$\frac{\lambda^2}{\sigma^2(t)}\exp(-\zeta\frac{\lambda^2}{\sigma^2(t)}) \lesssim \underline{\alpha}^6\frac{k^3}{p^3}\frac{1}{(\overline{\alpha}^2 k+\sigma^2 m)(\lambda^2 m+\lambda\sqrt{m}\sqrt{\overline{\alpha}^2 k+\sigma^2 m}+\overline{\alpha}^2 k+\sigma^2 m)^2}$$

*(6)* ***[Sample complexity]:*** $N(n) \lesssim \min\left\{\frac{\underline{\alpha}^2}{\overline{\alpha}^2}\frac{k}{p}; \frac{\lambda^2}{\sigma^2(t)}\exp(-\frac{\zeta}{2}\frac{\lambda^2}{\sigma^2(t)})\right\}$.

*Consider some $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ and $\|\mathbf{v}\|_2 = 1$ parameterizing $\mathbf{D}(u)$. Under Assumptions (1)-(6), $u \mapsto F_n(\mathbf{D}(u))$ admits a local minimum in $[-t, t]$ with probability exceeding $1 - c_0/n$, for some universal constant $c_0 > 0$.*

We discuss in the next section the main building blocks of this result and give a high-level structure of the proof.

The statement of Theorem 5 presents the precise conditions under which the existence of a local minimum for sparse coding can be proved to happen with high probability. We now derive a corollary that constitutes a simplified version of Theorem 5 and that makes it possible to better understand how the different quantities involved in our problem, such that the coherence $\mu_0$ and the noise $\sigma$, are allowed to vary with $(n, p, k, m)$.

**Corollary 6** (Local minimum of sparse coding, specific scaling)
*Consider $n$ independent signals following the generative model from Section 6.2.2, with the reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$. Denoting by $N(n) \triangleq \log^2(n)/\sqrt{n}$, let us introduce the following set of assumptions for any $\beta > 3$:*

*(1)* ***[Coherence]:*** $\mu_0 \lesssim \frac{\underline{\alpha}^4}{\overline{\alpha}^4} \frac{1}{\sqrt{\log(p)kp}}$

*(2)* ***[Noise level]:*** $\sigma \lesssim \frac{1}{\beta}\frac{\underline{\alpha}}{p}$

*(3)* ***[Regularization]:*** $\lambda = O\Big(\overline{\alpha}\frac{\sqrt{\log(p)}}{p}\Big)$

*(4)* ***[Sparsity level]:*** $\log(p)\,\max\Big\{\frac{m}{p^2}; \frac{\overline{\alpha}^4}{\underline{\alpha}^4}\Big\} \lesssim k$

*(5)* ***[Minimum signal intensity]:*** $\Big(\frac{\log(p)}{p^{\beta-3}}\Big)^{\frac{1}{6}} \lesssim \frac{\underline{\alpha}}{\overline{\alpha}}$

*(6)* ***[Sample complexity]:*** $N(n) \lesssim \min\Big\{\frac{\underline{\alpha}^4}{\overline{\alpha}^4}\frac{k}{p\sqrt{\log(p)}}; \frac{\log(p)}{p^{\beta/2}}\Big\}$.

*Consider some $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ and $\|\mathbf{v}\|_2 = 1$ parameterizing $\mathbf{D}(u)$. Under Assumptions (1)-(6), $u \mapsto F_n(\mathbf{D}(u))$ admits a local minimum in $[-1/p, 1/p]$ with probability exceeding $1 - c_0/n$, for some universal constant $c_0 > 0$.*

Some comments and remarks are in order:

*Assumption (1)*: While we hoped for a coherence scaling in $O(1/k)$, the relationship between the first, second and third derivatives of $\Phi_n$ (see Assumption (4) in Theorem 5) has led us to consider a coherence in $O(1/\sqrt{kp\log(p)})$, which unfortunately brings into play the full size of the dictionary, as opposed to just the sparsity level as in the noiseless setting of Geng et al. (2011). We could improve this weakness of our analysis by refining the upper bound of the third derivative of $\Phi_n$ (which is currently scaling in $O(\overline{\alpha}^2 k)$).

*Assumption (2)*: We observe a quite natural scaling of the noise which should decrease in $O(1/p)$ as the number of atoms increases. The noiseless setting can be easily handled and leads to the same scalings as those displayed in Corollary 6. In fact, it suffices to follow the proof of Corollary 6, noticing that $\sigma(t)$ becomes equal to $\overline{\alpha}|t|$ when $\sigma$ is set to zero. In both the noisy and noiseless settings, the perturbation of the reference dictionary is chosen to be in the order of $|t| = O(1/p)$.

*Assumption (3)*: The choice of the regularization parameter is driven by the constraints imposed in Proposition 10 and the fact that we want the perturbation of the Taylor expansion—which is proportional to $\frac{\lambda^2}{\sigma^2(t)}\exp(-\zeta\frac{\lambda^2}{\sigma^2(t)})$, to remain small.

*Assumption (4)*: Perhaps surprisingly, we need to impose a lower bound on the

sparsity level $k$. We do so in order to guarantee enough curvature for the function $\Phi_n$. In fact, the second derivative of $t \mapsto \Phi_n(\mathbf{D}(t)|\text{sign}(\mathbf{A}_0))$ evaluated at $t = 0$ is proportional to $k/p$, as proved in Lemmas 10 and 11.

*Assumption (5)*: Our study relies on the exact recovery of the supports of $\{\boldsymbol{\alpha}_0^i\}_{i\in[\![1;n]\!]}$. This is only possible when the minimum magnitude of the signals $\underline{\alpha}$ is bounded sufficiently away from zero, as quantify by the inequality $\left(\frac{\log(p)}{p^{\beta-3}}\right)^{1/6} \lesssim \frac{\underline{\alpha}}{\overline{\alpha}}$. Note that this inequality points out that it is likely to extend our result to the case of distributions of $\boldsymbol{\alpha}_0$ with some small enough mass at zero. Interestingly, and more generally, a question of interest is whether we can conduct a similar analysis without having to go through exact recovery, since, in the end, we essentially care about the dictionary and the function $F_n$, regardless of the behavior of the learned coefficients $\boldsymbol{\alpha}$.

*Assumption (6)*: The sample complexity we obtain indicates we need in the order of $O(\log^2(p)p^3) \approx O(k^2p^3)$ signals to be able to prove the existence of a local minimum, compared to $O(kp^3)$ in Geng et al. (2011). We believe that by refining the concentration bounds, it is possible to replace the term $N(n) = \log^2(n)/\sqrt{n}$ by $\log(n)/\sqrt{n}$, which would thus lead to the same scaling as in Geng et al. (2011).

We now look at a more detailed description of the proof of Theorem 5.

### 6.3.2 Architecture of the proof

The proof is based upon three building blocks that we now present. The main idea consists in first focusing on the study of the function $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$, before looking at the behavior of $t \mapsto F_n(\mathbf{D}(t))$ whose properties will stem from the concentration of the average of the independent random variables $\{f_{\mathbf{x}^i}(\mathbf{D}(t))\}_{i\in[\![1;n]\!]}$. We give below some details regarding the three steps we follow:

(1) Proving that the functions $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ and $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$ coincide for suitable values of $t \in \mathcal{T}_{\text{coincide}} \subseteq \mathbb{R}$, and that this event happens with high probability. The consequence of this point is that we can study a smooth function with an explicit representation instead of dealing with the intricate function $f_{\mathbf{x}}$. This step is fully described in Section 6.3.3.

(2) Computing the second-order Taylor expansion of $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ at $t = 0$, along with a uniform upper bound on the third derivative of this function. Again, the terms in this expansion and the upper bound should be viewed as random variables whose value need to be controlled with high probability. Importantly, since $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ and $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$ may differ, we need to account for this event (even though it happens with low probability). This points results in a *perturbed* Taylor development for the function $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$, as analyzed in Lemma 7. Details of this second ingredient are given in Section 6.3.4.

(3) Aggregating the results for $n$ (independent) signals, based on the computation of step (2). For all the random quantities, we proceed in the same way: we first show they are bounded with high probability, and conditionally to these events, we apply Hoeffding's and Bernstein's inequalities for a concentration around the

expectation: We then have the expression of the *perturbed* Taylor development for the function $t \mapsto F_n(\mathbf{D}(t))$, which is valid with high probability.

The final conclusion follows by invoking Lemma 7 that guarantees the existence of a local minimum within a set $\mathcal{T}_{\text{min}} \subseteq \mathbb{R}$. It finally remains to check that $\mathcal{T}_{\text{min}} \subseteq \mathcal{T}_{\text{coincide}}$. This third building block is presented in details in Section 6.3.5.

The presence of noise in our analysis has led us to consider a structure of proof radically different from the schemes proposed in the related work Gribonval and Schnass (2010); Geng et al. (2011). In particular, our formulation in problem (6.3) differs from that of Gribonval and Schnass (2010); Geng et al. (2011) where the $\ell_1$-norm of $\mathbf{A}$ is minimized over the *equality* constraint $\mathbf{AD} = \mathbf{X}$ and the dictionary normalization $\mathbf{D} \in \mathcal{D}$. Optimality is then characterized through the linearization of the equality constraint, a technique that could not be easily extended to the noisy case.

The remainder of this section is dedicated to the description of the three key ingredients we referred to above.

### 6.3.3 Exact sign recovery for perturbed dictionaries

The objective of this section is to determine the conditions under which the two functions $t \mapsto \Phi_n(\mathbf{D}(t)|\text{sign}(\mathbf{A}_0))$ and $t \mapsto F_n(\mathbf{D}(t))$ coincide. As briefly exposed in Section 6.2.1, it turns out that this question comes down to studying exact recovery for some $\ell_1$-regularized least-squares problems. Let us momentarily assume that

$$f_{\mathbf{x}}(\mathbf{D}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

has a solution $\hat{\boldsymbol{\alpha}}$ with support J, and that the matrix $\mathbf{D}_{\text{J}}^\top \mathbf{D}_{\text{J}}$ is invertible. In this case, and according to Definition 6.4 of $\phi_{\mathbf{x}}$, we have the equality

$$f_{\mathbf{x}}(\mathbf{D}) = \phi_{\mathbf{x}}(\mathbf{D}|\text{sign}(\hat{\boldsymbol{\alpha}})).$$

Within our probabilistic model for the signal $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}$, the sign vector $\text{sign}(\hat{\boldsymbol{\alpha}})$ is a random variable that depends on both the coefficients $\boldsymbol{\alpha}_0$ *and* the noise $\boldsymbol{\varepsilon}$, which is obviously difficult to control. It is therefore interesting to study when $\text{sign}(\hat{\boldsymbol{\alpha}}) = \text{sign}(\boldsymbol{\alpha}_0)$ holds.

Exact sign recovery in the problem associated with $f_{\mathbf{x}}(\mathbf{D}_0)$ has already been well-studied (e.g., see Fuchs, 2005; Zhao and Yu, 2006; Wainwright, 2009). In particular, exact recovery is guaranteed to happen with high probability provided that (a) the dictionary $\mathbf{D}_0$ satisfies the so-called *irrepresentability condition*—as assumed in Eq. (6.7), and that (b) the non-zero coefficients of $\boldsymbol{\alpha}_0$ are far enough from the noise level.

However, in our context, we need the same conclusion to hold not only at the dictionary $\mathbf{D}_0$, but also at $\mathbf{D}(t) \neq \mathbf{D}_0$ for some values of $t \in \mathbb{R}$. We make this statement precise in the following proposition:

**Proposition 10** (Exact recovery for perturbed dictionaries)
*Let us consider some* $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ *and some normalized vector* $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_2 = 1$,

*parametrizing* $\mathbf{D}(t)$. *Assume that the signal* $\mathbf{x} \in \mathbb{R}^m$ *is generated according to the model from Section 6.2.2. Let* $\eta \in (0, \eta_0)$. *Let* $\mathcal{T}_{coincide} \subseteq \mathbb{R}$ *be the set*

$$\mathcal{T}_{coincide} \triangleq \Big\{ t \in \mathbb{R};\ |t| \leq \min\Big\{1, \frac{1}{3}\Big\lfloor \frac{1-\eta}{2-\eta} - \frac{1-\eta_0}{2-\eta_0} \Big\rceil \frac{1}{k} \Big\} \Big\}.$$

*Define for any* $t \in \mathcal{T}_{coincide}$ *the regularization parameter*

$$\lambda \geq 3 \frac{\max\{\sigma, \overline{\alpha}|t|\}}{\eta} \sqrt{6 \log(p)} \quad \textit{and} \quad \text{threshold}(\lambda) \triangleq \frac{(2\sqrt{2}+1)\lambda}{1 - k\mu(t)}.$$

*Moreover, consider the noise condition*

$$\text{threshold}(\lambda) < \underline{\alpha}.$$

*There exists a universal constant* $\zeta > 0$ *such that for any* $t \in \mathcal{T}_{coincide}$ *with probability exceeding*

$$1 - 8 \exp\Big( - \zeta \frac{\lambda^2}{\max\{\sigma^2, \overline{\alpha}^2 t^2\}} \Big),$$

*the vector* $\hat{\boldsymbol{\alpha}}(t) \in \mathbb{R}^p$ *defined by*

$$\hat{\boldsymbol{\alpha}}(t) = \begin{pmatrix} \big[ [\mathbf{D}(t)]_{\mathrm{J}}^\top [\mathbf{D}(t)]_{\mathrm{J}} \big]^{-1} \big[ [\mathbf{D}(t)]_{\mathrm{J}}^\top \mathbf{x} - \lambda \text{sign}([\boldsymbol{\alpha}_0]_{\mathrm{J}}) \big] \\ \mathbf{0} \end{pmatrix},$$

*is the unique solution of* $\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} [\frac{1}{2}\|\mathbf{x} - \mathbf{D}(t)\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1]$, *and exact sign recovery holds, namely,* $\text{sign}(\hat{\boldsymbol{\alpha}}(t)) = \text{sign}(\boldsymbol{\alpha}_0)$.

The structure of the proof closely follows that of Wainwright (2009), except that additional terms measuring the perturbation between $\mathbf{D}_0$ and $\mathbf{D}(t)$ have to be controlled. Interestingly, this perturbation measured by the parameter $|t|$ acts as a second source of noise, where the variance $\sigma^2$ is comparable to $\overline{\alpha}^2 t^2$. In addition, the result from Proposition 10 indicates that the perturbation should remain small—that is, $\mathbf{D}(t)$ close to $\mathbf{D}_0$ with $|t|$ in the order of $O(1/k)$—if we want to guarantee exact recovery at the dictionary $\mathbf{D}(t)$. In the same vein, some precaution needs to be taken to prevent the non-zero coefficients of $\boldsymbol{\alpha}_0$ from being too small.

In the light of Definition 6.4, the previous proposition shows that for suitable values of $t$, the functions $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ and $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$ coincide with probability greater than $1 - 8\exp(-\zeta \frac{\lambda^2}{\max\{\sigma^2, \overline{\alpha}^2 t^2\}})$; we therefore refer to this event as $\mathcal{E}_{\text{coincide}}$. In other words, with probability $\Pr(\mathcal{E}_{\text{coincide}})$, we can equivalently study the smooth function $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ in lieu of $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$, which constitutes one of the building block of our approach.

Importantly, the event $\mathcal{E}_{\text{coincide}}$ is only concerned with a single signal; when we consider a collection of $n$ independent signals, we should instead study the event $\bigcap_{i=1}^n \mathcal{E}_{\text{coincide}}^i$ to guarantee that $\Phi_n$ and $F_n$ do coincide. However, as the number of observations $n$ becomes large, it is unrealistic and not possible to ensure that exact recovery will hold *simultaneously*—and with high probability—for the $n$ signals. To get around this issue,

we will exploit the fact that the quantities we need to control are averages over $n$ random variables.

It is now natural to consider the local behavior of $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ through its Taylor expansion around $t = 0$.

### 6.3.4 Computation of the Taylor expansion of $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$

We want to characterize some local minima of $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$ via the function $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$. The assumptions we make through the coherence $\mu_0$ and the perturbation parameter $t$ guarantee that for any support $\mathrm{J} \subseteq [\![1; k]\!]$, the matrix $[\mathbf{D}(t)]_{\mathrm{J}}^{\top}[\mathbf{D}(t)]_{\mathrm{J}}$ is invertible; under these conditions, and according to Definition 6.4, $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ is a smooth function for which we can compute a Taylor expansion. As precisely described in Lemma 15 and Section 6.8, we then obtain an expression of the form

$$\left|\phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0)) - \left(\phi_{\mathbf{x}}(\mathbf{D}_0|\text{sign}(\boldsymbol{\alpha}_0)) + a_{\mathbf{x}}t + b_{\mathbf{x}}t^2\right)\right| \leq L_{\mathbf{x}}|t|^3, \tag{6.9}$$

which is valid for any $t$ that preserves the smoothness of $\phi_{\mathbf{x}}$. The quantities $a_{\mathbf{x}}$, $b_{\mathbf{x}}$ and $L_{\mathbf{x}}$ are random variables that are combinations of linear, quadratic and bilinear forms with respect to $\boldsymbol{\alpha}_0$ and/or $\boldsymbol{\varepsilon}$. These random variables can be shown to concentrate around their expectation, as further discussed in Propositions 11, 12 and 13.

On the event $\mathcal{E}_{\text{coincide}}$ (see Section 6.3.3), the functions $t \mapsto f_{\mathbf{x}}(\mathbf{D}(t))$ and $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\text{sign}(\boldsymbol{\alpha}_0))$ coincide, so that the development in Eq. (6.9) directly applies for $f_{\mathbf{x}}$. When the previous property is not true anymore, i.e., on the event $\mathcal{E}_{\text{coincide}}^c$, we use the fact that the difference between $f_{\mathbf{x}}$ and $\phi_{\mathbf{x}}$ can always be upper bounded. More specifically, we can notice from the definitions of both $f_{\mathbf{x}}$ and $\phi_{\mathbf{x}}$ that for any $\mathbf{D} \in \mathbb{R}^{m\times p}$ and any $\mathbf{s} \in \{-1, 0, 1\}^p$ such that $\phi_{\mathbf{x}}$ is properly defined, we have

$$f_{\mathbf{x}}(\mathbf{D}) \leq \frac{1}{2}\|\mathbf{x}\|_2^2 \quad \text{and} \quad \phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) \leq \frac{1}{2}\|\mathbf{x}\|_2^2.$$

We thus end up with the following *perturbed* Taylor expansion

$$\left|f_{\mathbf{x}}(\mathbf{D}(t)) - \left(f_{\mathbf{x}}(\mathbf{D}_0) + a_{\mathbf{x}}t + b_{\mathbf{x}}t^2\right)\right| \leq L_{\mathbf{x}}|t|^3 + r_{\mathbf{x}}, \text{ with } r_{\mathbf{x}} = 2\mathbb{1}_{\mathcal{E}_{\text{coincide}}^c}\|\mathbf{x}\|_2^2. \tag{6.10}$$

In order to prove the existence of a local minimum around $\mathbf{D}_0$, we will subsequently have to probabilistically control a upper bound of the gradient $a_{\mathbf{x}}$, a lower bound of the curvature $b_{\mathbf{x}}$, along with upper bounds on the third derivative $L_{\mathbf{x}}$ and the perturbation $r_{\mathbf{x}}$.

### 6.3.5 Aggregation of $n$ signals and existence of a local minimum

Starting from the expansion in Eq. (6.10) for $n$ independent signals $\{\mathbf{x}^i\}_{i\in[\![1;n]\!]}$, the final step of the proof consists in concentrating the random variables $\frac{1}{n}\sum_{i=1}^n a_{\mathbf{x}^i}$, $\frac{1}{n}\sum_{i=1}^n b_{\mathbf{x}^i}$, $\frac{1}{n}\sum_{i=1}^n L_{\mathbf{x}^i}$ and $\frac{1}{n}\sum_{i=1}^n r_{\mathbf{x}^i}$. To this end, we make use of Hoeffding's and Bernstein's inequalities conditioned to an appropriate event where the random variables at stake are bounded. The discussion of the existence of a local minimum then follows from the next lemma:

**Lemma 7** (Local minimum from perturbed Taylor expansion)
*Let $f$ be a continuous real-valued function. Assume there exist $(a, b, L, r)$ such that $|a|, b, L$ and $r$ are strictly positive, with*

$$\theta \triangleq \frac{|a|L}{b^2} \in (0, 1).$$

*Let us define*

$$t_0 \triangleq \max\left\{\frac{|a|}{b}, \frac{1}{\sqrt{1-\theta}}\sqrt{\frac{r}{b}}\right\},$$

*and assume that for all $t \in [-t_0, t_0]$,*

$$|f(t) - (f(0) + at + 2bt^2)| \leq L|t|^3 + r.$$

*If the following condition holds*

$$r < \frac{b^3}{L^2}[1-\theta],$$

*then $f$ admits a local minimum in $\mathcal{T}_{\min} \triangleq (-t_0, t_0)$.*

After enforcing the conditions required by Lemma 7, it remains to check a last inequality. Indeed, while Proposition 10 says we cannot go too far from $\mathbf{D}_0$ to guarantee exact recovery, we need at the same time, and according to Lemma 7, to go far enough to ensure the existence of a local minimum. This results in testing the inclusion $\mathcal{T}_{\min} \subseteq \mathcal{T}_{\text{coincide}}$, or equivalently, the inequality (with the notation from Lemma 7):

$$\max\left\{\frac{|a|}{b}, \frac{1}{\sqrt{1-\theta}}\sqrt{\frac{r}{b}}\right\} \leq |t| \leq \min\left\{1, \frac{1}{3}\left[\frac{1-\eta}{2-\eta} - \frac{1-\eta_0}{2-\eta_0}\right]\frac{1}{k}\right\}.$$

## 6.4 Some experimental validations

In this section, we try to illustrate the results from Section 6.3.1 through some simulations. Although we do not manage to exactly highlight the scalings found in Corollary 6, our experiments still underline the main interesting trends put forward by our results.

In the setting we consider, we take $p = 20, m = 10$ and the reference dictionary $\mathbf{D}_0$ has its entries generated i.i.d. according to a standard Gaussian distribution. Moreover, the sparsity level $k$ is fixed to 4, while the non-zero coefficients of $\boldsymbol{\alpha}_0$ are uniformly drawn over the segment $[-\overline{\alpha}, -\underline{\alpha}] \cup [\underline{\alpha}, \overline{\alpha}]$, with $\overline{\alpha} = 1$ and $\underline{\alpha} = 0.1$. The level of noise is set according to Corollary 6, i.e., $\sigma = \overline{\alpha}/p$, whereas the number of signals is equal to $n = 2500$.

We study the behavior of the solutions of problem (6.3), by looking at the variations with respect to one parameter while keeping the other ones fixed (to the values mentioned above). This protocol leads to the series of results displayed in Figure 6.2, indexed by the letters (a) to (e).

We detail two important aspects of the experiment, namely, the choice of the regularization parameter $\lambda$, and how we deal with the invariances of problem (6.3) (see

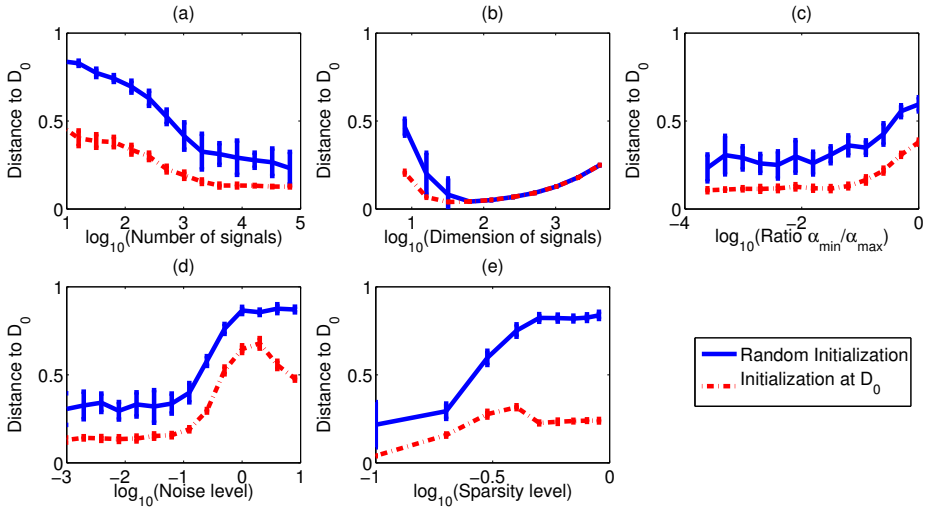



Figure 6.2: Estimation error, i.e., the normalized Frobenius distance between $\mathbf{D}_0$ and the solution of problem (6.3), versus some varying parameters. The curves represent the average error and standard deviation based on 10 runs, for random and fixed initializations. Details about the setting can be found in the text.

details in the end of Section 6.2.2). On the one hand, since our analysis relies on exact recovery and we know in advance the correct sparsity level (i.e., the "oracle" value of $k$), we first tune $\lambda$ over a logarithmic grid of values so that the learned $\hat{\boldsymbol{\alpha}}$'s match this sparsity level. Note that this tuning step is performed over an auxiliary set of signals.

On the other hand, we know that the dictionary $\hat{\mathbf{D}}$ that we learn by minimizing problem (6.3) may differ from $\mathbf{D}_0$ up to sign flips and atom permutations. To account for these possible transformations, we solve, as a post-processing step, the following problem

$$\min_{\substack{\mathbf{s}\in\{-1,1\}^p \\ \boldsymbol{\Sigma}\in\mathcal{P}}} \left\|\hat{\mathbf{D}} - \mathbf{D}_0\boldsymbol{\Sigma}\mathrm{Diag}(\mathbf{s})\right\|_{\mathrm{F}}^2,$$

where $\boldsymbol{\Sigma}$ belongs to the set of permutation matrices $\mathcal{P}$. Let us denote by $\boldsymbol{\Sigma}(j)$ the result of the permutation of the index $j$ by $\boldsymbol{\Sigma}$. Since both $\hat{\mathbf{D}}$ and $\mathbf{D}_0$ have normalized atoms, the previous problem is equivalent to

$$\max_{\substack{\mathbf{s}\in\{-1,1\}^p \\ \boldsymbol{\Sigma}\in\mathcal{P}}} \sum_{j=1}^{p} \mathbf{s}_j[\hat{\mathbf{d}}^j]^\top \mathbf{d}_0^{\boldsymbol{\Sigma}(j)} = \max_{\boldsymbol{\Sigma}\in\mathcal{P}} \sum_{j=1}^{p} \left|[\hat{\mathbf{d}}^j]^\top \mathbf{d}_0^{\boldsymbol{\Sigma}(j)}\right|.$$

We recognize here an assignment problem based on the absolute correlation matrix $\hat{\mathbf{D}}^\top\mathbf{D}_0$, which can be efficiently solved using the Hungarian algorithm (Kuhn, 1955).

To solve problem (6.3), we use the stochastic algorithm from Mairal et al. (2010a) [1] where the batch size is fixed to 512, while the number of epochs is chosen so as to pass

1. The code is available at `http://www.di.ens.fr/willow/SPAMS/`.

over each signal 25 times (on average). We consider two types of initialization, one starting from a random dictionary, and the other one using the correct dictionary $\mathbf{D}_0$ as a starting point.

Let us comment on Figure 6.2. The measure we consider in the experiments is the normalized Frobenius distance between $\mathbf{D}_0$ and the (transformed) solution of problem (6.3). The results from Figure 6.2 show the error averaged over 10 runs, corresponding to 10 different initial dictionaries in the random initialization case. [2] Overall, it is interesting to see that the curves based on the random and fixed initialization roughly have the same behavior. Plot (a) shows that the number of signals is indeed beneficial to obtain a local minimum around $\mathbf{D}_0$. The scaling in $O(p^3)$ advocated by Corollary 6 (in this case, about 8000 signals) seems reasonable, though slightly pessimisitc, in that the error does not decrease much and stabilizes from about 4000 signals.

Plot (b) shows that, when the coherence is small enough ($\mu_0$ decreases as $m$ increases with $p$ fixed, and hence the problem becomes easier), the initialization does not seem to matter anymore. However, we still observe an increasing error since the number of signals remains fixed to 2500.

As for plot (c), it is interesting to see that the quality of the solution does not seem to be affected as $\underline{\alpha}$ decreases. This contradicts Corollary 6 and confirms the fact that we should probably not go through exact recovery to prove our results. Moreover, notice that the case $\underline{\alpha} = \overline{\alpha} = 1$ corresponding to $\boldsymbol{\alpha}_0$ distributed on the hypercube appears as more complex, which our theory is not able to predict so far.

We now turn to plot (d). This simulation confirms that the noise is an important factor; we observe a sharp transition around values close to $1/k$, which may indicate that our scaling in $O(1/p)$ is too pessimistic.

Finally, the figure in plot (e) seems to point out that the sparser the signals, the better the learned dictionary. It may be interesting to better understand in future work why dense signals harm the identifiability of the dictionary. One potential explanation would be that as $k$ increases, we need more signals, while the simulation kept $n$ fixed.

## 6.5 Conclusion

We have conducted a non-asymptotic analysis of the local minima of sparse coding in the presence of noise, thus extending prior work which focused on noiseless settings (Gribonval and Schnass, 2010; Geng et al., 2011). Within a probabilistic model of sparse signals, we have shown that a local minimum exists with high probability along some curves passing through the reference dictionary. The natural next step to undertake is to prove the existence of a local minimum in a neighborhood of the reference dictionary, i.e., within the entire manifold of normalized dictionaries, as opposed to just along single curves. Arguments based on (1) concentration of the measure uniformly over all such curves, and (2) the discretization of this set of curves through $\epsilon$-nets (Massart, 2003) shall be useful to this end.

2. Note that due to the stochastic optimization tool used here, even the curves for the fixed initialization exhibit some small variations.

Our study can be further developed in multiple ways. On the one hand, while we have assumed *deterministic* coherence-based conditions scaling in $O(1/k)$, it may interesting to consider non-deterministic assumptions (Tropp, 2008; Candès and Plan, 2009), which are likely to lead to improved scalings.

On the other hand, we may also use more realistic generative models for $\boldsymbol{\alpha}_0$, for instance, spike and slab models (Ishwaran and Rao, 2005), or signals with compressible priors (Cevher, 2008). Similarly, it may be interesting to look at misspecified models, e.g., with a certain fraction of non-sparse signals.

Moreover, a more ambitious question is about the possible characterization of a global minimum for sparse coding (modulo the inevitable invariances of the problem).

Finally, it is also of interest to study the case of sparse structured dictionary learning (e.g., see Jenatton et al., 2011c, and references therein), where the $\ell_1$-norm would be replaced by other sparsity-inducing norms capable of modelling classes of structured signals.

## 6.6 Proofs of the main results

This section contains the detailed proofs of the main results of the chapter.

### 6.6.1 Proof of Theorem 5

The proof is based on the different building blocks presented in Section 6.3.2. Note that the assumption of the theorem on $k\mu_0$ makes it possible to upper bound the terms depending only on $k\mu_0$ by universal constants.

To begin with, we consider the event $\mathcal{E}_{\text{taylor}}$ over which we control the first-, second-, third-order and perturbation terms displayed in Eq. (6.10) for each of the $n$ signals. These terms are respectively controlled in Propositions 11, 12, 13 and 14. Using the union bound and setting $\tau = 2\log(n)$, the probability of the event of interest is lower bounded by

$$\Pr(\mathcal{E}_{\text{taylor}}) \geq 1 - \sum_{y \in \left\{c^{(0)}_{a_{\mathbf{x}}}, c^{(0)}_{b_{\mathbf{x}}}, c^{(0)}_{L_{\mathbf{x}}}, 2\right\}} 1 - \left[1 - \frac{y}{n^2}\right]^{n+1} \geq 1 - \frac{c_0}{n},$$

for some universal constant $c_0 > 0$.

On the event $\mathcal{E}_{\text{taylor}}$, and according to Propositions 11, 13 and Lemmas 8, 12, we have that

$$\Big|\frac{1}{n}\sum_{i=1}^{n} a_{\mathbf{x}^i}\Big| \lesssim \lambda \max\Big\{\lambda; \overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0; \overline{\alpha}N(n)\Big\},$$

and

$$\Big|\frac{1}{n}\sum_{i=1}^{n} L_{\mathbf{x}^i}\Big| \lesssim \max\{1, N(n)\}(\lambda^2 m + \lambda\sqrt{m}\sqrt{\overline{\alpha}^2 k + \sigma^2 m} + \overline{\alpha}^2 k + \sigma^2 m).$$

Still on the event $\mathcal{E}_{\text{taylor}}$, and since the theorem assumes that $N(n) \lesssim \frac{\underline{\alpha}^2}{\overline{\alpha}^2}\frac{k}{p}$ and $\lambda \lesssim \underline{\alpha}\frac{\underline{\alpha}}{\overline{\alpha}}\frac{k}{p}$, we also have $\lambda^2 \lesssim \underline{\alpha}\lambda \lesssim \underline{\alpha}^2\frac{\underline{\alpha}}{\overline{\alpha}}\frac{k}{p} \lesssim \underline{\alpha}^2\frac{k}{p}$, so that we obtain the following lower bound for

the curvature term (see Proposition 12 and Lemma 10):

$$\frac{1}{n}\sum_{i=1}^{n} b_{\mathbf{x}^i} \gtrsim \underline{\alpha}^2 \frac{k}{p}.$$

It now remains to handle the perturbation term. To this end, we apply Proposition 14 by considering the random variables $\{\mathbb{1}_{[\mathcal{E}^i_{\text{coincide}}]^c}\|\mathbf{x}^i\|_2^2\}_{i\in[\![1;n]\!]}$. Combining the results from both Proposition 10 and Lemma 14, we have the following upper bounds of the expectation and the variance

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{[\mathcal{E}^i_{\text{coincide}}]^c}\|\mathbf{x}^i\|_2^2\Big] &\lesssim (\overline{\alpha}^2 k + \sigma^2 m)\frac{\lambda^2}{\sigma^2(t)}\exp\Big(-\zeta\frac{\lambda^2}{\sigma^2(t)}\Big)\\
\mathbb{E}\Big[\mathbb{1}_{[\mathcal{E}^i_{\text{coincide}}]^c}\|\mathbf{x}^i\|_2^4\Big] &\lesssim (\overline{\alpha}^2 k + \sigma^2 m)^2\frac{\lambda^4}{\sigma^4(t)}\exp\Big(-\zeta\frac{\lambda^2}{\sigma^2(t)}\Big),
\end{aligned}$$

where the universal constant $\zeta > 0$ comes from Proposition 14. In addition, since we assume that the sample complexity $N(n)$ satisfies $N(n) \lesssim \frac{\lambda^2}{\sigma^2(t)}\exp(-\frac{\zeta}{2}\frac{\lambda^2}{\sigma^2(t)})$, we can apply Proposition 14 by setting $\tau = \log(n)$ and using for $\varsigma^2$ the upper bound above, which leads to

$$\frac{1}{n}\sum_{i=1}^{n} r_{\mathbf{x}^i} \lesssim (\overline{\alpha}^2 k + \sigma^2 m)\frac{\lambda^2}{\sigma^2(t)}\exp\Big(-\zeta\frac{\lambda^2}{\sigma^2(t)}\Big).$$

With the previous elements in place, we can apply Lemma 7 which explicits the relationships that the first-, second-, third-order and perturbation terms must satisfy to guarantee the existence of a local minimum. Writing directly these relationships leads to the inequalities numbered (4) and (5) in the theorem. Note that the term $\theta$ in Lemma 7 can be guaranteed to be no greater than $1/2$ by setting the multiplicative constant of $\lambda$ small enough, so that terms in $1/(1-\theta)$ can be seen as universal constants.

The remaining conditions, that is, inequalities (2) and (3), result, on the one hand, from the constraints imposed on $\lambda$ and $t$ by Proposition 10, and, on the other hand, from the test $\mathcal{T}_{\text{min}} \subseteq \mathcal{T}_{\text{coincide}}$, as explained in Section 6.3.5.

### 6.6.2 Proof of Corollary 6

The proof of Corollary 6 mostly consists in checking that the proposed choices of scalings satisfy the conditions required by Theorem 5.

We first discuss the choice of $\lambda = O(\overline{\alpha}\sqrt{\log(p)}/p)$. Let $\beta$ be a scalar greater than 3. At the cost of having a more stringent constraint on the noise $\sigma(t)$, we can always have the equality $\beta\log(p) = \zeta\frac{\lambda^2}{\sigma^2(t)}$, which leads in turn to the advertised scaling of $\sigma(t)$ and the choice $|t| = O(1/p)$. Note that we will have to check that the conditions imposed by $t_{\text{min}}$ in Theorem 5 are valid with $|t| = O(1/p)$. Moreover, since we constrain the sparsity level by $\log(p)\frac{\overline{\alpha}^4}{\underline{\alpha}^4} \lesssim k$, we indeed satisfy condition (3) from Theorem 5.

We now turn to the coherence condition (1) from Theorem 5: We need to check the inequality

$$\frac{\underline{\alpha}^4}{\overline{\alpha}^4}\frac{1}{\sqrt{\log(p)kp}} \lesssim \frac{p-2k}{3kp}.$$

To this end, it suffices to notice that $\frac{\alpha}{\overline{\alpha}} \leq 1$ and $k < \sqrt{kp} < \sqrt{\log(p)kp}$, along with the fact that for some universal constant $z \in (0,1)$, we have $z\frac{1}{k} \leq \frac{p-2k}{3kp}$.

In the light of the previous points, and since we ask for $\log(p)\frac{m}{p^2} \lesssim k$, we obtain the following simplifications:

$$\begin{aligned}
(\overline{\alpha}^2 k + \sigma^2 m) &\lesssim \overline{\alpha}^2 k \\
(\lambda^2 m + \lambda\sqrt{m}\sqrt{\overline{\alpha}^2 k + \sigma^2 m} + \overline{\alpha}^2 k + \sigma^2 m) &\lesssim \overline{\alpha}^2 k \\
\frac{\lambda^2}{\sigma^2(t)} \exp\Big(-\zeta\frac{\lambda^2}{\sigma^2(t)}\Big) &\lesssim \frac{\log(p)}{p^\beta} \\
\lambda &\lesssim \overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0.
\end{aligned}$$

In addition, the constraint of the corollary on the sample complexity $N(n)$ implies that

$$N(n) \lesssim \frac{\underline{\alpha}^4}{\overline{\alpha}^4}\frac{k}{p\sqrt{\log(p)}} = O\Big(\sqrt{\frac{k}{p}}k\mu_0\Big) \lesssim \frac{\underline{\alpha}^2}{\overline{\alpha}^2}\frac{k}{p},$$

as required by condition (6) in Theorem 5. Similarly, the inequality $N(n) \lesssim \frac{\lambda^2}{\sigma^2(t)}\exp(-\frac{\zeta}{2}\frac{\lambda^2}{\sigma^2(t)})$ from condition (6) in Theorem 5 exactly corresponds to condition (6) in the corollary where $N(n) \lesssim \log(p)/p^{\beta/2}$.

At this stage of the proof, it remains to check conditions (2), (4) and (5) from Theorem 5:

**Condition (2):** On the one hand, we have

$$\frac{\lambda p}{\underline{\alpha}^2 k}\max\Big\{\lambda; \overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0; \overline{\alpha}N(n)\Big\} \lesssim \overline{\alpha}\frac{\sqrt{\log(p)}}{p}\frac{p}{\underline{\alpha}^2 k}\overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0 \lesssim \frac{\underline{\alpha}^2}{\overline{\alpha}^2}\frac{1}{p}$$

which is indeed less than $|t| = 1/p$. On the other hand, the second constraint of $t_{\min}$ reads

$$\Big[\frac{p}{\underline{\alpha}^2 k}(\overline{\alpha}^2 k + \sigma^2 m)\frac{\lambda^2}{\sigma^2(t)}\exp(-\zeta\frac{\lambda^2}{\sigma^2(t)})\Big]^{\frac{1}{2}} \lesssim \Big[\frac{p}{\underline{\alpha}^2 k}\overline{\alpha}^2 k\frac{\log(p)}{p^\beta}\Big]^{\frac{1}{2}} \lesssim \frac{\overline{\alpha}}{\underline{\alpha}}\frac{\sqrt{\log(p)}}{p^{(\beta-1)/2}}.$$

Having $\frac{\overline{\alpha}}{\underline{\alpha}}\frac{\sqrt{\log(p)}}{p^{(\beta-1)/2}} \lesssim |t| = \frac{1}{p}$ is then equivalent to $\Big(\frac{\log(p)}{p^{\beta-3}}\Big)^{1/2} \lesssim \frac{\alpha}{\overline{\alpha}}$, which is true thanks to condition (5) in the corollary.

**Condition (4):** We have

$$\begin{aligned}
\lambda\max\Big\{\lambda; \overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0; \overline{\alpha}N(n)\Big\}\Big(\lambda^2 m + \lambda\sqrt{m}\sqrt{\overline{\alpha}^2 k + \sigma^2 m} + \overline{\alpha}^2 k + \sigma^2 m\Big) \\
\lesssim \lambda\overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0\overline{\alpha}^2 k \lesssim \underline{\alpha}^4\frac{k^2}{p^2},
\end{aligned}$$

as required by Theorem 5.

**Condition (5):** We first have

$$\underline{\alpha}^6 \frac{k^3}{p^3} \frac{1}{(\overline{\alpha}^2 k + \sigma^2 m)(\lambda^2 m + \lambda\sqrt{m}\sqrt{\overline{\alpha}^2 k + \sigma^2 m} + \overline{\alpha}^2 k + \sigma^2 m)^2} \gtrsim \frac{\underline{\alpha}^6}{\overline{\alpha}^6} \frac{1}{p^3},$$

so that it suffices to check that

$$\frac{\lambda^2}{\sigma^2(t)} \exp\Big(-\zeta \frac{\lambda^2}{\sigma^2(t)}\Big) = \frac{\log(p)}{p^\beta} \lesssim \frac{\underline{\alpha}^6}{\overline{\alpha}^6} \frac{1}{p^3},$$

which is in turn guaranteed thanks to condition (5) in the corollary.
We have examined all the conditions required by Theorem 5 for which the scalings proposed in the corollary apply; this concludes the proof.

### 6.6.3 Proof of Proposition 10

First and foremost, we look at the assumption made on the parameter $t$. For any $t \in \mathcal{T}_{\text{coincide}}$, we have

$$|t| \leq \frac{1}{3}\Big[\frac{1-\eta}{2-\eta} - \frac{1-\eta_0}{2-\eta_0}\Big]\frac{1}{k} \leq \frac{1}{3}\Big[\frac{1-\eta}{2-\eta}\frac{1}{k} - \mu_0\Big],$$

which, by Eq. (6.6) and (6.7), leads to

$$\mu(t) \leq \frac{1-\eta}{2-\eta}\frac{1}{k} < \frac{1}{k}.$$

This inequality shows that the matrix $[\mathbf{D}(t)]_{\mathrm{J}}^\top [\mathbf{D}(t)]_{\mathrm{J}}$ is invertible for any $\mathrm{J} \subseteq [\![1;p]\!]$ with $|\mathrm{J}| \leq k$, and

$$\big\|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top [\mathbf{D}(t)]_{\mathrm{J}} \big[[\mathbf{D}(t)]_{\mathrm{J}}^\top [\mathbf{D}(t)]_{\mathrm{J}}\big]^{-1}\big\|_\infty \leq \frac{k\mu(t)}{1-k\mu(t)} \leq 1-\eta.$$

The core of the proof relies on the control of four probabilistic events:

$$\begin{aligned}
\mathcal{E}^1 &= \{\boldsymbol{\varepsilon} \in \mathbb{R}^m;\ \|[\mathbf{D}(t)]_{\mathrm{J}}^\top \boldsymbol{\varepsilon}\|_\infty \leq \sqrt{2}\lambda\},\\
\mathcal{E}^2 &= \{\boldsymbol{\varepsilon} \in \mathbb{R}^m;\ \|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top (\mathbf{I} - [\mathbf{P}(t)]_{\mathrm{J}})\boldsymbol{\varepsilon}\|_\infty \leq \lambda\frac{\eta}{3}\},\\
\mathcal{E}^3 &= \{\boldsymbol{\alpha}_0 \in \mathbb{R}^p;\ \|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top (\mathbf{I} - [\mathbf{P}(t)]_{\mathrm{J}})[\mathbf{D}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty \leq \lambda\frac{\eta}{3}\},\\
\mathcal{E}^4 &= \{\boldsymbol{\alpha}_0 \in \mathbb{R}^p;\ \|[\mathbf{D}(t)]_{\mathrm{J}}^\top ([\mathbf{D}(t)]_{\mathrm{J}} - [\mathbf{D}_0]_{\mathrm{J}})[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty \leq \sqrt{2}\lambda\}.
\end{aligned}$$

First, notice that given our noise assumption, Lemma 18 shows that

$$\|[\hat{\boldsymbol{\alpha}}(t) - \boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty \leq \mathrm{threshold}(\lambda) < \underline{\alpha} \leq \min_{j\in\mathrm{J}} |[\boldsymbol{\alpha}_0]_j|,$$

which implies in turn that $\mathrm{sign}(\hat{\boldsymbol{\alpha}}(t)) = \mathrm{sign}(\boldsymbol{\alpha}_0)$. It remains to prove that $\hat{\boldsymbol{\alpha}}(t)$ is the unique solution of the Lasso program. To this end, we take advantage of Lemma 19.

On the event $\bigcap_{i=1}^4 \mathcal{E}^i$, we have

$$\|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top(\mathbf{I}-[\mathbf{P}(t)]_{\mathrm{J}})\mathbf{x}\|_\infty + \lambda\|\!|[\mathbf{D}(t)_{\mathrm{J}^c}]^\top[\mathbf{D}(t)]_{\mathrm{J}}[[\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{D}(t)]_{\mathrm{J}}]^{-1}\|\!|_\infty$$

which is first upper bounded by

$$\|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top(\mathbf{I}-[\mathbf{P}(t)]_{\mathrm{J}})\boldsymbol{\varepsilon}\|_\infty + \|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top(\mathbf{I}-[\mathbf{P}(t)]_{\mathrm{J}})[\mathbf{D}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty + \lambda(1-\eta),$$

and then by

$$\lambda\frac{\eta}{3} + \lambda\frac{\eta}{3} + \lambda(1-\eta) = \lambda(1-\frac{\eta}{3}) < \lambda.$$

Putting together the pieces with $\mathrm{sign}(\hat{\boldsymbol{\alpha}}(t)) = \mathrm{sign}(\boldsymbol{\alpha}_0)$, Lemma 19 leads to the desired conclusion.

The remainder of the proof is dedicated to the control of the four events $\mathcal{E}^i$, $i$ in $[\![1;4]\!]$. Note that we always proceed in the same way: conditionally on the draw of J, we compute uniform bounds, i.e., that depend only on $k$, leading to the conclusion without conditioning.

**Event $\mathcal{E}^1$**: Consider the centered Gaussian variable $\mathbf{u}_j = \boldsymbol{\varepsilon}^\top[\mathbf{d}(t)]^j$ for $j$ in J. Since $\mathbf{D}(t)$ has unit $\ell_2$-norm columns, the variance of $\mathbf{u}_j$ is $\sigma^2$. Applying Lemma 6 along with the union bound, we have

$$\Pr(\mathcal{E}^1) \geq 1 - 2k\exp\Big(-\frac{\lambda^2}{\sigma^2}\Big).$$

**Event $\mathcal{E}^2$**: We proceed similarly and consider $\mathbf{u}_j = \boldsymbol{\varepsilon}^\top(\mathbf{I}-[\mathbf{P}(t)]_{\mathrm{J}})[\mathbf{d}(t)]^j$ for $j$ in $\mathrm{J}^c$. Since $\mathbf{D}(t)$ is normalized and the spectral norm of a projector is bounded by one, the variance of $\mathbf{u}_j$ is upper bounded by $\sigma^2$. We obtain,

$$\Pr(\mathcal{E}^2) \geq 1 - 2(p-k)\exp\Big(-\frac{\eta^2}{18}\frac{\lambda^2}{\sigma^2}\Big).$$

**Events $\mathcal{E}^3$ and $\mathcal{E}^4$**: For the treatment of these events, we simply invoke Lemma 17.

Finally, by independence of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\varepsilon}$ combined with the union bound, we have

$$\Pr(\cap_{i=1}^4\mathcal{E}^i) = \Pr(\mathcal{E}^1\cap\mathcal{E}^2)\Pr(\mathcal{E}^3\cap\mathcal{E}^4) \geq \Big[1-\Pr([\mathcal{E}^1]^c)-\Pr([\mathcal{E}^2]^c)\Big]\Big[1-\Pr([\mathcal{E}^3]^c)-\Pr([\mathcal{E}^4]^c)\Big].$$

Putting the different pieces together, the probability of exact recovery is lower bounded by

$$\begin{aligned}\Pr(\cap_{i=1}^4\mathcal{E}^i) &\geq 1 - 8\max\{p-k,k\}\exp\Big(-\min\Big\{1,\frac{\eta^2}{18},\frac{1}{2},\frac{\eta^2}{54}\Big\}\frac{\lambda^2}{\max\{\sigma^2,\overline{\alpha}^2t^2\}}\Big)\\ &= 1 - 8\exp\Big(-\frac{\eta^2}{54}\frac{\lambda^2}{\max\{\sigma^2,\overline{\alpha}^2t^2\}} + \log\big(\max\{p-k,k\}\big)\Big),\end{aligned}$$

so that there exists a universal constant $\zeta > 0$ such that

$$\Pr(\cap_{i=1}^4\mathcal{E}^i) \geq 1 - 8\exp\Big(-\zeta\frac{\lambda^2}{\max\{\sigma^2,\overline{\alpha}^2t^2\}}\Big).$$

### 6.6.4 Proof of Lemma 7

The function $f$ is lower and upper bounded by the functions $\phi_l$ and $\phi_u$ defined as

$$\phi_l(t) = f(0) + at + 2bt^2 - L|t|^3 - r \text{ and } \phi_u(t) = f(0) + at + 2bt^2 + L|t|^3 + r.$$

Let us first assume that $a < 0$. We recall that $\theta = |a|L/b^2$ and it is assumed to belong to $(0, 1)$. Observe that

$$\phi_l(|a|/b) - f(0) = -r + a\frac{|a|}{b} + 2b\frac{|a|^2}{b^2} - L\frac{|a|^3}{b^3} = -r + L\frac{|a|^3}{b^3}\Big[\frac{b^2}{|a|L} - 1\Big] = -r + \frac{|a|^2}{b}(1-\theta).$$

As a result, if the condition $r < \frac{|a|^2}{b}(1-\theta)$ is satisfied, we have $\phi_l(|a|/b) > f(0)$. Let us now consider that $r \geq \frac{|a|^2}{b}(1-\theta)$ holds, or equivalently,

$$t' \triangleq \frac{1}{\sqrt{1-\theta}}\sqrt{\frac{r}{b}} > \frac{|a|}{b}. \tag{6.11}$$

We study the sign of $\phi_l(t') - f(0)$, that is,

$$\begin{aligned}\phi_l(t') - f(0) = -r + at' + 2b[t']^2 - L[t']^3 &= \frac{r}{\nu(1-\theta)}\Big[-\theta\sqrt{1-\theta} + (1+\theta)\nu - \frac{1}{\sqrt{1-\theta}}\nu^2\Big] \\ &= \frac{r}{\nu(1-\theta)}\psi_\theta(\nu),\end{aligned}$$

where we have introduced the quantity $\nu \triangleq L\sqrt{r}/b^{3/2}$. The second-order polynomial function $\psi_\theta$ admits two distinct positive roots

$$u_1 = \theta\sqrt{1-\theta} \quad \text{and} \quad u_2 = \sqrt{1-\theta},$$

and for any $u \in (u_1, u_2)$, $\psi_\theta(u)$ is strictly positive. Moreover, using Eq. (6.11), we already know that $\nu \geq u_1$. It is then sufficient to notice that $\nu \leq u_2$ is equivalent to imposing the condition

$$r \leq \frac{b^3}{L^2}(1-\theta).$$

To summarize, we have proved so far that, under the conditions of the lemma, it holds that $\phi_l(t_0) > f(0)$. Now, notice that $\phi_l(-t_0) - \phi_l(t_0) = -2a\,t_0$, so that for any $a < 0$, we end up with the following chains of inequalities

$$f(0) < \phi_l(t_0) \leq f(t_0) \quad \text{and} \quad f(0) < \phi_l(t_0) < \phi_l(-t_0) \leq f(-t_0).$$

On the compact set $[-t_0; t_0]$, since $f$ is continuous, it has a global minimum. Moreover, given that $f(0) < \min\{f(-t_0), f(t_0)\}$ and $0 \in [-t_0; t_0]$, this minimum lies in the interior of $\mathcal{T}_{\min} = (-t_0; t_0)$.

Eventually, the case $a > 0$ can be dealt with in a similar fashion.

## 6.7 Control of the Taylor expansion

This section of the appendix is dedicated to the intermediate results necessary to probabilistically control the Taylor expansion of the function $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\mathrm{sign}(\boldsymbol{\alpha}_0))$. We begin with the control of the first order term. Note that the concrete computation of the Taylor expansion is postponed to Section 6.8.

### 6.7.1 Concentration of the first-order terms

**Lemma 8** (Upper bound of the first-order term $a_{\mathbf{x}}$)
*There exist a universal constant $c^{(0)}_{a_{\mathbf{x}}}$ and coefficients $\{c^{(j)}_{a_{\mathbf{x}}}\}_{j\in[\![1;3]\!]}$ depending only on $k\mu_0$ such that for any $\tau \geq 2$,*

$$\Pr\Big(|a_{\mathbf{x}}| \leq c^{(1)}_{a_{\mathbf{x}}} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}\}\tau\Big) \geq 1 - c^{(0)}_{a_{\mathbf{x}}} \exp(-\tau),$$

*and*

$$|\mathbb{E}[a_{\mathbf{x}}]| \leq c^{(2)}_{a_{\mathbf{x}}} \lambda\overline{\alpha}\sqrt{\frac{k}{p}}k\mu_0 + \lambda^2 c^{(3)}_{a_{\mathbf{x}}}.$$

*Proof.* The first part of the proof simply consists in gathering the concentration results of all the lemmas from Section 6.7.1, and applying the union bound. The second conclusion follows by putting together the non-vanishing expectation terms appearing in the lemmas from Section 6.7.1, and using the upper bound from Lemma 9. □

**Lemma 9** (Upper bound of the leading expectation term)
*For any $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ and for any vector $\mathbf{v}$ with unit $\ell_2$-norm, we have*

$$\Big|\mathbb{E}\Big[\mathrm{Tr}\big([[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}]^{-1}[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\big)\Big]\Big| \leq \sqrt{\frac{k}{p}}\frac{k\mu_0}{1-k\mu_0}\sqrt{1+k\mu_0}.$$

*Proof.* By definition of $\mathbf{W}$, we know that $\mathrm{diag}(\mathbf{W}^\top\mathbf{D}_0) = \mathbf{0}$, and since $\mathrm{Diag}(\mathbf{v})$ is diagonal, it holds that $\mathrm{diag}(\mathbf{D}_0^\top\mathbf{W}\mathrm{Diag}(\mathbf{v})) = \mathbf{0}$. As a consequence, we can freely subtract the identity matrix in the trace, and we have

$$\begin{aligned}
\mathrm{Tr}\big([[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}]^{-1}[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\big) &= \mathrm{Tr}\big((\mathbf{I} - [[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}]^{-1})[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\big) \\
&\leq \|\mathbf{I} - [[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}]^{-1}\|_{\mathrm{F}}\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}} \\
&\leq \frac{k\mu_0}{1-k\mu_0}\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}} \\
&\leq \frac{k\mu_0}{1-k\mu_0}\|[\mathbf{D}_0]_{\mathrm{J}}[\mathbf{D}_0]_{\mathrm{J}}^\top\|_2^{1/2}\|[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}} \\
&\leq \frac{k\mu_0}{1-k\mu_0}\sqrt{1+k\mu_0}\|[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}},
\end{aligned}$$

where we have repeatedly used Lemma 16. Now, we have by definition of the sampling procedure and the normalization of both $\mathbf{v}$ and $\mathbf{W}$,

$$\mathbb{E}[\|[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2] = \mathbb{E}\Big[\sum_{j=1}^{p} \delta(j)\mathbf{v}_j^2 \|\mathbf{w}^j\|_2^2\Big] = \frac{k}{p}.$$

The conclusion follows by Jensen's inequality. □

**Proposition 11** (Concentration of $\frac{1}{n}\sum_{i=1}^n a_{\mathbf{x}^i}$)
*There exist a universal constant $c_{a_{\mathbf{x}}}^{(0)}$ and a coefficient $c_{a_{\mathbf{x}}}^{(1)}$ depending only on $k\mu_0$ such that for any $\tau \geq 2$, we have*

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^n a_{\mathbf{x}^i} - \mathbb{E}[a_{\mathbf{x}}]\Big| \leq c_{a_{\mathbf{x}}}^{(1)} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}\}\frac{\tau^2}{\sqrt{n}}\Big) \geq \big[1 - c_{a_{\mathbf{x}}}^{(0)} \exp(-\tau)\big]^{n+1}.$$

*Proof.* The proof consists in conditioning for each independent signal $\mathbf{x}^i$ to the event defined in Lemma 8 where the first order terms we want to concentrate are bounded. For any $\tau_2 \geq 2$, and with probability exceeding $\big[1 - c_{a_{\mathbf{x}}}^{(0)} \exp(-\tau_2)\big]^n$, we thus have for all $i \in [\![1; n]\!]$

$$|a_{\mathbf{x}^i}| \leq c_{a_{\mathbf{x}}}^{(1)} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}\}\tau_2.$$

We then consider the collection of independent, bounded variables $\{a_{\mathbf{x}^i}\}_{i\in[\![1;n]\!]}$, and the conclusion follows by applying Hoeffding's inequality (see Lemma 23), that is,

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^n a_{\mathbf{x}^i} - \mathbb{E}[a_{\mathbf{x}}]\Big| \leq c_{a_{\mathbf{x}}}^{(1)} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}\}\frac{\tau_1\tau_2}{\sqrt{n}}\Big) \geq \big[1 - 2\exp(-\tau_1^2)\big]\big[1 - c_{a_{\mathbf{x}}}^{(0)} \exp(-\tau_2)\big]^n,$$

for any $\tau_1 \geq 0$ and $\tau_2 \geq 2$. We can further simplifying the right-hand side since for any $\tau \geq 1$,

$$\big[1 - 2\exp(-\tau^2)\big]\big[1 - \exp(-\tau)\big]^n \geq \big[1 - 2\exp(-\tau)\big]^{n+1}.$$

□

### Details of the concentration of the first-order terms

This section contains all the intermediate results necessary to concentrate the first-order term of the Taylor expansion. In particular, we will have to make use of the concentration inequalities from Lemmas 22, 25 and 26. In any case, we need to control bilinear or quadratic forms that bring into play the Frobenius norm of a matrix. Upper bounds of this norm are obtained by exploiting the following facts

- According to Lemma 16, we have $\|[\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}]^{-1}\|_{\mathrm{F}} \leq 1/(1 - k\mu_0)$ and $\|\mathbf{D}_{\mathrm{J}}\mathbf{D}_{\mathrm{J}}^\top\|_2 \leq 1 + k\mu$.
- Projectors have their spectral norm bounded by one.
- The dictionary $\mathbf{D}_0$ and the matrix $\mathbf{W}$ have unit $\ell_2$-norm columns, so that in combination with the fact that the vector $\mathbf{v}$ is also normalized, we have $\|\mathbf{D}_0\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}} = 1$ and $\|\mathbf{W}\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}} = 1$.

The names of the next Lemma refer to the notation introduced in Lemma 15.

**Lemma** (Concentration of $A_{\boldsymbol{\varepsilon\alpha}}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $A_{\boldsymbol{\varepsilon\alpha}}$ is upper bounded by $\sigma\overline{\alpha}\tau$, with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $A_{\mathbf{s\alpha}}$)
*For any $\tau \geq 2$, we have*

$$\left|A_{\mathbf{s}\boldsymbol{\alpha}} - \lambda\mathbb{E}[|[\boldsymbol{\alpha}_0]_1|]\,\mathrm{Tr}\big([[\mathbf{D}_0]_J^\top[\mathbf{D}_0]_J]^{-1}[\mathbf{D}_0]_J^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_J\big)\right| \leq 64\lambda\overline{\alpha}k\mu_0\frac{\sqrt{1+k\mu_0}}{1-k\mu_0}\tau$$

*with probability exceeding $1 - 4\exp(-\tau)$.*

**Lemma** (Concentration of $A_{\boldsymbol{\varepsilon}\mathbf{s},1}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $A_{\boldsymbol{\varepsilon}\mathbf{s},1}$ is upper bounded by*

$$\lambda\sigma\frac{1}{1-k\mu_0}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $A_{\boldsymbol{\varepsilon}\mathbf{s},2}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $A_{\boldsymbol{\varepsilon}\mathbf{s},2}$ is upper bounded by*

$$\lambda\sigma\frac{\sqrt{1+k\mu_0}}{(1-k\mu_0)^{3/2}}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $A_{\mathbf{ss}}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $A_{\mathbf{ss}}$ is upper bounded by*

$$2\lambda^2\frac{\sqrt{1+k\mu_0}}{(1-k\mu_0)^2}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $A_{\boldsymbol{\varepsilon\varepsilon}}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $A_{\boldsymbol{\varepsilon\varepsilon}}$ is upper bounded by*

$$\sigma^2\frac{1}{\sqrt{1-k\mu_0}}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

### 6.7.2 Concentration of the second-order terms

**Lemma 10** (Upper bound of the second-order term $b_{\mathbf{x}}$)
*There exist a universal constant $c_{b_{\mathbf{x}}}^{(0)}$ and coefficients $\{c_{b_{\mathbf{x}}}^{(j)}\}_{j\in[\![1;3]\!]}$ depending only on $k\mu_0$*

*such that for any* $\tau \geq 1$,

$$\Pr\Big(|b_{\mathbf{x}}| \leq c_{b_{\mathbf{x}}}^{(1)} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}, \overline{\alpha}^2\}\tau\Big) \geq 1 - c_{b_{\mathbf{x}}}^{(0)} \exp(-\tau),$$

*and*

$$\underline{\alpha}^2 \frac{k}{p}\Big[1 - \frac{k}{p}\frac{1+p\mu_0}{1-k\mu_0}\Big] - \lambda\overline{\alpha}\, c_{b_{\mathbf{x}}}^{(2)} - \lambda^2\, c_{b_{\mathbf{x}}}^{(3)} \leq \mathbb{E}[b_{\mathbf{x}}] \leq \overline{\alpha}^2 \frac{k}{p} + \lambda\,\overline{\alpha}\, c_{b_{\mathbf{x}}}^{(2)} + \lambda^2\, c_{b_{\mathbf{x}}}^{(3)}.$$

*Proof.* The first part of the proof simply consists in gathering the concentration results of all the lemmas from Section 6.7.2, and applying the union bound. The second conclusion follows by putting together the non-vanishing expectation terms appearing in the lemmas from Section 6.7.2, and using the lower/upper bounds from Lemma 11. □

**Lemma 11** (Upper bound of the leading expectation term)
*Let* $\mathbf{P}_0$ *be the orthogonal projector that projects onto the span of* $[\mathbf{D}_0]_{\mathrm{J}}$. *For any* $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ *and for any vector* $\mathbf{v}$ *with unit* $\ell_2$*-norm, we have*

$$\frac{k}{p}\Big[1 - \frac{k}{p}\frac{1+p\mu_0}{1-k\mu_0}\Big] \leq \mathbb{E}\Big[\mathrm{Tr}\big([\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}^\top(\mathbf{I}-\mathbf{P}_0)[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\big)\Big] \leq \frac{k}{p}.$$

*Proof.* First, by definition of the sampling procedure and the normalization of both $\mathbf{v}$ and $\mathbf{W}$, we have

$$\mathbb{E}[\|[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2] = \mathbb{E}\Big[\sum_{j=1}^p \delta(j)\mathbf{v}_j^2\|\mathbf{w}^j\|_2^2\Big] = \frac{k}{p}.$$

As a result, the right-hand side is easily proved by upper bounding the spectral norm of $\mathbf{I}-\mathbf{P}_0$ by one. We now turn to the lower bound. We start with the following inequality

$$\begin{aligned}\mathrm{Tr}([\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}^\top\mathbf{P}_0[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}) &\leq |\!|\!|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}]^{-1}|\!|\!|_2\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2 \\ &\leq \frac{1}{1-k\mu_0}\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2,\end{aligned}$$

where we have used Lemma 16. Developing the Frobenius norm with the property $\mathrm{Diag}(\mathbf{W}^\top\mathbf{D}_0) = \mathbf{0}$, we obtain

$$\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2 = \sum_{j=1}^p\sum_{\substack{i=1\\i\neq j}}^p \delta(j)\delta(i)\mathbf{v}_j^2([\mathbf{w}^j]^\top\mathbf{d}_0^i)^2,$$

so that

$$\mathbb{E}\Big[\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2\Big] = \frac{k(k-1)}{p(p-1)}\|\mathbf{D}_0^\top\mathbf{W}\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}}^2.$$

Since $\frac{k(k-1)}{p(p-1)} \leq \frac{k^2}{p^2}$ and $\|\mathbf{W}\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}}^2 = 1$, the expectation above is upper bounded by

$$\mathbb{E}\Big[\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}\|_{\mathrm{F}}^2\Big] \leq \frac{k^2}{p^2}|\!|\!|\mathbf{D}_0\mathbf{D}_0^\top|\!|\!|_2.$$

Applying Lemma 16, we have $|\!|\!|\mathbf{D}_0\mathbf{D}_0^\top|\!|\!|_2 \leq 1+p\mu_0$, and the advertised conclusion follows. □

**Proposition 12** (Concentration of $\frac{1}{n}\sum_{i=1}^{n} b_{\mathbf{x}^i}$)
*There exist a universal constant* $c_{b_{\mathbf{x}}}^{(0)}$ *and a coefficient* $c_{b_{\mathbf{x}}}^{(1)}$ *depending only on* $k\mu_0$ *such that for any* $\tau \geq 2$*, we have*

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} b_{\mathbf{x}^i} - \mathbb{E}[b_{\mathbf{x}}]\Big| \leq c_{b_{\mathbf{x}}}^{(1)} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}, \overline{\alpha}^2\}\frac{\tau^2}{\sqrt{n}}\Big) \geq \big[1 - c_{b_{\mathbf{x}}}^{(0)} \exp(-\tau)\big]^{n+1}.$$

*Proof.* The proof essentially follows that of Proposition 11. It uses Lemma 10 in place of Lemma 8, which leads to

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} b_{\mathbf{x}^i} - \mathbb{E}[b_{\mathbf{x}}]\Big| \leq c_{b_{\mathbf{x}}}^{(1)} \max\{\lambda^2, \lambda\overline{\alpha}, \lambda\sigma, \sigma^2, \sigma\overline{\alpha}, \overline{\alpha}^2\}\frac{\tau_1\tau_2}{\sqrt{n}}\Big) \geq \big[1 - 2\exp(-\tau_1^2)\big]\big[1 - c_{b_{\mathbf{x}}}^{(0)} \exp(-\tau_2)\big]^n,$$

for any $\tau_1 \geq 0$ and $\tau_2 \geq 2$. We apply simplifications similar to those found in the proof of Proposition 11. □

### Details of the concentration of the second-order terms

This section contains all the intermediate results necessary to concentrate the second-order term of the Taylor expansion. In particular, we will have to make use of the concentration inequalities from Lemmas 22, 25 and 26. In any case, we need to control bilinear or quadratic forms that bring into play the Frobenius norm of a matrix. Upper bounds of this norm are obtained by exploiting the following facts

- According to Lemma 16, we have $\|[\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}]^{-1}\|_{\mathrm{F}} \leq 1/(1 - k\mu_0)$ and $\|\mathbf{D}_{\mathrm{J}}\mathbf{D}_{\mathrm{J}}^\top\|_2 \leq 1 + k\mu$.
- Projectors have their spectral norm bounded by one.
- The dictionary $\mathbf{D}_0$ and the matrix $\mathbf{W}$ have unit $\ell_2$-norm columns, so that in combination with the fact that the vector $\mathbf{v}$ is also normalized, we have $\|\mathbf{D}_0\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}} = 1$ and $\|\mathbf{W}\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}} = 1$.

The names of the next Lemma refer to the notation introduced in Lemma 15.

**Lemma** (Concentration of $B_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$)
*Let* $\mathbf{P}_0$ *be the orthogonal projector that projects onto the span of* $[\mathbf{D}_0]_{\mathrm{J}}$*. There exist universal constants* $c_0, c_1 > 0$ *such that for any* $\tau \geq 1$*, we have*

$$\Big|B_{\boldsymbol{\alpha}\boldsymbol{\alpha}} - \mathbb{E}[([\boldsymbol{\alpha}_0]_1)^2]\,\mathrm{Tr}([\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}}^\top(\mathbf{I} - \mathbf{P}_0)[\mathbf{W}\mathrm{Diag}(\mathbf{v})]_{\mathrm{J}})\Big| \leq \overline{\alpha}^2\tau$$

*with probability exceeding* $1 - c_0\exp(-c_1\tau)$.

**Lemma** (Concentration of $B_{\mathbf{s}\boldsymbol{\alpha},1}$)
*For any* $\tau \geq 2$*, the term* $B_{\mathbf{s}\boldsymbol{\alpha},1}$ *is upper bounded by*

$$33\lambda\overline{\alpha}\tau$$

*with probability exceeding* $1 - 4\exp(-\tau)$.

**Lemma** (Concentration of $B_{\mathbf{s}\boldsymbol{\alpha},2}$)
*For any $\tau \geq 2$, the term $B_{\mathbf{s}\boldsymbol{\alpha},2}$ is upper bounded by*

$$65\lambda\overline{\alpha}\frac{1}{1-k\mu_0}\tau$$

*with probability exceeding $1 - 4\exp(-\tau)$.*

**Lemma** (Concentration of $B_{\mathbf{s}\boldsymbol{\alpha},3}$)
*For any $\tau \geq 2$, the term $B_{\mathbf{s}\boldsymbol{\alpha},3}$ is upper bounded by*

$$65\lambda\overline{\alpha}\frac{\sqrt{1+k\mu_0}}{(1-k\mu_0)^2}\tau$$

*with probability exceeding $1 - 4\exp(-\tau)$.*

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon}\boldsymbol{\alpha},1}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $B_{\boldsymbol{\varepsilon}\boldsymbol{\alpha},1}$ is upper bounded by*

$$\sigma\overline{\alpha}\frac{1}{\sqrt{1-k\mu_0}}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon}\boldsymbol{\alpha},2}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $B_{\boldsymbol{\varepsilon}\boldsymbol{\alpha},2}$ is upper bounded by*

$$\sigma\overline{\alpha}\frac{\sqrt{1+k\mu_0}}{1-k\mu_0}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon}\mathbf{s},1}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $B_{\boldsymbol{\varepsilon}\mathbf{s},1}$ is upper bounded by*

$$\frac{1}{2}\lambda\sigma\frac{1}{\sqrt{1-k\mu_0}}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon}\mathbf{s},2}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $B_{\boldsymbol{\varepsilon}\mathbf{s},2}$ is upper bounded by*

$$2\lambda\sigma\frac{\sqrt{1+k\mu_0}}{(1-k\mu_0)^2}\tau$$

*with probability exceeding $1 - c_0\exp(-c_1\tau)$.*

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon}\mathbf{s},3}$)
*There exist universal constants $c_0, c_1 > 0$ such that for any $\tau \geq 1$, the term $B_{\boldsymbol{\varepsilon}\mathbf{s},3}$ is upper bounded by*

$$\lambda\sigma\frac{1}{(1-k\mu_0)^{3/2}}\tau$$

*with probability exceeding* $1 - c_0 \exp(-c_1 \tau)$.

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon s},4}$)
*There exist universal constants* $c_0, c_1 > 0$ *such that for any* $\tau \geq 1$, *the term* $B_{\boldsymbol{\varepsilon s},4}$ *is upper bounded by*

$$\lambda \sigma \frac{1}{(1 - k\mu_0)^{5/2}} \tau$$

*with probability exceeding* $1 - c_0 \exp(-c_1 \tau)$.

**Lemma** (Concentration of $B_{\mathbf{ss}}$)
*There exist universal constants* $c_0, c_1 > 0$ *such that for any* $\tau \geq 1$, *the term* $B_{\mathbf{ss}}$ *is upper bounded by*

$$12\lambda^2 \frac{\sqrt{1 + k\mu_0}}{(1 - k\mu_0)^3} \tau$$

*with probability exceeding* $1 - c_0 \exp(-c1\tau)$.

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon\varepsilon},1}$)
*There exist universal constants* $c_0, c_1 > 0$ *such that for any* $\tau > 1$, *the term* $B_{\boldsymbol{\varepsilon\varepsilon},1}$ *is upper bounded by*

$$\sigma^2 \frac{1}{1 - k\mu_0} \tau$$

*with probability exceeding* $1 - c_0 \exp(-c_1 \tau)$.

**Lemma** (Concentration of $B_{\boldsymbol{\varepsilon\varepsilon},2}$)
*There exist universal constants* $c_0, c_1 > 0$ *such that for any* $\tau \geq 1$, *the term* $B_{\boldsymbol{\varepsilon\varepsilon},2}$ *is upper bounded by*

$$2\sigma^2 \frac{1 + k\mu_0}{(1 - k\mu_0)^{3/2}} \tau$$

*with probability exceeding* $1 - c_0 \exp(-c_1 \tau)$.

### 6.7.3 Control of the third derivative

**Proposition 13** (Upper bound on the third derivative of $\Phi_n$)
*There exist a constant* $c_{L_{\mathbf{x}}} > 0$ *depending only on* $k\mu_0$ *such that for any sign matrix* $\mathbf{S} \in \{-1, 0, 1\}^{p \times n}$ *and for all* $|t| \leq 1$ *such that* $t \mapsto \Phi_n(\mathbf{D}(t)|\mathbf{S})$ *is three times differentiable, we have*

$$\Big| \frac{1}{n} \sum_{i=1}^{n} \nabla^3 \big[\phi_{\mathbf{x}^i}(\mathbf{D}(t)|\mathbf{s}^i)\big] - \big|\mathbb{E}\big[\nabla^3 \big[\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})\big]\big]\big| \Big| \leq$$
$$c_{L_{\mathbf{x}}} \big[\lambda^2 m + \lambda \sqrt{m} (\overline{\alpha}^2 k + \sigma^2 m)^{1/2} \sqrt{\tau} + (\overline{\alpha}^2 k + \sigma^2 m)\tau\big] \tfrac{\tau}{\sqrt{n}},$$

*with probability exceeding* $[1 - 2\exp(-\tau)]^{n+1}$, *for any* $\tau \geq 1$.

*Proof.* The proof essentially follows that of Proposition 11. It uses Lemma 12 in place of Lemma 8, which leads to

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\nabla^3[\phi_{\mathbf{x}^i}(\mathbf{D}(t)|\mathbf{s}^i)] - |\mathbb{E}[\nabla^3[\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})]]\Big| \leq$$
$$c_{L_{\mathbf{x}}}[\lambda^2 m + \lambda\sqrt{m}(\overline{\alpha}^2 k + \sigma^2 m)^{1/2}\sqrt{\tau_2} + (\overline{\alpha}^2 k + \sigma^2 m)\tau_2]\tfrac{\tau_1}{\sqrt{n}},$$

with probability exceeding $[1-2\exp(\tau_1^2)][1-\exp(-\tau_2)]^n$, for any $\tau_1 \geq 0$ and any $\tau_2 \geq 1$. We apply simplifications similar to those found in the proof of Proposition 11. □

**Lemma 12** (Upper bound on the third derivative of $\phi_{\mathbf{x}}$)
*There exist a constant $c_{L_{\mathbf{x}}} > 0$ depending only on $k\mu_0$ such that for any sign vector $\mathbf{s} \in \{-1,0,1\}^p$ and for all $|t| \leq 1$ with $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})$ three times differentiable, we have*

$$|\nabla^3[\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})]| \leq c_{L_{\mathbf{x}}}[\lambda^2 m + \lambda\sqrt{m}(\overline{\alpha}^2 k + \sigma^2 m)^{1/2}\sqrt{\tau} + (\overline{\alpha}^2 k + \sigma^2 m)\tau].$$

*with probability exceeding $1-\exp(-\tau)$, for any $\tau \geq 1$. Moreover, we have*

$$|\mathbb{E}[\nabla^3[\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})]]| \leq c_{L_{\mathbf{x}}}[\lambda^2 m + \lambda\sqrt{m}(\overline{\alpha}^2 k + \sigma^2 m)^{1/2} + (\overline{\alpha}^2 k + \sigma^2 m)].$$

*Proof.* The core of the proof consists of upper bounding the derivative of the three parts $T_{\mathbf{xx}}$, $T_{\mathbf{sx}}$ and $T_{\mathbf{ss}}$ composing $\phi_{\mathbf{x}}$ (see Section 6.8 for their complete expressions). We proceed by using Cauchy Schwartz's inequality, upper bounding the $\ell_2$-norm of sign vectors by $\sqrt{m}$, and applying the bounds from Section 6.8, which leads to

$$|\nabla^3[\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})]| \leq c_{L_{\mathbf{x}}}\Big(\lambda^2 m + \lambda\sqrt{m}\|\mathbf{x}\|_2 + \|\mathbf{x}\|_2^2\Big),$$

for some positive constant $c_{L_{\mathbf{x}}}$ depending only on $k\mu_0$. Now, combined with Lemma 13, we obtain the first desired conclusion. Finally, since

$$\mathbb{E}[\|\mathbf{x}\|_2^2] \leq (\overline{\alpha}^2 k + \sigma^2 m),$$

we obtain by Jensen's inequality $\mathbb{E}[\|\mathbf{x}\|_2] \leq (\overline{\alpha}^2 k + \sigma^2 m)^{1/2}$, and the second conclusion follows. □

### 6.7.4 Control of the residual term

**Proposition 14**
*Consider $n$ independent draws $\{(\boldsymbol{\alpha}_0^i, \boldsymbol{\varepsilon}^i)\}_{i\in[\![1;n]\!]}$ following the generative model from Section 6.2.2. Consider also $n$ independent events $\{\mathcal{E}^i\}_{i\in[\![1;n]\!]}$ defined on the same probability space as that of $(\boldsymbol{\alpha}_0, \varepsilon)$. Let us define $n$ non-negative scalars $\{\varsigma_i\}_{i\in[\![1;n]\!]}$ such that for any $i \in [\![1;n]\!]$, $\varsigma_i^2$ is greater than, or equal to, the variance of $\mathbb{1}_{\mathcal{E}^i}\|\mathbf{x}^i\|_2^2$. Moreover, introduce $\varsigma^2 \triangleq \frac{1}{n}\sum_{i=1}^n \varsigma_i^2$. For any $\tau \geq 1$, if we have*

$$\tau^2 \leq \frac{3}{20}\frac{\varsigma}{(\overline{\alpha}^2 k + \sigma^2 m)}\sqrt{n},$$

*then it holds that*

$$\Pr\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathcal{E}^i}\|\mathbf{x}^i\|_2^2 - \mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^2] \leq 2\varsigma\frac{\tau}{\sqrt{n}}\Big) \geq \big[1-2\exp(-\tau)\big]^{n+1}.$$

*Proof.* The proof consists in conditioning to the event (see Lemma 13)

$$\bigcap_{i=1}^{n}\Big\{(\boldsymbol{\alpha}_0^i,\boldsymbol{\varepsilon}^i)\in\mathbb{R}^p\times\mathbb{R}^m;\ \|\mathbf{x}^i\|_2^2\leq 5(\overline{\alpha}^2k+\sigma^2m)\tau\Big\},$$

and then considering the collection of independent, zero-mean, bounded variables $\{\mathbb{1}_{\mathcal{E}^i}\|\mathbf{x}^i\|_2^2 - \mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^2]\}_{i\in[\![1;n]\!]}$. The conclusion follows by applying Bernstein's inequality (see Lemma 24), that is, for any $\tau_1\geq 0$ and for any $\tau\geq 1$, if it holds that $\tau_1\leq \frac{3}{2}\frac{\varsigma}{10(\overline{\alpha}^2k+\sigma^2m)\tau}\sqrt{n}$, we then have

$$\Pr\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\mathcal{E}^i}\|\mathbf{x}^i\|_2^2 - \mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^2] \leq 2\varsigma\frac{\tau_1}{\sqrt{n}}\Big) \geq \big[1-2\exp(-\tau_1^2)\big]\big[1-\exp(-\tau)\big]^n.$$

Setting $\tau_1=\tau$, and further simplifying the right-hand side since for any $\tau\geq 1$,

$$\big[1-2\exp(-\tau^2)\big]\big[1-\exp(-\tau)\big]^n \geq \big[1-2\exp(-\tau)\big]^{n+1},$$

we obtain the desired conclusion. □

**Lemma 13** (Control of the $\ell_2$-norm of a signal)
*Let* $\mathbf{x}$ *be a signal following the generative model from Section 6.2.2. For any* $\tau\geq 1$*, we have*

$$\Pr(\|\mathbf{x}\|_2^2\leq 5(\overline{\alpha}^2k+\sigma^2m)\tau)\geq 1-\exp(-\tau).$$

*Proof.* The result is a direct application of Lemma 20. We recall that we have $\mathbf{x} = [\mathbf{D}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}+\boldsymbol{\varepsilon}$, and that the norm of $\mathbf{x}$ can be expressed as follows

$$\|\mathbf{x}\|_2 = \Big\|\big[\,\overline{\alpha}[\mathbf{D}_0]_{\mathrm{J}}\ \sigma\mathbf{I}\,\big]\begin{pmatrix}\frac{1}{\overline{\alpha}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}\\ \frac{1}{\sigma}\boldsymbol{\varepsilon}\end{pmatrix}\Big\|_2.$$

□

**Lemma 14** (Truncated expectation of the $\ell_2$-norm of a signal)
*Let* $\mathbf{x}$ *be a signal following the generative model from Section 6.2.2. Consider an event* $\mathcal{E}$ *defined on the same probability space as that of* $(\boldsymbol{\alpha}_0,\boldsymbol{\varepsilon})$*. For any* $\tau\geq 2$*, we have*

$$\mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^2]\leq 5(\overline{\alpha}^2k+\sigma^2m)\tau\Big[\Pr(\mathcal{E})+9\exp(-\tau)\Big].$$

*Moreover, it holds that*

$$\mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^4]\leq 25(\overline{\alpha}^2k+\sigma^2m)^2\tau^2\Big[\Pr(\mathcal{E})+33\exp(-\tau)\Big].$$

*Proof.* Let fix some $\tau \geq 2$. We introduce the event

$$\mathcal{K} \triangleq \Big\{(\boldsymbol{\alpha}_0, \boldsymbol{\varepsilon}) \in \mathbb{R}^p \times \mathbb{R}^m;\ \|\mathbf{x}\|_2^2 \leq 5(\overline{\alpha}^2 k + \sigma^2 m)\tau\Big\},$$

and define $l_\tau$ as the largest integer such that $\tau \in [l_\tau, l_\tau + 1)$. We can then "discretize" the event $\mathcal{K}^c$ as

$$\mathcal{K}^c \subseteq \bigcup_{l=l_\tau}^{\infty} \mathcal{K}_l^c, \quad \text{with } \mathcal{K}_l^c = \Big\{(\boldsymbol{\alpha}_0, \boldsymbol{\varepsilon}) \in \mathbb{R}^p \times \mathbb{R}^m;\ \frac{\|\mathbf{x}\|_2^2}{5(\overline{\alpha}^2 k + \sigma^2 m)} \in [l, l+1)\Big\}.$$

We now have

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^2] &= \mathbb{E}[\mathbb{1}_{\mathcal{E}\cap\mathcal{K}}\|\mathbf{x}\|_2^2] + \mathbb{E}[\mathbb{1}_{\mathcal{E}\cap\mathcal{K}^c}\|\mathbf{x}\|_2^2] \\
&\leq 5(\overline{\alpha}^2 k + \sigma^2 m)\Pr(\mathcal{E})\tau + \sum_{l=l_\tau}^{\infty} \mathbb{E}[\mathbb{1}_{\mathcal{E}\cap\mathcal{K}_l^c}\|\mathbf{x}\|_2^2] \\
&\leq 5(\overline{\alpha}^2 k + \sigma^2 m)\Big[\Pr(\mathcal{E})\tau + \sum_{l=l_\tau}^{\infty}(l+1)\mathbb{E}[\mathbb{1}_{\mathcal{E}\cap\mathcal{K}_l^c}]\Big] \\
&\leq 5(\overline{\alpha}^2 k + \sigma^2 m)\Big[\Pr(\mathcal{E})\tau + \sum_{l=l_\tau}^{\infty}(l+1)\mathbb{E}[\mathbb{1}_{\{(\boldsymbol{\alpha}_0,\boldsymbol{\varepsilon})\in\mathbb{R}^p\times\mathbb{R}^m;\ \|\mathbf{x}\|_2^2 \geq 5(\overline{\alpha}^2 k+\sigma^2 m)l\}}]\Big].
\end{aligned}$$

Applying Lemma 13, we continue with the upper bound

$$\mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^2] \leq 5(\overline{\alpha}^2 k + \sigma^2 m)\Big[\Pr(\mathcal{E})\tau + \sum_{l=l_\tau}^{\infty}(l+1)\exp(-l)\Big].$$

We recognize here the derivative of the Taylor series of $u \mapsto u^{l_\tau+1}/(1-u)$ for $u \in (0,1)$, which leads to

$$\sum_{l=l_\tau}^{\infty}(l+1)u^l = \frac{l_\tau - l_\tau u + 1}{(1-u)^2}u^{l_\tau}.$$

When setting $u = 1/e$, we obtain with $\tau \in [l_\tau, l_\tau + 1)$

$$\sum_{l=l_\tau}^{\infty}(l+1)\exp(-l) \leq 2\frac{l_\tau}{1-1/e}\exp(-l_\tau) \leq 2\frac{e\tau}{1-1/e}\exp(-\tau) \leq 9\tau\exp(-\tau).$$

We thus reach the first advertised conclusion. The second result follows along the same lines, except that we end up with

$$\mathbb{E}[\mathbb{1}_{\mathcal{E}}\|\mathbf{x}\|_2^4] \leq 25(\overline{\alpha}^2 k + \sigma^2 m)^2\Big[\Pr(\mathcal{E})\tau^2 + \sum_{l=l_\tau}^{\infty}(l+1)^2\exp(-l)\Big].$$

As above, and noticing that $(1+l)^2 = (1+l)(2+l) - (1+l)$, we can recognize the difference of the second and first derivatives of the Taylor series of $u \mapsto u^{l_\tau+1}/(1-u)$ for $u \in (0,1)$, which gives

$$\sum_{l=l_\tau}^{\infty}(l+1)^2 u^l = \frac{l_\tau^2(1-u)^2 + 2l_\tau(1-u) + u + 1}{(1-u)^3}u^{l_\tau}.$$

Again, when setting $u = 1/e$, and with similar upper bounds as before, we obtain

$$\sum_{l=l_\tau}^{\infty} (l+1)^2 \exp(-l) \leq 33\tau^2 \exp(-\tau),$$

which concludes the proof.

□

## 6.8 Computation of the Taylor expansion

For any sign vector $\mathbf{s}$ with support J, the function $\mathbf{D} \mapsto \phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s})$ as defined in (6.4) is smooth as long as the matrix $\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}$ is invertible. In this section, we derive a second-order Taylor expansion of the composition $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})$ around the value $t = 0$, assuming that the smoothness conditions described above are satisfied for values of $t$ small enough. Moreover, we are interested in this expansion in the specific setting from Section 6.2.2 where each signal $\mathbf{x}$ is equal to $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}$, and where the sign vector $\mathbf{s}$ is such that $\mathbf{s} = \mathrm{sign}(\boldsymbol{\alpha}_0)$.

Throughout the different computations, we will need to refer to the orthogonal projector onto the span of $[\mathbf{D}_0]_{\mathrm{J}}$, which we denote by $\mathbf{P}_0$. In addition, the matrix $\boldsymbol{\Pi}_{\mathrm{J}} \in \mathbb{R}^{p \times |\mathrm{J}|}$ that projects onto the columns indexed by J, i.e., $\mathbf{D}_{\mathrm{J}} = \mathbf{D}\boldsymbol{\Pi}_{\mathrm{J}}$, is represented by $\boldsymbol{\Pi}$ to ease notation. For short, we refer to $\mathrm{Diag}(\mathbf{v})$, $\mathrm{Diag}(\cos(\mathbf{v}t))$ and $\mathrm{Diag}(\sin(\mathbf{v}t))$ as respectively $\mathbf{V}$, $\mathbf{C}$ and $\mathbf{S}$. Finally, we denote by $\boldsymbol{\Theta}_0$ the inverse of $[\mathbf{D}_0]_{\mathrm{J}}^\top [\mathbf{D}_0]_{\mathrm{J}}$.

**Lemma 15** (Taylor expansion of $\phi_{\mathbf{x}}(.|\mathbf{s})$)
*Assuming the invertibility conditions for $\mathbf{D}(t)$ around $t = 0$, the function $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})$ admits a second-order Taylor expansion around $t = 0$, and we have*

$$\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s}) = \phi_{\mathbf{x}}(\mathbf{D}_0|\mathbf{s}) + a_{\mathbf{x}} t + b_{\mathbf{x}} t^2 + o(t^2),$$

*with the following expressions for $a_{\mathbf{x}}$ and $b_{\mathbf{x}}$:*

$$\begin{aligned} a_{\mathbf{x}} \;=\; & -\underbrace{\boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\alpha}_0}_{A_{\boldsymbol{\varepsilon}\boldsymbol{\alpha}}} - \lambda \underbrace{\mathbf{s}^\top \boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathbf{D}_0^\top \mathbf{W}\mathbf{V}\boldsymbol{\alpha}_0}_{A_{\mathbf{s}\boldsymbol{\alpha}}} \\ & + \lambda \underbrace{\boldsymbol{\varepsilon}^\top (\mathbf{I} - \mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathbf{s}}_{A_{\boldsymbol{\varepsilon}\mathbf{s},1}} - \lambda \underbrace{\boldsymbol{\varepsilon}^\top \mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathbf{V}\mathbf{W}^\top \mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathbf{s}}_{A_{\boldsymbol{\varepsilon}\mathbf{s},2}} \\ & + \lambda^2 \underbrace{\mathbf{s}^\top \boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathrm{sym}(\mathbf{V}\mathbf{W}^\top \mathbf{D}_0)\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathbf{s}}_{A_{\mathbf{s}\mathbf{s}}} \\ & - \underbrace{\boldsymbol{\varepsilon}^\top \mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top \mathbf{V}\mathbf{W}^\top (\mathbf{I} - \mathbf{P}_0)\boldsymbol{\varepsilon}}_{A_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}}, \end{aligned}$$

*and*

$$
\begin{aligned}
b_{\mathbf{x}} \;=\; & \underbrace{\boldsymbol{\alpha}_0^\top \mathbf{V}\mathbf{W}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\alpha}_0}_{B_{\boldsymbol{\alpha}\boldsymbol{\alpha}}} \\
& +\lambda\Big[\underbrace{\frac{1}{2}\mathbf{s}^\top\mathbf{V}^2\boldsymbol{\alpha}_0}_{B_{\mathbf{s}\boldsymbol{\alpha},1}} - \underbrace{\mathbf{s}^\top\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}\mathbf{W}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\alpha}_0}_{B_{\mathbf{s}\boldsymbol{\alpha},2}} + \underbrace{\mathbf{s}^\top(\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{D}_0^\top\mathbf{W}\mathbf{V})^2\boldsymbol{\alpha}_0}_{B_{\mathbf{s}\boldsymbol{\alpha},3}}\Big] \\
& + \underbrace{\boldsymbol{\varepsilon}^\top\mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}\mathbf{W}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\alpha}_0}_{B_{\boldsymbol{\varepsilon}\boldsymbol{\alpha},1}} + \underbrace{\boldsymbol{\varepsilon}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{D}_0^\top\mathbf{W}\mathbf{V}\boldsymbol{\alpha}_0}_{B_{\boldsymbol{\varepsilon}\boldsymbol{\alpha},2}} \\
& +\lambda\Big[\underbrace{\frac{1}{2}\boldsymbol{\varepsilon}^\top\mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}^2\mathbf{s}}_{B_{\boldsymbol{\varepsilon}\mathbf{s},1}} - 2\underbrace{\boldsymbol{\varepsilon}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathrm{sym}(\mathbf{V}\mathbf{W}^\top\mathbf{D}_0)\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{s}}_{B_{\boldsymbol{\varepsilon}\mathbf{s},2}} \\
& - \underbrace{\boldsymbol{\varepsilon}^\top\mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}\mathbf{W}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{s}}_{B_{\boldsymbol{\varepsilon}\mathbf{s},3}} + \underbrace{\boldsymbol{\varepsilon}^\top(\mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}\mathbf{W}^\top)^2\mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{s}}_{B_{\boldsymbol{\varepsilon}\mathbf{s},4}}\Big] \\
& -\frac{\lambda^2}{2}\underbrace{\mathbf{s}^\top\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top[\mathbf{V}^2\mathbf{D}_0^\top\mathbf{D}_0 + \mathbf{D}_0^\top\mathbf{D}_0\mathbf{V}^2 - 2\mathbf{V}\mathbf{W}^\top\mathbf{W}\mathbf{V} + 8\mathrm{sym}(\mathbf{V}\mathbf{W}^\top\mathbf{D}_0)\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathrm{sym}(\mathbf{V}\mathbf{W}^\top\mathbf{D}_0)]\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{s}}_{B_{\mathbf{s}\mathbf{s}}} \\
& + \underbrace{\boldsymbol{\varepsilon}^\top((\mathbf{P}_0-\mathbf{I})\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}\mathbf{W}^\top(\mathbf{I}-\mathbf{P}_0) + \mathbf{D}_0\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{V}\mathbf{W}^\top(\mathbf{I}-\mathbf{P}_0)\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{D}_0^\top)\boldsymbol{\varepsilon}}_{B_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon},1}} \\
& + \underbrace{2\boldsymbol{\varepsilon}^\top(\mathbf{I}-\mathbf{P}_0)(\mathbf{W}\mathbf{V}\boldsymbol{\Pi}\boldsymbol{\Theta}_0\boldsymbol{\Pi}^\top\mathbf{D}_0^\top)^2\boldsymbol{\varepsilon}}_{B_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon},2}}.
\end{aligned}
$$

**Some details about the computation of the Taylor expansion**

In this section, we detail the computations underlying Lemma 15. For convenience, we rewrite $\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})$ as

$$
\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \underbrace{\frac{1}{2}\mathbf{x}^\top\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top\mathbf{x}}_{T_{\mathbf{x}\mathbf{x}}} + \underbrace{\lambda\mathbf{s}_{\mathrm{J}}^\top\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top\mathbf{x}}_{T_{\mathbf{s}\mathbf{x}}} - \underbrace{\frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^\top\boldsymbol{\Sigma}(t)^{-1}\mathbf{s}_{\mathrm{J}}}_{T_{\mathbf{s}\mathbf{s}}}.
$$

where we have introduced $\boldsymbol{\Sigma}(t) \triangleq \boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}$.

**Derivatives of $t \mapsto \boldsymbol{\Sigma}(t)^{-1}$.** We gather here the formula related to the derivatives of $\boldsymbol{\Sigma}(t)$ and $\boldsymbol{\Sigma}(t)^{-1}$.

$$
\begin{aligned}
\nabla\boldsymbol{\Sigma}(t) &= \boldsymbol{\Pi}_{\mathrm{J}}^{\top}\Big[ -\mathbf{V}\mathbf{S}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{C} - \mathbf{C}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V} + \mathbf{V}\mathbf{C}\mathbf{W}^{\top}\mathbf{W}\mathbf{S}\mathbf{V} + \mathbf{V}\mathbf{S}\mathbf{W}^{\top}\mathbf{W}\mathbf{C}\mathbf{V} \\
&\quad -\mathbf{V}\mathbf{S}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{S} + \mathbf{C}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{C}\mathbf{V} - \mathbf{S}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V} + \mathbf{V}\mathbf{C}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{C}\Big]\boldsymbol{\Pi}_{\mathrm{J}} \\
\nabla^2\boldsymbol{\Sigma}(t) &= \boldsymbol{\Pi}_{\mathrm{J}}^{\top}\Big[ -\mathbf{V}^2\mathbf{C}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{C} + 2\mathbf{V}\mathbf{S}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V} - \mathbf{C}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{C}\mathbf{V}^2 \\
&\quad -\mathbf{V}^2\mathbf{S}\mathbf{W}^{\top}\mathbf{W}\mathbf{S} + 2\mathbf{V}\mathbf{C}\mathbf{W}^{\top}\mathbf{W}\mathbf{C}\mathbf{V} - \mathbf{S}\mathbf{W}^{\top}\mathbf{W}\mathbf{S}\mathbf{V}^2 \\
&\quad -\mathbf{V}^2\mathbf{C}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{S} - 2\mathbf{V}\mathbf{S}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{C}\mathbf{V} - \mathbf{C}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{S}\mathbf{V}^2 \\
&\quad -\mathbf{S}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{C}\mathbf{V}^2 - 2\mathbf{V}\mathbf{S}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{C}\mathbf{V} - \mathbf{V}^2\mathbf{S}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{C}\Big]\boldsymbol{\Pi}_{\mathrm{J}} \\
\nabla^3\boldsymbol{\Sigma}(t) &= \boldsymbol{\Pi}_{\mathrm{J}}^{\top}\Big[\mathbf{V}^3\mathbf{S}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{C} + 3\mathbf{V}^2\mathbf{C}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V} + 3\mathbf{V}\mathbf{S}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{C}\mathbf{V}^2 + \mathbf{C}\mathbf{D}_0^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V}^3 \\
&\quad -\mathbf{V}^2\mathbf{C}\mathbf{W}^{\top}\mathbf{W}\mathbf{S} - 3\mathbf{V}^2\mathbf{S}\mathbf{W}^{\top}\mathbf{W}\mathbf{C}\mathbf{V} - 3\mathbf{V}\mathbf{C}\mathbf{W}^{\top}\mathbf{W}\mathbf{S}\mathbf{V}^2 - \mathbf{S}\mathbf{W}^{\top}\mathbf{W}\mathbf{C}\mathbf{V}^2 \\
&\quad \mathbf{V}^3\mathbf{S}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{S} - 3\mathbf{V}^2\mathbf{C}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{C}\mathbf{V} + 3\mathbf{V}\mathbf{S}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{S}\mathbf{V}^2 - \mathbf{C}\mathbf{D}_0^{\top}\mathbf{W}\mathbf{C}\mathbf{V}^3 \\
&\quad -\mathbf{V}^3\mathbf{C}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{C} + 3\mathbf{V}^2\mathbf{S}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V} + -3\mathbf{V}\mathbf{C}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{C}\mathbf{V}^2 + \mathbf{S}\mathbf{W}^{\top}\mathbf{D}_0\mathbf{S}\mathbf{V}^3\Big]\boldsymbol{\Pi}_{\mathrm{J}}
\end{aligned}
$$

and

$$
\begin{aligned}
\nabla[\boldsymbol{\Sigma}(t)^{-1}] &= -\boldsymbol{\Sigma}(t)^{-1}\nabla\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}(t)^{-1} \\
\nabla^2[\boldsymbol{\Sigma}(t)^{-1}] &= -\nabla[\boldsymbol{\Sigma}(t)^{-1}]\nabla\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}(t)^{-1} - \boldsymbol{\Sigma}(t)^{-1}\nabla^2\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}(t)^{-1} - \boldsymbol{\Sigma}(t)^{-1}\nabla\boldsymbol{\Sigma}(t)\nabla[\boldsymbol{\Sigma}(t)^{-1}] \\
\nabla^3[\boldsymbol{\Sigma}(t)^{-1}] &= -\boldsymbol{\Sigma}(t)^{-1}\nabla^3\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}(t)^{-1} - 2\nabla[\boldsymbol{\Sigma}(t)^{-1}]\nabla^2\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}(t)^{-1} - 2\boldsymbol{\Sigma}(t)^{-1}\nabla^2\boldsymbol{\Sigma}(t)\nabla[\boldsymbol{\Sigma}(t)^{-1}] \\
&\quad -2\nabla[\boldsymbol{\Sigma}(t)^{-1}]\nabla\boldsymbol{\Sigma}(t)\nabla[\boldsymbol{\Sigma}(t)^{-1}] - \nabla^2[\boldsymbol{\Sigma}(t)^{-1}]\nabla\boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}(t)^{-1} \\
&\quad -\boldsymbol{\Sigma}(t)^{-1}\nabla\boldsymbol{\Sigma}(t)\nabla^2[\boldsymbol{\Sigma}(t)^{-1}].
\end{aligned}
$$

**Derivatives of $t \mapsto \mathbf{D}(t)$.** We gather here the formula related to the derivatives of $\mathbf{D}(t)$:

$$
\begin{aligned}
\nabla\mathbf{D}(t) &= [-\mathbf{D}_0\mathrm{Diag}(\sin(\mathbf{v}t)) + \mathbf{W}\mathrm{Diag}(\cos(\mathbf{v}t))]\mathrm{Diag}(\mathbf{v}) \\
\nabla^2\mathbf{D}(t) &= -\mathbf{D}(t)\mathrm{Diag}(\mathbf{v})^2 \\
\nabla^3\mathbf{D}(t) &= [\mathbf{D}_0\mathrm{Diag}(\sin(\mathbf{v}t)) - \mathbf{W}\mathrm{Diag}(\cos(\mathbf{v}t))]\mathrm{Diag}(\mathbf{v})^3.
\end{aligned}
$$

**Derivatives of $t \mapsto \phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s})$.** The different derivatives of the three terms $T_{\mathbf{xx}}$, $T_{\mathbf{sx}}$ and $T_{\mathbf{ss}}$ can be computed by applying the product rule, based on the derivatives of $\mathbf{D}(t)$ and $\boldsymbol{\Sigma}(t)^{-1}$ (see above). We only give the formula for the first derivatives, namely

$$
\begin{aligned}
\nabla T_{\mathbf{xx}} &= \mathbf{x}^{\top}(\nabla\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}\mathbf{D}(t)^{\top} + \mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\nabla[\boldsymbol{\Sigma}(t)^{-1}]\boldsymbol{\Pi}_{\mathrm{J}}\mathbf{D}(t)^{\top} \\
&\quad +\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}[\nabla\mathbf{D}(t)]^{\top})\mathbf{x} \\
\nabla T_{\mathbf{sx}} &= \lambda\mathbf{s}_{\mathrm{J}}^{\top}(\nabla[\boldsymbol{\Sigma}(t)^{-1}]\boldsymbol{\Pi}_{\mathrm{J}}^{\top}\mathbf{D}(t)^{\top} + \boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}^{\top}[\nabla\mathbf{D}(t)]^{\top})\mathbf{x} \\
\nabla T_{\mathbf{ss}} &= \frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^{\top}\nabla[\boldsymbol{\Sigma}(t)^{-1}]\mathbf{s}_{\mathrm{J}}.
\end{aligned}
$$

We obtain the results advertised in Lemma 15 by evaluating the different derivatives at $t = 0$. Simplifications arise by making use of the following facts:

$$\begin{aligned} \mathbf{x} &= \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} \\ \mathbf{I} &= \boldsymbol{\Pi}_{\mathrm{J}}^\top \boldsymbol{\Pi}_{\mathrm{J}} \\ \boldsymbol{\alpha}_0 &= \boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Pi}_{\mathrm{J}}^\top \boldsymbol{\alpha}_0 \\ \mathbf{P}_0\mathbf{D}_0\boldsymbol{\Pi}_{\mathrm{J}} &= \mathbf{D}_0\boldsymbol{\Pi}_{\mathrm{J}}. \end{aligned}$$

**Some details about the upper bound of the third derivative**

We need to upper bound quantities of the form $\|\mathbf{D}_0\mathbf{C}\mathbf{V}^h\boldsymbol{\Pi}_{\mathrm{J}}\|_{\mathrm{F}}$ and $\|\mathbf{D}_0\mathbf{S}\mathbf{V}^h\boldsymbol{\Pi}_{\mathrm{J}}\|_{\mathrm{F}}$ for $h$ in $[\![0;3]\!]$. Thanks to the normalization of the vector $\mathbf{v}$ and the columns of $\mathbf{D}_0$, it easy to obtain upper bounds by constant terms when $h \geq 1$. Moreover, since $|\sin(t)| \leq |t|$, it is also simple to derive upper bounds scaling in $|t|$ as soon as the sine term (i.e., $\mathbf{S}$) is present. In the case of $\|\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}\|_{\mathrm{F}}$, we can notice that for any matrix $\mathbf{A}$

$$\begin{aligned} \|\mathbf{A}\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}\|_{\mathrm{F}}^2 &= \mathrm{Tr}(\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{C}^\top\mathbf{D}_0^\top\mathbf{A}^\top\mathbf{A}\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}) \\ &= \mathrm{Tr}(\mathbf{A}\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{C}^\top\mathbf{D}_0^\top\mathbf{A}^\top) \\ &\leq \mathrm{Tr}(\mathbf{A}\mathbf{D}_0\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}_0^\top\mathbf{A}^\top) \\ &\leq \|\mathbf{D}_0\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}_0^\top\|_2\|\mathbf{A}\|_{\mathrm{F}}^2 \\ &\leq (1+k\mu_0)\|\mathbf{A}\|_{\mathrm{F}}^2 \end{aligned}$$

In the light of the remarks above, a little calculation leads to the following upper bounds:

$$\begin{aligned} \|\nabla\boldsymbol{\Sigma}(t)\|_{\mathrm{F}} &\leq 2(\sqrt{1+k\mu_0} + |t|\sqrt{1+k\mu_0} + |t| + t^2) \overbrace{\leq}^{|t|\leq 1} 8\sqrt{1+k\mu_0} \\ \|\nabla^2\boldsymbol{\Sigma}(t)\|_{\mathrm{F}} &\leq 2(\sqrt{1+k\mu_0} + 1 + |t|\sqrt{1+k\mu_0} + 3|t| + 2t^2) \overbrace{\leq}^{|t|\leq 1} 16\sqrt{1+k\mu_0} \\ \|\nabla^3\boldsymbol{\Sigma}(t)\|_{\mathrm{F}} &\leq 2(\sqrt{1+k\mu_0} + 3 + |t|\sqrt{1+k\mu_0} + 7|t| + 4t^2) \overbrace{\leq}^{|t|\leq 1} 32\sqrt{1+k\mu_0}, \end{aligned}$$

and for $|t| \leq 1$,

$$\begin{aligned} \|\nabla[\boldsymbol{\Sigma}(t)^{-1}]\|_{\mathrm{F}} &\leq \frac{8\sqrt{1+k\mu_0}}{(1-k\mu_0)^2} \\ \|\nabla^2[\boldsymbol{\Sigma}(t)^{-1}]\|_{\mathrm{F}} &\leq \frac{144(1+k\mu_0)}{(1-k\mu_0)^3} \\ \|\nabla^3[\boldsymbol{\Sigma}(t)^{-1}]\|_{\mathrm{F}} &\leq \frac{3872(1+k\mu_0)^{3/2}}{(1-k\mu_0)^4}. \end{aligned}$$

Based on these upper bounds, we can further control the terms appearing in $T_{\mathbf{sx}}$ and $T_{\mathbf{xx}}$, namely

$$\begin{aligned}\|\nabla^3[\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top]\|_{\mathrm{F}} &\leq 8\|\nabla^3[\boldsymbol{\Sigma}(t)^{-1}]\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top\|_{\mathrm{F}} \\ &\leq \frac{30976(1+k\mu_0)^{3/2}((1+k\mu_0)^{1/2}+1)}{(1-k\mu_0)^4},\end{aligned}$$

and

$$\begin{aligned}\|\nabla^3[\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top]\|_{\mathrm{F}} &\leq 27\|\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\nabla^3[\boldsymbol{\Sigma}(t)^{-1}]\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{D}(t)^\top\|_{\mathrm{F}} \\ &\leq \frac{104544(1+k\mu_0)^{3/2}((1+k\mu_0)^{1/2}+1)^2}{(1-k\mu_0)^4},\end{aligned}$$

where we use the fact there are respectively 8 and 27 terms appearing in $\nabla^3 T_{\mathbf{sx}}$ and $\nabla^3 T_{\mathbf{xx}}$. Moreover, the upper bound by $((1+k\mu_0)^{1/2}+1)$ exploits that for any matrix $\mathbf{A}$

$$\begin{aligned}\|\mathbf{D}(t)\boldsymbol{\Pi}_{\mathrm{J}}\mathbf{A}\|_{\mathrm{F}} &\leq \|\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}\mathbf{A}\|_{\mathrm{F}} + \|\mathbf{W}\mathbf{S}\boldsymbol{\Pi}_{\mathrm{J}}\mathbf{A}\|_{\mathrm{F}} \\ &\leq \mathrm{Tr}(\mathbf{A}^\top\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{C}^\top\mathbf{D}_0^\top\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}\mathbf{A}^\top)^{1/2} + \|\mathbf{W}\mathbf{S}\boldsymbol{\Pi}_{\mathrm{J}}\|_{\mathrm{F}}\|\mathbf{A}\|_{\mathrm{F}} \\ &\leq (\|\boldsymbol{\Pi}_{\mathrm{J}}^\top\mathbf{C}\mathbf{D}_0^\top\mathbf{D}_0\mathbf{C}\boldsymbol{\Pi}_{\mathrm{J}}\|_2^{1/2}+1)\|\mathbf{A}\|_{\mathrm{F}} \\ &\leq ((1+k\mu_0)^{1/2}+1)\|\mathbf{A}\|_{\mathrm{F}},\end{aligned}$$

where we can notice in the last inequality that the coherence of $\mathbf{D}_0\mathbf{C}$ is lower than that of $\mathbf{D}_0$, and we can therefore make use of Lemma 16.

## 6.9 Technical lemmas

The final section of this appendix gathers different technical lemmas required by the main results of the chapter.

### 6.9.1 Properties related to the mutual coherence

**Lemma 16**
*Let $\mathbf{D} \in \mathbb{R}^{m\times p}$ be a dictionary with coherence $\mu$ and normalized columns (i.e., with unit $\ell_2$-norm). For any $\mathrm{J} \subseteq [\![1;p]\!]$ with $|\mathrm{J}| \leq k$, We have*

$$\|\mathbf{D}_{\mathrm{J}}\mathbf{D}_{\mathrm{J}}^\top\|_2 = \|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\|_2 \leq 1+k\mu.$$

*Similarly, it holds*

$$\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\|_\infty \leq 1+k\mu \quad \textit{and} \quad \|\mathbf{D}_{\mathrm{J}^{\mathrm{c}}}^\top\mathbf{D}_{\mathrm{J}}\|_\infty \leq k\mu.$$

*Moreover, if we further have $k\mu < 1$, then $\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}$ is invertible and*

$$\max\left\{\|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}\|_\infty, \|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}\|_2, \|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}\|_{\mathrm{F}}\right\} \leq \frac{1}{1-k\mu},$$

*with, in addition,*

$$\max\left\{\|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}-\mathbf{I}\|_\infty, \|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}-\mathbf{I}\|_2, \|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}-\mathbf{I}\|_{\mathrm{F}}\right\} \leq \frac{k\mu}{1-k\mu}.$$

*Proof.* These properties are already well-known (see, e.g. Fuchs, 2005). We briefly prove them. First, we introduce $\mathbf{H} = \mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}} - \mathbf{I}$. A straightforward elementwise upper bound leads to $\|\mathbf{H}\|_{\mathrm{F}} \leq |\mathrm{J}|\mu \leq k\mu$. As a consequence, we have

$$\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\|_2 \leq 1 + \|\mathbf{H}\|_2 \leq 1 + \|\mathbf{H}\|_{\mathrm{F}} \leq 1 + k\mu.$$

By definition of $\|.\|_\infty$, we also have

$$\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\|_\infty \leq 1 + \|\mathbf{H}\|_\infty = 1 + \max_{i\in\mathrm{J}} \sum_{j\in\mathrm{J}, j\neq i} |[\mathbf{d}^i]^\top\mathbf{d}^j| \leq 1 + k\mu.$$

Note that for $\|\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}\|_\infty$, there are no diagonal terms to take into account.

Now, if $k\mu < 1$ holds, then we have $\max\{\|\mathbf{H}\|_\infty, \|\mathbf{H}\|_2, \|\mathbf{H}\|_{\mathrm{F}}\} < 1$ and there are convergent series expansion of $[\mathbf{I}+\mathbf{H}]^{-1}$ in each of these norms (Horn and Johnson, 1990). By sub-multiplicativity, we obtain

$$\|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}\| = \|\sum_{t=0}^{\infty}(-1)^t\mathbf{H}^t\| \leq \sum_{t=0}^{\infty}\|\mathbf{H}\|^t \leq 1/(1-k\mu),$$

where $\|.\|$ stands for one the three aforementioned matrix norms. The last result lies in the fact that $\|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1} - \mathbf{I}\| = \|\sum_{t=1}^{\infty}(-1)^t\mathbf{H}^t\| \leq k\mu/(1-k\mu)$. □

### 6.9.2 Lemmas related to the study of problem (6.1)

**Lemma 17** (Control of the perturbation of $\mathbf{D}_0$)
*For any* $\mathrm{J} \subseteq [\![1;p]\!]$*, we introduce the orthogonal projector*

$$[\mathbf{P}(t)]_{\mathrm{J}} \triangleq [\mathbf{D}(t)]_{\mathrm{J}}[[\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{D}(t)]_{\mathrm{J}}]^{-1}[\mathbf{D}(t)]_{\mathrm{J}}^\top$$

*that projects onto the span of* $[\mathbf{D}(t)]_{\mathrm{J}}$*.*
*Fix* $\tau > 0$*, and consider a random vector* $\boldsymbol{\alpha}_0$ *that follows the generative model from Section 6.2.2. For any* $\mathrm{J} \subseteq [\![1;p]\!]$ *with* $|\mathrm{J}| = k$*, and for all* $|t| \leq 1$ *such that* $[\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{D}(t)]_{\mathrm{J}}$ *is invertible, we have*

$$\Pr(\|[\mathbf{D}(t)]_{\mathrm{J}^c}^\top(\mathbf{I} - [\mathbf{P}(t)]_{\mathrm{J}})[\mathbf{D}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty > \tau) \leq 2(p-k)\exp\left(\frac{-\tau^2}{6\overline{\alpha}^2t^2}\right),$$

*and*

$$\Pr(\|[\mathbf{D}(t)]_{\mathrm{J}}^\top([\mathbf{D}(t)]_{\mathrm{J}} - [\mathbf{D}_0]_{\mathrm{J}})[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty > \tau) \leq 2k\exp\left(\frac{-\tau^2}{4\overline{\alpha}^2t^2}\right).$$

*Proof.* We focus on the first inequality. Consider for $j \in \mathrm{J}^c$ the zero-mean sub-Gaussian variable $u_j \triangleq [\mathbf{d}(t)^j]^\top (\mathbf{I} - [\mathbf{P}(t)]_\mathrm{J})[\mathbf{D}_0]_\mathrm{J}[\boldsymbol{\alpha}_0]_\mathrm{J}$. The variance of $u_j$ is upper bounded by

$$\begin{aligned}
\mathbb{E}[u_j^2] &\leq \overline{\alpha}^2 [\mathbf{d}(t)^j]^\top (\mathbf{I} - [\mathbf{P}(t)]_\mathrm{J})[\mathbf{D}_0]_\mathrm{J}[\mathbf{D}_0]_\mathrm{J}^\top (\mathbf{I} - [\mathbf{P}(t)]_\mathrm{J})\mathbf{d}(t)^j \\
&= \overline{\alpha}^2 \|[\mathbf{D}_0]_\mathrm{J}^\top (\mathbf{I} - [\mathbf{P}(t)]_\mathrm{J})\mathbf{d}(t)^j\|_2^2 \\
&\leq \overline{\alpha}^2 \|\mathrm{Diag}(\tan(\mathbf{v}_\mathrm{J} t))\mathbf{W}_\mathrm{J}^\top (\mathbf{I} - [\mathbf{P}(t)]_\mathrm{J})\mathbf{d}(t)^j\|_2^2 \\
&\leq \overline{\alpha}^2 \|\mathbf{W}_\mathrm{J}\mathrm{Diag}(\tan(\mathbf{v}_\mathrm{J} t))\|_2^2 \\
&\leq \overline{\alpha}^2 \|\mathbf{W}_\mathrm{J}\mathrm{Diag}(\tan(\mathbf{v}_\mathrm{J} t))\|_\mathrm{F}^2 \\
&= \overline{\alpha}^2 \sum_{j \in \mathrm{J}} \tan^2(\mathbf{v}_j t),
\end{aligned}$$

where the first inequality exploits the definition of $\mathbf{D}_0$, i.e., $\mathbf{D}_0 = \mathrm{Diag}(\cos(\mathbf{v}t))^{-1}\mathbf{D}(t) - \mathbf{W}\mathrm{Diag}(\tan(\mathbf{v}t))$, combined with the relation $(\mathbf{I} - [\mathbf{P}(t)]_\mathrm{J})[\mathbf{D}(t)]_\mathrm{J} = \mathbf{0}$, while the other upper bounds use the normalization of $\mathbf{d}(t)$ and $\mathbf{W}_\mathrm{J}$, along with the fact that projectors have their spectral norm bounded by one. The final result follows by applying Lemma 6 and the union bound over $|\mathrm{J}^c| = p - k$ terms. Moreover, if $|t| \leq 1$, we have $|\mathbf{v}_j t| \leq 1$ and $\tan^2(\mathbf{v}_j t) \leq 3(\mathbf{v}_j t)^2$, which leads to $\sum_{j \in \mathrm{J}} \tan^2(\mathbf{v}_j t) \leq 3\|\mathbf{v}\|_2^2 t^2 = 3t^2$.

For the second inequality, we proceed similarly with $u_j \triangleq [\mathbf{d}(t)^j]^\top ([\mathbf{D}(t)]_\mathrm{J} - [\mathbf{D}_0]_\mathrm{J})[\boldsymbol{\alpha}_0]_\mathrm{J}$ for $j \in \mathrm{J}$. In this case, the variance of $u_j$ is upper bounded by

$$\begin{aligned}
\mathbb{E}[u_j^2] &\leq \overline{\alpha}^2 [\mathbf{d}(t)^j]^\top ([\mathbf{D}(t)]_\mathrm{J} - [\mathbf{D}_0]_\mathrm{J})([\mathbf{D}(t)]_\mathrm{J} - [\mathbf{D}_0]_\mathrm{J})^\top \mathbf{d}(t)^j \\
&= \overline{\alpha}^2 \|([\mathbf{D}(t)]_\mathrm{J} - [\mathbf{D}_0]_\mathrm{J})^\top \mathbf{d}(t)^j\|_2^2 \\
&\leq \overline{\alpha}^2 \|[\mathbf{D}(t)]_\mathrm{J} - [\mathbf{D}_0]_\mathrm{J}\|_2^2 \\
&\leq \overline{\alpha}^2 \|[\mathbf{D}_0]_\mathrm{J}(\mathrm{Diag}(\cos(\mathbf{v}_\mathrm{J} t)) - \mathbf{I}) + \mathbf{W}_\mathrm{J}\mathrm{Diag}(\sin(\mathbf{v}_\mathrm{J} t))\|_2^2 \\
&\leq 2\overline{\alpha}^2 \|[\mathbf{D}_0]_\mathrm{J}(\mathrm{Diag}(\cos(\mathbf{v}_\mathrm{J} t)) - \mathbf{I})\|_\mathrm{F}^2 + \|\mathbf{W}_\mathrm{J}\mathrm{Diag}(\sin(\mathbf{v}_\mathrm{J} t))\|_\mathrm{F}^2 \\
&= 2\overline{\alpha}^2 \sum_{j \in \mathrm{J}} (\cos(\mathbf{v}_j t) - 1)^2 + \sin^2(\mathbf{v}_j t) = 2\overline{\alpha}^2 \sum_{j \in \mathrm{J}} 2(1 - \cos(\mathbf{v}_j t)) \\
&\leq 2\overline{\alpha}^2 \sum_{j \in \mathrm{J}} (\mathbf{v}_j t)^2 \leq 2\overline{\alpha}^2 t^2
\end{aligned}$$

where we have used the same arguments as above, along with the inequality $2(1 - \cos(z)) \leq z^2$. The conclusion results from the application of Lemma 6 along with the union bound over $|\mathrm{J}| = k$ terms. □

**Lemma 18**
*Let $\mathrm{J} \subseteq [\![1; p]\!]$ and $\mathbf{s} \in \{-1, 0, 1\}^{|\mathrm{J}|}$. Consider a dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ such that $\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}$ is invertible. Consider also the vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ defined by*

$$\boldsymbol{\alpha} = \begin{pmatrix} [\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}[\mathbf{D}_\mathrm{J}^\top \mathbf{x} - \lambda \mathbf{s}] \\ \mathbf{0} \end{pmatrix},$$

*with $\mathbf{x} \in \mathbb{R}^m$ and $\lambda$ a nonnegative scalar. If the vector $\mathbf{x}$ can be written as $\mathbf{x} = [\mathbf{D}_0]_\mathrm{J}[\boldsymbol{\alpha}_0]_\mathrm{J} + \boldsymbol{\varepsilon}$ for some $(\mathbf{D}_0, \boldsymbol{\alpha}_0, \boldsymbol{\varepsilon}) \in \mathbb{R}^{m \times p} \times \mathbb{R}^p \times \mathbb{R}^m$, then we have*

$$\|[\boldsymbol{\alpha} - \boldsymbol{\alpha}_0]_\mathrm{J}\|_\infty \leq \|[\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}\|_\infty \Big[\lambda + \|\mathbf{D}_\mathrm{J}^\top (\mathbf{D}_\mathrm{J} - [\mathbf{D}_0]_\mathrm{J})[\boldsymbol{\alpha}_0]_\mathrm{J}\|_\infty + \|\mathbf{D}_\mathrm{J}^\top \boldsymbol{\varepsilon}\|_\infty\Big].$$

*Proof.* The proof consists of simple algebraic manipulations. We first plug the expression of $\mathbf{x}$ into that of $\boldsymbol{\alpha}$ and then use the triangle inequality for $\|.\|_\infty$, along with the definition and the sub-multiplicativity of $|||.|||_\infty$. □

**Lemma 19**
*Let $\mathbf{x} \in \mathbb{R}^m$ be a signal. Consider $\mathrm{J} \subseteq [\![1;p]\!]$ and a dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ such that $\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}$ is invertible. Consider also a sign vector $\mathbf{s} \in \{-1,1\}^{|\mathrm{J}|}$ and define $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ by*

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} [\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}[\mathbf{D}_\mathrm{J}^\top \mathbf{x} - \lambda \mathbf{s}] \\ \mathbf{0} \end{pmatrix},$$

*for some regularization parameter $\lambda \geq 0$. Let us introduce the projector $\mathbf{P}_\mathrm{J} \triangleq \mathbf{D}_\mathrm{J}[\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}\mathbf{D}_\mathrm{J}^\top$ that projects onto the span of $\mathbf{D}_\mathrm{J}$. If the following two conditions hold*

$$\begin{cases} \mathrm{sign}\Big([\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}[\mathbf{D}_\mathrm{J}^\top \mathbf{x} - \lambda \mathbf{s}]\Big) = \mathbf{s}, \\ \|\mathbf{D}_{\mathrm{J}^c}^\top (\mathbf{I} - \mathbf{P}_\mathrm{J})\mathbf{x}\|_\infty + \lambda |||\mathbf{D}_{\mathrm{J}^c}^\top \mathbf{D}_\mathrm{J}[\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}|||_\infty < \lambda, \end{cases}$$

*then $\hat{\boldsymbol{\alpha}}$ is the unique solution of $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} [\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1]$ and we have $\mathrm{sign}(\hat{\boldsymbol{\alpha}}_\mathrm{J}) = \mathbf{s}$.*

*Proof.* We first check that $\hat{\boldsymbol{\alpha}}$ is a solution of the Lasso program. It is well-known (e.g., see Fuchs, 2005; Wainwright, 2009) that this statement is equivalent to the existence of a subgradient $\mathbf{z} \in \partial\|\hat{\boldsymbol{\alpha}}\|_1$ such that $-\mathbf{D}^\top(\mathbf{x} - \mathbf{D}\hat{\boldsymbol{\alpha}}) + \lambda \mathbf{z} = 0$, where $\mathbf{z}_j = \mathrm{sign}(\hat{\boldsymbol{\alpha}}_j)$ if $\hat{\boldsymbol{\alpha}}_j \neq 0$, and $|\mathbf{z}_j| \leq 1$ otherwise.

We now build from $\mathbf{s}$ such a subgradient. Given the definition of $\hat{\boldsymbol{\alpha}}$ and the assumption made on its sign, we can take $\mathbf{z}_\mathrm{J} \triangleq \mathbf{s}$. It now remains to find a subgradient on $\mathrm{J}^c$ that agrees with the fact that $\hat{\boldsymbol{\alpha}}_{\mathrm{J}^c} = \mathbf{0}$. More precisely, we define $\mathbf{z}_{\mathrm{J}^c}$ by

$$\lambda \mathbf{z}_{\mathrm{J}^c} \triangleq \mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{x} - \mathbf{D}\hat{\boldsymbol{\alpha}}) = \mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{I} - \mathbf{P}_\mathrm{J})\mathbf{x} + \lambda \mathbf{D}_{\mathrm{J}^c}^\top \mathbf{D}_\mathrm{J}[\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}\mathbf{s}.$$

Using our assumption, we have $\|\mathbf{z}_{\mathrm{J}^c}\|_\infty < 1$. We have therefore proved that $\hat{\boldsymbol{\alpha}}$ is a solution of the Lasso program. The uniqueness comes from Lemma 1 in Wainwright (2009). □

### 6.9.3 Some useful concentration results

In this section, we first recall some known concentration results before proving more specific lemmas.

#### Compilation of some known results

**Definition 6** (Sub-Gaussian variables, e.g., see Buldygin and Kozachenko (2000); Vershynin (2010))
*A zero-mean random variable $z$ is sub-Gaussian with parameter $\sigma > 0$ if and only if for any $t \geq 0$*

$$\mathbb{E}[\exp(tz)] \leq \exp\Big(\frac{\sigma^2 t^2}{2}\Big).$$

*We notably have for any $t \geq 0$ the tail probability*

$$\Pr(|z| > t) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big).$$

*Moreover, if $\{z_j\}_{j\in[\![1;p]\!]}$ is a collection of $p$ independent sub-Gaussian variables with parameters $\sigma_j$, then $\sum_{j=1}^p z_j$ is also sub-Gaussian, with parameter $\sum_{j=1}^p \sigma_j^2$.*

**Lemma 20** (From Hsu et al. (2011))
*Let us consider $\mathbf{z} \in \mathbb{R}^m$ a random vector of independent sub-Gaussian variables with parameters upper bounded by $\sigma > 0$. Let $\mathbf{A} \in \mathbb{R}^{m\times p}$ be a fixed matrix. For all $t > 0$, it holds*

$$\Pr\Big(\|\mathbf{A}\mathbf{z}\|_2^2 > \sigma^2(\|\mathbf{A}\|_{\mathrm{F}}^2 + 2\sqrt{\mathrm{Tr}[(\mathbf{A}^\top\mathbf{A})^2]t} + 2|\!|\!|\mathbf{A}^\top\mathbf{A}|\!|\!|_2 t)\Big) \leq \exp(-t).$$

*In particular, for any $t \geq 1$, we have*

$$\Pr\Big(\|\mathbf{A}\mathbf{z}\|_2^2 > 5\sigma^2\|\mathbf{A}\|_{\mathrm{F}}^2 t\Big) \leq \exp(-t).$$

**Lemma 21** (From Krahmer and Ward (2010))
*Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ be a matrix with all its diagonal terms equal to zero. Consider a vector $\mathbf{y}$ with independent Rademacher entries. We have for all $t > 0$*

$$\Pr(|\mathbf{y}^\top\mathbf{M}\mathbf{y}| > t) \leq 2\exp\Big(-\frac{1}{64}\min\Big\{\frac{\frac{96}{65}t}{|\!|\!|\mathbf{M}|\!|\!|_2}, \frac{t^2}{\|\mathbf{M}\|_{\mathrm{F}}^2}\Big\}\Big).$$

**Lemma 22** (From Hanson and Wright (1971), as stated in Nelson (2010))
*Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ be a symmetric matrix. Consider a vector $\mathbf{y}$ with i.i.d. sub-Gaussian entries with parameter upper bounded by $\sigma$. There exist universal constants $c_0, c_1 > 0$ such that we have for all $t > 0$*

$$\Pr(|\mathbf{y}^\top\mathbf{M}\mathbf{y} - \mathrm{Tr}(\mathbf{M})| > t) \leq c_0\exp\Big(-c_1\min\Big\{\frac{t}{\sigma^2\|\mathbf{M}\|_{\mathrm{F}}}, \frac{t^2}{\sigma^4\|\mathbf{M}\|_{\mathrm{F}}^2}\Big\}\Big).$$

**Lemma 23** (Hoeffding's Inequality, e.g., see Boucheron et al. (2004))
*Let $\{z_j\}_{j\in[\![1;p]\!]}$ be a collection of independent, bounded random variables such that for any $j \in [\![1;p]\!]$, it holds that $z_j \in [\underline{z}_j, \overline{z}_j]$ almost surely. For any $t \geq 0$, we have*

$$\Pr\Big(\Big|\sum_{j=1}^p z_j - \sum_{j=1}^p \mathbb{E}[z_j]\Big| > t\Big) \leq 2\exp\Big(-\frac{2t^2}{\sum_{j=1}^p(\overline{z}_j - \underline{z}_j)^2}\Big).$$

**Lemma 24** (Bernstein's Inequality, e.g., see Boucheron et al. (2004))
*Let $\{z_j\}_{j\in[\![1;p]\!]}$ be a collection of independent, zero-mean and bounded random variables such that for any $j \in [\![1;p]\!]$, it holds that $|z_j| \leq \overline{z}$ almost surely. Let us denote by $\sigma_j^2$ the variance of $z_j$ and define $\sigma^2 \triangleq \frac{1}{p}\sum_{j=1}^p \sigma_j^2$. For any $t \geq 0$, we have*

$$\Pr\Big(\frac{1}{p}\sum_{j=1}^p z_j > t\Big) \leq \exp\Big(-\frac{t^2}{2\sigma^2 + 2/3\overline{z}t}\Big).$$

*In particular, for any* $t \leq \frac{3\sigma}{2\bar{z}}\sqrt{p}$*, we have*

$$\Pr\Big(\frac{1}{p}\sum_{j=1}^{p} z_j > \frac{2\sigma t}{\sqrt{p}}\Big) \leq \exp\big(-t^2\big).$$

**Some more specific tail bounds**

**Lemma 25**
*Let* $\mathbf{M} \in \mathbb{R}^{n\times p}$ *be a matrix. Consider two independent sub-Gaussian vectors* $\mathbf{y} \in \mathbb{R}^n$ *and* $\mathbf{x} \in \mathbb{R}^m$ *with i.i.d. entries of respective parameters* $\sigma_{\mathbf{x}}$ *and* $\sigma_{\mathbf{y}}$*. There exist universal constants* $c_0, c_1 > 0$ *such that we have for all* $t \geq 1$

$$\Pr(|\mathbf{x}^\top \mathbf{M}\mathbf{y}| \leq \sigma_{\mathbf{x}}\sigma_{\mathbf{y}}\|\mathbf{M}\|_{\mathrm{F}} t) \geq 1 - c_0\exp(-c_1 t).$$

*Proof.* We simply apply the result from Lemma 22 on the quadratic form

$$\begin{bmatrix} 1/\sigma_{\mathbf{x}}\,\mathbf{x} \\ 1/\sigma_{\mathbf{y}}\,\mathbf{y} \end{bmatrix}^\top \frac{1}{2}\begin{bmatrix} \mathbf{0} & \sigma_{\mathbf{x}}\sigma_{\mathbf{y}}\mathbf{M} \\ \sigma_{\mathbf{x}}\sigma_{\mathbf{y}}\mathbf{M}^\top & \mathbf{0} \end{bmatrix}\begin{bmatrix} 1/\sigma_{\mathbf{x}}\,\mathbf{x} \\ 1/\sigma_{\mathbf{y}}\,\mathbf{y} \end{bmatrix}.$$

Notice that the Frobenius norm of the augmented matrix is bounded by $\sqrt{2}/2\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}\|\mathbf{M}\|_{\mathrm{F}}$, with $\sqrt{2}/2 < 1$. □

**Lemma 26**
*Let* $\mathbf{M} \in \mathbb{R}^{n\times n}$ *be a matrix. Consider a random vector* $\mathbf{y} \in \mathbb{R}^n$ *with i.i.d. entries. Assume the distribution of these entries to be symmetric and bounded, with* $\|\mathbf{y}\|_\infty \leq \sigma$ *almost surely. For all* $t \geq 2$*, it holds that*

$$\Pr\Big(|\mathbf{y}^\top \mathbf{M}\,\mathrm{sign}(\mathbf{y}) - \mathbb{E}[|\mathbf{y}_1|]\mathrm{Tr}(\mathbf{M})| \leq 64\sigma\|\mathbf{M}\|_{\mathrm{F}} t\Big) \geq 1 - 4\exp(-t).$$

*Proof.* We start by splitting $\mathbf{y}^\top \mathbf{M}\,\mathrm{sign}(\mathbf{y})$ into

$$\sum_{i=1}^{n} \mathbf{M}_{ii}|\mathbf{y}_i| + \sum_{i=1}^{n} \mathrm{sign}(\mathbf{y}_i)|\mathbf{y}_i| \sum_{k\neq i} \mathrm{sign}(\mathbf{y}_k)\mathbf{M}_{ik}.$$

The expectation of the first term is equal to $\mathbb{E}[|\mathbf{y}_1|]\mathrm{Tr}(\mathbf{M})$, while the second term is centered. Since the first term consists of a sum of independent bounded random variables, we can apply Hoeffding's inequality (see Lemma 23), that is, for all $t \geq 0$,

$$\Pr(|\sum_{i=1}^{n} \mathbf{M}_{ii}(\mathbf{y}_i - \mathbb{E}[|\mathbf{y}_i|])| \geq t) \leq 2\exp\Big(-\frac{2t^2}{\sigma^2\|\mathrm{diag}(\mathbf{M})\|_2^2}\Big),$$

where we have used the fact that $|\mathbf{y}_i| \in [0, \sigma]$ almost surely. Let us focus on the second residual term. Notice that

$$\sum_{i=1}^{n} \mathrm{sign}(\mathbf{y}_i)|\mathbf{y}_i| \sum_{k\neq i} \mathrm{sign}(\mathbf{y}_k)\mathbf{M}_{ik} = \mathrm{sign}(\mathbf{y})^\top \mathrm{Diag}(|\mathbf{y}|)\mathbf{M}_{\mathrm{off}}\mathrm{sign}(\mathbf{y}),$$

where $\mathbf{M}_{\text{off}}$ stands for the matrix $\mathbf{M}$ with its diagonal terms set to zero. By symmetry, the magnitude of $\mathbf{y}_i$ is *independent* of its sign. So, conditionally on the values of $|\mathbf{y}_i|$, we can apply Lemma 21, that is, for all $t \geq 2$,

$$\Pr(|\text{sign}(\mathbf{y})^\top \text{Diag}(|\mathbf{y}|)\mathbf{M}_{\text{off}}\text{sign}(\mathbf{y})| \leq 45\sigma\|\mathbf{M}_{\text{off}}\|_{\text{F}} t) \geq 1 - 2\exp(-t),$$

where we have used the fact $64 \times 65/96 \leq 45$, and $\|\text{Diag}(|\mathbf{y}|)\mathbf{M}_{\text{off}}\|_{\text{F}} \leq \sigma\|\mathbf{M}_{\text{off}}\|_{\text{F}}$. Putting the pieces together with the union bound, the inequality $\|\mathbf{M}_{\text{off}}\|_{\text{F}} + \|\text{diag}(\mathbf{M})\|_2 \leq \sqrt{2}[\|\mathbf{M}_{\text{off}}\|_{\text{F}}^2 + \|\text{diag}(\mathbf{M})\|_2^2]^{1/2} = \sqrt{2}\|\mathbf{M}\|_{\text{F}}$, and the simplification $45\sqrt{2} \leq 64$, we obtain the desired conclusion. □

# 7

# Conclusion

The main thread of this thesis is structured sparsity which we have studied from various angles, with statistical, algorithmic and applied considerations. We have first introduced structured sparsity-inducing norms. They are capable of encoding high-order structural information about the problem at hand, unlike standard sparsity-promoting penalties based on cardinality. We have precisely characterized what type of prior knowledge they can model, and have derived analysis for consistent structured variable selection, in both low- and high-dimensional settings.

Our second contribution lies in the use of structured sparsity-inducing norms within the framework of dictionary learning. We have shown how the nonzero patterns of the decompositions and/or dictionary elements can self-organize to best adapt to the considered class of signals. In addition, we have proposed an efficient and simple optimization tool built upon nested block-coordinate descent procedures. These points are eventually illustrated by an application to face recognition where we learn localized features more robust to occlusions.

The third section of the thesis is dedicated to convex optimization. It proves that problems regularized by structured sparsity-inducing norms can be efficiently solved by proximal methods. In particular, we have shown how the proximity operator can be computed exactly in various settings, based on dual formulations. This algorithmic development notably paves the way for numerous large-scale applications, ranging from topic models to background subtraction.

The fourth chapter discusses at greater length two applications of structured sparsity to neuroimaging, in both supervised and unsupervised settings. We have first considered the inter-subject prediction of sizes of objects from fMRI signals. A sparse hierarchical regularization is shown to well capture the multi-scale aspects of the data, while dealing properly with the inter-subject variability. Furthermore, we have studied brain resting-state time series where structured dictionary learning has proved to be a promising tool.

The last chapter of this thesis is slightly next to the scope of structured sparsity developed throughout the four previous chapters. It studies the local minima of the standard non-convex formulation of sparse coding. In particular, we have proposed a non-asymptotic analysis within a probabilistic model of sparse noisy signals, thus extending earlier work limited to noiseless settings and/or under-complete dictionaries. Our main result notably shows that, with high probability, a local minimum indeed exists along some curves passing through the reference dictionary generating the signals.

This thesis has explored several questions related to structured sparsity, hence providing with some answers, either full or partial. These answers raise in turn new challenges and new directions for future research which we now discuss.

Throughout this thesis, we make the key assumption that we have at disposal some structural prior knowledge that justifies the use of our norms. This point can of course be debatable. Even though we have seen that dictionary learning relaxes in some sense this constraint, it is rather rare to have an *exact* match between the structure we *can* actually model and what the real data dictate. This remark suggests that an important question is to be able to *learn from the data* the adequate structure. This is an obviously difficult question, notably due to the combinatorial and discrete nature of the problem. Further exploring submodular tools (Bach, 2010a,b) in this case seems an interesting avenue of research.

Moreover, this thesis focuses on the use of structured sparsity only via the regularization. It would also be of interest to consider structured data-fitting terms, for instance, to model specific forms of noise and outliers (Liu et al., 2010a).

On the applied side, the last few years have been witnessing an increasing number of applications of structured sparsity in various fields. Only recently, the domain of natural language processing (NLP) has been considered (Martins et al., 2011). Further developments of structured sparsity in NLP seem natural since it is a field where a great deal of structural information is available, e.g., hierarchical in $n$-grams models. Moreover, the large-scale problems typical from NLP could benefit from recent stochastic extensions of proximal methods (Duchi and Singer, 2009; Hu et al., 2009; Xiao, 2010). The contribution from this thesis regarding the computation of proximal operators for structured norms could be directly applied.

# A

# Proofs

This section of the appendix gathers the proofs and some technical elements encountered throughout the manuscript.

## A.1 Proofs and Technical Elements of Chapter 1

### A.1.1 Proof of Proposition 5

We recall that $L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{w}^\top \mathbf{x}_i)$. Since $\mathbf{w} \mapsto \Omega(\mathbf{w})$ is convex and goes to infinity when $\|\mathbf{w}\|_2$ goes to infinity, and since $L$ is lower bounded, by Weierstrass' theorem, the problem in (2.2) admits at least one global solution.

• *First case:* $\mathbf{Q}$ *invertible.* The Hessian of $L$ is

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \frac{\partial^2 \ell}{\partial y'^2}(y_i, \mathbf{w}^\top \mathbf{x}_i).$$

It is positive definite since $\mathbf{Q}$ is positive definite and $\min_{i\in\{1,\dots,n\}} \frac{\partial^2 \ell}{\partial y'^2}(y_i, \mathbf{w}^\top \mathbf{x}_i) > 0$. So $L$ is strictly convex. Consequently the objective function $L + \mu\Omega$ is strictly convex, hence the uniqueness of its minimizer.

• *Second case:* $\{1, \dots, p\}$ *belongs to* $\mathcal{G}$. We prove the uniqueness by contradiction. Assume that the problem in (2.2) admits two different solutions $\mathbf{w}$ and $\tilde{\mathbf{w}}$. Then one of the two solutions is different from 0, say $\mathbf{w} \neq 0$.

By convexity, it means that any point of the segment $[\mathbf{w}, \tilde{\mathbf{w}}] = \{a\mathbf{w} + (1-a)\tilde{\mathbf{w}};\ a \in [0,1]\}$ also minimizes the objective function $L + \mu\Omega$. Since both $L$ and $\mu\Omega$ are convex functions, it means that they are both linear when restricted to $[\mathbf{w}, \tilde{\mathbf{w}}]$.

Now, $\mu\Omega$ is only linear on segments of the form $[\mathbf{v}, t\mathbf{v}]$ with $\mathbf{v} \in \mathbb{R}^p$ and $t > 0$. So we necessarily have $\tilde{\mathbf{w}} = t\mathbf{w}$ for some positive $t$. We now show that $L$ is strictly convex on $[\mathbf{w}, t\mathbf{w}]$, which will contradict that it is linear on $[\mathbf{w}, \tilde{\mathbf{w}}]$. Let $E = \mathrm{Span}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ and $E^\perp$ be the orthogonal of $E$ in $\mathbb{R}^p$. The vector $\mathbf{w}$ can be decomposed in $\mathbf{w} = \mathbf{w}' + \mathbf{w}''$ with $\mathbf{w}' \in E$ and $\mathbf{w}'' \in E^\perp$. Note that we have $\mathbf{w}' \neq 0$ (since if it was equal to 0, $\mathbf{w}''$ would be the minimizer of $\mu\Omega$, which would imply $\mathbf{w}'' = 0$ and contradict $\mathbf{w} \neq 0$). We thus have $(\mathbf{w}^\top \mathbf{x}_1, \dots, \mathbf{w}^\top \mathbf{x}_n) = (\mathbf{w}'^\top \mathbf{x}_1, \dots, \mathbf{w}'^\top \mathbf{x}_n) \neq 0$.

This implies that the function $s \mapsto \ell(y_i, s\mathbf{w}^\top \mathbf{x}_i)$ is a polynomial of degree 2. So it is not linear. This contradicts the existence of two different solutions, and concludes the proof of uniqueness.

**Remark 7**
*Still by using that a sum of convex functions is constant on a segment if and only if the functions are linear on this segment, the proof can be extended in order to replace the alternative assumption "$\{1,\dots,p\}$ belongs to $\mathcal{G}$" by the weaker but more involved assumption: for any $(j,k) \in \{1,\dots,p\}^2$, there exists a group $g \in \mathcal{G}$ which contains both $j$ and $k$.*

### A.1.2 Proof of Theorem 1

For $\mathbf{w} \in \mathbb{R}^p$, we denote by $Z(w)$ its zero pattern (i.e., the indices of zero-components of $\mathbf{w}$). To prove the result, it suffices to prove that for any set $I \subset \{1,\dots,p\}$ with $I^c \notin \mathcal{Z}$ and $|I| \leq k-1$, the probability of

$$\mathcal{E}_I = \{Y \in \mathbb{R}^n\text{: there exists } \mathbf{w} \text{ solution of the problem in (2.2) with } Z(w) = I^c\}$$

is equal to 0. We will prove this by contradiction: assume that there exists a set $I \subset \{1,\dots,p\}$ with $I^c \notin \mathcal{Z}$, $|I| \leq k-1$ and $\Pr(\mathcal{E}_I) > 0$. Since $I^c \notin \mathcal{Z}$, there exists $\alpha \in \mathrm{Hull}(I) \setminus I$. Let $J = I \cup \{\alpha\}$ and $\mathcal{G}_I = \{g \in \mathcal{G} : g \cap I \neq \emptyset\}$ be the set of active groups. Define $\mathbb{R}^J = \{\mathbf{w} \in \mathbb{R}^p : \mathbf{w}_{J^c} = 0\}$. The restrictions $L_J : \mathbb{R}^J \to \mathbb{R}$ and $\Omega_J : \mathbb{R}^J \to \mathbb{R}$ of $L$ and $\Omega$ are continuously differentiable functions on $\{\mathbf{w} \in \mathbb{R}^J : \mathbf{w}_I \neq 0\}$ with respective gradients

$$\nabla L_J(\mathbf{w}) = \left(\frac{\partial L_J}{\partial \mathbf{w}_j}(\mathbf{w})\right)_{j\in J}^\top \quad \text{and} \quad \nabla \Omega_J(\mathbf{w}) = \left(\mathbf{w}_j\Big(\sum_{\substack{g\in\mathcal{G}_I,\\ g\ni j}} (\omega_j^g)^2 \|\omega^g \circ \mathbf{w}\|_2^{-1}\Big)\right)_{j\in J}^\top.$$

Let $f(\mathbf{w}, Y) = \nabla L_J(\mathbf{w}) + \mu \nabla \Omega_J(\mathbf{w})$, where the dependence in $Y$ of $f(\mathbf{w}, Y)$ is hidden in $\nabla L_J(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i)_J \frac{\partial \ell}{\partial y'}(y_i, \mathbf{w}^\top \mathbf{x}_i)$.

For $Y \in \mathcal{E}_I$, there exists $\mathbf{w} \in \mathbb{R}^J$ with $Z(\mathbf{w}) = I^c$, which minimizes the convex function $L_J + \mu\Omega_J$. The vector $\mathbf{w}$ satisfies $f(\mathbf{w}, Y) = 0$. So we have proved $\mathcal{E}_I \subset \mathcal{E}'_I$, where

$$\mathcal{E}'_I = \{Y \in \mathbb{R}^n : \text{ there exists } \mathbf{w} \in \mathbb{R}^J \text{ with } Z(\mathbf{w}) = I^c \text{ and } f(\mathbf{w}, Y) = 0\}.$$

Let $\tilde{y} \in \mathcal{E}_I$. Consider the equation $f(\mathbf{w}, \tilde{y}) = 0$ on $\{\mathbf{w} \in \mathbb{R}^J : \mathbf{w}_j \neq 0 \text{ for any } j \in I\}$. By construction, we have $|J| \leq k$, and thus, by assumption, the matrix

$$\mathbf{X}_J = \left((\mathbf{x}_1)_J, ..., (\mathbf{x}_n)_J\right)^\top \in \mathbb{R}^{n\times|J|}$$

has rank $|J|$. As in the proof of Proposition A.1.1, this implies that the function $L_J$ is strictly convex, and then, the uniqueness of the minimizer of $L_J + \mu\Omega$, and also the uniqueness of the point at which the gradient of this function vanishes. So the equation $f(\mathbf{w}, \tilde{y}) = 0$ on $\{\mathbf{w} \in \mathbb{R}^J : \mathbf{w}_j \neq 0 \text{ for any } j \in I\}$ has a unique solution, which we will write $\mathbf{w}_J^{\tilde{y}}$.

On a small enough ball around $(\mathbf{w}_J^{\tilde{y}}, \tilde{y})$, $f$ is continuously differentiable since none of the norms vanishes at $\mathbf{w}_J^{\tilde{y}}$. Let $(f_j)_{j\in J}$ be the components of $f$ and $\mathbf{H}_{JJ} = (\frac{\partial f_j}{\partial \mathbf{w}_k})_{j\in J, k\in J}$. The matrix $\mathbf{H}_{JJ}$ is actually the sum of:

a) the Hessian of $L_J$, which is positive definite (still from the same argument as in the proof of Theorem A.1.1),

b) the Hessian of the norm $\Omega_J$ around $(\mathbf{w}_J^{\tilde{y}}, \tilde{y})$ that is positive semidefinite on this small ball according to the Hessian characterization of convexity (Borwein and Lewis, 2006, Theorem 3.1.11).

Consequently, $\mathbf{H}_{JJ}$ is invertible. We can now apply the implicit function theorem to obtain that for $Y$ in a neighborhood of $\tilde{y}$,

$$\mathbf{w}^Y = \psi(Y),$$

with $\psi = (\psi_j)_{j \in J}$ a continuously differentiable function satisfying the matricial relation

$$(\dots, \nabla\psi_j, \dots)\mathbf{H}_{JJ} + (\dots, \nabla_y f_j, \dots) = 0.$$

Let $\mathbf{c}_\alpha$ denote the column vector of $\mathbf{H}_{JJ}^{-1}$ corresponding to the index $\alpha$, and let $\mathbf{D}$ the diagonal matrix whose $(i,i)$-th element is $\frac{\partial^2 \ell}{\partial y \partial y'}(y_i, \mathbf{x}_i^\top \mathbf{w}^Y)$. Since $(\dots, \nabla_y f_j, \dots) = \frac{1}{n}\mathbf{D}\mathbf{X}_J$, we have

$$\nabla\psi_\alpha = -\frac{1}{n}\mathbf{D}\mathbf{X}_J\mathbf{c}_\alpha.$$

Now, since $\mathbf{X}_J$ has full rank, $\mathbf{c}_\alpha \neq \mathbf{0}$ and none of the diagonal elements of $\mathbf{D}$ is null (by assumption on $\ell$), we have $\nabla\psi_\alpha \neq \mathbf{0}$. Without loss of generality, we may assume that $\partial\psi_\alpha/\partial y_1 \neq 0$ on a neighborhood of $\tilde{y}$.

We can apply again the implicit function theorem to show that on an open ball in $\mathbb{R}^n$ centered at $\tilde{y}$, say $\mathcal{B}_{\tilde{y}}$, the solution to $\psi_\alpha(Y) = 0$ can be written $y_1 = \varphi(y_2, \dots, y_n)$ with $\varphi$ a continuously differentiable function.

By Fubini's theorem and by using the fact that the Lebesgue measure of a singleton in $\mathbb{R}^n$ equals zero, we get that the set $A(\tilde{y}) = \{Y \in \mathcal{B}_{\tilde{y}} : \psi_\alpha(Y) = 0\}$ has thus zero probability. Let $\mathcal{S} \subset \mathcal{E}_I$ be a compact set. We thus have $\mathcal{S} \subset \mathcal{E}_I'$.

By compacity, the set $\mathcal{S}$ can be covered by a finite number of ball $\mathcal{B}_{\tilde{y}}$. So there exist $\tilde{y}_1, \dots, \tilde{y}_m$ such that we have $\mathcal{S} \subset A(\tilde{y}_1) \cup \dots \cup A(\tilde{y}_m)$. Consequently, we have $\Pr(\mathcal{S}) = 0$.

Since this holds for any compact set in $\mathcal{E}_I$ and since the Lebesgue measure is regular, we have $\Pr(\mathcal{E}_I) = 0$, which contradicts the definition of $I$, and concludes the proof.

### A.1.3 Proof of the minimality of the Backward procedure (see Algorithm 1)

There are essentially two points to show: (1) $\mathcal{G}$ spans $\mathcal{Z}$, and (2) $\mathcal{G}$ is minimal.

The first point can be shown by a proof by recurrence on the depth of the DAG. At step $t$, the base $\mathcal{G}^{(t)}$ verifies $\{\bigcup_{g \in \mathcal{G}'} G,\ \forall \mathcal{G}' \subseteq \mathcal{G}^{(t)}\} = \{g \in \mathcal{Z}, |g| \leq t\}$ because an element $g \in \mathcal{Z}$ is either the union of itself or the union of elements strictly smaller. The initialization $t = \min_{g \in \mathcal{Z}} |g|$ is easily verified, the leafs of the DAG being necessarily in $\mathcal{G}$.

As for the second point, we proceed by contradiction. If there exists another base $\mathcal{G}^*$ that spans $\mathcal{Z}$ such that $\mathcal{G}^* \subset \mathcal{G}$, then

$$\exists\, e \in \mathcal{G},\ e \notin \mathcal{G}^*.$$

By definition of the set $\mathcal{Z}$, there exists in turn $\mathcal{G}' \subseteq \mathcal{G}^*$, $\mathcal{G}' \neq \{e\}$ (otherwise, $e$ would belong to $\mathcal{G}^*$), verifying $e = \bigcup_{g \in \mathcal{G}'} g$, which is impossible by construction of $\mathcal{G}$ whose members cannot be the union of elements of $\mathcal{Z}$.

### A.1.4 Proof of Proposition 6

The proposition comes from a classic result of Fenchel Duality (Borwein and Lewis, 2006, Theorem 3.3.5 and Exercise 3.3.9) when we consider the convex function

$$h_J : \mathbf{w}_J \mapsto \frac{\lambda}{2} \left[\Omega_J(\mathbf{w}_J)\right]^2,$$

whose Fenchel conjugate $h_J^*$ is given by $\boldsymbol{\kappa}_J \mapsto \frac{1}{2\lambda} \left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2$ (Boyd and Vandenberghe, 2004, example 3.27). Since the set

$$\{\mathbf{w}_J \in \mathbb{R}^{|J|};\ h_J(\mathbf{w}_J) < \infty\} \cap \{\mathbf{w}_J \in \mathbb{R}^{|J|};\ L_J(\mathbf{w}_J) < \infty \text{ and } L_J \text{ is continuous at } \mathbf{w}_J\}$$

is not empty, we get the first part of the proposition. Moreover, the primal-dual variables $\{\mathbf{w}_J, \boldsymbol{\kappa}_J\}$ is optimal if and only if

$$\begin{cases} -\boldsymbol{\kappa}_J & \in & \partial L_J(\mathbf{w}_J), \\ \boldsymbol{\kappa}_J & \in & \partial[\frac{\lambda}{2} \left[\Omega_J(\mathbf{w}_J)\right]^2] = \lambda \Omega_J(\mathbf{w}_J) \partial \Omega_J(\mathbf{w}_J), \end{cases}$$

where $\partial \Omega_J(\mathbf{w}_J)$ denotes the subdifferential of $\Omega_J$ at $\mathbf{w}_J$. The differentiability of $L_J$ at $\mathbf{w}_J$ then gives $\partial L_J(\mathbf{w}_J) = \{\nabla L_J(\mathbf{w}_J)\}$. It now remains to show that

$$\boldsymbol{\kappa}_J \in \lambda \Omega_J(\mathbf{w}_J) \partial \Omega_J(\mathbf{w}_J) \tag{A.1}$$

is equivalent to

$$\mathbf{w}_J^\top \boldsymbol{\kappa}_J = \frac{1}{\lambda} \left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 = \lambda \left[\Omega_J(\mathbf{w}_J)\right]^2. \tag{A.2}$$

As a starting point, the Fenchel-Young inequality (Borwein and Lewis, 2006, Proposition 3.3.4) gives the equivalence between (A.1) and

$$\frac{\lambda}{2} \left[\Omega_J(\mathbf{w}_J)\right]^2 + \frac{1}{2\lambda} \left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 = \mathbf{w}_J^\top \boldsymbol{\kappa}_J. \tag{A.3}$$

In addition, we have (Rockafellar, 1997)

$$\partial \Omega_J(\mathbf{w}_J) = \{\mathbf{u}_J \in \mathbb{R}^{|J|}; \mathbf{u}_J^\top \mathbf{w}_J = \Omega_J(\mathbf{w}_J) \ \text{ and } \ \Omega_J^*(\mathbf{u}_J) \leq 1\}. \tag{A.4}$$

Thus, if $\boldsymbol{\kappa}_J \in \lambda \Omega_J(\mathbf{w}_J) \partial \Omega_J(\mathbf{w}_J)$ then $\mathbf{w}_J^\top \boldsymbol{\kappa}_J = \lambda \left[\Omega_J(\mathbf{w}_J)\right]^2$. Combined with (A.3), we obtain $w_J^\top \boldsymbol{\kappa}_J = \frac{1}{\lambda} \left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2$.

Reciprocally, starting from (A.2), we notably have

$$\mathbf{w}_J^\top \boldsymbol{\kappa}_J = \lambda \left[\Omega_J(\mathbf{w}_J)\right]^2 .$$

In light of (A.4), it suffices to check that $\Omega_J^*(\boldsymbol{\kappa}_J) \leq \lambda \Omega_J(\mathbf{w}_J)$ in order to have Eq. (A.1). Combining Eq. (A.2) with the definition of the dual norm, it comes

$$\frac{1}{\lambda}\left[\Omega_J^*(\boldsymbol{\kappa}_J)\right]^2 = \mathbf{w}_J^\top \boldsymbol{\kappa}_J \leq \Omega_J^*(\boldsymbol{\kappa}_J)\Omega_J(\mathbf{w}_J),$$

which concludes the proof of the equivalence between Eq. (A.1) and Eq. (A.2).

### A.1.5 Proofs of Propositions 7 and 8

In order to check that the reduced solution $\mathbf{w}_J$ is optimal for the full problem in Eq. (2.5), we complete with zeros on $J^c$ to define $\mathbf{w}$, compute $\boldsymbol{\kappa} = -\nabla L(\mathbf{w})$, which is such that $\boldsymbol{\kappa}_J = -\nabla L_J(\mathbf{w}_J)$, and get a duality gap for the full problem equal to

$$\frac{1}{2\lambda}\left(\left[\Omega^*(\boldsymbol{\kappa})\right]^2 - \lambda \mathbf{w}_J^\top \boldsymbol{\kappa}_J\right).$$

By designing upper and lower bounds for $\Omega^*(\boldsymbol{\kappa})$, we get sufficient and necessary conditions.

#### Proof of Proposition 7

Let us suppose that $\mathbf{w}^* = \binom{\mathbf{w}_J^*}{\mathbf{0}_{J^c}}$ is optimal for the full problem in Eq. (2.5). Following the same derivation as in Lemma 32 (up to the squaring of the regularization $\Omega$), we have that $\mathbf{w}^*$ is a solution of Eq. (2.5) if and only if for all $\mathbf{u} \in \mathbb{R}^p$,

$$\mathbf{u}^\top \nabla L(\mathbf{w}^*) + \lambda \Omega(\mathbf{w}^*)(\mathbf{u}_J^\top \mathbf{r}_J + (\Omega_J^c)[\mathbf{u}_{J^c}]) \geq 0,$$

with

$$\mathbf{r} = \sum_{g \in \mathcal{G}_J} \frac{\omega^g \circ \omega^g \circ \mathbf{w}^*}{\|\omega^g \circ \mathbf{w}^*\|_2}.$$

We project the optimality condition onto the variables that can possibly enter the active set, i.e., the variables in $\Pi_{\mathcal{P}}(J)$. Thus, for each $K \in \Pi_{\mathcal{P}}(J)$, we have for all $\mathbf{u}_{K\backslash J} \in \mathbb{R}^{|K\backslash J|}$,

$$\mathbf{u}_{K\backslash J}^\top \nabla L(\mathbf{w}^*)_{K\backslash J} + \lambda \Omega(\mathbf{w}^*) \sum_{g \in \mathcal{G}_{K\backslash J} \cap (\mathcal{G}_J)^c} \|\omega_{K\backslash J}^g \circ \mathbf{u}_{g \cap K\backslash J}\|_2 \geq 0.$$

By combining Lemma 31 and the fact that $\mathcal{G}_{K\backslash J} \cap (\mathcal{G}_J)^c = \mathcal{G}_K \backslash \mathcal{G}_J$, we have for all $G \in \mathcal{G}_K \backslash \mathcal{G}_J$, $K\backslash J \subseteq g$ and therefore $\mathbf{u}_{g \cap K\backslash J} = \mathbf{u}_{K\backslash J}$. Since we cannot compute the dual norm of $\mathbf{u}_{K\backslash J} \mapsto \|\omega_{K\backslash J}^g \circ \mathbf{u}_{K\backslash J}\|_2$ in closed-form, we instead use the following upperbound

$$\|\omega_{K\backslash J}^g \circ u_{K\backslash J}\|_2 \leq \|\omega_{K\backslash J}^g\|_\infty \|\mathbf{u}_{K\backslash J}\|_2,$$

so that we get for all $\mathbf{u}_{K\setminus J} \in \mathbb{R}^{|K\setminus J|}$,

$$\mathbf{u}_{K\setminus J}^\top \nabla L(\mathbf{w}^*)_{K\setminus J} + \lambda\Omega(\mathbf{w}^*) \sum_{g\in\mathcal{G}_K\setminus\mathcal{G}_J} \|\omega^g_{K\setminus J}\|_\infty \|\mathbf{u}_{K\setminus J}\|_2 \geq 0.$$

Finally, Proposition 6 gives $\lambda\Omega(\mathbf{w}^*) = \{-\lambda \mathbf{w}^{*\top}\nabla L(\mathbf{w}^*)\}^{\frac{1}{2}}$, which leads to the desired result.

**Proof of Proposition 8**

The goal of the proof is to upper bound the dual norm $\Omega^*(\boldsymbol{\kappa})$ by taking advantage of the structure of $\mathcal{G}$; we first show how we can upper bound $\Omega^*(\boldsymbol{\kappa})$ by $(\Omega^c_J)^*[\boldsymbol{\kappa}_{J^c}]$. We indeed have:

$$\begin{aligned}
\Omega^*(\boldsymbol{\kappa}) &= \max_{\sum_{g\in\mathcal{G}_J}\|\omega^g\circ\mathbf{v}\|_2+\sum_{g\in(\mathcal{G}_J)^c}\|\omega^g\circ\mathbf{v}\|_2\leq 1} \mathbf{v}^\top\boldsymbol{\kappa} \\
&\leq \max_{\sum_{g\in\mathcal{G}_J}\|\omega^g_J\circ\mathbf{v}_J\|_2+\sum_{g\in(\mathcal{G}_J)^c}\|\omega^g\circ\mathbf{v}\|_2\leq 1} \mathbf{v}^\top\boldsymbol{\kappa} \\
&= \max_{\Omega_J(\mathbf{v}_J)+(\Omega^c_J)(\mathbf{v}_{J^c})\leq 1} \mathbf{v}^\top\boldsymbol{\kappa} \\
&= \max\left\{\Omega^*_J(\boldsymbol{\kappa}_J), (\Omega^c_J)^*[\boldsymbol{\kappa}_{J^c}]\right\},
\end{aligned}$$

where in the last line, we use Lemma 33. Thus the duality gap is less than

$$\frac{1}{2\lambda}\left([\Omega^*(\boldsymbol{\kappa})]^2 - [\Omega^*_J(\boldsymbol{\kappa}_J)]^2\right) \leq \frac{1}{2\lambda}\max\{0, [(\Omega^c_J)^*[\boldsymbol{\kappa}_{J^c}]]^2 - [\Omega^*_J(\boldsymbol{\kappa}_J)]^2\},$$

and a sufficient condition for the duality gap to be smaller than $\varepsilon$ is

$$(\Omega^c_J)^*[\boldsymbol{\kappa}_{J^c}] \leq (2\lambda\varepsilon + [\Omega^*_J(\boldsymbol{\kappa}_J)]^2)^{\frac{1}{2}}.$$

Using Proposition 6, we have $-\lambda\mathbf{w}^\top\nabla L(\mathbf{w}) = [\Omega^*_J(\boldsymbol{\kappa}_J)]^2$ and we get the right-hand side of Proposition 8. It now remains to upper bound $(\Omega^c_J)^*[\boldsymbol{\kappa}_{J^c}]$. To this end, we call upon Lemma 29 to obtain:

$$(\Omega^c_J)^*[\boldsymbol{\kappa}_{J^c}] \leq \max_{g\in(\mathcal{G}_J)^c}\left\{\sum_{j\in g}\left\{\frac{\boldsymbol{\kappa}_j}{\sum_{h\in j, h\in(\mathcal{G}_J)^c}\omega^h_j}\right\}^2\right\}^{\frac{1}{2}}.$$

Among all groups $g \in (\mathcal{G}_J)^c$, the ones with the maximum values are the largest ones, i.e., those in the fringe groups $\mathcal{F}_J = \{g \in (\mathcal{G}_J)^c \;;\; \nexists g' \in (\mathcal{G}_J)^c, g \subseteq g'\}$. This argument leads to the result of Proposition 8.

### A.1.6 Proof of Theorem 2

*Necessary condition:* We mostly follow the proof of Zou (2006); Bach (2008b). Let $\hat{\mathbf{w}} \in \mathbb{R}^p$ be the unique solution of

$$\min_{\mathbf{w}\in\mathbb{R}^p} L(\mathbf{w}) + \mu\,\Omega(\mathbf{w}) = \min_{\mathbf{w}\in\mathbb{R}^p} F(\mathbf{w}).$$

The quantity $\hat{\Delta} = (\hat{\mathbf{w}} - \mathbf{w}^\star)/\mu$ is the minimizer of $\tilde{F}$ defined as

$$\tilde{F}(\Delta) = \frac{1}{2}\Delta^\top \mathbf{Q}\Delta - \mu^{-1}\mathbf{q}^\top\Delta + \mu^{-1}\left[\Omega(\mathbf{w}^\star + \mu\Delta) - \Omega(\mathbf{w}^\star)\right],$$

where $\mathbf{q} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{x}_i$. The random variable $\mu^{-1}\mathbf{q}^\top\Delta$ is a centered Gaussian with variance $\sqrt{\Delta^\top\mathbf{Q}\Delta}/(n\mu^2)$. Since $\mathbf{Q} \to \mathbf{Q}^\star$, we obtain that for all $\Delta \in \mathbb{R}^p$,

$$\mu^{-1}\mathbf{q}^\top\Delta = o_p(1).$$

Since $\mu \to 0$, we also have by taking the directional derivative of $\Omega$ at $\mathbf{w}$ in the direction of $\Delta$

$$\mu^{-1}\left[\Omega(\mathbf{w}^\star + \mu\Delta) - \Omega(\mathbf{w}^\star)\right] = [\mathbf{r}^\star_{\mathrm{J}^\star}]^\top\Delta_{\mathrm{J}^\star} + \Omega_{[\mathrm{J}^\star]^c}(\Delta_{[\mathrm{J}^\star]^c}) + o(1),$$

so that for all $\Delta \in \mathbb{R}^p$

$$\tilde{F}(\Delta) = \Delta^\top\mathbf{Q}^\star\Delta + [\mathbf{r}^\star_{\mathrm{J}^\star}]^\top\Delta^\star_{\mathrm{J}} + \Omega_{[\mathrm{J}^\star]^c}(\Delta_{[\mathrm{J}^\star]^c}) + o_p(1) = \tilde{F}_{\lim}(\Delta) + o_p(1).$$

The limiting function $\tilde{F}_{\lim}$ being stricly convex (because $\mathbf{Q}^\star \succ 0$) and $\tilde{F}$ being convex, we have that the minimizer $\hat{\Delta}$ of $\tilde{F}$ tends in probability to the unique minimizer of $\tilde{F}_{\lim}$ (Fu and Knight, 2000) referred to as $\Delta^*$.

By assumption, with probability tending to one, we have $\mathrm{J}^\star = \{j \in [\![1;p]\!], \hat{\mathbf{w}}_j \neq 0\}$, hence for any $j \in [\mathrm{J}^\star]^c$ $\mu\hat{\Delta}_j = (\hat{\mathbf{w}} - \mathbf{w}^\star)_j = 0$. This implies that the nonrandom vector $\Delta^*$ verifies $\Delta^*_{[\mathrm{J}^\star]^c} = 0$.

As a consequence, $\Delta^*_{\mathrm{J}^\star}$ minimizes $\Delta^\top_{\mathrm{J}^\star}\mathbf{Q}^\star_{\mathrm{J}^\star\mathrm{J}^\star}\Delta_{\mathrm{J}^\star} + [\mathbf{r}^\star_{\mathrm{J}^\star}]^\top\Delta^\star_{\mathrm{J}}$, hence $\mathbf{r}^\star_{\mathrm{J}^\star} = -\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}\Delta^*_{\mathrm{J}^\star}$. Besides, since $\Delta^*$ is the minimizer of $\tilde{F}_{\lim}$, by taking the directional derivatives as in the proof of Lemma 32, we have

$$(\Omega_{[\mathrm{J}^\star]^c})^*[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\Delta^*_{\mathrm{J}^\star}] \leq 1.$$

This gives the necessary condition.

*Sufficient condition:* We turn to the sufficient condition. We first consider the problem reduced to the hull $\mathrm{J}^\star$,

$$\min_{\mathbf{w}\in\mathbb{R}^{|\mathrm{J}^\star|}} L^\star_{\mathrm{J}}(\mathbf{w}_{\mathrm{J}^\star}) + \mu\Omega_{\mathrm{J}}(\mathbf{w}_{\mathrm{J}^\star}).$$

that is strongly convex since $\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}$ is positive definite and thus admits a unique solution $\hat{\mathbf{w}}_{\mathrm{J}^\star}$. With similar arguments as the ones used in the necessary condition, we can show that $\hat{\mathbf{w}}_{\mathrm{J}^\star}$ tends in probability to the true vector $\mathbf{w}_{\mathrm{J}^\star}$. We now consider the vector $\hat{\mathbf{w}} \in \mathbb{R}^p$ which is the vector $\hat{\mathbf{w}}_{\mathrm{J}^\star}$ padded with zeros on $[\mathrm{J}^\star]^c$. Since, from Theorem 1, we almost surely have $\mathrm{Hull}(\{j \in [\![1;p]\!], \hat{\mathbf{w}}_j \neq 0\}) = \{j \in [\![1;p]\!], \hat{\mathbf{w}}_j \neq 0\}$, we have already that the vector $\hat{\mathbf{w}}$ consistently estimates the hull of $\mathbf{w}^\star$ and we have that $\hat{\mathbf{w}}$ tends in probability

to $\mathbf{w}^\star$. From now on, we consider that $\hat{\mathbf{w}}$ is sufficiently close to $\mathbf{w}^\star$, so that for any $g \in \mathcal{G}_{\mathrm{J}^\star}$, $\|\omega^g \circ \hat{\mathbf{w}}\|_2 \neq 0$. We may thus introduce

$$\hat{\mathbf{r}} = \sum_{g \in \mathcal{G}_{\mathrm{J}^\star}} \frac{\omega^g \circ \omega^g \circ \hat{\mathbf{w}}}{\|\omega^g \circ \hat{\mathbf{w}}\|_2}.$$

It remains to show that $\hat{\mathbf{w}}$ is indeed optimal for the full problem (that admits a unique solution due to the positiveness of $\mathbf{Q}$). By construction, the optimality condition (see Lemma 32) relative to the active variables $\mathrm{J}^\star$ is already verified. More precisely, we have

$$\nabla L(\hat{\mathbf{w}})_{\mathrm{J}^\star} + \mu\, \hat{\mathbf{r}}_{\mathrm{J}^\star} = \mathbf{Q}_{\mathrm{J}^\star \mathrm{J}^\star}(\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}) - \mathbf{q}_{\mathrm{J}^\star} + \mu\, \hat{\mathbf{r}}_{\mathrm{J}^\star} = 0.$$

Moreover, for all $\mathbf{u}_{[\mathrm{J}^\star]^c} \in \mathbb{R}^{|[\mathrm{J}^\star]^c|}$, by using the previous expression and the invertibily of $\mathbf{Q}$, we have

$$\mathbf{u}^\top_{[\mathrm{J}^\star]^c} \nabla L(\hat{\mathbf{w}})_{[\mathrm{J}^\star]^c} = u^\top_{[\mathrm{J}^\star]^c} \left\{ -\mu\, \mathbf{Q}_{[\mathrm{J}^\star]^c \mathrm{J}^\star} \mathbf{Q}^{-1}_{\mathrm{J}^\star \mathrm{J}^\star} \hat{\mathbf{r}}_{\mathrm{J}^\star} + \mathbf{Q}_{[\mathrm{J}^\star]^c \mathrm{J}^\star} \mathbf{Q}^{-1}_{\mathrm{J}^\star \mathrm{J}^\star} \mathbf{q}_{\mathrm{J}^\star} - \mathbf{q}_{[\mathrm{J}^\star]^c} \right\}.$$

The terms related to the noise vanish, having actually $\mathbf{q} = o_p(1)$. Since $\mathbf{Q} \to \mathbf{Q}^\star$ and $\hat{\mathbf{r}}_{\mathrm{J}^\star} \to \mathbf{r}^\star_{\mathrm{J}^\star}$, we get for all $\mathbf{u}_{[\mathrm{J}^\star]^c} \in \mathbb{R}^{|[\mathrm{J}^\star]^c|}$

$$\mathbf{u}^\top_{[\mathrm{J}^\star]^c} \nabla L(\hat{\mathbf{w}})_{[\mathrm{J}^\star]^c} = -\mu\, \mathbf{u}^\top_{[\mathrm{J}^\star]^c} \left\{ \mathbf{Q}^\star_{[\mathrm{J}^\star]^\star \mathrm{J}^\star} [\mathbf{Q}^\star_{\mathrm{J}^\star \mathrm{J}^\star}]^{-1} \mathbf{r}^\star_{\mathrm{J}^\star} \right\} + o_p(\mu).$$

Since we assume $(\Omega_{[\mathrm{J}^\star]^c})^*[\mathbf{Q}^\star_{[\mathrm{J}^\star]^c \mathrm{J}^\star} [\mathbf{Q}^\star_{\mathrm{J}^\star \mathrm{J}^\star}]^{-1} \mathbf{r}^\star_{\mathrm{J}^\star}] < 1$, we obtain

$$-\mathbf{u}^\top_{[\mathrm{J}^\star]^c} \nabla L(\hat{\mathbf{w}})_{[\mathrm{J}^\star]^c} < \mu(\Omega^c_{\mathrm{J}^\star})[\mathbf{u}_{[\mathrm{J}^\star]^c}] + o_p(\mu),$$

which proves the optimality condition of Lemma 32 relative to the inactive variables: $\hat{\mathbf{w}}$ is therefore optimal for the full problem.

### A.1.7 Proof of Theorem 3

Since our analysis takes place in a finite-dimensional space, all the norms defined on this space are equivalent. Therefore, we introduce the equivalence parameters $a(\mathrm{J}^\star), A(\mathrm{J}^\star) > 0$ such that

$$\forall \mathbf{u} \in \mathbb{R}^{|\mathrm{J}^\star|},\ a(\mathrm{J}^\star)\|\mathbf{u}\|_1 \leq \Omega_{\mathrm{J}^\star}[\mathbf{u}] \leq A(\mathrm{J}^\star)\|\mathbf{u}\|_1.$$

We similarly define $a([\mathrm{J}^\star]^c), A([\mathrm{J}^\star]^c) > 0$ for the norm $(\Omega^c_{\mathrm{J}^\star})$ on $\mathbb{R}^{|[\mathrm{J}^\star]^c|}$. In addition, we immediatly get by order-reversing:

$$\forall \mathbf{u} \in \mathbb{R}^{|\mathrm{J}^\star|},\ A(\mathrm{J}^\star)^{-1}\|\mathbf{u}\|_\infty \leq (\Omega_{\mathrm{J}^\star})^*[\mathbf{u}] \leq a(\mathrm{J}^\star)^{-1}\|\mathbf{u}\|_\infty.$$

For any matrix $\mathbf{\Gamma}$, we also introduce the operator norm $\|\mathbf{\Gamma}\|_{m,s}$ defined as

$$\|\mathbf{\Gamma}\|_{m,s} = \sup_{\|\mathbf{u}\|_s \leq 1} \|\mathbf{\Gamma}\mathbf{u}\|_m.$$

Moreover, our proof will rely on the control of the *expected dual norm for isonormal vectors*: $\mathbb{E}\left[(\Omega^c_{\mathrm{J}^\star})^*(\mathbf{v})\right]$ with $\mathbf{v}$ a centered Gaussian random variable with unit covariance matrix. In the case of the Lasso, it is of order $(\log p)^{1/2}$.

Following Bach (2008b) and Nardi and Rinaldo (2008), we consider the reduced problem on $\mathrm{J}^\star$,

$$\min_{\mathbf{w}\in\mathbb{R}^p} L_{\mathrm{J}^\star}(\mathbf{w}_{\mathrm{J}^\star}) + \mu\Omega_{\mathrm{J}}(\mathbf{w}_{\mathrm{J}^\star})$$

with solution $\hat{\mathbf{w}}_{\mathrm{J}^\star}$, which can be extended to $\mathrm{J}^c$ with zeros. From optimality conditions (see Lemma 32), we know that

$$(\Omega^\star_{\mathrm{J}})^*[\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}(\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}) - \mathbf{q}_{\mathrm{J}^\star}] \leq \mu, \tag{A.5}$$

where the vector $\mathbf{q}\in\mathbb{R}^p$ is defined as $\mathbf{q} \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{x}_i$. We denote by

$$\nu \triangleq \min\{|\mathbf{w}^\star_j|;\ \mathbf{w}^\star_j \neq 0\}$$

the smallest nonzero components of $\mathbf{w}^\star$. We first prove that we must have with high probability $\|\hat{\mathbf{w}}_g\|_\infty > 0$ for all $g\in\mathcal{G}_{\mathrm{J}^\star}$, proving that the hull of the active set of $\hat{\mathbf{w}}_{\mathrm{J}^\star}$ is exactly $\mathrm{J}^\star$ (i.e., no active group is missing).
We have

$$\begin{aligned}\|\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty &\leq \|\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\|_{\infty,\infty}\|\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}(\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star})\|_\infty \\ &\leq |\mathrm{J}^\star|^{1/2}\kappa^{-1}\left(\|\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}(\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}) - \mathbf{q}_{\mathrm{J}^\star}\|_\infty + \|\mathbf{q}_{\mathrm{J}^\star}\|_\infty\right),\end{aligned}$$

hence from (A.5) and the definition of $A(\mathrm{J}^\star)$,

$$\|\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty \leq |\mathrm{J}^\star|^{1/2}\kappa^{-1}\left(\mu A(\mathrm{J}^\star) + \|\mathbf{q}_{\mathrm{J}^\star}\|_\infty\right). \tag{A.6}$$

Thus, if we assume $\mu \leq \frac{\kappa\nu}{3|\mathrm{J}^\star|^{1/2}A(\mathrm{J}^\star)}$ and

$$\|\mathbf{q}_{\mathrm{J}^\star}\|_\infty \leq \frac{\kappa\nu}{3|\mathrm{J}^\star|^{1/2}}, \tag{A.7}$$

we get

$$\|\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty \leq 2\nu/3, \tag{A.8}$$

so that for all $g\in\mathcal{G}_{\mathrm{J}^\star}$, $\|\hat{\mathbf{w}}_g\|_\infty \geq \frac{\nu}{3}$, hence the hull is indeed selected.

This also ensures that $\hat{\mathbf{w}}_{\mathrm{J}^\star}$ satisfies the equation (see Lemma 32)

$$\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}\left(\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\right) - \mathbf{q}_{\mathrm{J}^\star} + \mu\hat{\mathbf{r}}_{\mathrm{J}^\star} = 0, \tag{A.9}$$

where

$$\hat{\mathbf{r}} = \sum_{g\in\mathcal{G}_{\mathrm{J}^\star}} \frac{\omega^g\circ\omega^g\circ\hat{\mathbf{w}}}{\|\omega^g\circ\hat{\mathbf{w}}\|_2}.$$

We now prove that the $\hat{\mathbf{w}}$ padded with zeros on $[\mathrm{J}^\star]^c$ is indeed optimal for the full problem with high probability. According to Lemma 32, since we have already proved (A.9), it suffices to show that

$$(\Omega_{[\mathrm{J}^\star]^c})^*[\nabla L(\hat{\mathbf{w}})_{[\mathrm{J}^\star]^c}] \leq \mu.$$

Defining $\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star} \triangleq \mathbf{q}_{[\mathrm{J}^\star]^c} - \mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}\mathbf{q}_{\mathrm{J}^\star}$, we can write the gradient of $L$ on $[\mathrm{J}^\star]^c$ as

$$\begin{aligned}\nabla L(\hat{\mathbf{w}})_{[\mathrm{J}^\star]^c} &= -\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star} - \mu\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}\hat{\mathbf{r}}_{\mathrm{J}^\star} \\ &= -\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star} - \mu\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}(\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}) - \mu\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}\mathbf{r}^\star_{\mathrm{J}^\star},\end{aligned}$$

which leads us to control the difference $\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}$. Using Lemma 30, we get

$$\|\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}\|_1 \leq \|\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty \left(\sum_{g\in\mathcal{G}_{\mathrm{J}^\star}} \frac{\|\omega^g_{\mathrm{J}^\star}\|_2^2}{\|\omega^g\circ\mathbf{w}\|_2} + \sum_{g\in\mathcal{G}_{\mathrm{J}^\star}} \frac{\|\omega^g\circ\omega^g\circ\mathbf{w}\|_1^2}{\|\omega^g\circ w\|_2}^3\right),$$

where $\mathbf{w} = t_0\hat{\mathbf{w}} + (1-t_0)\mathbf{w}^\star$ for some $t_0 \in (0,1)$.

Let $\overline{\mathbf{J}} = \{k \in \mathrm{J}^\star : \mathbf{w}^\star_k \neq 0\}$ and let $\varphi$ be defined as

$$\varphi = \sup_{\substack{\mathbf{u}\in\mathbb{R}^p:\overline{\mathbf{J}}\subset\{k\in\mathrm{J}^\star:\mathbf{u}_k\neq 0\}\subset\mathrm{J}^\star \\ g\in\mathcal{G}_{\mathrm{J}^\star}}} \frac{\|\omega^g\circ\omega^g\circ\mathbf{u}\|_1}{\|\omega^g_{\overline{\mathbf{J}}}\circ\omega^g_{\overline{\mathbf{J}}}\circ\mathbf{u}_{\overline{\mathbf{J}}}\|_1} \geq 1.$$

The term $\varphi$ basically measures how close $\mathrm{J}^\star$ and $\overline{\mathbf{J}}$ are, i.e., how relevant the prior encoded by $\mathcal{G}$ about the hull $\mathrm{J}^\star$ is. By using Eq. (A.8), we have

$$\|\omega^g\circ\mathbf{w}\|_2^2 \geq \|\omega^g_{\overline{\mathbf{J}}}\circ\mathbf{w}_{\overline{\mathbf{J}}}\|_2^2 \geq \|\omega^g_{\overline{\mathbf{J}}}\circ\omega^g_{\overline{\mathbf{J}}}\circ\mathbf{w}_{\overline{\mathbf{J}}}\|_1\frac{\nu}{3} \geq \|\omega^g\circ\omega^g\circ\mathbf{w}\|_1\frac{\nu}{3\varphi},$$

along with

$$\|\omega^g\circ\mathbf{w}\|_2 \geq \|\omega^g_{\overline{\mathbf{J}}}\circ\mathbf{w}_{\overline{\mathbf{J}}}\|_2 \geq \|\omega^g_{\overline{\mathbf{J}}}\|_2\frac{\nu}{3} \geq \|\omega^g_{\mathrm{J}^\star}\|_2\frac{\nu}{3\sqrt{\varphi}}$$

and

$$\|\mathbf{w}\|_\infty \leq \frac{5}{3}\|\mathbf{w}^\star\|_\infty.$$

Therefore we have

$$\begin{aligned}\|\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}\|_\infty &\leq \|\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty \sum_{g\in\mathcal{G}_{\mathrm{J}^\star}}\left(\frac{\|\omega^g_{\mathrm{J}^\star}\|_2^2}{\|\omega^g\circ\mathbf{w}\|_\infty} + \frac{5\varphi}{\nu}\frac{\|\mathbf{w}^\star\|_\infty\|\omega^g_{\mathrm{J}^\star}\circ\omega^g_{\mathrm{J}^\star}\|_1}{\|\omega^g\circ\mathbf{w}\|_2}\right) \\ &\leq \frac{3\sqrt{\varphi}\|\hat{\mathbf{w}}_{\mathrm{J}^\star}-\mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty}{\nu}\left(1+\frac{5\varphi\|\mathbf{w}^\star\|_\infty}{\nu}\right)\sum_{g\in\mathcal{G}_{\mathrm{J}^\star}}\|\omega^g_{\mathrm{J}^\star}\|_2.\end{aligned}$$

Introducing $\alpha = \frac{18\varphi^{3/2}\|\mathbf{w}^\star\|_\infty}{\nu^2}\sum_{g\in\mathcal{G}_{\mathrm{J}^\star}}\|\omega^g_{\mathrm{J}^\star}\|_2$, we thus have proved

$$\|\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}\|_1 \leq \alpha\|\hat{\mathbf{w}}_{\mathrm{J}^\star} - \mathbf{w}^\star_{\mathrm{J}^\star}\|_\infty. \tag{A.10}$$

By writing the Schur complement of $\mathbf{Q}$ on the block matrices $\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c}$ and $\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}$, the positiveness of $\mathbf{Q}$ implies that the diagonal terms $\mathrm{diag}(\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}\mathbf{Q}_{\mathrm{J}^\star[\mathrm{J}^\star]^c})$ are less than one, which results in $\|\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1/2}\|_{\infty,2} \leq 1$. We then have

$$\begin{aligned}\|\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}(\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star})\|_\infty &= \|\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1/2}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1/2}(\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star})\|_\infty && \text{(A.11)} \\ &\leq \|\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1/2}\|_{\infty,2}\|\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1/2}\|_2\|\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}\|_2 && \text{(A.12)} \\ &\leq \kappa^{-1/2}\|\hat{\mathbf{r}}_{\mathrm{J}^\star} - \mathbf{r}^\star_{\mathrm{J}^\star}\|_1 && \text{(A.13)} \\ &\leq \kappa^{-3/2}\alpha|\mathrm{J}^\star|^{1/2}\left(\mu A(\mathrm{J}^\star) + \|\mathbf{q}_{\mathrm{J}^\star}\|_\infty\right), && \text{(A.14)}\end{aligned}$$

where the last line comes from Eq. (A.6) and (A.10). We get

$$(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}(\hat{\mathbf{r}}_{\mathrm{J}^\star}-\mathbf{r}^\star_{\mathrm{J}^\star})] \leq \frac{\alpha|\mathrm{J}^\star|^{1/2}}{\kappa^{3/2}a([\mathrm{J}^\star]^c)}\left(\mu A(\mathrm{J}^\star)+\|\mathbf{q}_{\mathrm{J}^\star}\|_\infty\right).$$

Thus, if the following inequalities are verified

$$\frac{\alpha|\mathrm{J}^\star|^{1/2}A(\mathrm{J}^\star)}{\kappa^{3/2}a([\mathrm{J}^\star]^c)}\mu \leq \tfrac{\tau}{4}, \tag{A.15}$$

$$\frac{\alpha|\mathrm{J}^\star|^{1/2}}{\kappa^{3/2}a([\mathrm{J}^\star]^c)}\|\mathbf{q}_{\mathrm{J}^\star}\|_\infty \leq \tfrac{\tau}{4}, \tag{A.16}$$

$$(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}] \leq \tfrac{\mu\tau}{2}, \tag{A.17}$$

we obtain

$$\begin{aligned}(\Omega^c_{\mathrm{J}^\star})^*[\nabla L(\hat{\mathbf{w}})_{[\mathrm{J}^\star]^c}] &\leq (\Omega^c_{\mathrm{J}^\star})^*[-\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}-\mu\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{r}^\star_{\mathrm{J}^\star}]\\ &\leq (\Omega^c_{\mathrm{J}^\star})^*[-\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}]+\mu(1-\tau)+\mu\tau/2\leq\mu,\end{aligned}$$

i.e., $\mathrm{J}^\star$ is exactly selected.

Combined with earlier constraints, this leads to the first part of the desired proposition.

We now need to make sure that the conditions Eq. (A.7), (A.16) and (A.17) hold with high probability. To this end, we upperbound, using Gaussian concentration inequalities, two tail-probabilities. First, $\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}$ is a centered Gaussian random vector with covariance matrix

$$\begin{aligned}\mathbb{E}\left[\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}\mathbf{q}^\top_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}\right] &= \mathbb{E}\Big[\mathbf{q}_{[\mathrm{J}^\star]^c}\mathbf{q}^\top_{[\mathrm{J}^\star]^c}-\mathbf{q}_{[\mathrm{J}^\star]^c}\mathbf{q}^\top_{\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star[\mathrm{J}^\star]^c}\\ &\quad -\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{q}_{\mathrm{J}^\star}\mathbf{q}^\top_{[\mathrm{J}^\star]^c}+\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{q}_{\mathrm{J}^\star}\mathbf{q}^\top_{\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star[\mathrm{J}^\star]^c}\Big]\\ &= \frac{\sigma^2}{n}\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star},\end{aligned}$$

where $\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star} \triangleq \mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c}-\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star[\mathrm{J}^\star]^c}$. In particular, $(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}]$ has the same distribution as $\psi(\mathbf{v})$, with $\psi:\mathbf{u}\mapsto(\Omega^c_{\mathrm{J}^\star})^*(\sigma n^{-1/2}\mathbf{Q}^{1/2}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star}\mathbf{u})$ and $\mathbf{v}$ a centered Gaussian random variable with unit covariance matrix.

Since for any $\mathbf{u}$ we have $\mathbf{u}^\top\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star}\mathbf{u}\leq\mathbf{u}^\top\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c}\mathbf{u}\leq\|\mathbf{Q}^{1/2}\|_2^2\|\mathbf{u}\|_2^2$, by using Sudakov-Fernique inequality (Adler, 1990, Theorem 2.9), we get:

$$\begin{aligned}\mathbb{E}[(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}] = \mathbb{E}\sup_{(\Omega^c_{\mathrm{J}^\star})(\mathbf{u})\leq 1}\mathbf{u}^\top\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star} &\leq \sigma n^{-1/2}\|\mathbf{Q}\|_2^{1/2}\mathbb{E}\sup_{(\Omega^c_{\mathrm{J}^\star})(\mathbf{u})\leq 1}\mathbf{u}^\top\mathbf{v}\\ &\leq \sigma n^{-1/2}\|\mathbf{Q}\|_2^{1/2}\mathbb{E}[(\Omega^c_{\mathrm{J}^\star})^*(\mathbf{v})].\end{aligned}$$

In addition, we have

$$|\psi(\mathbf{u})-\psi(\mathbf{s})|\leq\psi(\mathbf{u}-\mathbf{s})\leq\sigma n^{-1/2}a([\mathrm{J}^\star]^c)^{-1}\|\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star}\|_\infty^{1/2}(\mathbf{u}-\mathbf{s}).$$

On the other hand, since $\mathbf{Q}$ has unit diagonal and $\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}^{-1}\mathbf{Q}_{\mathrm{J}^\star[\mathrm{J}^\star]^c}$ has diagonal terms less than one, $\mathbf{Q}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star}$ also has diagonal terms less than one, which implies that $\|\mathbf{Q}^{1/2}_{[\mathrm{J}^\star]^c[\mathrm{J}^\star]^c|\mathrm{J}^\star}\|_{\infty,2} \leq 1$. Hence $\psi$ is a Lipschitz function with Lipschitz constant upper bounded by $\sigma n^{-1/2} a([\mathrm{J}^\star]^c)^{-1}$. Thus by concentration of Lipschitz functions of multivariate standard random variables (Massart, 2003, Theorem 3.4), we have for $t > 0$:

$$\Pr\Big[(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}] \geq t + \sigma n^{-1/2}\|\mathbf{Q}\|_2^{1/2}\mathbb{E}\left[(\Omega^c_{\mathrm{J}^\star})^*(\mathbf{v})\right]\Big] \leq \exp\left(-\frac{nt^2 a([\mathrm{J}^\star]^c)^2}{2\sigma^2}\right).$$

Applied for $t = \mu\tau/2 \geq 2\sigma n^{-1/2}\|\mathbf{Q}\|_2^{1/2}\mathbb{E}\left[(\Omega^c_{\mathrm{J}^\star})^*(\mathbf{v})\right]$, we get (because $(u-1)^2 \geq u^2/4$ for $u \geq 2$):

$$\Pr\Big[(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{q}_{[\mathrm{J}^\star]^c|\mathrm{J}^\star}] \geq t\Big] \leq \exp\left(-\frac{n\mu^2\tau^2 a([\mathrm{J}^\star]^c)^2}{32\sigma^2}\right).$$

It finally remains to control the term $\Pr(\|\mathbf{q}_{\mathrm{J}^\star}\|_\infty \geq \xi)$, with

$$\xi = \frac{\kappa\nu}{3}\min\left\{1, \frac{3\tau\kappa^{1/2}a([\mathrm{J}^\star]^c)}{4\alpha\nu}\right\}.$$

We can apply classical inequalities for standard random variables (Massart, 2003, Theorem 3.4) that directly lead to

$$\Pr(\|\mathbf{q}_{\mathrm{J}^\star}\|_\infty \geq \xi) \leq 2|\mathrm{J}^\star| \exp\left(-\frac{n\xi^2}{2\sigma^2}\right).$$

To conclude, Theorem 3 holds with

$$C_1(\mathcal{G}, \mathrm{J}^\star) = \frac{a([\mathrm{J}^\star]^c)^2}{16}, \tag{A.18}$$

$$C_2(\mathcal{G}, \mathrm{J}^\star) = \left(\frac{\kappa\nu}{3}\min\left\{1, \frac{\tau\kappa^{1/2}a([\mathrm{J}^\star]^c)\nu}{24\varphi^{3/2}\|\mathbf{w}^\star\|_\infty \sum_{g\in\mathcal{G}_{\mathrm{J}^\star}}\|\omega^g_{\mathrm{J}^\star}\|_2}\right\}\right)^2, \tag{A.19}$$

$$C_3(\mathcal{G}, \mathrm{J}^\star) = 4\|\mathbf{Q}\|_2^{1/2}\mathbb{E}\left[(\Omega^c_{\mathrm{J}^\star})^*(\mathbf{v})\right], \tag{A.20}$$

and

$$C_4(\mathcal{G}, \mathrm{J}^\star) = \frac{\kappa\nu}{3A(\mathrm{J}^\star)}\min\left\{1, \frac{\tau\kappa^{1/2}a([\mathrm{J}^\star]^c)\nu}{24\varphi^{3/2}\|\mathbf{w}^\star\|_\infty \sum_{g\in\mathcal{G}_{\mathrm{J}^\star}}\|\omega^g_{\mathrm{J}^\star}\|_\infty}\right\},$$

where we recall the definitions: $\mathbf{v}$ a centered Gaussian random variable with unit covariance matrix, $\overline{\mathbf{J}} = \{j \in \mathrm{J}^\star : \mathbf{w}^\star_j \neq 0\}$, $\nu = \min\{|\mathbf{w}^\star_j|;\ j \in \overline{\mathbf{J}}\}$,

$$\varphi = \sup_{\substack{\mathbf{u}\in\mathbb{R}^p : \overline{\mathbf{J}} \subset \{k\in\mathrm{J}^\star : \mathbf{u}_k \neq 0\} \subset \mathrm{J}^\star \\ g\in\mathcal{G}_{\mathrm{J}^\star}}} \frac{\|\omega^g \circ \omega^g \circ \mathbf{u}\|_1}{\|\omega^g_{\overline{\mathbf{J}}} \circ \omega^g_{\overline{\mathbf{J}}} \circ \mathbf{u}_{\overline{\mathbf{J}}}\|_1},$$

$\kappa = \lambda_{\min}(\mathbf{Q}_{\mathrm{J}^\star\mathrm{J}^\star}) > 0$ and $\tau > 0$ such that $(\Omega^c_{\mathrm{J}^\star})^*[\mathbf{Q}_{[\mathrm{J}^\star]^c\mathrm{J}^\star}\mathbf{Q}^{-1}_{\mathrm{J}^\star\mathrm{J}^\star}\mathbf{r}^\star] < 1 - \tau$.

### A.1.8 A first order approach to solve Eq. 2.2 and Eq. 2.5

Both regularized minimization problems Eq. 2.2 and Eq. 2.5 (that just differ in the squaring of $\Omega$) can be solved by using generic toolboxes for second-order cone programming (SOCP) (Boyd and Vandenberghe, 2004). We propose here a first order approach that takes up ideas from Micchelli and Pontil (2006); Rakotomamonjy et al. (2008) and that is based on the following variational equalities: for $\mathbf{x} \in \mathbb{R}^p$, we have

$$\|\mathbf{x}\|_1^2 = \min_{\substack{\mathbf{z} \in \mathbb{R}_+^p, \\ \sum_{j=1}^p \mathbf{z}_j \leq 1}} \sum_{j=1}^p \frac{\mathbf{x}_j^2}{\mathbf{z}_j},$$

whose minimum is uniquely attained for $\mathbf{z}_j = |\mathbf{x}_j|/\|\mathbf{x}\|_1$. Similarly, we have

$$2\|\mathbf{x}\|_1 = \min_{\mathbf{z} \in \mathbb{R}_+^p} \sum_{j=1}^p \frac{\mathbf{x}_j^2}{\mathbf{z}_j} + \|\mathbf{z}\|_1,$$

whose minimun is uniquely obtained for $\mathbf{z}_j = |\mathbf{x}_j|$. Thus, we can equivalently rewrite Eq. 2.2 as

$$\min_{\substack{\mathbf{w} \in \mathbb{R}^p, \\ (\boldsymbol{\eta}^g)_{g \in \mathcal{G}} \in \mathbb{R}_+^{|\mathcal{G}|}}} + \frac{1}{n} \sum_{i=1}^n \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) \frac{\mu}{2} \sum_{j=1}^p \mathbf{w}_j^2 \boldsymbol{\zeta}_j^{-1} + \frac{\mu}{2} \|(\boldsymbol{\eta}^g)_{g \in \mathcal{G}}\|_1, \tag{A.21}$$

with $\boldsymbol{\zeta}_j = (\sum_{g \ni j} (\omega_j^g)^2 (\boldsymbol{\eta}^g)^{-1})^{-1}$. In the same vein, Eq. 2.5 is equivalent to

$$\min_{\substack{\mathbf{w} \in \mathbb{R}^p, \\ (\boldsymbol{\eta}^g)_{g \in \mathcal{G}} \in \mathbb{R}_+^{|\mathcal{G}|}, \\ \sum_{g \in \mathcal{G}} \boldsymbol{\eta}^g \leq 1}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \mathbf{w}^\top \mathbf{x}_i) \frac{\lambda}{2} \sum_{j=1}^p \mathbf{w}_j^2 \boldsymbol{\zeta}_j^{-1}, \tag{A.22}$$

where $\boldsymbol{\zeta}_j$ is defined as above. The reformulations Eq. A.21 and Eq. A.22 are *jointly* convex in $\{\mathbf{w}, (\boldsymbol{\eta}^g)_{g \in \mathcal{G}}\}$ and lend themselves well to a simple alternating optimization scheme between $\mathbf{w}$ (for instance, $\mathbf{w}$ can be computed in closed-form when the square loss is used) and $(\boldsymbol{\eta}^g)_{g \in \mathcal{G}}$ (whose optimal value is always a closed-form solution). If the variables $(\boldsymbol{\eta}^g)_{g \in \mathcal{G}} \in \mathbb{R}_+^{|\mathcal{G}|}$ are bounded away from zero by a smoothing parameter, the convergence of this scheme is guaranteed by standard results about block coordinate descent procedures (Bertsekas, 1999).

This first order approach is computationally appealing since it allows *warm-restart*, which can dramatically speed up the computation over regularization paths. Moreover, it does not make any assumptions on the nature of the family of groups $\mathcal{G}$.

### A.1.9 Technical lemmas

In this last section of the appendix, we give several technical lemmas. We consider $I \subseteq [\![1; p]\!]$ and $\mathcal{G}_I = \{g \in \mathcal{G};\ g \cap I \neq \varnothing\} \subseteq \mathcal{G}$, i.e., the set of active groups when the variables $I$ are selected.

We begin with a dual formulation of $\Omega^*$ obtained by conic duality (Boyd and Vandenberghe, 2004):

**Lemma 27**
*Let* $\mathbf{u}_I \in \mathbb{R}^{|I|}$. *We have*

$$\begin{aligned}(\Omega_I)^*[\mathbf{u}_I] \quad &= \quad \min_{(\boldsymbol{\xi}_I^g)_{g\in\mathcal{G}_I}} \max_{g\in\mathcal{G}_I} \|\boldsymbol{\xi}_I^g\|_2 \\ &\quad s.t. \quad \mathbf{u}_j + \sum_{g\in\mathcal{G}_I, G\ni j} \omega_j^g \boldsymbol{\xi}_j^g = 0 \;\; \textit{and} \;\; \boldsymbol{\xi}_j^g = 0 \;\; \textit{if} \;\; j \notin G.\end{aligned}$$

*Proof.* By definiton of $(\Omega_I)^*[\mathbf{u}_I]$, we have

$$(\Omega_I)^*[\mathbf{u}_I] = \max_{\Omega_I(\mathbf{v}_I)\leq 1} \mathbf{u}_I^\top \mathbf{v}_I.$$

By introducing the primal variables $(\boldsymbol{\alpha}_g)_{g\in\mathcal{G}_I} \in \mathbb{R}^{|\mathcal{G}_I|}$, we can rewrite the previous maximization problem as

$$(\Omega_I)^*[\mathbf{u}_I] = \max_{\mathbf{v}, \sum_{g\in\mathcal{G}_I}\boldsymbol{\alpha}_g\leq 1} \mathbf{u}_I^\top v_I, \quad \text{s.t.} \quad \forall\, g \in \mathcal{G}_I,\; \|\omega_I^g \circ \mathbf{u}_{G\cap I}\|_2 \leq \boldsymbol{\alpha}_g,$$

which is a second-order cone program (SOCP) with $|\mathcal{G}_I|$ second-order cone constraints. This primal problem is convex and satisfies Slater's conditions for generalized conic inequalities, which implies that strong duality holds (Boyd and Vandenberghe, 2004). We now consider the Lagrangian $\mathcal{L}$ defined as

$$\mathcal{L}(\mathbf{v}_I, \boldsymbol{\alpha}_g, \gamma, \boldsymbol{\tau}_g, \boldsymbol{\xi}_I^g) = \mathbf{u}_I^\top \mathbf{v}_I + \gamma(1 - \sum_{g\in\mathcal{G}_I} \boldsymbol{\alpha}_g) + \sum_{g\in\mathcal{G}_I} \begin{pmatrix} \boldsymbol{\alpha}_g \\ \omega_I^g \circ \mathbf{u}_{G\cap I} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\tau}_g \\ \boldsymbol{\xi}_I^g \end{pmatrix},$$

with the dual variables $\{\gamma, (\boldsymbol{\tau}_g)_{g\in\mathcal{G}_I}, (\boldsymbol{\xi}_I^g)_{g\in\mathcal{G}_I}\} \in \mathbb{R}_+ \times \mathbb{R}^{|\mathcal{G}_I|} \times \mathbb{R}^{|I|\times|\mathcal{G}_I|}$ such that for all $g \in \mathcal{G}_I$, $\boldsymbol{\xi}_j^g = 0$ if $j \notin G$ and $\|\boldsymbol{\xi}_I^g\|_2 \leq \boldsymbol{\tau}_g$. The dual function is obtained by taking the derivatives of $\mathcal{L}$ with respect to the primal variables $\mathbf{v}_I$ and $(\boldsymbol{\alpha}_g)_{g\in\mathcal{G}_I}$ and equating them to zero, which leads to

$$\begin{aligned} \forall j \in I, \quad \mathbf{u}_j + \textstyle\sum_{g\in\mathcal{G}_I, G\ni j} \omega_j^g \boldsymbol{\xi}_j^g \quad &= 0 \\ \forall g \in \mathcal{G}_I, \qquad \gamma - \boldsymbol{\tau}_g \quad &= 0. \end{aligned}$$

After simplifying the Lagrangian, the dual problem then reduces to

$$\min_{\gamma, (\boldsymbol{\xi}_I^g)_{g\in\mathcal{G}_I}} \gamma \quad \text{s.t.} \quad \begin{cases} \forall j \in I, \mathbf{u}_j + \sum_{g\in\mathcal{G}_I, G\ni j} \omega_j^g \boldsymbol{\xi}_j^g = 0 \;\text{ and }\; \boldsymbol{\xi}_j^g = 0 \;\text{ if }\; j \notin G, \\ \forall g \in \mathcal{G}_I, \|\boldsymbol{\xi}_I^g\|_2 \leq \gamma, \end{cases}$$

which is equivalent to the displayed result. □

Since we cannot compute in closed-form the solution of the previous optimization problem, we focus on a different *but closely related* problem, i.e., when we replace the objective $\max_{g\in\mathcal{G}_I} \|\boldsymbol{\xi}_I^g\|_2$ by $\max_{g\in\mathcal{G}_I} \|\boldsymbol{\xi}_I^g\|_\infty$, to obtain a *meaningful* feasible point:

**Lemma 28**
*Let* $\mathbf{u}_I \in \mathbb{R}^{|I|}$. *The following problem*

$$\begin{aligned} \min_{(\boldsymbol{\xi}_I^g)_{g\in\mathcal{G}_I}} \quad & \max_{g\in\mathcal{G}_I} \|\boldsymbol{\xi}_I^g\|_\infty \\ s.t. \quad & \mathbf{u}_j + \sum_{g\in\mathcal{G}_I, g\ni j} \omega_j^g \boldsymbol{\xi}_j^g = 0 \ \text{ and } \ \boldsymbol{\xi}_j^g = 0 \ \text{ if } \ j \notin g, \end{aligned}$$

*is minimized for* $(\boldsymbol{\xi}_j^g)^* = -\dfrac{\mathbf{u}_j}{\sum_{h\in j, h\in\mathcal{G}_I}\omega_j^h}$.

*Proof.* We proceed by contradiction. Let us assume there exists $(\boldsymbol{\xi}_I^g)_{g\in\mathcal{G}_I}$ such that

$$\begin{aligned} \max_{g\in\mathcal{G}_I} \|\boldsymbol{\xi}_I^g\|_\infty &< \max_{g\in\mathcal{G}_I} \|(\boldsymbol{\xi}_I^g)^*\|_\infty \\ &= \max_{g\in\mathcal{G}_I} \max_{j\in g} \frac{|\mathbf{u}_j|}{\sum_{h\in j, h\in\mathcal{G}_I}\omega_j^h} \\ &= \frac{|\mathbf{u}_{j_0}|}{\sum_{g\in j_0, g\in\mathcal{G}_I}\omega_{j_0}^h}, \end{aligned}$$

where we denote by $j_0$ an argmax of the latter maximization. We notably have for all $g \ni j_0$:

$$|\boldsymbol{\xi}_{j_0}^g| < \frac{|\mathbf{u}_{j_0}|}{\sum_{h\in j_0, h\in\mathcal{G}_I}\omega_{j_0}^h}.$$

By multiplying both sides by $\omega_{j_0}^g$ and by summing over $g \ni j_0$, we get

$$|\mathbf{u}_{j_0}| = |\sum_{g\in\mathcal{G}_I, g\ni j_0} \omega_{j_0}^g \boldsymbol{\xi}_{j_0}^g| \leq \sum_{g\ni j_0} \omega_{j_0}^g |\boldsymbol{\xi}_{j_0}^g| < |\mathbf{u}_{j_0}|,$$

which leads to a contradiction. □

We now give an upperbound on $\Omega^*$ based on Lemma 27 and Lemma 28:

**Lemma 29**
*Let* $\mathbf{u}_I \in \mathbb{R}^{|I|}$. *We have*

$$(\Omega_I)^*[\mathbf{u}_I] \leq \max_{g\in\mathcal{G}_I} \left\{ \sum_{j\in g} \left\{ \frac{\mathbf{u}_j}{\sum_{h\in j, h\in\mathcal{G}_I}\omega_j^h} \right\}^2 \right\}^{\frac{1}{2}}.$$

*Proof.* We simply plug the minimizer obtained in Lemma 28 into the problem of Lemma 27. □

We now derive a lemma to control the difference of the gradient of $\Omega_J$ evaluated in two points:

**Lemma 30**
*Let $\mathbf{u}_J, \mathbf{v}_J$ be two nonzero vectors in $\mathbb{R}^{|J|}$. Let us consider the mapping $\mathbf{w}_J \mapsto \mathbf{r}(\mathbf{w}_J) = \sum_{g\in\mathcal{G}_J} \frac{\omega_J^g \circ \omega_J^g \circ \mathbf{w}_J}{\|\omega_J^g \circ \mathbf{w}_J\|_2} \in \mathbb{R}^{|J|}$. There exists $\mathbf{z}_J = t_0\mathbf{u}_J + (1-t_0)\mathbf{v}_J$ for some $t_0 \in (0,1)$ such that*

$$\|\mathbf{r}(\mathbf{u}_J) - \mathbf{r}(\mathbf{v}_J)\|_1 \leq \|\mathbf{u}_J - \mathbf{v}_J\|_\infty \left( \sum_{g\in\mathcal{G}_J} \frac{\|\omega_J^g\|_2^2}{\|\omega_J^g \circ \mathbf{z}_J\|_2} + \sum_{g\in\mathcal{G}_J} \frac{\|\omega_J^g \circ \omega_J^g \circ \mathbf{z}_J\|_1^2}{\|\omega_J^g \circ \mathbf{z}_J\|_2^3} \right).$$

*Proof.* For $j,k \in J$, we have

$$\frac{\partial \mathbf{r}_j}{\partial \mathbf{w}_k}(\mathbf{w}_J) = \sum_{g\in\mathcal{G}_J} \frac{(\omega_j^g)^2}{\|\omega_J^g \circ \mathbf{w}_J\|_2} \mathbb{I}_{j=k} - \sum_{g\in\mathcal{G}_J} \frac{(\omega_j^g)^2 \mathbf{w}_j}{\|\omega_J^g \circ \mathbf{w}_J\|_2^3} (\omega_k^g)^2 \mathbf{w}_k,$$

with $\mathbb{I}_{j=k} = 1$ if $j=k$ and 0 otherwise. We then consider $t \in [0,1] \mapsto h_j(t) = r_j(t\mathbf{u}_J + (1-t)\mathbf{v}_J)$. The mapping $h_j$ being continuously differentiable, we can apply the mean-value theorem: there exists $t_0 \in (0,1)$ such that

$$h_j(1) - h_j(0) = \frac{\partial h_j(t)}{\partial t}(t_0).$$

We then have

$$\begin{aligned} |\mathbf{r}_j(\mathbf{u}_J) - \mathbf{r}_j(\mathbf{v}_J)| &\leq \sum_{k\in J} \left| \frac{\partial \mathbf{r}_j}{\partial \mathbf{w}_k}(\mathbf{z}) \right| |\mathbf{u}_k - \mathbf{v}_k| \\ &\leq \|\mathbf{u}_J - \mathbf{v}_J\|_\infty \left( \sum_{g\in\mathcal{G}_J} \frac{(\omega_j^g)^2}{\|\omega_J^g \circ \mathbf{z}_J\|_2} + \sum_{k\in J} \sum_{g\in\mathcal{G}_J} \frac{(\omega_j^g)^2 |z_j|}{\|\omega_J^g \circ \mathbf{z}_J\|_2^3} (\omega_k^g)^2 |\mathbf{z}_k| \right), \end{aligned}$$

which leads to

$$\|\mathbf{r}(\mathbf{u}_J) - \mathbf{r}(\mathbf{v}_J)\|_1 \leq \|\mathbf{u}_J - \mathbf{v}_J\|_\infty \left( \sum_{g\in\mathcal{G}_J} \frac{\|\omega_J^g\|_2^2}{\|\omega_J^g \circ \mathbf{z}_J\|_2} + \sum_{g\in\mathcal{G}_J} \frac{\|\omega_J^g \circ \omega_J^g \circ \mathbf{z}_J\|_1^2}{\|\omega_J^g \circ \mathbf{z}_J\|_2^3} \right).$$

□

Given an active set $J \subseteq [\![1;p]\!]$ and a direct parent $K \in \Pi_{\mathcal{P}}(J)$ of $J$ in the DAG of nonzero patterns, we have the following result:

**Lemma 31**
*For all $g \in \mathcal{G}_K \backslash \mathcal{G}_J$, we have $K \backslash J \subseteq g$.*

*Proof.* We proceed by contradiction. We assume there exists $g_0 \in \mathcal{G}_K \backslash \mathcal{G}_J$ such that $K \backslash J \nsubseteq g_0$. Given that $K \in \mathcal{P}$, there exists $\mathcal{G}' \subseteq \mathcal{G}$ verifying $K = \bigcap_{g\in\mathcal{G}'} g^c$. Note that $g_0 \notin \mathcal{G}'$ since by definition $g_0 \cap K \neq \varnothing$.

We can now build the pattern $\tilde{K} = \bigcap_{g\in\mathcal{G}'\cup\{g_0\}} g^c = K \cap g_0^c$ that belongs to $\mathcal{P}$. Moreover, $\tilde{K} = K \cap g_0^c \subset K$ since we assumed $g_0^c \cap K \neq \varnothing$. In addition, we have that $J \subset K$ and $J \subset g_0^c$ because $K \in \Pi_{\mathcal{P}}(J)$ and $g_0 \in \mathcal{G}_K \backslash \mathcal{G}_J$. This results in $J \subset \tilde{K} \subset K$, which is impossible by definition of $K$. □

We give below an important Lemma to characterize the solutions of (2.2).

**Lemma 32**
*The vector* $\hat{\mathbf{w}} \in \mathbb{R}^p$ *is a solution of*

$$\min_{\mathbf{w}\in\mathbb{R}^p} L(\mathbf{w}) + \mu\,\Omega(\mathbf{w})$$

*if and only if*

$$\begin{cases} \nabla L(\hat{\mathbf{w}})_{\hat{\mathrm{J}}} + \mu\,\hat{\mathbf{r}}_{\hat{\mathrm{J}}} = 0 \\ (\Omega^c_{\hat{\mathrm{J}}})^*[\nabla L(\hat{\mathbf{w}})_{\hat{\mathrm{J}}^c}] \leq \mu, \end{cases}$$

*with* $\hat{\mathrm{J}}$ *the hull of* $\{j \in [\![1;p]\!], \hat{\mathbf{w}}_j \neq 0\}$ *and the vector* $\hat{\mathbf{r}} \in \mathbb{R}^p$ *defined as*

$$\hat{\mathbf{r}} = \sum_{g\in\mathcal{G}_{\hat{\mathrm{J}}}} \frac{\omega^g \circ \omega^g \circ \hat{\mathbf{w}}}{\|\omega^g \circ \hat{\mathbf{w}}\|_2}.$$

*In addition, the solution* $\hat{\mathbf{w}}$ *satisfies*

$$\Omega^*[\nabla L(\hat{\mathbf{w}})] \leq \mu.$$

*Proof.* The problem

$$\min_{\mathbf{w}\in\mathbb{R}^p} L(\mathbf{w}) + \mu\,\Omega(\mathbf{w}) = \min_{\mathbf{w}\in\mathbb{R}^p} F(\mathbf{w})$$

being convex, the directional derivative optimality condition are necessary and sufficient (Borwein and Lewis, 2006, Propositions 2.1.1-2.1.2). Therefore, the vector $\hat{\mathbf{w}}$ is a solution of the previous problem if and only if for all directions $\mathbf{u} \in \mathbb{R}^p$, we have

$$\lim_{\substack{\varepsilon\to 0 \\ \varepsilon>0}} \frac{F(\hat{\mathbf{w}} + \varepsilon u) - F(\hat{\mathbf{w}})}{\varepsilon} \geq 0.$$

Some algebra leads to the following equivalent formulation

$$\forall \mathbf{u} \in \mathbb{R}^p,\ \mathbf{u}^\top \nabla L(\hat{\mathbf{w}}) + \mu\,\mathbf{u}^\top_{\hat{\mathrm{J}}}\hat{\mathbf{r}}_{\hat{\mathrm{J}}} + \mu\,(\Omega^c_{\hat{\mathrm{J}}})[\mathbf{u}_{\hat{\mathrm{J}}^c}] \geq 0. \tag{A.23}$$

The first part of the lemma then comes from the projections on $\hat{\mathrm{J}}$ and $\hat{\mathrm{J}}^c$.

An application of the Cauchy-Schwartz inequality on $\mathbf{u}^\top_{\hat{\mathrm{J}}}\hat{\mathbf{r}}_{\hat{\mathrm{J}}}$ gives for all $\mathbf{u} \in \mathbb{R}^p$

$$\mathbf{u}^\top_{\hat{\mathrm{J}}}\hat{\mathbf{r}}_{\hat{\mathrm{J}}} \leq (\Omega_{\hat{\mathrm{J}}})[\mathbf{u}_{\hat{\mathrm{J}}}].$$

Combined with the equation (A.23), we get $\forall \mathbf{u} \in \mathbb{R}^p,\ \mathbf{u}^\top \nabla L(\hat{\mathbf{w}}) + \mu\,\Omega(\mathbf{u}) \geq 0$, hence the second part of the lemma. □

We end up with a lemma regarding the dual norm of the sum of two *disjoint* norms (see Rockafellar, 1997):

**Lemma 33**
*Let* $A$ *and* $B$ *be a partition of* $[\![1;p]\!]$*, i.e.,* $A \cap B = \varnothing$ *and* $A \cup B = [\![1;p]\!]$*. We consider two norms* $\mathbf{u}_A \in \mathbb{R}^{|A|} \mapsto \|\mathbf{u}_A\|_A$ *and* $\mathbf{u}_B \in \mathbb{R}^{|B|} \mapsto \|\mathbf{u}_B\|_B$*, with dual norms* $\|\mathbf{v}_A\|^*_A$ *and* $\|\mathbf{v}_B\|^*_B$*. We have*

$$\max_{\|\mathbf{u}_A\|_A + \|\mathbf{u}_B\|_B \leq 1} \mathbf{u}^\top \mathbf{v} = \max\left\{\|\mathbf{v}_A\|^*_A,\ \|\mathbf{v}_B\|^*_B\right\}.$$

## A.2 Proofs and Technical Elements of Chapter 2

### A.2.1 Links with Tree-Structured Nonconvex Regularization

We present in this section an algorithm introduced by Donoho (1997) in the more general context of approximation from dyadic partitions (see Section 6 in Donoho, 1997) for solving the following problem

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \delta^g(\mathbf{v}), \tag{A.24}$$

where the $\mathbf{u}$ in $\mathbb{R}^p$ is given, $\lambda$ is a regularization parameter, $\mathcal{G}$ is a set of tree-structured groups in the sense of Definition 1, and the functions $\delta^g$ are defined as in Eq. (4.4)—that is, $\delta^g(\mathbf{v}) = 1$ if there exists $j$ in $g$ such that $\mathbf{v}_j \neq 0$, and 0 otherwise. This problem can be viewed as a proximal operator for the nonconvex regularization $\sum_{g\in\mathcal{G}} \delta^g(\mathbf{v})$. As we will show, it can be solved efficiently, and in fact it can be used to obtain approximate solutions of the nonconvex problem presented in Eq. (4.1), or to solve tree-structured wavelet decompositions as done by Baraniuk et al. (2010).

We now briefly show how to derive the dynamic programming approach introduced by Donoho (1997). Given a group $g$ in $\mathcal{G}$, we use the same notations root$(g)$ and children(g) introduced in Section 4.3.5. It is relatively easy to show that finding a solution of Eq. (A.24) amounts to finding the support $S \subseteq \{1,\dots,p\}$ of its solution and that the problem can be equivalently rewritten

$$\min_{S\subseteq\{1,\dots,p\}} -\frac{1}{2}\|\mathbf{u}_S\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \delta^g(S), \tag{A.25}$$

with the abusive notation $\delta^g(S) = 1$ if $g \cap S \neq \emptyset$ and 0 otherwise. We now introduce the quantity

$$\psi_g(S) \triangleq \begin{cases} 0 & \text{if } g\cap S = \emptyset \\ -\frac{1}{2}\|\mathbf{u}_{\text{root}(g)}\|_2^2 + \lambda + \sum_{h\in\text{children}(g)} \psi_h(S) & \text{otherwise.} \end{cases}$$

After a few computations, solving Eq. (A.25) can be shown to be equivalent to minimizing $\psi_{g_0}(S)$ where $g_0$ is the root of the tree. It is then easy to prove that for any group $g$ in $\mathcal{G}$, we have

$$\min_{S\subseteq\{1,\dots,p\}} \psi_g(S) = \min\Big(0, -\frac{1}{2}\|\mathbf{u}_{\text{root}(g)}\|_2^2 + \lambda + \sum_{h\in\text{children}(g)} \min_{S'\subseteq\{1,\dots,p\}} \psi_h(S')\Big),$$

which leads to the following dynamic programming approach presented in Algorithm 10. This algorithm shares several conceptual links with Algorithm 7 and 8. It traverses the tree in the same order, has a complexity in $O(p)$, and it can be shown that the whole procedure actually performs a sequence of thresholding operations on the variable $\mathbf{v}$.

**Algorithm 10** Computation of the Proximal Operator for the Nonconvex Approach

Inputs: $\mathbf{u} \in \mathbb{R}^p$, a tree-structured set of groups $\mathcal{G}$ and $g_0$ (root of the tree).
Outputs: $\mathbf{v}$ (primal solution).
Initialization: $\mathbf{v} \leftarrow \mathbf{u}$.
Call `recursiveThresholding`$(g_0)$.

**Procedure** `recursiveThresholding`$(g)$

$\eta \leftarrow \min\Big(0, -\frac{1}{2}\|\mathbf{u}_{\text{root}(g)}\|_2^2 + \lambda + \sum_{h \in \text{children}(g)} \texttt{recursiveThresholding}(h)\Big)$.
**if** $\eta = 0$ **then**
  $\mathbf{v}_g \leftarrow 0$.
**end if**
**return** $\eta$.

### A.2.2 Proofs

#### Proof of Lemma 2

*Proof.* The proof relies on tools from conic duality (Boyd and Vandenberghe, 2004). Let us introduce the cone $\mathcal{C} \triangleq \{(\mathbf{v}, z) \in \mathbb{R}^{p+1};\ \|\mathbf{v}\| \leq z\}$ and its dual counterpart $\mathcal{C}^* \triangleq \{(\boldsymbol{\xi}, \tau) \in \mathbb{R}^{p+1};\ \|\boldsymbol{\xi}\|_* \leq \tau\}$. These cones induce generalized inequalities for which Lagrangian duality also applies. We refer the interested readers to Boyd and Vandenberghe (2004) for further details.

We can rewrite problem (4.7) as

$$\min_{\mathbf{v} \in \mathbb{R}^p, \mathbf{z} \in \mathbb{R}^{|\mathcal{G}|}} \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \omega_g z_g, \text{ such that } (\mathbf{v}_{|g}, z_g) \in \mathcal{C},\ \forall g \in \mathcal{G},$$

by introducing the primal variables $\mathbf{z} = (z_g)_{g \in \mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|}$, with the additional $|\mathcal{G}|$ conic constraints $(\mathbf{v}_{|g}, z_g) \in \mathcal{C}$, for $g \in \mathcal{G}$.

This primal problem is convex and satisfies Slater's conditions for generalized conic inequalities (i.e., existence of a feasible point in the interior of the domain), which implies that strong duality holds (Boyd and Vandenberghe, 2004). We now consider the Lagrangian $\mathcal{L}$ defined as

$$\mathcal{L}(\mathbf{v}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \omega_g z_g - \sum_{g \in \mathcal{G}} \begin{pmatrix} z_g \\ \mathbf{v}_{|g} \end{pmatrix}^\top \begin{pmatrix} \tau_g \\ \boldsymbol{\xi}^g \end{pmatrix},$$

with the dual variables $\boldsymbol{\tau} = (\tau_g)_{g \in \mathcal{G}}$ in $\mathbb{R}^{|\mathcal{G}|}$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}^g)_{g \in \mathcal{G}}$ in $\mathbb{R}^{p \times |\mathcal{G}|}$, such that for all $g \in \mathcal{G}$, $\boldsymbol{\xi}^g_j = 0$ if $j \notin g$ and $(\boldsymbol{\xi}^g, \tau_g) \in \mathcal{C}^*$.

The dual function is obtained by minimizing out the primal variables. To this end, we take the derivatives of $\mathcal{L}$ with respect to the primal variables $\mathbf{v}$ and $\mathbf{z}$ and set them to zero, which leads to

$$\mathbf{v} - \mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g = 0 \quad \text{and} \quad \forall g \in \mathcal{G},\ \lambda \omega_g - \tau_g = 0.$$

After simplifying the Lagrangian and flipping (without loss of generality) the sign of $\boldsymbol{\xi}$, we obtain the dual problem in Eq. (4.8). We derive the optimality conditions from the Karush–Kuhn–Tucker conditions for generalized conic inequalities (Boyd and Vandenberghe, 2004). We have that $\{\mathbf{v}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\xi}\}$ are optimal if and only if

$$\begin{aligned} & & \forall g \in \mathcal{G}, z_g \tau_g - \mathbf{v}_{|g}^\top \boldsymbol{\xi}^g &= 0, \quad \text{(Complementary slackness)} \\ \forall g \in \mathcal{G}, (\mathbf{v}_{|g}, z_g) \in \mathcal{C}, & & \forall g \in \mathcal{G}, \lambda \omega_g - \tau_g &= 0, \\ \forall g \in \mathcal{G}, (\boldsymbol{\xi}^g, \tau_g) \in \mathcal{C}^*, & & \mathbf{v} - \mathbf{u} + \textstyle\sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g &= 0. \end{aligned}$$

Combining the complementary slackness with the definition of the dual norm, we have

$$\forall g \in \mathcal{G}, \ z_g \tau_g = \mathbf{v}_{|g}^\top \boldsymbol{\xi}^g \leq \|\mathbf{v}_{|g}\| \|\boldsymbol{\xi}^g\|_*.$$

Furthermore, using the fact that $\forall g \in \mathcal{G}, \ (\mathbf{v}_{|g}, z_g) \in \mathcal{C}$ and $(\boldsymbol{\xi}^g, \tau_g) = (\boldsymbol{\xi}^g, \lambda \omega_g) \in \mathcal{C}^*$, we obtain the following chain of inequalities

$$\forall g \in \mathcal{G}, \ \lambda z_g \omega_g = \mathbf{v}_{|g}^\top \boldsymbol{\xi}^g \leq \|\mathbf{v}_{|g}\| \|\boldsymbol{\xi}^g\|_* \leq z_g \|\boldsymbol{\xi}^g\|_* \leq \lambda z_g \omega_g,$$

for which equality must hold. In particular, we have $\mathbf{v}_{|g}^\top \boldsymbol{\xi}^g = \|\mathbf{v}_{|g}\| \|\boldsymbol{\xi}^g\|_*$ and $z_g \|\boldsymbol{\xi}^g\|_* = \lambda z_g \omega_g$. If $\mathbf{v}_{|g} \neq 0$, then $z_g$ cannot be equal to zero, which implies in turn that $\|\boldsymbol{\xi}^g\|_* = \lambda \omega_g$. Eventually, applying Lemma 34 gives the advertised optimality conditions.

Conversely, starting from the optimality conditions of Lemma 2, and making use again of Lemma 34, we can derive the Karush–Kuhn–Tucker conditions displayed above. More precisely, we define for all $g \in \mathcal{G}$,

$$\tau_g \triangleq \lambda \omega_g \quad \text{and} \quad z_g \triangleq \|\mathbf{v}_{|g}\|.$$

The only condition that needs to be discussed is the complementary slackness condition. If $\mathbf{v}_{|g} = 0$, then it is easily satisfied. Otherwise, combining the definitions of $\tau_g$, $z_g$ and the fact that

$$\mathbf{v}_{|g}^\top \boldsymbol{\xi}^g = \|\mathbf{v}_{|g}\| \|\boldsymbol{\xi}^g\|_* \text{ and } \|\boldsymbol{\xi}^g\|_* = \lambda \omega_g,$$

we end up with the desired complementary slackness. □

### Optimality condition for the projection on the dual ball

**Lemma 34** (Projection on the dual ball)

*Let* $\mathbf{w} \in \mathbb{R}^p$ *and* $t > 0$. *We have* $\boldsymbol{\kappa} = \Pi_{\|.\|_* \leq t}(\mathbf{w})$ *if and only if*

$$\begin{cases} \textit{if } \|\mathbf{w}\|_* \leq t, & \boldsymbol{\kappa} = \mathbf{w}, \\ \textit{otherwise}, & \|\boldsymbol{\kappa}\|_* = t \ \textit{ and } \ \boldsymbol{\kappa}^\top (\mathbf{w} - \boldsymbol{\kappa}) = \|\boldsymbol{\kappa}\|_* \|\mathbf{w} - \boldsymbol{\kappa}\|. \end{cases}$$

*Proof.* When the vector $\mathbf{w}$ is already in the ball of $\|.\|_*$ with radius $t$, i.e., $\|\mathbf{w}\|_* \leq t$, the situation is simple, since the projection $\Pi_{\|.\|_* \leq t}(\mathbf{w})$ obviously gives $\mathbf{w}$ itself. On the other hand, a necessary and sufficient optimality condition for having $\boldsymbol{\kappa} = \Pi_{\|.\|_* \leq t}(\mathbf{w}) =$

$\arg\min_{\|\mathbf{y}\|_* \leq t} \|\mathbf{w}-\mathbf{y}\|_2$ is that the residual $\mathbf{w}-\boldsymbol{\kappa}$ lies in the normal cone of the constraint set (Borwein and Lewis, 2006), that is, for all $\mathbf{y}$ such that $\|\mathbf{y}\|_* \leq t$, $(\mathbf{w}-\boldsymbol{\kappa})^\top(\mathbf{y}-\boldsymbol{\kappa}) \leq 0$. The displayed result then follows from the definition of the dual norm, namely $\|\boldsymbol{\kappa}\|_* = \max_{\|\mathbf{z}\|\leq 1} \mathbf{z}^\top \boldsymbol{\kappa}$. □

**Proof of Lemma 3**

*Proof.* First, notice that the conclusion $\boldsymbol{\xi}^h = \Pi_{\|.\|_* \leq \lambda\omega_h}(\mathbf{v}_{|h} + \boldsymbol{\xi}^h)$ simply comes from the definition of $\boldsymbol{\xi}^h$ and $\mathbf{v}$, along with the fact that $\boldsymbol{\xi}^g = \boldsymbol{\xi}^g_{|h}$ since $g \subseteq h$. We now examine $\boldsymbol{\xi}^g$.

The proof mostly relies on the optimality conditions characterizing the projection onto a ball of the dual norm $\|\cdot\|_*$. Precisely, by Lemma 34, we need to show that either

$$\boldsymbol{\xi}^g = \mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g}, \text{ if } \|\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g}\|_* \leq t_g,$$

or

$$\|\boldsymbol{\xi}^g\|_* = t_g \quad \text{and} \quad \boldsymbol{\xi}^{g\top}(\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g} - \boldsymbol{\xi}^g) = \|\boldsymbol{\xi}^g\|_* \|\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g} - \boldsymbol{\xi}^g\|.$$

Note that the feasibility of $\boldsymbol{\xi}^g$, i.e., $\|\boldsymbol{\xi}^g\|_* \leq t_g$, holds by definition of $\kappa^g$.

Let us first assume that $\|\boldsymbol{\xi}^g\|_* < t_g$. We necessarily have that $\mathbf{u}_{|g}$ also lies in the interior of the ball of $\|.\|_*$ with radius $t_g$, and it holds that $\boldsymbol{\xi}^g = \mathbf{u}_{|g}$. Since $g \subseteq h$, we have that the vector $\mathbf{u}_{|h} - \boldsymbol{\xi}^g = \mathbf{u}_{|h} - \mathbf{u}_{|g}$ has only zero entries on $g$. As a result, $\boldsymbol{\xi}^h_g = 0$ (or equivalently, $\boldsymbol{\xi}^h_{|g} = 0$) and we obtain

$$\boldsymbol{\xi}^g = \mathbf{u}_{|g} = \mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g},$$

which is the desired conclusion. From now on, we assume that $\|\boldsymbol{\xi}^g\|_* = t_g$. It then remains to show that

$$\boldsymbol{\xi}^{g\top}(\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g} - \boldsymbol{\xi}^g) = \|\boldsymbol{\xi}^g\|_* \|\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g} - \boldsymbol{\xi}^g\|.$$

We now distinguish two cases, according to the norm used.

$\ell_2$-norm: As a consequence of Lemma 34, the optimality condition reduces to the conditions for equality in the Cauchy-Schwartz inequality, i.e., when the vectors have same signs and are linearly dependent. Applying these conditions to individual projections we get that there exists $\rho_g, \rho_h > 0$ such that

$$\rho_g \boldsymbol{\xi}^g = \mathbf{u}_{|g} - \boldsymbol{\xi}^g \quad \text{and} \quad \rho_h \boldsymbol{\xi}^h = \mathbf{u}_{|h} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h. \tag{A.26}$$

Note that the case $\rho_h = 0$ leads to $\mathbf{u}_{|h} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h = 0$, and therefore $\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g} = 0$ since $g \subseteq h$, which directly yields the result. The case $\rho_g = 0$ implies $\mathbf{u}_{|g} - \boldsymbol{\xi}^g = 0$ and therefore $\boldsymbol{\xi}^h_{|g} = 0$, yielding the result as well. Now, we can therefore assume $\rho_h > 0$ and $\rho_g > 0$. From the first equality of (A.26), we have $\boldsymbol{\xi}^g = \boldsymbol{\xi}^g_{|g}$ since $(\rho_g + 1)\boldsymbol{\xi}^g = \mathbf{u}_{|g}$. Further using the fact that $g \subseteq h$ in the second equality of (A.26), we obtain

$$(\rho_h + 1)\boldsymbol{\xi}^h_{|g} = \mathbf{u}_{|g} - \boldsymbol{\xi}^g = \rho_g \boldsymbol{\xi}^g.$$

This implies that $\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g} = \rho_g \boldsymbol{\xi}^g - \frac{\rho_g}{\rho_h+1}\boldsymbol{\xi}^g$, which eventually leads to

$$\boldsymbol{\xi}^g = \frac{\rho_h + 1}{\rho_g \rho_h}(\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}).$$

The desired conclusion follows $\boldsymbol{\xi}^{g\top}(\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}) = \|\boldsymbol{\xi}^g\|_2 \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}\|_2$.

$\ell_\infty$-norm: In this case, the optimality corresponds to the conditions for equality in the $\ell_\infty$-$\ell_1$ Hölder inequality. Specifically, $\boldsymbol{\xi}^g = \Pi_{\|.\|_* \leq t_g}(\mathbf{u}_{|g})$ holds if and only if for all $\boldsymbol{\xi}^g_j \neq 0, j \in g$, we have

$$\mathbf{u}_j - \boldsymbol{\xi}^g_j = \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g\|_\infty \mathrm{sign}(\boldsymbol{\xi}^g_j).$$

Looking at the same condition for $\boldsymbol{\xi}^h$, we have that $\boldsymbol{\xi}^h = \Pi_{\|.\|_* \leq t_h}(\mathbf{u}_{|h} - \boldsymbol{\xi}_g)$ holds if and only if for all $\boldsymbol{\xi}^h_j \neq 0, j \in h$, we have

$$\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j = \|\mathbf{u}_{|h} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h\|_\infty \mathrm{sign}(\boldsymbol{\xi}^h_j).$$

From those relationships we notably deduce that for all $j \in g$ such that $\boldsymbol{\xi}^g_j \neq 0$, $\mathrm{sign}(\boldsymbol{\xi}^g_j) = \mathrm{sign}(\mathbf{u}_j) = \mathrm{sign}(\boldsymbol{\xi}^h_j) = \mathrm{sign}(\mathbf{u}_j - \boldsymbol{\xi}^g_j) = \mathrm{sign}(\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j)$. Let $j \in g$ such that $\boldsymbol{\xi}^g_j \neq 0$. At this point, using the equalities we have just presented,

$$|\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j| = \begin{cases} \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g\|_\infty & \text{if } \boldsymbol{\xi}^h_j = 0 \\ \|\mathbf{u}_{|h} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h\|_\infty & \text{if } \boldsymbol{\xi}^h_j \neq 0. \end{cases}$$

Since $\|\mathbf{u}_{|g} - \boldsymbol{\xi}^g\|_\infty \geq \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}\|_\infty$ (which can be shown using the sign equalities above), and $\|\mathbf{u}_{|h} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h\|_\infty \geq \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}\|_\infty$ (since $g \subseteq h$), we have

$$\|\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}\|_\infty \geq |\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j| \geq \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}\|_\infty,$$

and therefore for all $\boldsymbol{\xi}^g_j \neq 0$, $j \in g$, we have $\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j = \|\mathbf{u}_{|g} - \boldsymbol{\xi}^g - \boldsymbol{\xi}^h_{|g}\|_\infty \mathrm{sign}(\boldsymbol{\xi}^g_j)$, which yields the result. □

**Proof of Lemma 5**

*Proof.* Notice first that the procedure `computeSqNorm` is called exactly once for each group $g$ in $\mathcal{G}$, computing a set of scalars $(\rho_g)_{g\in\mathcal{G}}$ in an order which is compatible with the convergence in one pass of Algorithm 6—that is, the children of a node are processed prior to the node itself. Following such an order, the update of the group $g$ in the original Algorithm 6 computes the variable $\boldsymbol{\xi}^g$ which updates implicitly the primal variable as follows

$$\mathbf{v}_{|g} \leftarrow (1 - \frac{\lambda\omega_g}{\|\mathbf{v}_{|g}\|_2})_+ \mathbf{v}_{|g}.$$

It is now possible to show by induction that for all group $g$ in $\mathcal{G}$, after a call to the procedure `computeSqNorm`$(g)$, the auxiliary variable $\eta_g$ takes the value $\|\mathbf{v}_{|g}\|_2^2$ where $\mathbf{v}$ has the same value as during the iteration $g$ of Algorithm 6. Therefore, after calling the

procedure `computeSqNorm`$(g_0)$, where $g_0$ is the root of the tree, the values $\rho_g$ correspond to the successive scaling factors of the variable $\mathbf{v}_{|g}$ obtained during the execution of Algorithm 6. After having computed all the scaling factors $\rho_g$, $g \in \mathcal{G}$, the procedure `recursiveScaling` ensures that each variable $j$ in $\{1, \ldots, p\}$ is scaled by the product of all the $\rho_h$, where $h$ is an ancestor of the variable $j$.

The complexity of the algorithm is easy to characterize: Each procedure `computeSqNorm` and `recursiveScaling` is called $p$ times, each call for a group $g$ has a constant number of operations plus as many operations as the number of children of $p$. Since each child can be called at most one time, the total number of operation of the algorithm is $O(p)$. □

### Sign conservation by projection

The next lemma specifies a property for projections when $\|.\|$ is further assumed to be a $\ell_q$-norm (with $q \geq 1$). We recall that in that case, $\|.\|_*$ is simply the $\ell_{q'}$-norm, with $q' = (1 - 1/q)^{-1}$.

**Lemma 35** (Projection on the dual ball and sign property)

*Let $\mathbf{w} \in \mathbb{R}^p$ and $t > 0$. Let us assume that $\|.\|$ is a $\ell_q$-norm (with $q \geq 1$). Consider also a diagonal matrix $\mathbf{S} \in \mathbb{R}^{p \times p}$ whose diagonal entries are in $\{-1, 1\}$. We have $\Pi_{\|.\|_* \leq t}(\mathbf{w}) = \mathbf{S}\Pi_{\|.\|_* \leq t}(\mathbf{S}\mathbf{w})$.*

*Proof.* Let us consider $\boldsymbol{\kappa} = \Pi_{\|.\|_* \leq t}(\mathbf{w})$. Using essentially the same argument as in the proof of Lemma 34, we have for all $\mathbf{y}$ such that $\|\mathbf{y}\|_{q'} \leq t$, $(\mathbf{w} - \boldsymbol{\kappa})^\top (\mathbf{y} - \boldsymbol{\kappa}) \leq 0$. Noticing that $\mathbf{S}^\top \mathbf{S} = \mathbf{I}$ and $\|\mathbf{y}\|_{q'} = \|\mathbf{S}\mathbf{y}\|_{q'}$, we further obtain $(\mathbf{S}\mathbf{w} - \mathbf{S}\boldsymbol{\kappa})^\top (\mathbf{y}' - \mathbf{S}\boldsymbol{\kappa}) \leq 0$ for all $\mathbf{y}'$ with $\|\mathbf{y}'\|_{q'} \leq t$. This implies in turn that $\mathbf{S}\Pi_{\|.\|_* \leq t}(\mathbf{w}) = \Pi_{\|.\|_* \leq t}(\mathbf{S}\mathbf{w})$, which is equivalent to the advertised conclusion. □

Based on this lemma, note that we can assume without loss of generality that the vector we want to project (in this case, $\mathbf{w}$) has only nonnegative entries. Indeed, it is sufficient to store beforehand the signs of that vector, compute the projection of the vector with nonnegative entries, and assign the stored signs to the result of the projection.

### Non-negativity constraint for the proximal operator

The next lemma shows how we can easily add a non-negativity constraint on the proximal operator when the norm $\Omega$ is *absolute* (Stewart and Sun, 1990, Definition 1.2), that is, a norm for which the relation $\Omega(\mathbf{u}) \leq \Omega(\mathbf{w})$ holds for any two vectors $\mathbf{w}$ and $\mathbf{u} \in \mathbb{R}^p$ such that $|\mathbf{u}_j| \leq |\mathbf{w}_j|$ for all $j$.

**Lemma 36** (Non-negativity constraint for the proximal operator)

*Let $\boldsymbol{\kappa} \in \mathbb{R}^p$ and $\lambda > 0$. Consider an absolute norm $\Omega$. We have*

$$\operatorname*{arg\,min}_{\mathbf{z} \in \mathbb{R}^p} \left[ \frac{1}{2} \|[\boldsymbol{\kappa}]_+ - \mathbf{z}\|_2^2 + \lambda \Omega(\mathbf{z}) \right] = \operatorname*{arg\,min}_{\mathbf{z} \in \mathbb{R}^p_+} \left[ \frac{1}{2} \|\boldsymbol{\kappa} - \mathbf{z}\|_2^2 + \lambda \Omega(\mathbf{z}) \right]. \tag{A.27}$$

*Proof.* Let us denote by $\hat{\mathbf{z}}^+$ and $\hat{\mathbf{z}}$ the unique solutions of the left- and right-hand side of (A.27) respectively. Consider the normal cone $\mathcal{N}_{\mathbb{R}^p_+}(\mathbf{z}_0)$ of $\mathbb{R}^p_+$ at the point $\mathbf{z}_0$ (Borwein and Lewis, 2006) and decompose $\boldsymbol{\kappa}$ into its positive and negative parts, $\boldsymbol{\kappa} = [\boldsymbol{\kappa}]_+ + [\boldsymbol{\kappa}]_-$. We can now write down the optimality conditions for the two convex problems above (Borwein and Lewis, 2006): $\hat{\mathbf{z}}^+$ is optimal if and only if there exists $\mathbf{w} \in \partial\Omega(\hat{\mathbf{z}}^+)$ such that $\hat{\mathbf{z}}^+ - [\boldsymbol{\kappa}]_+ + \lambda\mathbf{w} = \mathbf{0}$. Similarly, $\hat{\mathbf{z}}$ is optimal if and only if there exists $(\mathbf{s}, \mathbf{u}) \in \partial\Omega(\hat{\mathbf{z}}) \times \mathcal{N}_{\mathbb{R}^p_+}(\hat{\mathbf{z}})$ such that $\hat{\mathbf{z}} - \boldsymbol{\kappa} + \lambda\mathbf{s} + \mathbf{u} = \mathbf{0}$. We now prove that $[\boldsymbol{\kappa}]_- = \boldsymbol{\kappa} - [\boldsymbol{\kappa}]_+$ belongs to $\mathcal{N}_{\mathbb{R}^p_+}(\hat{\mathbf{z}}^+)$. We proceed by contradiction. Let us assume that there exists $\mathbf{z} \in \mathbb{R}^p_+$ such that $[\boldsymbol{\kappa}]_-^\top(\mathbf{z} - \hat{\mathbf{z}}^+) > 0$. This implies that there exists $j \in \{1, \dots, p\}$ for which $[\boldsymbol{\kappa}_j]_- < 0$ and $\mathbf{z}_j - \hat{\mathbf{z}}^+_j < 0$. In other words, we have $0 \leq \mathbf{z}_j = \mathbf{z}_j - [\boldsymbol{\kappa}_j]_+ < \hat{\mathbf{z}}^+_j = \hat{\mathbf{z}}^+_j - [\boldsymbol{\kappa}_j]_+$. With the assumption made on $\Omega$ and replacing $\hat{\mathbf{z}}^+_j$ by $\mathbf{z}_j$, we have found a solution to the left-hand side of (A.27) with a stricly smaller cost function than the one evaluated at $\hat{\mathbf{z}}^+$, hence the contradiction. Putting the pieces together, we now have

$$\hat{\mathbf{z}}^+ - [\boldsymbol{\kappa}]_+ + \lambda\mathbf{w} = \hat{\mathbf{z}}^+ - \boldsymbol{\kappa} + \lambda\mathbf{w} + [\boldsymbol{\kappa}]_- = \mathbf{0}, \text{ with } (\mathbf{w}, [\boldsymbol{\kappa}]_-) \in \partial\Omega(\hat{\mathbf{z}}^+) \times \mathcal{N}_{\mathbb{R}^p_+}(\hat{\mathbf{z}}^+),$$

which shows that $\hat{\mathbf{z}}^+$ is the solution of the right-hand side of (A.27). □

### A.2.3 Counterexample for $\ell_q$-norms, with $q \notin \{1, 2, \infty\}$.

The result we have proved in Proposition 9 in the specific setting where $\|.\|$ is the $\ell_2$- or $\ell_\infty$-norm does not hold more generally for $\ell_q$-norms, when $q$ is not in $\{1, 2, \infty\}$. Let $q > 1$ satisfying this condition. We denote by $q' \triangleq (1 - q^{-1})^{-1}$ the norm parameter dual to $q$. We keep the same notation as in Lemma 3 and assume from now on that $\|\mathbf{u}_{|g}\|_{q'} > t_g$ and $\|\mathbf{u}_{|h}\|_{q'} > t_g + t_h$. These two inequalities guarantee that the vectors $\mathbf{u}_{|g}$ and $\mathbf{u}_{|h} - \boldsymbol{\xi}^g$ do not lie in the interior of the $\ell_{q'}$-norm balls, of respective radius $t_g$ and $t_h$.

We show in this section that there exists a setting for which the conclusion of Lemma 3 does not hold anymore. We first focus on a necessary condition of Lemma 3:

**Lemma 37** (Necessary condition of Lemma 3)

*Let $\|.\|$ be a $\ell_q$-norm, with $q \notin \{1, 2, \infty\}$. If the conclusion of Lemma 3 holds, then the vectors $\boldsymbol{\xi}^g_{|g}$ and $\boldsymbol{\xi}^h_{|g}$ are linearly dependent.*

*Proof.* According to our assumptions on $\mathbf{u}_{|g}$ and $\mathbf{u}_{|h} - \boldsymbol{\xi}^g$, we have that $\|\boldsymbol{\xi}^g\|_{q'} = t_g$ and $\|\boldsymbol{\xi}^h\|_{q'} = t_h$. In this case, we can apply the second optimality conditions of Lemma 34, which states that equality holds in the $\ell_q$-$\ell_{q'}$ Hölder inequality. As a result, there exists $\rho_g, \rho_h > 0$ such that for all $j$ in $g$:

$$|\boldsymbol{\xi}^g_j|^{q'} = \rho_g |\mathbf{u}_j - \boldsymbol{\xi}^g_j|^q \quad \text{and} \quad |\boldsymbol{\xi}^h_j|^{q'} = \rho_h |\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j|^q. \tag{A.28}$$

If the conclusion of Lemma 3 holds—that is, we have $\boldsymbol{\xi}^g = \Pi_{\|.\|_* \leq t_g}(\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g})$, notice that it is not possible to have the following scenarios, as proved below by contradiction:

- If $\|\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g}\|_{q'} < t_g$, then we would have $\boldsymbol{\xi}^g = \mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g}$, which is impossible since $\|\boldsymbol{\xi}^g\|_{q'} = t_g$.
- If $\|\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g}\|_{q'} = t_g$, then we would have for all $j$ in $g$, $|\boldsymbol{\xi}^h_j|^{q'} = \rho_h |\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j|^q = 0$, which implies that $\boldsymbol{\xi}^h_{|g} = 0$ and $\|\mathbf{u}_{|g}\|_{q'} = t_g$. This is impossible since we assumed $\|\mathbf{u}_{|g}\|_{q'} > t_g$.

We therefore have $\|\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g}\|_{q'} > t_g$ and using again the second optimality conditions of Lemma 34, there exists $\rho > 0$ such that for all $j$ in $g$, $|\boldsymbol{\xi}^g_j|^{q'} = \rho |\mathbf{u}_j - \boldsymbol{\xi}^g_j - \boldsymbol{\xi}^h_j|^q$. Combined with the previous relation on $\boldsymbol{\xi}^h_{|g}$, we obtain for all $j$ in $g$, $|\boldsymbol{\xi}^g_j|^{q'} = \frac{\rho}{\rho_h} |\boldsymbol{\xi}^h_j|^{q'}$. Since we can assume without loss of generality that $\mathbf{u}$ only has nonnegative entries (see Lemma 35), the vectors $\boldsymbol{\xi}^g$ and $\boldsymbol{\xi}^h$ can also be assumed to have nonnegative entries, hence the desired conclusion. □

We need another intuitive property of the projection $\Pi_{\|.\|_* \leq t}$ to derive our counterexample:

**Lemma 38** (Order-preservation by projection)

*Let $\|.\|$ be a $\ell_q$-norm, with $q \notin \{1, \infty\}$ and $q' \triangleq 1/(1 - q^{-1})$. Let us consider the vectors $\boldsymbol{\kappa}, \mathbf{w} \in \mathbb{R}^p$ such that $\boldsymbol{\kappa} = \Pi_{\|.\|_* \leq t}(\mathbf{w}) = \arg\min_{\|\mathbf{y}\|_{q'} \leq t} \|\mathbf{y} - \mathbf{w}\|_2$, with the radius $t$ satisfying $\|\mathbf{w}\|_{q'} > t$. If we have $\mathbf{w}_i < \mathbf{w}_j$ for some $(i, j)$ in $\{1, \ldots, p\}^2$, then it also holds that $\boldsymbol{\kappa}_i < \boldsymbol{\kappa}_j$.*

*Proof.* Let us first notice that given the assumption on $t$, we have $\|\boldsymbol{\kappa}\|_{q'} = t$. The Lagrangian $\mathcal{L}$ associated with the convex minimization problem underlying the definition of $\Pi_{\|.\|_* \leq t}$ can be written as

$$\mathcal{L}(\mathbf{y}, \alpha) = \frac{1}{2}\|\mathbf{y} - \mathbf{w}\|_2^2 + \alpha\big[\|\mathbf{y}\|_{q'}^{q'} - t^{q'}\big], \text{ with the Lagrangian parameter } \alpha \geq 0.$$

At optimality, the stationarity condition for $\boldsymbol{\kappa}$ leads to

$$\forall\, j \in \{1, \ldots, p\},\ \boldsymbol{\kappa}_j - \mathbf{w}_j + \alpha q' |\boldsymbol{\kappa}_j|^{q'-1} = 0.$$

We can assume without loss of generality that $\mathbf{w}$ only has nonnegative entries (see Lemma 35). Since the components of $\boldsymbol{\kappa}$ and $\mathbf{w}$ have the same signs (see Lemma 35), we therefore have $|\boldsymbol{\kappa}_j| = \boldsymbol{\kappa}_j \geq 0$, for all $j$ in $\{1, \ldots, p\}$. Note that $\alpha$ cannot be equal to zero because of $\|\boldsymbol{\kappa}\|_{q'} = t < \|\mathbf{w}\|_{q'}$.

Let us consider the continuously differentiable function $\varphi_w : \kappa \mapsto \kappa - w + \alpha q' \kappa^{q'-1}$ defined on $(0, \infty)$. Since $\varphi_w(0) = -w < 0$, $\lim_{\kappa \to \infty} \varphi_w(\kappa) = \infty$ and $\varphi_w$ is strictly nondecreasing, there exists a unique $\kappa^*_w > 0$ such that $\varphi_w(\kappa^*_w) = 0$. If we now take $w < v$, we have

$$\varphi_v(\kappa^*_w) = \varphi_w(\kappa^*_w) + w - v = w - v < 0 = \varphi_v(\kappa^*_v).$$

With $\varphi_v$ being strictly nondecreasing, we thus obtain $\kappa^*_w < \kappa^*_v$. The desired conclusion stems from the application of the previous result to the stationarity condition of $\boldsymbol{\kappa}$. □

Based on the two previous lemmas, we are now in position to present our counterexample:

**Proposition 15** (Counterexample)

*Let $\|.\|$ be a $\ell_q$-norm, with $q \notin \{1, 2, \infty\}$ and $q' \triangleq 1/(1-q^{-1})$. Let us consider $\mathcal{G} = \{g, h\}$, with $g \subseteq h \subseteq \{1, \dots, p\}$ and $|g| > 1$. Let $\mathbf{u}$ be a vector in $\mathbb{R}^p$ that has at least two different nonzero entries in $g$, i.e., there exists $(i, j)$ in $g \times g$ such that $0 < |\mathbf{u}_i| < |\mathbf{u}_j|$. Let us consider the successive projections*

$$\boldsymbol{\xi}^g \triangleq \Pi_{\|.\|_* \leq t_g}(\mathbf{u}_{|g}) \quad \textit{and} \quad \boldsymbol{\xi}^h \triangleq \Pi_{\|.\|_* \leq t_h}(\mathbf{u}_{|h} - \boldsymbol{\xi}^g)$$

*with $t_g, t_h > 0$ satisfying $\|\mathbf{u}_{|g}\|_{q'} > t_g$ and $\|\mathbf{u}_{|h}\|_{q'} > t_g + t_h$. Then, the conclusion of Lemma 3 does not hold.*

*Proof.* We apply the same rationale as in the proof of Lemma 38. Writing the stationarity conditions for $\boldsymbol{\xi}^g$ and $\boldsymbol{\xi}^h$, we have for all $j$ in $g$

$$\boldsymbol{\xi}^g_j + \alpha q' (\boldsymbol{\xi}^g_j)^{q'-1} - \mathbf{u}_j = 0, \quad \text{and} \quad \boldsymbol{\xi}^h_j + \beta q' (\boldsymbol{\xi}^h_j)^{q'-1} - (\mathbf{u}_j - \boldsymbol{\xi}^g_j) = 0, \tag{A.29}$$

with Lagrangian parameters $\alpha, \beta > 0$. We now proceed by contradiction and assume that $\boldsymbol{\xi}^g = \Pi_{\|.\|_* \leq t_g}(\mathbf{u}_{|g} - \boldsymbol{\xi}^h_{|g})$. According to Lemma 37, there exists $\rho > 0$ such that for all $j$ in $g$, $\boldsymbol{\xi}^h_j = \rho \boldsymbol{\xi}^g_j$. If we combine the previous relations on $\boldsymbol{\xi}^g$ and $\boldsymbol{\xi}^h$, we obtain for all $j$ in $g$,

$$\boldsymbol{\xi}^g_j = C (\boldsymbol{\xi}^g_j)^{q'-1}, \text{ with } C \triangleq \frac{q'(\alpha - \beta \rho^{q'-1})}{\rho}.$$

If $C < 0$, then we have a contradiction, since the entries of $\boldsymbol{\xi}^g$ and $\mathbf{u}_{|g}$ have the same signs. Similarly, the case $C = 0$ leads a contradiction, since we would have $\mathbf{u}_{|g} = 0$ and $\|\mathbf{u}_{|g}\|_{q'} > t_g$. As a consequence, it follows that $C > 0$ and for all $j$ in $g$, $\boldsymbol{\xi}^g_j = \exp\{\frac{\log(C)}{2-q'}\}$, which means that all the entries of the vector $\boldsymbol{\xi}^g_g$ are identical. Using Lemma 38, since there exists $(i, j) \in g \times g$ such that $\mathbf{u}_i < \mathbf{u}_j$, we also have $\boldsymbol{\xi}^g_i < \boldsymbol{\xi}^g_j$, which leads to a contradiction. □

# Bibliography


P. A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

R. J. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. IMS, 1990.

M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

R. K. Ahuja, T. L. Magnanti, and J. Orlin. *Network Flows*. Prentice Hall, 1993.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

J. Y. Audibert. Pac-bayesian aggregation and multi-armed bandits, 2010. Habilitation thesis.

M. Babenko and A. V. Goldberg. Experimental evaluation of a parametric flow algorithm. Technical report, Microsoft Research, 2006. MSR-TR-2006-77.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008a.

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008b.

F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008c.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2009.

F. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, 2010a.

F. Bach. Shaping level sets with submodular functions. Technical report, Preprint arXiv:1012.1501, 2010b.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical report, Preprint arXiv:0812.1869, 2008.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*. MIT press, 2011. To appear.

R. Baraniuk. Optimal tree approximation with wavelets. *Wavelet Applications in Signal and Image Processing VII*, 3813:206214, 1999.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

R. G. Baraniuk, R. A. DeVore, G. Kyriazis, and X. M. Yu. Near best tree approximation. *Advances in Computational Mathematics*, 16(4):357–373, 2002.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56:1982–2001, 2010.

A. I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete and Computational Geometry*, 13(1):189–202, 1995.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

S. Becker, J. Bobin, and E. Candes. NESTA: A Fast and Accurate First-order Method for Sparse Recovery. Technical report, Preprint arXiv:0904.3367, 2009.

C. F. Beckmann and S. M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23 (2):137–152, 2004.

C.F. Beckmann, M. DeLuca, J.T. Devlin, and S.M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1001, 2005.

Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 2009.

D. P. Bertsekas. *Linear network optimization: algorithms and codes*. The MIT Press, 1991.

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

J. Bien, Y. Xu, and M. W. Mahoney. CUR from a Sparse Optimization Viewpoint. In *Advances in Neural Information Processing Systems*, 2010.

B. B. Biswal, M. Mennes, X. N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734, 2010.

D. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2008.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2): 1–30, 2010.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. *Advanced Lectures on Machine Learning*, pages 208–240, 2004.

Y.L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends ® in Machine Learning*, 3:1–124, 2011.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

D. M. Bradley and J. A. Bagnell. Differentiable sparse coding. In *Advances in Neural Information Processing Systems*, 2009a.

D. M. Bradley and J. A. Bagnell. Convex coding. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009b.

P. Brucker. An O(n) algorithm for quadratic knapsack problems. *Operations Research Letters*, 3:163–166, 1984.

V. V. Buldygin and I. U.V . Kozachenko. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Society, 2000.

W. L. Buntine. Variational Extensions to EM and Multinomial PCA. In *Proceedings of the European Conference on Machine Learning (ECML)*, 2002.

D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized non-negative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. To appear.

P. J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, 1994.

E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009.

E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

M. K. Carroll, G. A. Cecchi, I. Rish, R. Garg, and A. R. Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112 – 122, 2009.

G. A. Cecchi, I. Rish, B. Thyreau, B. Thirion, M. Plaze, M. L. Paillere-Martinot, C. Martelli, J. L. Martinot, and J.B. Poline. Discriminative network models of schizophrenia. In *Advances in Neural Information Processing Systems*. 2009.

V. Cevher. Learning with compressible priors. In *Advances in Neural Information Processing Systems*, 2008.

V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery using markov random fields. In *Advances in Neural Information Processing Systems*, 2008.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

X. Chen, Q. Lin, S. Kim, J. Peña, J. G. Carbonell, and E. P. Xing. An efficient proximal-gradient method for single and multi-task regression with structured sparsity. Technical report, Preprint arXiv:1005.4717, 2010.

D. B. Chklovskii and A. A. Koulakov. Maps in the brain: What can we learn from them? *Annual Review of Neuroscience*, 27(1):369–392, 2004.

F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *American Mathematical Society*, 22(1):211–231, 2009.

R. R. Coifman and D. L. Donoho. Translation-invariant de-noising. *Lectures notes in statistics*, pages 125–125, 1995.

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, 2010.

P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2006.

D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003.

M. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, 1998.

N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2006.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. DÁrdenne, W. Richter, J. D. Cohen, and J. Haxby. Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences*, 106(26):10415, 2009.

I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

S. Dehaene, G. Le Clec'H, L. Cohen, J.-B. Poline, P.-F. van de Moortele, and D. Le Bihan. Inferring behavior from functional brain images. *Nature Neuroscience*, 1:549, 1998.

J. P. Doignon and J. C. Falmagne. *Knowledge Spaces*. Springer-Verlag, 1998.

D. L. Donoho. CART and best-ortho-basis: a connection. *Annals of Statistics*, pages 1870–1911, 1997.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), 1995.

C. Dossal. A necessary and sufficient condition for exact recovery by l1 minimization. Technical report, HAL-00164738:1, 2007.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the $\ell_1$-ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–451, 2004.

E. Eger, C. Kell, and A. Kleinschmidt. Graded size sensitivity of object exemplar evoked activity patterns in human loc subregions. *J. Neurophysiol.*, 100(4):2038-47, 2008.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.

C. Févotte, N. Bertin, and J. L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21 (3):793–830, 2009.

G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache, and J.-B. Poline. Improved detection sensitivity in functional MRI data using a brain parcelling technique. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'02)*, pages 467–474, 2002.

L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Classic papers in combinatorics*, pages 243–248, 1987.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. Technical report, Preprint arXiv:1001.0736, 2010.

K. J. Friston, A. P. Holmes, K. J. Worsley, J. B. Poline, C. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

W. Fu and K. Knight. Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28 (5):1356–1378, 2000.

W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

J. J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18:30, 1989.

Q. Geng, H. Wang, and J. Wright. On the Local Correctness of L1 Minimization for Dictionary Learning. Technical report, Preprint arXiv:1101.5672, 2011.

A. V. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988.

A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *IEEE International Symposium on Biomedical Imaging*, 2009.

R. Gribonval and K. Schnass. Dictionary identification—sparse matrix-factorization via $\ell_1$-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228, 2004.

H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European journal of operational research*, 54(2):227–236, 1991.

L. Grosenick, S. Greer, and B. Knutson. Interpretable classifiers for FMRI improve prediction of purchases. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(6):539–548, 2009. ISSN 1534-4320.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

L. He and L. Carin. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497, 2009.

D. S. Hochbaum and S. P. Hong. About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical programming*, 69 (1):269–309, 1995.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. Technical report, Available at `stat.wharton.upenn.edu/~skakade/papers/Working/vector.pdf`, 2011.

C. Hu, J. T. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, 2009.

J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4): 1978–2004, 2010.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010b.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011a.

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion. Multi-scale mining of fMRI data with hierarchical structured sparsity. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2011b.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011c.

S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

I. T. Jolliffe. *Principal component analysis*, volume 1 of *Series in Statistics*. Springer, 1986.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

A. Juditski and A. S. Nemirovski. Accuracy guarantees for l1-recovery. Technical report, Preprint arXiv:1008.3651, 2010.

K. Kavukcuoglu, M. A. Ranzato, R. Fergus, and Y. Le-Cun. Learning invariant features through topographic filter maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. Technical report, Preprint arXiv:1009.0744, 2010.

A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu. Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, 26(2):317 – 329, 2005.

S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, 2008.

J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, 2007.

A. Lefèvre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the itakura-saito divergence. Technical report, Preprint arXiv:1106.419, 2011a.

A. Lefèvre, F. Bach, and C. Févotte. Itakura-saito nonnegative matrix factorization with group sparsity. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011b.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Verlag, 2005.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*, 2(1):245–263, 2008.

G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010a.

H. Liu. *Nonparametric Learning in High Dimensions*. PhD thesis, Carnegie Mellon University, 2010.

H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. Technical report, Preprint arXiv:1006.3316, 2010b.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

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

K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical report, Preprint arXiv:0903.1468, 2009.

L. Mackey. Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems*, 2009.

N. Maculan and J. R. G. Galdino de Paula. A linear-time median-finding algorithm for projecting a vector on the simplex of Rn. *Operations Research Letters*, 8(4):219–222, 1989.

M. W. Mahoney and P. Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697, 2009.

J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, École normale supérieure de Cachan - ENS Cachan, 2010. Available at `http://tel.archives-ouvertes.fr/tel-00595312/fr/`.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems*, 2009a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009b.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1):19–60, 2010a.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010b.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

S. G. Mallat. *A wavelet tour of signal processing*. Academic Press, 1999.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2001.

A. M. Martinez and A. C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43(1):44 – 58, 2008.

A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo. Structured sparsity in structured prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

A. Maurer and M. Pontil. $k$-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B*, 70:53–71, 2008.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society. Series B*, 72(4):417–473, 2010.

S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6(2):1099, 2006.

C. A. Micchelli, J. M. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.

V. Michel, E. Eger, C. Keribin, J.-B. Poline, and B. Thirion. A supervised clustering approach for extracting predictive information from brain activation images. *MMBIA'10*, 2010.

V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fMRI-based prediction of behaviour. *Medical Imaging, IEEE Transactions on*, PP(99):1, 2011.

B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, 2006.

J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.*, 255:2897–2899, 1962.

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24:227, 1995.

D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, 2009.

J. Nelson. Johnson-Lindenstrauss notes. Technical report, MIT-CSAIL, Available at `web.mit.edu/minilek/www/jl_notes.pdf`, 2010.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 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*, 20(3):389, 2000b.

G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, pages 339–358, 1998.

G. Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

Z. Qin and D. Goldfarb. Structured sparsity via alternating directions methods. Technical report, Preprint arXiv:1105.0728, 2011.

A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An efficient projection for $\ell_1/\ell_\infty$ regularization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

P. Ramachandran and G. Varoquaux. Mayavi: 3d visualization of scientific data. *Computing in Science Engineering*, 13(2):40 –51, march-april 2011.

N. S. Rao, R. D. Nowak, S. J. Wright, and N. G. Kingsbury. Convex approaches to model wavelet sparsity patterns. In *International Conference on Image Processing (ICIP)*, 2011.

F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

P. Ravikumar, G. Raskutti, M. Wainwright, and B. Yu. Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. In *Advances in Neural Information Processing Systems*, 2008.

J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, and D. Van De Ville. Decoding brain states from fMRI connectivity graphs. *NeuroImage*, 2010.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

J. Rissman, H. T. Greely, and A. D. Wagner. Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences*, 107(21):9849–9854, 2010.

R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1997.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35(3):1012–1030, 2007.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

S. Ryali, K. Supekar, D. A. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752 – 764, 2010.

M. Schmidt and K. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

M. Schmidt, D. Kim, and S. Sra. Projected newton-type methods in machine learning. In *Optimization for Machine Learning*. Springer, 2011a. To appear.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, 2011b.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

A. P. Singh and G. J. Gordon. A Unified View of Matrix Factorization Models. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases-Part II*, 2008.

P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer, 2003.

P. Sprechmann, I. Ramirez, P. Cancela, and G. Sapiro. Collaborative sources identification in mixed signals via hierarchical sparse modeling. Technical report, Preprint arXiv:1010.4893, 2010a.

P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar. Collaborative hierarchical sparse modeling. Technical report, Preprint arXiv:1003.0400, 2010b.

G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory (Computer Science and Scientific Computing)*. Academic Press, 1990.

M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.

M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux. Hierarchical penalization. In *Advances in Neural Information Processing Systems*, 2007.

B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.-B. Poline. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.*, 27(8):678–693, 2006.

A. Tibau Puig, A. Wiesel, A. Zaas, C. Woods, G. Ginsburg, G. Fleury, and A. Hero. Order-preserving factor analysis-application to longitudinal gene expression. *IEEE Transactions on Signal Processing*, 99(99):1–1, 2011.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

K. C. Toh, M. J. Todd, and R. H. Tütüncü. SDPT3–a MATLAB software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11(1): 545–581, 1999.

I. Tosic and P. Frossard. Dictionary learning. *Signal Processing Magazine*, 28(2):27–38, 2011.

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3), 2006.

J. A. Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathématique Académie des sciences*, 346(23-24):1271–1274, 2008.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

A. Tucholka, B. Thirion, M. Perrot, P. Pinel, J. F. Mangin, and J. B. Poline. Probabilistic anatomo-functional parcellation of the cortex: how many regions? *MICCAI 2008*, pages 399–406, 2008.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95(2):189–217, 2003.

K. Ugurbil, L. Toth, and D. Kim. How accurate is magnetic resonance imaging of brain function? *Trends in Neurosciences*, 26(2):108 – 114, 2003.

D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. Technical report, Preprint arXiv:1011.5395, 2010.

S. Van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

A. Vanhaudenhuyse, Q. Noirhomme, L. J. F. Tshibanda, M. A. Bruno, P. Boveroux, C. Schnakers, A. Soddu, V. Perlbarg, D. Ledoux, J. F. Brichant, et al. Default network connectivity reflects level of consciousness in non-communicative brain-damaged patients. *Brain*, 133:161, 2010.

G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, and B. Thirion. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In *MICCAI*. 2010a.

G. Varoquaux, A. Gramfort, J. B. Poline, and B. Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems*. 2010b.

G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. Sparse structured dictionary learning for brain resting-state activity modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010c.

G. Varoquaux, M. Keller, J.B. Poline, P. Ciuciu, and B. Thirion. ICA-based sparse features recovery from fMRI datasets. In *International Symposium on Biomedical Imaging*, pages 1177–1180. IEEE, 2010d.

G. Varoquaux, S. Sadaghiani, P. Pinel, A. Kleinschmidt, J. B. Poline, and B. Thirion. A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, 51(1): 288–299, 2010e.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, Preprint arXiv:1011.3027, 2010.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using $\ell_1$- constrained quadratic programming. *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using $\ell_1$-regularized logistic regression. In *Advances in Neural Information Processing Systems*, 2007.

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515, 2009.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 210–227, 2008.

S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. Technical report, UW-Madison, 2010.

S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

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

Z. J. Xiang, Y. T. Xi, U. Hasson, and P. J. Ramadge. Boosting with spatial regularization. In *Advances in Neural Information Processing Systems*, 2009.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, 42(4):1414 – 1429, 2008.

K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *Advances in Neural Information Processing Systems*, 2009.

G. X. Yuan, K. W. Chang, C. J. Hsieh, and C. J. Lin. Comparison of optimization methods and software for large-scale l1-regularized linear classification. *Journal of Machine Learning Research*, 11:3183–3234, 2010.

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

R. Zass and A. Shashua. Nonnegative sparse PCA. In *Advances in Neural Information Processing Systems*, 2007.

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

Y. Zhang, A. d'Aspremont, and L. El Ghaoui. Sparse pca: Convex relaxations, algorithms and applications. In M. Anjos and J. B. Lasserre, editors, *Handbook on Semidefinite, Cone and Polynomial Optimization: theory, algorithms, software and applications*. 2011.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.

M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems*, 2009.

S. Zhou. Restricted eigenvalue conditions on subgaussian random matrices. Technical report, Preprint arXiv:0912.4045, 2009.

J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: maximum margin supervised topic models for regression and classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

D. Zoran and Y. Weiss. The "tree-dependent components" of natural scenes are edge filters. In *Advances in Neural Information Processing Systems*, 2009.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320, 2005.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.